Quantitative & Behavioral Genetics for Social Scientists

J.C. Barnes
University of Cincinnati

&

Kevin M. Beaver
Florida State University

August 16, 2018

Contents

List of Tables

List of Figures

Foreword

We dedicate this book to...

J.C. Barnes, Cincinnati, OH

Kevin M. Beaver, Tallahassee, FL

Date

Preface

Social scientists study human behavior by looking into what makes humans differ from one another. For the most part, this leads researchers to investigate the social world for obvious reasons: humans—much more so than any other species—are social animals. We rely on one another for all manner of things; from the planting and harvesting of crops, to the manufacture of clothes, to the physical labor that is required for our entertainment in sporting competitions. There is no other animal quite like *Homo sapiens*, and it seems safe to assume that those differences must arise due to our social environments.

This may be so, but it is only part of the story. Over the past 60 years or so, research into quantitative genetics has repeatedly shown that genetic factors also play a role in the etiology of most human differences. This, therefore, requires that we start with a different assumption about the causes of human behavior. Social science teaches us that human behavior is the result of environmental exposures. To the extent that genes matter, they can be considered a nuisance parameter that is easily ignored without any loss of information. Quantitative genetics, however, takes the exact opposite approach. As we will discuss in the chapters that follow (especially in Chapter 3), quantitative geneticists assume that human behavior is the result of genetic influences and an environmental nuisance parameter.

As you can see, there is a fundamental disagreement separating the quantitative geneticists from the social scientists. It may, at first, appear to be minor academic banter. "What's the big deal?" you might be thinking. Why doesn't one side just agree to accept the other's perspective? The reason is that the two approaches lead down very different paths. Moreover, social scientists have often assumed that genetic influences are ignorable noise (Udry, 1995). Yet, quantitative geneticists have always understood and argued the opposite. All human behaviors result from a mix of genetic and environmental influences. While quantitative geneticists often introduce the environmental influence as a deviation or nuisance parameter, this does not mean that they have ignored the environmental influences. As you will see, quantifying and conceptually identifying environmental parameters is one important element of quantitative genetic research.

Against the above backdrop, this book attempts to reconcile some of the most heated debates in the social sciences. Specifically, we hope to build a bridge between the environmental and the genetic for those who study human behavior. This is, admittedly, an overly-ambitious goal. It might be argued that scholars much smarter than us have tried

and failed to achieve these very aims. We recognize the work that lies ahead.

In order to ease the burden on the reader and to make the work more tractable for us, we have adopted several conventions that will weave through the content that follows. These are, generally, organizing principles or notational conventions that will be used throughout the text. It is our hope that by adopting these conventions, we can lighten the reading experience so that there is less time spent "Googling" terms and symbols and more time spent thinking about the concepts and what they mean for one's own research agenda.

Three general points are worth noting before you dive into the content that follows. First, although the title of this text blends *quantitative* and *behavioral* genetics—which might suggest they are one and the same—it is our position that the former (quantitative genetics) refers broadly to the topics that are covered in Chapter 3. Specifically, for the purposes of this text, quantitative genetics covers topics related to the translation of genotypic values into quantitative phenotypes (i.e., those that are not qualitative, either/or traits) using the principles of Mendelian inheritance. Behavioral genetics, for the purposes of this text, refers to the study of how—and/or the degree to which—genetic variance is associated with phenotypic variance.

The second point to keep in mind while reading this text is that we will generally differentiate between two approaches to behavioral genetics research: 1) biometrical model fitting and 2) genomics studies. These are, admittedly, broadly and somewhat loosely defined categories. For this reason, we adopt them only as a heuristic and organizing principle. This is not to say that studies are either one or the other. In fact, some of the best research available capitalizes on both types of designs and, increasingly, the thin barrier between the two is being dissolved as technological innovations expand at an accelerating rate. But, for the student who is learning many of these techniques for the first time, it will be useful to categorize and compartmentalize the various approaches based on their analytic framework. With this in mind, we broadly classify studies that attempt to estimate latent variance components like heritability (i.e., h^2), the shared environment (c^2), and the nonshared environment (e^2) as *biometrical models*. As you can see from the table of contents, these model fitting strategies are covered in chapters 5 and 6. They will also be featured in portions of chapters 9 and 10.

Genomics is the term we have adopted to capture all other research designs in behavioral genetics. As the term implies, these types of research designs will attempt to identify specific genetic variants that are related to the phenotype of focus. In this respect, one might prefer to call these studies quantitative trait loci analyses (QTL), but we refrain from using this phrase because it has—in some arenas—been used to refer to a specific type of genomics research strategy. Thus, to avoid confusion, we will refer broadly to studies that attempt to isolate specific genetic variants as predictors of a phenotype simply as *genomics studies*. These will be covered in chapters 7 and 8 and will be featured in portions of chapters 9 and 10.

This sort of classification scheme is not unique to the present text. In fact, Kendler (2005) developed a similar (although his included 4 approaches) scheme. What is unique to this

text is the way in which we tie together biometrical models with genomics. At the risk of "jumping the gun", we will take this opportunity to briefly introduce to you our integrative approach.

The casual consumer of behavioral genetic research might be excused for thinking that research fits into either the biometrical camp or the genomics camp (see, broadly, Tabery, 2014). But, as it turns out, both camps are really just two sides of the same coin. To see how this is so, let us introduce the central equation that will run through the entire text. The reader can think of this equation as forming the backbone of our approach. That equation is:

$$P = \Psi(G, E)$$

where P is a phenotypic score, G is the genotypic value, E is the environmental value, and Ψ represents an arbitrary function where G and E can be combined in various ways to form P. That is to say, Ψ reveals that P is not simply an additive outcome where G is added to E. Of course, in some cases, P *may* be best described as an additive function of G and E. But, we need not assume this, so we use Ψ as a general reminder that G and E can combine in various ways to form P. Thus, Ψ affords the model flexibility to account for any arbitrary correlations and interactions between G and E.[1]

Do not be discouraged from pressing forward if this information does not quite "sink in" at this point. We will more formally develop the above model throughout the text, beginning with chapter 3. This equation will then—beginning with chapter 5 and continuing through chapter 10—provide the starting point for our discussion of each and every modeling strategy. We will show, for instance, how biometrical models (e.g., the ACE model) inform the above equation by attempting to estimate the degree to which variance in P is attributable to variance in G and/or E. Moreover, we will show how genomics work (e.g., candidate gene studies and genome-wide association analyses) can be thought of as methods that inform which specific genes go into the "global" G that appears in the equation above. Approaching the study of behavioral genetics in this way will allow one to see the parallels between the different strategies that so often appear to be at odds with one another in academic research (Tabery, 2014).

Finally, the third point to keep in mind while reading is that we have attempted to minimize the burden of relying on mathematical proofs as much as possible. The substantive modeling chapters (i.e., chapters 5 through 10) each begin with a conceptual discussion of the model(s) of focus. Here, we offer a description of the modeling strategy of focus, we offer hypothetical examples to emphasize scenarios where one might be inclined to rely on the focal design, and we cover the mathematical principles that are necessary to grasp the logic of the design under consideration. This is not a mathematics/statistics text, nor is it a strict "how to" guide. It is somewhere in between. We hope, therefore, that readers from all backgrounds and with varying levels of comfort with the quantitative aspects of behavioral genetics will find our discussion tractable. We hope to offer helpful insight for those trying to

[1]We must acknowledge the work of Golan et al. (2014) and Zuk et al. (2012). Their papers were some of the first we encountered that made this general model an explicit focus and, as such, we rely on their notation to develop the model in this text.

grasp the principles of the various designs that have become popular in behavioral genetics research over the past few decades. We cannot cover all of the designs that have been used, of course, but we can offer insight into some of those that are used most frequently. And this is precisely what we have set out to do. In essence, we hope that readers will gain from our discussion some sense of when, where, and for which research questions certain designs are appropriate.

Related to the above point, we have attempted to minimize the use of esoteric notation. Whenever possible, we utilize notation that is typically seen in social science research. We will generally refer to an outcome variable as either a phenotype P or simply as Y. We will often refer to a predictor variable simply as X, unless it is given a substantive label (e.g., the specific variable of focus). Following the convention established by others before us (e.g., Falconer and MacKay [1989]), we will use P to identify the phenotype or the phenotypic value (but note that some sections of chapters 2 and 3 will use p to denote proportions), G will denote the genotype or a genotypic value, and E will be the environmental component. We will rely quite heavily on regression-based modeling strategies (see, generally, Wooldridge, 2010). Thus, we have adopted a regression-based notation convention that is consistent with most social science texts. We will sometimes refer to the expected value of a variable. Typically, an expected value is symbolized as E, but this may lead to confusion in the present context given that E refers to the environment. Thus, we have adopted the blackboard-bolded $\mathbb{E}$ as the symbol for the expected value. Similarly, we adopted the blackboard-bolded $\mathbb{P}$ to identify probabilities so as to avoid confusion with our use of P as the phenotype or the use of p for a proportion.

There are many approaches one could take to reading this text. In fact, one of the reasons we endeavored to write this book was to offer students and researchers a text that they could selectively read as needed. With this in mind, we envision most readers will want to pick and choose chapters based on their present interests. We encourage this approach, but with one caveat; the material covered in chapter 3 should be considered before moving to any of the chapters in part II (i.e., the modeling strategies). Chapter 3 will provide the conceptual (and, in some ways, mathematical) foundation necessary to truly internalize and understand the modeling strategies discussed in part II. Thus, we encourage those readers who intend to selectively jump from one topic to the next to begin with chapter 3 before moving to any others.

Part I

Introduction & Background

Chapter 1

The Importance of Genetics for Social Science Research

Genes matter. They matter for physical characteristics, emotional traits, and behavioral patterns. They matter for males and females. They matter for all races, all ages, and for persons living in every region of the world. They matter for why you are who you are and why somebody else is who they are. They matter for positive outcomes, negative outcomes, and for neutral outcomes. In fact, there are so few studies showing a lack of genetic effects that it is safe to assume that every human phenotype is under some level of genetic influence (Polderman et al., 2015). The genetic basis to human phenotypes is so well-established and so strongly supported by empirical research that the first law of behavioral genetics is that "All human behavioral traits are heritable" (Turkheimer, 2000, p. 160). We would be hard-pressed to identify any other finding in all of the social sciences that has been replicated so many times that it could be codified into a law.

Given how much research has been produced examining the genetic underpinnings to behavior, it seems reasonable to conclude that most behavioral and social science research would, in some capacity, integrate genetics findings. But this does not appear to be the case in most social science disciplines. Part of the reason for the omission may be due to the fact that most social scientists have little background in the methods and statistics that are used in quantitative and behavioral genetic research. This is problematic for at least three key reasons. First, without such background knowledge, scholars are in a position where they are unable to be informed consumers and critics of genetics research.

Second, given the backdrop of criticisms leveled against research that fails to take into account genetic influences (see, for example, Barnes et al., 2014a), many social scientists are placed in the precarious situation wherein their research findings may be confounded (at least partially), but they are ill-equipped to address the methodological critiques. In order to address the critiques effectively, scholars must be able to implement the statistical methods employed by quantitative genetic statistics.

Third, as classically trained social scientists, we understand there is a relative dearth of resources accessible to folks who lack a background in genetics. Certainly there are books that are dedicated to the key findings, tenets, and theories of quantitative and behavioral genetics research (e.g., Beaver, 2017; Knopik et al., 2017), just as there are published articles that employ quantitative genetic designs in social science journals. But these works do not provide information that walks the beginner—or, perhaps, the casual observer—through the types of statistics that are widely used by quantitative and behavioral geneticists. As a result, scholars may be forced to pursue other research interests.

Taken together, the most pressing obstacle to those interested in reading, understanding, and conducting quantitative and behavioral genetic research is learning the methods and statistics that are used in this body of literature. Against this backdrop, the overarching purpose of this book is to expose interested students and researchers to the methods, statistics, and modeling approaches used in quantitative and behavioral genetics. Before jumping into these topics, we will first provide some background information that will help couch the rest of the book within the larger literatures to which we will speak.

1.1 What are Genes and Why Do They Matter?

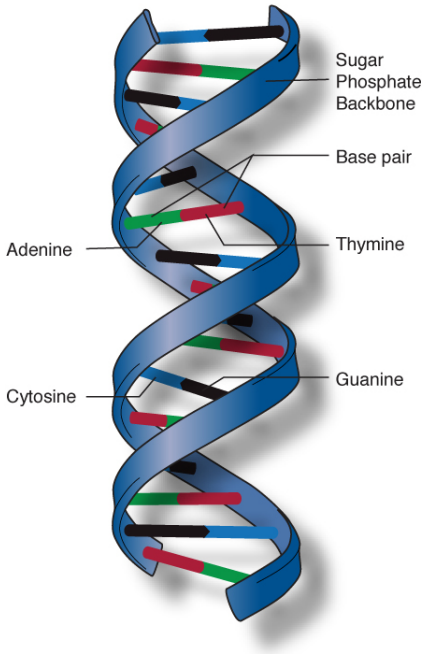
Although this book is largely about quantitative genetics, it is essential to have at least a cursory understanding of the basic mechanisms that govern molecular processes and, therefore, how deoxyribonucleic acid (DNA) can affect behavioral outcomes. More detailed discussions of DNA and its linkage to behaviors is available elsewhere (e.g., Beaver, 2017; Carey, 2003; Flint, Greenspan, & Kendler, 2010) and we encourage interested readers to consult those pieces of scholarship closely. What we will present here is only an introduction to some of the rudimentary points about DNA and more detailed information will be presented throughout the rest of this book as is necessary.

DNA is a chemical code that is the basic building block of all living organisms. Located in the nucleus of every cell (except for red blood cells), DNA provides the instructions needed for organisms to form, develop, and live. The information that is contained within DNA is responsible for creating many of the observable and unobservable differences among humans, including differences in physical characteristics, health outcomes, personality traits, and behaviors. DNA is passed from parents-to-children and thus part of the reason that biological relatives resemble each other on phenotypes (a phenotype is any measurable trait) is because they share some of the genetic material that is responsible for creating these phenotypes.

Figure 1.1 displays the well-known double-helix structure of DNA, wherein two genetic fibers—known as polynucleotides—are twisted around each other in the form of a spiral staircase. Along the inside of each polynucleotide protrudes nucleotides (also referred to as bases). There are four different nucleotides in the DNA alphabet and each nucleotide is referred to by its first letter, such that adenine $\rightarrow$ A, thymine $\rightarrow$ T, guanine $\rightarrow$ G,

and cytosine → C. The two polynucleotides are held together by the nucleotides on one polynucleotide bonding together with nucleotides on the other polynucleotide. The bonding of base pairs does not occur randomly, but rather follows specific base pair bonding rules, wherein A can only bond with T (and vice versa) and G can only bond with C (and vice versa).

Figure 1.1: The Double-Helix Structure of DNA



Suppose a series of nucleotides on one polynucleotide was arranged as follows:

ACCGGAATTACCTATAC

Using what is known about the bonding of base pairs, we would know that the arrangement of base pair sequences on the other polynucleotide would be as follows:

TGGCCTTAATGGATATG

Just keep in mind that it is always easy to figure out the base pairs of one polynucleotide when the sequence of the other polynucleotide is known; all that has to be done is to substitute an A for a T (and vice versa) and a G for a C (and vice versa). As a result, it becomes somewhat redundant to list out the base pair sequences for both polynucleotides because once the DNA sequences for one polynucleotide are presented the other is known by default.

On certain stretches of DNA, contiguous base pairs work together. These stretches of DNA are genes and there are about 20,000-23,000 genes found in the human genome. On average,

1,000 or more base pairs are included in a single gene. But, exactly what do genes do? This is where there is a lot of confusion and misperceptions regarding genes. Although genes are often portrayed as having magical and omnipotent effects, genes are "only" responsible for coding for the production of proteins. Proteins are organic compounds and there are two key types of proteins: structural proteins and functional proteins. Structural proteins are responsible for providing structure and form to the human body. Fingernails, hair, and tendons, for instance, are comprised of structural proteins. Functional proteins, in contrast, are involved in various processes in humans, such as in the process of neurotransmission. Functional proteins, such as enzymes, are thought to be key in the DNA-behavior nexus.

Although genes are responsible for the production of proteins, it is important to realize that genes do not directly create proteins. Genes contain the instructions needed for proteins to be manufactured. The instructions contained in genes is ultimately converted into a protein by a process known as the central dogma of biology. There are two steps in the central dogma of biology: transcription and translation. It is not necessary to understand the central dogma of biology for this book, but briefly, in the process of transcription DNA is transformed into ribonucleic acid (RNA). With translation, RNA (technically mRNA) works in conjunction with ribosomes to produce amino acids. As amino acids are produced, they are linked to together in a growing protein chain which is referred to as a polypeptide chain. After the polypeptide chain is completed it is known as a protein. Once the protein has been manufactured, it performs its specialized duty.

What is important to realize is that different proteins can have different effects on the human body and brain. Some proteins may be more efficient at the transmission of information in the brain, some proteins may buffer against the effects of stress, and still other proteins may result in chemical imbalances in the brain. It is the effects that these proteins have that ultimately are responsible for producing behavioral outcomes. Moreover, depending on the precise DNA sequences that you inherited, your body might code for the production of a protein that is different from somebody else who inherited different DNA sequences. The end result could be that two persons are differentially predisposed to certain behaviors because they have unique genotypes that produce different proteins. These differences in the proteins might bring about emotional and behavioral outcomes.

What is important to keep in mind for now is that genes do not have a one-to-one correspondence with behavioral outcomes. In other words, genes are neither necessary nor sufficient conditions for a particular behavior to emerge; all that genes can do is increase or decrease the probability that a behavior will ultimately emerge. Also, bear in mind that the probabilities attached to different genes can vary, such that one genetic variant might have a larger influence on compared to another one. Overall, though, the consensus is that single genes will have relatively small effects on behavioral phenotypes, typically accounting for less than 1% of the variance. What this suggests, therefore, is that there are likely hundreds of genes that are involved in the etiology of human complex traits. To date, the empirical evidence supports such a claim, with a wide array of genes being linked to various forms of human activity, but with most of these genes having relatively small effects (Chabris et al., 2014). When aggregated, though, the small effects of these genes can account for a large

proportion of behavioral variance. We will go into more detail on these points in Chapters 3 and 4.

1.2 A Fundamental Question: Are Genes Causal Agents or Attributes?

Before moving any further, it is important that we pause and consider a critical question. The question is whether causal inferences can be drawn from candidate gene research. Put differently, is it appropriate to assign a causal effect to a genetic variant? Imagine a study finds that persons who carry the *w1* allele of gene W have a higher probability of developing medical condition M compared to those who carry the *w2* allele.[1] In this case, would it be appropriate to say "gene W causes the medical condition M"? Colloquially, it seems that this phrase is appropriate and if you were to say this to a random passerby, s/he is likely to understand what you mean. But colloquial usages of many scientific terms do not always line up perfectly with their scientific meaning. It is our impression that the word *cause*—and, of course, all of its derivates such as *causality*—represents an important case in point. Scholars and philosophers have spent nearly 300 years debating the meaning of the word *cause* and, more importantly, under which conditions one can reasonably claim to have found a cause (Hume, 1739; Pearl, 2009; Rothman and Greenland, 2005; VanderWeele, 2015). In a very general sense, social scientists have "settled" on a broad definition of the word *cause* to capture the *influence* of one thing on another (see generally, Pearl, 2009).

This general definition, of course, leads to an all important follow-up question. A question that has been the source of anxiety for scholars worldwide since the time of Hume (1700s). That question, of course, is "how can we identify a cause?" Whether explicitly stated or not, the over-arching goal of most social science research is to identify causal relationships. And it turns out that one of the most popular approaches to thinking through this issue is the counterfactual/potential outcomes framework (Holland, 1986; Rubin, 1974; VanderWeele, 2015). Although this text is not intended to provide an introduction to causal inference or counterfactual/potential outcomes reasoning, a brief discussion is warranted because it may help to illuminate important issues that should be kept in mind whenever one embarks on a candidate gene study. Readers who desire a more thorough treatment of causal inference and/or the counterfactual/potential outcomes framework are encourage to consult the excellent overview by Holland (1986), the accessible and thorough text by VanderWeele (2015), or the expansive work of Imbens and Rubin (2015).

In essence, the counterfactual/potential outcomes framework forces one to think of the relationship between a cause S and its effect Y. S represents a treatment or intervention

[1] This reveals an important point that is necessary to define the cause of one thing on another. Specifically, causal agents must have at least one reference group. Put a different way, it makes no sense to say that X causes Y if the former only has one condition. This would be akin to trying to predict a variable with a constant.

that is delivered ($S = t$) or not ($S = c$) and Y_t reveals the potential outcome under t and Y_c is the potential outcome under c. The most straightforward manner to observing a causal effect would be to observe Y_t and *simultaneously* observe Y_c; which is to say that the most direct route to specifying a causal effect of S on Y is to observe both potential outcomes and calculate the difference between the two:

$$\tau = Y_t - Y_c$$

Notice that there is no subscript identifying the participants who contribute information to the Ys. This is intended to reflect the fact that there is only *one* participant in our study. In other words, we only need to have a single participant in a study if we can *simultaneously* observe him/her under two different states: when the treatment is applied (i.e., $S = t$) and when there is no treatment (i.e., $S = c$).

It takes no great insight, though, to see the big shortcoming; what Holland (1986) referred to as "the fundamental problem of causal inference." Specifically, we cannot generate a value for τ because it is impossible to observe a single participant in two *different* states simultaneously. Outside of a Nobel Prize worthy discovery of a wormhole connecting us to the multiverse (see, generally, Greene, XXXX), we only have access to one state of affairs. Put differently, we can only observe person i in the condition t or in the condition c. This is not to say that person i cannot occupy condition t at one time and condition c at a later time. Indeed, it is entirely possible to dream up an experiment (or a natural observation) where a person transitions from one state to another. This, however, is *not* sufficient to satisfy the counterfactual/potential outcomes requirement of simultaneity. Thus is the reason longitudinal data cannot—in and of itself—satisfy the criteria of causality that are laid out below.

Now, given that we have established the physical impossibility of *simultaneously* observing the phenotype under two states (Y_t and Y_c), the question becomes: "is it even possible to make causal inferences?" The short answer is "yes." And the logic is relatively straightforward. We cannot directly calculate τ, but we can *estimate* it (i.e., $\hat{\tau}$) by conceptualizing the unobserved counterfactual outcome (i.e., the state that we do not observe) as a missing data problem. Thought of in this way, the unobserved counterfactual is now something that can be imputed. The most elegant/robust solution to this missing data problem is to conduct a randomized controlled trial (RCT) with two groups of participants.

Imagine you have $n = 100$ participants and they are randomly assigned to two conditions t and c with equal probability, such that $\mathbb{P}(S = t) = 0.50$ and $\mathbb{P}(S = c) = 0.50$. Because the treatment exposure S is randomly assigned, it can be considered an ignorable condition, meaning it is independent/exogenous of the potential outcomes Y_t and Y_c. This allows one to impute the "missing" information (i.e., the unobserved potential outcome) by relying on expected values estimated at the group-level. Let $\mathbb{E}(Y|S = t)$ represent the expected value of the phenotype for participants who receive the treatment and $\mathbb{E}(Y|S = c)$ represent the expected value of the phenotype for participants in the control condition (i.e., those who did

not receive the treatment). Under the assumptions spelled out in Holland (1986), then:

$$\hat{\tau} = \mathbb{E}(Y|S = t) - \mathbb{E}(Y|S = c)$$

Thus, we can *estimate* the causal effect of S on Y if we randomly assign participants to receive different conditions of S and calculate the difference between the expected values of Y observed under those conditions.

With these points in mind, let us now consider the *conceptual* meaning of a counterfactual/potential outcome. The counterfactual/potential outcome must be defined for each condition of S that is applicable. Thus, Y_t represents the value we would observe on Y if person i were given—perhaps contrary to fact—the treatment t. Y_c reveals the value we would observe on Y if—perhaps contrary to fact—person i were exposed to (non)treatment c. The implied assumption underlying this argument is that S represents a manipulable and dynamic factor; something that could, at least in theory, be targeted for intervention. If S represents something that is *not* manipulable, then it is said to be an *attribute* rather than a *cause*.

The distinction is more than semantic because Holland (1986: 954) argues that, "...causes are only those things that could, in principle, be treatments in experiments. The qualification 'in principle' is important because practical, ethical, and other considerations might make some experiments infeasible, that is, limit us to contemplating *hypothetical experiments*" (emphasis in original). Based on this passage, it seems as if Holland might be inclined to argue that a gene cannot act as a causal agent because it is not something that could—ethically, at least—be utilized as a treatment in an experiment. Although, the second sentenced quoted above appears to acquiesce, thereby allowing one room to argue that scholars can—even if only hypothetically—conceive of an experiment where a gene *is* a treatment variable. Of course, utilizing genes as treatment variables has a long tradition in animal research, especially in the drosophila (Falconer and Mackay, 1989).

Holland's (1986) argument is critically important to the quantitative genetics enterprise because it reveals the distinction between causal agents and attributes. Up to this point, we have *assumed* genes are causal agents. But, if they are considered attributes, then it, "...cannot be a cause in an experiment, because the notion of *potential exposability* does not apply to it" (Holland, 1986: 955; emphasis in original). In other words, attributes can only have *associations* with phenotypes. Thus, attributes carry no meaningful policy and/or real-world implications. In other words, the study of a link between an attribute and a phenotype provides little more than descriptive evidence.

It seems clear, though, that scholars across the globe consider genes causal agents. If not, then it becomes nearly impossible to justify the money and resources that have been applied to the study of human genetics over the past 50-60 years. With this in mind, it is our position that genes *do* act as causal agents and that they *can* be thought of in a counterfactual/potential outcomes framework. Although obvious ethical and technological

(though the latter is less important than the former) limitations exempt scholars from using a human gene as a treatment variable. It is, nonetheless, a hypothetical possibility. This would seem to satisfy Holland's (1986) "in principle" clause. Moreover, the very fact that neuro-genetics research has started to unpack the biological pathways that connect genetic variants to neurobiological substrates supports the argument that genes act as causal agents. As was reviewed in chapter 4, for instance, we now know that certain genetic loci code for neurtransmitter production. Differential levels of neurotransmitters have been linked to differential brain functioning, which has been linked to variance in behavioral outcomes. In other words, molecular genetics research *may* help to reveal more than associations. The identification of a causal pathway(s) between a gene(s) and a human behavioral outcome(s) is possible, meaning we can speak to the causal impact of a gene on a phenotype if the appropriate conditions are met.

The conditions for speaking to causality, of course, require the researcher to establish: 1) a relationship between the gene and the phenotype; 2) the direction of influence (i.e., that the gene affects the phenotype and not the reverse); and 3) that there are no alternative explanations for the observed association (Shadish, Cook, and Campbell, 2002). The first two criteria are rarely a problem; indeed, the demonstrations that follow will cover some of the ways that scholars can establish the first criterion of an association. The second criterion is a natural condition of the relationship between a gene and a phenotype. It cannot be that the phenotype has caused the genotype, thus the influence can only flow in one direction (but see VanderWeele et al. [2014] for a discussion of the various problems that might arise due to uncontrolled endogeneity).

The third criterion is much more difficult to satisfy. For a host of reasons, scholars have found it nearly impossible to rule out all the possible alternative explanations for why a gene may be associated with any given phenotype. In other words, it is not very difficult to come up with a host of reasons why a certain genotype may have a higher frequency in one population compared to another. In which case, any phenotype that also differs between those two populations is likely to show a spurious association between the gene and the phenotype. This concern is known as population stratification, and is covered in more detail in chapters ?? and ??.

For now, suffice it to say that genetics and genomics research do have the possibility of identifying causal relationships between genes and human behavior. This is not to say that any one study has been able to do so. But the possibility is there. And as we have learned from decades of behavioral economics research, humans have a tendency to overemphasize the probability that a rare outcome might actually occur (see generally, Kahneman, 2011). We raise this final point as a way to inspire caution in our readers. The studies discussed in this text have the possibility of identifying causal relationships. The chances of doing so with any single study are slim. But that does not mean such findings are useless. Just that they should be interpreted with appropriate levels of caution.

1.3 Overview of Relevant Findings

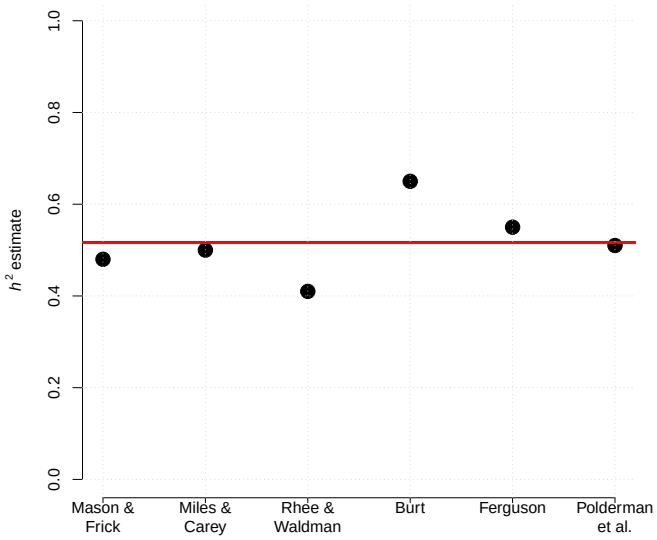
Findings from molecular genetic association studies have shown that specific genetic variants are associated with the propensity to engage in just about every human outcome that has been studied (Polderman et al., 2015). We have contributed some research to the study of genetic and environmental influences on antisocial behaviors, so we will briefly review that literature here. Note, however, that this book is not specifically directed at research on human antisocial behaviors. Rather, we intend this book to be widely applicable to most areas of behavioral and social science research. We simply chose to review the work on antisocial behavior here because that is the area we are most familiar with.

Much of the quantitative and behavioral genetics research on antisocial behavior seeks to estimate the proportion of phenotypic variance that is attributable to genetic and environmental influences. A detailed discussion of how this is accomplished will be provided in Chapters 3 and 4. For now, all that is essential to understand is that there are three variance components that are typically estimated in these studies: a heritability component (symbolized as h^2), a shared environmental component (c^2), and a nonshared environmental component (e^2). Heritability represents the proportion of phenotypic variance that is accounted for by genetic differences. The amount of variance that is not explained by genetic influences must be explained by environmental influences (and error) and biosocial research distinguishes between what is known as shared environments and nonshared environments. Shared environments are environmental factors that account for similarities between siblings whereas nonshared environments are environmental influences that produce differences between siblings. These three components account for 100% of the variance in any phenotype.

The various research designs that are capable of estimating these variance components, along with a detailed discussion of the mathematical foundations to them will be discussed in much greater detail later in this book. For now, we wish to focus on the findings that have emerged from studies that estimate the variance components. There have been hundreds of studies published that have estimated the genetic and environmental influences on a wide range of antisocial behaviors. The results generated from these studies have been relatively consistent despite the fact that they use different measurement strategies, analyze unique samples, and focus on different variations of antisocial phenotypes. Six meta-analyses on the heritability of antisocial behaviors have been published that summarize the findings from these studies (Burt, 2009; Ferguson, 2010; Mason & Frick, 1994; Miles & Carey, 1997; Polderman et al., 2015; Rhee & Waldman, 2002). The results from these studies are presented graphically in Figure 1.2. Overall, these meta-analyses have revealed that approximately 50% (the mean estimate from the meta-analyses is demarcated as the red horizontal line in the figure) of the variance in antisocial phenotypes is the result of genetic influences.

This heritability estimate is, of course, just an average and the value can increase or decrease depending on sample-specific characteristics. For instance, heritability estimates for antisocial behaviors tend to be highest in childhood (e.g., conduct disorder) and adulthood (e.g., criminal arrests in adulthood) and lower in adolescence (e.g., experimenting with alco-

Figure 1.2: h^2 Estimates from Meta-analyses



hol). Genetic effects also tend to be the strongest for the most serious and chronic types of antisocial behaviors and the lowest for the types of antisocial behaviors that are relatively minor and transient. Research findings have revealed, for example, that the heritability of career criminality and life-course-persistent offending is approximately 0.70 (Barnes, Beaver, & Boutwell, 2011) while the heritability of less severe offenses committed by adolescents is comparably lower. The short of it is that heritability estimates for antisocial behavior hover around 50% most of the time, but they can be higher or lower depending on the exact behavior that is being examined and the characteristics of the sample.

So a good starting point is that the heritability of antisocial behavior is 0.50, but even that can be a bit misleading. It permeates a belief that since DNA structure does not change—that is, the DNA sequences that you are born with are the ones that you will die with—genetic effects cannot change. This logic, however, is incorrect. Heritability estimates can fluctuate in response to different environmental conditions. What this means is that environmental and genetic influences can—and often do—work synergistically to produce variance in antisocial behaviors. To illustrate, genetic effects might be more pronounced for persons living in a disadvantaged and criminogenic environment compared to persons living in a privileged environment. There has been a great deal of research devoted to examining the behavioral phenotypes that are produced by the interaction of genes and the environment during the past decade. While there has emerged a great deal of support in favor of these types of gene-environment interactions, the mechanisms underlying how and why these types of interactions occur have been hotly contested in recent years (Belsky & Pluess, 2009; Moffitt & Beckley, 2015). We will revisit this issue and discuss it in much greater detail in Chapter 9.

To assume that only behaviors can be genetically influenced misses the mark; variance in environments can also be influenced by genetic factors in what has become known as a gene-environment correlation (rGE). Precisely how this is accomplished and the mathematical models available to estimate rGEs will be covered extensively in Chapter 9. The basic logic to testing for rGEs is to estimate the proportion of variance in an environmental measure that is accounted for by genetic influences. Studies have been conducted to estimate the genetic influence on most of the environments that are of interest to criminologists, including affiliating with delinquent peers, different parenting styles, and exposure to various types of stresses and strains. A large review of the literature has shown that most environments are accounted for by genetic effects, with the genetic effect being around 0.25 (Kendler & Baker, 2007). As a result, the heritability of environments, while still significant and pervasive, is not as strong as is it is on antisocial behaviors.

Why are these findings important? We offer two responses to this question. First, and perhaps the most obvious, is that they provide the rationale for writing this book. If there was no evidence of a genetic influence on human behaviors (we focused on antisocial behaviors here, but this pattern of findings is consistent with most human behaviors [see Polderman et al., 2015]), then this book would not be necessary. Our decision to write this book is multifaceted, but is anchored by the research findings showing beyond a shadow of a doubt that most human behaviors are influenced, in large part, by genetic factors.

Second, and perhaps most importantly, since most human outcomes have a genetic component, it is possible that the correlational findings produced by designs that do not rule out genetic influences will be biased. In other words, if genetic influences are widespread, then any study showing an association between behavior X and behavior Y runs a risk of being confounded if genetic influences that are shared between X and Y are not ruled out by design or with appropriate controls. Given that this is so central to the book and to the need to employ some type of quantitative genetic methodology, we devote the next section of this chapter to this important topic.

1.4 The Limits of Standard Social Science Methodologies (SSSMs)

Findings from empirical research have the capacity to shape how the future research gets done, directly impacting the methodological decisions that get made by the next wave of research. In some areas of study this is not a major concern because the choice of methodology has not obvious impact on the findings. But in other areas, the choice of methodology has the potential to greatly impact the substantive findings. As discussed above, one of the most consistent findings to emerge from the behavioral genetics literature is that about 50% of the variance in human behaviors is due to genetic effects. Yet, a large portion of the studies in many behavioral and social science disciplines (e.g., sociology, psychology, and criminology) ignore genetic influences by applying what has come to be known as the standard social

science methodology (SSSM).

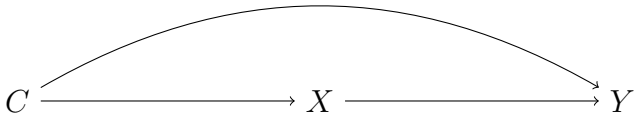
With SSSMs, there is a focal child/adolescent that is ultimately included in the sample; all other family members, including siblings, are not included in the study and thus no information is collected about them. The focal subject then is administered a battery of tests or questionnaires designed to measure whatever is the main topic of the study. For instance, if a researcher was interested in examining the association between parenting and the child's self-control, they would typically collect information from one adolescent per household and ask them to self-report on items related to the way their parents treat them and on items that are designed to measure individual variation in self-control. Statistical analyses would then be conducted to determine whether there is an association between parenting and self-control; if there is, then the assumption is that that same association would generalize to all family members within that household. The key point to remember is that with SSSMs, only one child per household is included in the sample.

Many social science studies are conducted using secondary data (i.e., data that has already been collected and is available for analysis), and there are a few datasets that have come to be quite popular, such as the Add Health and the NLSY. SSSMs are often used when analyzing these data, but doing so is not strictly necessary. Indeed, many of popular datasets like the Add Health and the NLSY include more than one sibling per family in the data, allowing one to carry out the types of analysis that will be discussed in chapters 4, 5, and 10.

So why is it even important that SSSMs are typified by including only one child per household in the sample? As will be discussed in great detail throughout the remainder of this book, it is required that at least two biological relatives (usually siblings) are included in the sample in order to estimate genetic influences or to control for them effectively. Without at least two biological relatives in the data, it is generally impossible to fully account for genetic effects in a straightforward and direct way. What this necessarily means is that any study that you write or that you read that only includes one child per household—that is, uses an SSSM—is unable to account for genetic influences. Some alternative strategies have been advanced, including longitudinal data that employ within-person designs, but these rest on strong assumptions (meaning they are unlikely to be met) regarding genetic effects. More specifically, the belief that within-person designs can control for all genetic effects is based on the premise that genetic influences (on the phenotype of interest) remain stable across the time frame that the data include. What research has shown consistently, however, is that genetic influences wax and wane across all different regions of the life course. Genes at one period of development might be trigged on whereas those same genes at a different developmental stage might be switched off. Layer in the complexity of gene-environment interactions and correlations, wherein these genes that are on might have different effects on behavioral phenotypes depending on exposure to environmental influences and it is easy to see that a within-person design does not live up to the analytic standard needed to fully examine genetic influences on antisocial phenotypes.

Perhaps the most pressing issue, though, is whether there are any significant consequences

that occur as a direct result of ignoring genetic effects by employing an SSSM. In short, does it even matter that a researcher might fail to employ genetically sensitive research designs? The answer to this question is most likely "yes." Failure to control for genetic influences effectively has the capacity to bias findings and produce entire bodies of knowledge that are misleading. To see how this could be the case, see the diagram below. In this diagram, you will see that there are three variables, labeled X, Y, and C. In much research, there is an interest in whether two variables, say X and Y, are associated even after partialling out the effects of confounder variables, such as C. If X and Y remain significantly related after controlling for all potential confounders (Cs), then that is often viewed as strong evidence that X and Y are related in a very defensible way. Of course, it is never possible to rule out all possible sources of confounding and thus the association between X and Y is rarely inferred to be causal. The best that most social science research can accomplish is to control for all *known* confounders and see whether the $X \rightarrow Y$ association remains significant.



The problem is that it is not always possible to know all of the potential confounders that should be included in C, which is why there is such a heavy reliance on the findings from empirical research to guide the inclusion of control variables. A primary point we would like you to consider as you read this book is whether genetic factors are a potential confounder for any given relationship of interest in your area of research. This book is designed to introduce to you everything you need to know in order to implement and apply designs that can account for such genetic confounding in your own work.

1.5 Conclusions & Aims of the Book

Behavioral and social science research has always been heavily quantitative, and this remains true for contemporary work. The vast majority of all studies published in leading social science journals is often quantitative in nature, with the types of statistical techniques being employed becoming more varied and more complex each year. This focus on quantitative methods has led to doctoral programs offering more classes devoted to advanced statistics, to more researchers attending national workshops devoted to learning the most cutting-edge statistical approaches, and to specialized courses on statistics being offered at the annual meetings of disciplinary organizations.

Anytime a new statistical approach is advanced, social scientists tend to clamor to available resources to learn about it so that they can ultimately use it in their own research. In the 1980s there was a focus on structural equation modeling, in the 1990s the attention shifted to multilevel modeling, in the 2000s the focus was placed on longitudinal growth models,

latent class analyses, and propensity score matching (PSM). For each statistic, scholars took the time to learn the technique and how it applies to their own research. Many times they were even tasked with the burden of locating new datasets that would allow them to use the statistic in an appropriate way. Despite the obstacles of learning a new technique, becoming familiar with software programs that could estimate this technique, and locating suitable data for it, scholars have not been discouraged and have risen to the challenge. The consensus would likely be that these techniques have opened up new avenues for research, they have helped to refine theories, and they have, in general, made statistical inferences stronger.

The history lesson in all of this is that social scientists tend to embrace statistical approaches that empower them to have more faith in their findings. We believe this is precisely what can be accomplished for those who learn, understand, and apply quantitative genetic techniques to their own research agenda. With this book, we intend to equip readers with a foundation of knowledge that can be used to understand quantitative genetics as it relates to their own area of research. In doing so, we hope to accomplish the following:

- Aim 1: Provide behavioral and social scientists with the background in Mendelian and quantitative genetics necessary to understand and conduct quantitative genetics research
- Aim 2: Review findings flowing from quantitative genetics research that applies broadly to human outcomes
- Aim 3: Discuss innovative strategies for estimating genetic influences on human outcomes
- Aim 4: Discuss innovative strategies for controlling genetic influences when one wants to isolate environmental influences on a human outcome(s)
- Aim 5: Outline and demonstrate the statistical tools used to conduct quantitative genetics research

For those readers who are interested and put forth the effort, this book will provide everything one needs to begin employing quantitative genetic methodologies in their own work. We are certain that those researchers will find their research is more defensible, is more accurate, and is perhaps better able to explain why humans develop and behave the way we do.

Chapter 2

Essential Statistics: Means, Variances, & Covariances

Three statistics form the backbone of quantitative genetics research. They are the mean, the variance, and the covariance. These three simple statistics—indeed, they can easily be computed with a hand calculator if the dataset is small enough—hold the key to nearly everything that will follow in this text. If you truly understand these three statistics, meaning you have a firm conceptual and computational grasp on them, then there is no reason you cannot also understand the more advanced and seemingly complicated analyses that appear in later chapters (e.g., the biometrical ACE model). Given this backdrop, then, it should come as no surprise that our introduction to quantitative genetics begins with an in-depth consideration of the most basic of all statistics, the mean.

2.1 The Mean

2.1.1 Computation

The mean, while being one of the most basic and easy-to-compute values, is utilized more often than perhaps any other statistic. This is not to say that quantitative genetics begins and ends with the mean. To be sure, researchers often employ a wide range of statistical tools when analyzing their data. What is important to keep in mind, though, is that the mean lies at the heart of nearly every parametric analysis (this is meant to distinguish the typical statistical analysis from its non-parametric counterparts that are discussed in later sections of this text). As you will see, the mean plays a central role in the calculation of variance and in the calculation of covariance. Since nearly everything that follows will build on the variance and the covariance, you should first be sure and gain a firm handle on the mean.

Typically, the mean is expressed as

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where $\bar{X}$—pronounced "x bar"—is a standard symbol used to represent the calculation that follows on the right-hand side of the equal sign; $\sum_{i=1}^{n}$ is the upper-case Greek letter sigma, and it is used when one wants to add up or sum a set of items starting with the first case (i.e., where $i = 1$) and continuing through the last case (i.e., where $i = n$ and n is the total number of cases in the dataset); X_i represents the items to be summed, indexed by the subscript i, which indicates that all items in the list should be summed one at a time; and n represents the total number of items that were summed together. An example will help to clarify.

Imagine you have conducted an experiment. It does not have to be the type of experiment that involves lab space, white coats, and sterile test tubes. On the contrary, an experiment can be as simple as flipping a coin 10 times or changing the running time on your in-ground sprinkler system. Nonetheless, let us imagine that you have collected data from a recent experiment and that the data points can be expressed as:

$$X = \{2, 5, 3, 8, 5\}$$

Several important points flow from this simple dataset. First, notice that the data are labeled "X", which is—with one exception—consistent with the numerator for the equation for the mean. The exception is that the numerator in the equation for the mean includes the subscript i and the symbol used to represent the dataset does not. This allows us to differentiate between X the dataset and X_i a specific value in that dataset. For instance, we could identify the value for the second case in the dataset as X_2, which is 5 (i.e., $X_2 = 5$).

Carrying out the calculation of the mean is straightforward if we break it into two parts: the numerator and the denominator. Let us begin with the numerator, which asks us to sum all the cases in the dataset:

$$\begin{aligned}\sum_{i=1}^{n} X_i &= 2 + 5 + 3 + 8 + 5 \\ &= 23\end{aligned}$$

When calculating the mean, the denominator simply represents the total number of cases in the dataset (i.e., the total number of cases that were summed in the numerator):

$$n = 5$$

Putting the two together results in:

$$\begin{aligned}\bar{X} &= \frac{23}{5} \\ &= 4.60\end{aligned}$$

Thus, the mean for the dataset X is 4.60.

2.1.2 Visualization

Computing the mean, as you have just seen, is straightforward and easy. In fact, most learn to do this very early in school. What we have found, however, is that students (and even researchers) understand how to calculate the mean and they even understand that it is a central component of modern statistics, but sometimes it is evident that they lack a firm grasp of the concept. In other words, it is not enough to be able to solve the equation. It is arguably more important that you understand, conceptually, what the mean is, what is tells us, and how it can be used in more advanced applications.

With this task in mind, it may help to visualize the mean. In order to do so, we will need to take a step backward and introduce a standard graphical technique known as a histogram (we invite readers who require a more formal introduction to data visualization to see Agresti & Franklin, 2013). Briefly, a histogram utilizes bars to represent the frequency of observations in the data. The taller the bar, the more often a certain value (or range of values, which is often the case because the histogram will "bin" values that are near one another into a single bar) is observed in the dataset. For example, let's say you draw 100 random numbers. They might look like this:[1]

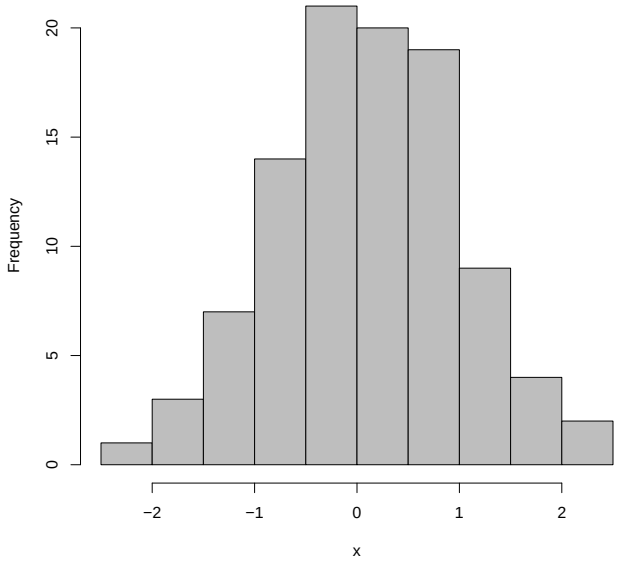
-0.63	1.51	0.92	1.36	-0.16	0.40	2.40	0.48	-0.57	-0.54
0.18	0.39	0.78	-0.10	-0.25	-0.61	-0.04	-0.71	-0.14	1.21
-0.84	-0.62	0.07	0.39	0.70	0.34	0.69	0.61	1.18	1.16
1.60	-2.21	-1.99	-0.05	0.56	-1.13	0.03	-0.93	-1.52	0.70
0.33	1.12	0.62	-1.38	-0.69	1.43	-0.74	-1.25	0.59	1.59
-0.82	-0.04	-0.06	-0.41	-0.71	1.98	0.19	0.29	0.33	0.56
0.49	-0.02	-0.16	-0.39	0.36	-0.37	-1.80	-0.44	1.06	-1.28
0.74	0.94	-1.47	-0.06	0.77	-1.04	1.47	0.00	-0.30	-0.57
0.58	0.82	-0.48	1.10	-0.11	0.57	0.15	0.07	0.37	-1.22
-0.31	0.59	0.42	0.76	0.88	-0.14	2.17	-0.59	0.27	-0.47

If we were to group closely related numbers together (i.e., "bin" them) and then plot the distribution (i.e., all of the observed data points) according to the frequency with which they

[1] the random numbers were generated in `R` with the following code: `set.seed(1); d<-matrix(rnorm(100),ncol=10)`

are observed, we would end up with something that looks like the histogram displayed in Figure 2.1.

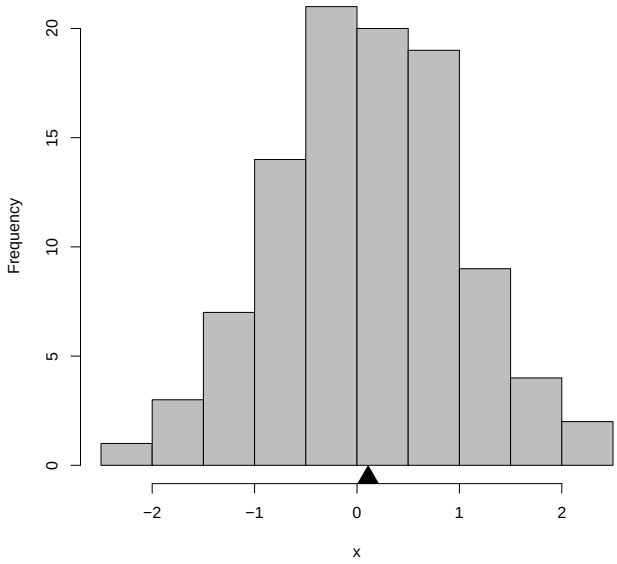
Figure 2.1: Histogram



Looking at Figure 2.1, imagine you were tasked with finding the balance point for the distribution of scores observed in the dataset x. It might help if you think of the histogram as a teeter-totter (a very oddly shaped one!). Now, imagine you have a triangular fulcrum that will be placed under the histogram so that the distribution displayed in Figure 2.1 is perfectly balanced. Where would you place it? If this were a real-world exercise you would probably try a few different positions until you got the balancing act just right. But, as it turns out, we can always find the perfect balance point for a distribution of data points. Just use the mean! That's right, the mean of a distribution of scores will locate the balancing point of the distribution so that an equal amount of density lies to the right and the left (properties required to balance any physical object). Figure 2.2 shows this relationship by adding a fulcrum precisely at the location of the mean.

While it may be somewhat difficult to visualize at first, imagine the distribution displayed in Figure 2.2 were a paperweight or any other dense object that had a finite amount of mass. (Such a paperweight might even be a good idea as a gift for your favorite quantitative geneticist!) If you were to hold the paperweight in your hand and attempt to balance it on your index finger, the balance point would be found precisely where the triangle is located; the mean.

Figure 2.2: Histogram with Fulcrum Representing the Mean



2.1.3 About Notation

Throughout this text you will notice short sub-sections like this one. While some will be longer than others, we will use the Quick Note sections as a way to direct your attention to important, albeit ancillary issues. This is the first of such sub-sections and, in this regard, it is a good example of the type of information you can expect from the Quick Notes.

Those who are learning statistics for the first (or the tenth!) time often find the symbols and naming conventions confusing. Throughout this text, we have adopted a consistent strategy for symbols and naming whenever possible. Yet, it is important to recognize that other texts may use symbols and names that are not presented here. We have attempted to identify the most commonly used alternatives whenever possible and we will often point these alternatives out to you in a short sub-section.

The first source of confusion in statistics often centers on the two most common symbols used to represent the mean: μ ("mu", which is pronounced like "mew") and the "bar" which was placed above X in the preceding discussion (i.e., $\bar{X}$). Typically, μ is reserved for cases when population data are being analyzed. Population data, in a general sense, represent everyone in the population of interest. Alternatively, sample data represent information drawn from a subset of the population. Whenever sample data are analyzed, $\bar{X}$ is the preferred symbol to represent the mean.

Only rarely—due almost exclusively to limited availability—are population data analyzed. Instead, most researchers interested in carrying out a quantitative genetic analyses will be

forced to rely on sample data. As a result, the symbols for sample statistics will be utilized throughout this text—as $\bar{X}$ was used in the preceding discussion. We will, however, point out the different notational conventions where appropriate.

2.2 Variance

2.2.1 Computation

Although the mean is useful and substantively important, in many contexts a researcher will be interested in the *differences* that are observed in the data. These differences are typically calculated and expressed as a statistic known as variance. Decomposing variance into genetic and environmental components forms the backbone of quantitative genetic analysis (Falconer & Mackay, 1989), so it is critical that we introduce the variance statistic in a way that will connect with the chapters that follow. In a general sense, variance is a statistic that numerically expresses the spread of scores around the mean. If there is no spread around the mean—think of a dataset where every case scores the same value—then the variance is zero. As individual cases begin to diverge from the mean (i.e., as differences emerge), the variance increases above zero with no theoretical limit. The calculation of variance can be expressed as:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

and the square root of the variance provides the standard deviation statistic, which can be thought of as a measure of the variance that has been placed back into a metric that is compatible with the observed data:

$$\sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

The numerator for both equations indicates that the mean is subtracted from each observed value of x (known as a mean deviation); these mean deviations are then squared; and finally they are summed across the entire dataset. The denominator takes the sum of the squared mean deviations and divides them by $n - 1$, essentially creating a measure of the average squared deviation observed in the data.

Carrying out the calculation for an example dataset may help to clarify any questions. Recall the simple five-item dataset presented earlier:

$$X = \{2, 5, 3, 8, 5\}$$

Recall also that $\bar{X} = 4.60$. Drawing on this information, we can calculate variance by subtracting the mean from each observation and then squaring that value:

Value (X_i)	Mean Deviation ($X_i - \bar{X}$)	Squared Mean Deviation $(X_i - \bar{X})^2$
2	$(2 - 4.60) = -2.60$	6.76
5	$(5 - 4.60) = 0.40$	0.16
3	$(3 - 4.60) = -1.60$	2.56
8	$(8 - 4.60) = 3.40$	11.56
5	$(5 - 4.60) = 0.40$	0.16

Before moving to the next step, it is important to note several points. Notice that the mean deviations tell us something meaningful. Specifically, the mean deviations reveal whether each case was above or below the mean by producing a positive mean deviation if the case was above the mean and by producing a negative mean deviation if the case was below the mean. Yet, by squaring these mean deviations, we lose this information because all values result in positive numbers. This may lead you to wonder why we would square the mean deviations. As it turns out, there is a very good reason why the mean deviations are squared.

Squaring the mean deviations solves a pesky computational problem. To see what we mean, add up the mean deviations before they are squared. Doing so results in the following:

$$\sum_{i=1}^{n}(-2.60 + 0.40 - 1.60 + 3.40 + 0.40) = 0.00$$

As you can see, adding up the mean deviations results in the value of 0.00. You may recall from your early mathematics training that zeros wreak havoc in equations. When they appear in the numerator of an equation, the resulting value is undetermined, which is another way of saying it is substantively meaningless. Think about it, what does it mean to say that you are going to divide 0 into $n - 1$ parts?

Unfortunately, this *in*convenient property will always hold. Summing the mean deviations in any dataset will always result in a value of 0.00. Thus, in order to avoid this roadblock, statisticians were forced to seek alternative approaches (Fisher, 1922). Squaring the mean deviations prior to summing them has become the most common solution because the squaring can be "undone" by taking the square root of the value at the end. Doing so will place the variance statistic back into the metric of the original variable (more on this in a moment).[2]

Once the mean deviations have been squared, the next step is to sum them up and, finally, to divide that summed value by $n - 1$. Doing so with our heuristic data results in a value of 5.30. But what does it mean to say that we have an observed variance of 5.30? As with our discussion of the mean, a visual depiction may help.

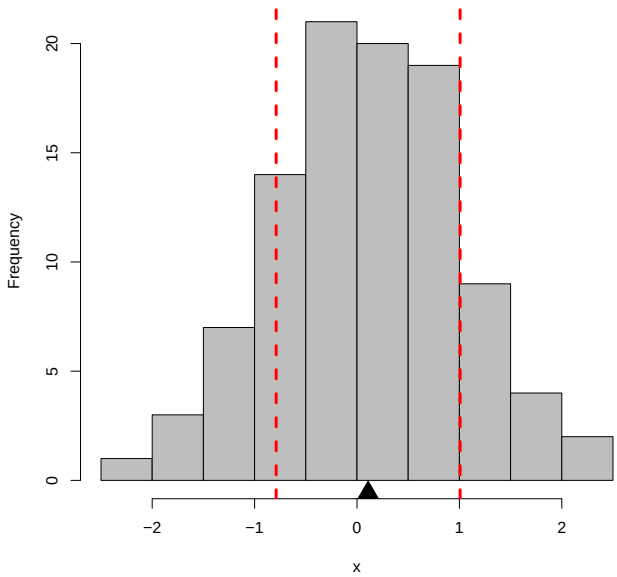
[2]Another solution is to take the absolute value but this is a less desirable strategy because it does not have a readily available way to "undo" the transformation.

2.2.2 Visualization

Let us return to the data that were presented visually in Figure 2.2. Recall that the histogram's bins provide information about the frequency of different values that are observed in the data. The taller the bin, the more often a certain value (or range of values) is observed. Moreover, a triangular fulcrum identifying the location of the mean was placed at the bottom of the distribution. With the preceding discussion of the computation of variance in mind, one point should become immediately obvious: the large majority of the cases in the data scored some value *other than* the mean. Put differently, most cases fall somewhere either above or below the mean, revealing precisely what is meant to say there is variance in a dataset.

The same data that were used to visualize the mean in Figure 2.2 have been reproduced in Figure 2.3. You will notice that the fulcrum again appears in the figure. In addition to the fulcrum, Figure 2.3 also includes two dashed red lines. These lines reveal the location of the first standard deviation interval (i.e., $\bar{x} \pm 1$ standard deviation). Recall from above that the standard deviation is nothing more than the square root of the variance. The standard deviation is sometimes preferred when descriptive information about a dataset needs to be relayed. The reason is that the standard deviation appears in the same metric as the original data. It does so by taking the square root of the *mean of the squared mean deviations* (i.e., the variance).

Figure 2.3: Histogram with the Mean as a Triangular Fulcrum and the Standard Deviation (s) as Broken Red Lines



2.2.3 About Variance Metrics

As you read the above discussion of the variance statistic, you may have been wondering about the two different notations that were introduced (i.e., s^2 and s) and you also may have wondered why $n-1$ appears in the denominator. To the first question, it is important to reiterate the points made above; s^2 denotes the variance and s denotes the standard deviation. There is only a slight computational difference between the two: the standard deviation is the square rooted form of the variance (i.e., $s = \sqrt{s^2}$). Just because a slight computational difference separates the two, do not assume they give the same information. As was pointed out above, the standard deviation (i.e., s) expresses the variance in a metric that is comparable to the original data (i.e., X). The variance expresses the differences observed in the data as squared mean deviations. In practice, this almost always results in the variance being larger (often much larger) than the standard deviation. Yet, for reasons that will be discussed later, the variance is the preferred measure of variation in quantitative genetics. Thus, unless otherwise noted, variation will be expressed as the variance throughout the remainder of the text.

Now, for the question of why $n-1$ appears in the denominator. You may have found yourself thinking, "why not n? Or $n-2$?s Or even $n-3$?" The answer, as it turns out, is based on the difference between sample data and population data. Recall that population data are rarely available. If they are, the variance equation can be adjusted to use n in the denominator rather than $n-1$. But since most analyses will utilize sample data, n must be corrected to account for the potential error that is involved in the mean statistic that is used in the numerator. Adjusting the denominator from n to $n-1$ results in a (slightly) larger variance statistic than if population data were analyzed. Think of it as a way to apologize (mathematically) for using an estimate of μ, $\bar{X}$.

Finally, similar to the mean equation, there are two symbols that are used to denote the variance. The first, which will be used throughout this text is s^2. This symbol is typically used whenever sample data are analyzed. When population data are analyzed, the lower case Greek letter for sigma is used: σ^2. Thus, in keeping with the theme of this text (that sample data are more likely to be analyzed than population data), we will prefer s^2 over σ^2.

2.3 Covariance

2.3.1 Computation

Now that we have introduced the mean and the variance, we have almost all of the components necessary to understand covariance. The only additional piece of information we need to keep in mind is that the covariance expresses the degree to which variance in one variable X is shared with variance in a second variable Y. In other words, the covariance is a bivariate statistic that will provide a numerical summary of the degree to which two

variables are associated with one another. In this sense, the covariance can be a positive or a negative number. Positive covariance between X and Y reveals that cases falling above the mean on one variable (e.g., X) tend to fall above the mean on the other variable (e.g., Y). Negative covariance reveals that cases falling above the mean on one variable (e.g., X) tend to fall below the mean on the other variable (e.g., Y).

In order to understand how covariance is computed, it is helpful to recall the equation for variance:

$$s_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

where the right-hand side of the equation is identical to the variance equation that was presented above. The only difference between this equation and the one presented in the previous section is that the present version has included X as a subscript on the left-hand side of the equation. This subscript was added so that we can distinguish between the variance of X and the variance of the second variable Y.

There is an interesting property about the variance equation that is important to notice before we introduce the covariance equation. Specifically, we can expand the numerator of the variance equation like so:

$$s_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{n-1}$$

This expanded equation is mathematically equivalent to the previous version. But, it reveals something interesting about the way variance is calculated. Specifically, the variance of X can be thought of as mean deviations observed in X multiplied by mean deviations observed in X.

If we wanted instead to observe whether X co-varied with a second variable, we could simply substitute one of the mean deviation statements for X with the corresponding mean deviation statement for Y:

$$cov_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

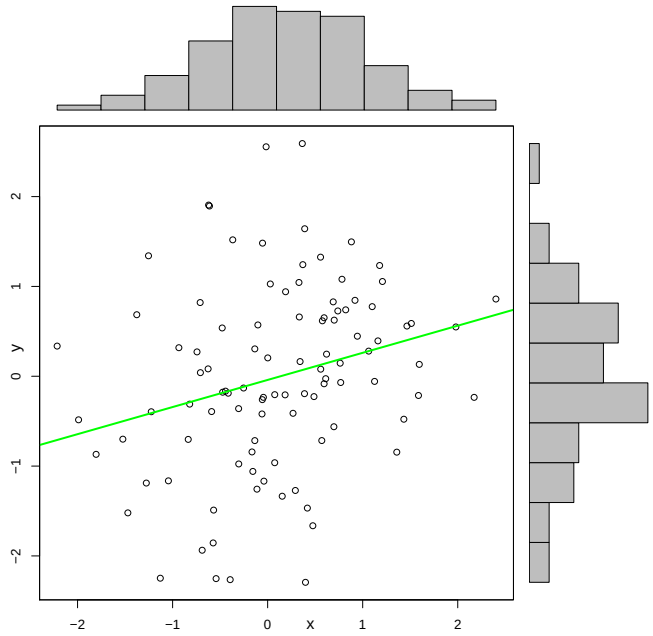
The right-hand side of the equation now multiplies mean deviations in X by mean deviations in Y.

2.3.2 Visualization

There are a number of ways to display the covariance between two variables. Perhaps the most useful way is with a scatterplot. A scatterplot between the observed cases for X and

the observed cases for Y is presented in Figure 2.4 (note that X is the same data that has been utilized up to this point. The cases for Y were generated by randomly drawing from a normal distribution that had a built-in association with X). Several points jump out when you first look at Figure 2.4. First, the primary plotting space reveals open circles. Each circle represents one case. You can identify each case's value for X and Y by referencing the relevant axis. Second, you will notice that marginal histograms are presented. These histograms are unconditional, meaning they reveal the distribution of X (at the top) and the distribution of Y (on the right) without making adjustments for any covariance between the two. Finally, you will notice that a bright green line appears in the primary graphical space. This line reveals the best-fitting linear association between X and Y. The best-fitting linear relationship is, in essence, what the value observed in the covariance equation is telling us. In other words, the covariance equation will produce a value that estimates the association between X and Y. That covariance can be displayed graphically like we see in Figure 2.4. Note, that Figure 2.4 reveals a positive association between X and Y. If the association were negative, it would look like the association revealed in Figure 2.5.

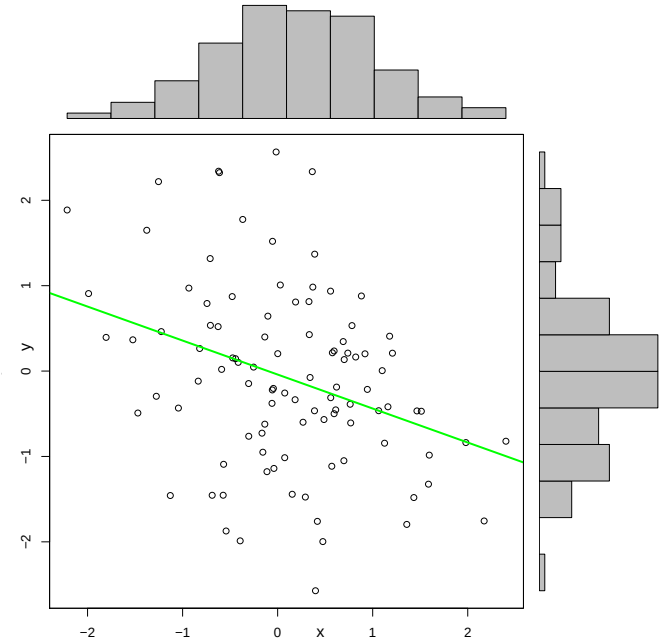
Figure 2.4: Scatterplot of the Positive Association between X and Y with Marginal Histograms



2.3.3 Matrices

We have now introduced you to the statistics that will form the foundation of everything else that follows in this text. There are, however, several additional pieces of information that need to be set in place before we can jump into some of the more advanced biometrical and

Figure 2.5: Scatterplot of the Negative Association between X and Y with Marginal Histograms



genetic modeling strategies. One such piece of information deals with the variance-covariance matrix that will be used in later sections, especially when introducing the biometrical models such as the ACE model. In a general sense, the variance-covariance matrix is a convenient and systematic tool used to store and relay information about the observed (or expected [more on that later]) variance of X, the observed variance of Y, and the covariance between X and Y.

For example, the data presented in Figure 2.5 can be expressed in a variance-covariance matrix like so:

$$\mathbf{A} = \begin{bmatrix} 0.807 & -0.321 \\ -0.321 & 1.131 \end{bmatrix}$$

Matrices like this are usually labeled with an upper-case, bolded letter. We chose **A** here simply for heuristic purposes. We just as easily could have named the matrix **C** or **Z**. Aside from the naming convention, there are two features of the matrix that should be kept in mind as you move through this text. First, a variance-covariance matrix will always present the variance values on the diagonal: $s^2_X = 0.807$ and $s^2_Y = 1.131$. The covariances are listed in the off-diagonal spaces: $cov_{XY} = -0.321$. Second, matrix **A** is a "full" matrix because all four cells of the matrix have numerical values. We could have simplified this matrix, however, to a "lower" (or "upper") matrix because the off-diagonal elements are identical. In other words, we do not need to present the upper off-diagonal number(s) because it is redundant with the lower off-diagonal value(s). For this reason, it is common practice to

eliminate the upper off-diagonal elements like so:

$$\mathbf{A} = \begin{bmatrix} 0.807 & \\ -0.321 & 1.131 \end{bmatrix}$$

2.3.4 Correlation

Although the covariance statistic and the variance-covariance matrix is very useful and provides information that will be used later in this text, it is also common practice to convert the covariance statistic into a standardized value that can be more easily interpreted. As was noted above, the variance statistic can take on any real value from 0.00 to positive infinity (theoretically, at least). In much the same way, the covariance statistic can take on any real number, both negative and positive. This property makes it difficult to gauge the strength of an association simply by observing the covariance statistic. In essence, knowing the covariance of X and Y absent any information about their variances or their standard deviations makes it difficult to say whether the covariance is substantively meaningful.

Enter the correlation coefficient, which is typically denoted as r. Recall that the variance statistic is based on the squared mean deviations and that the standard deviation puts the variance in a metric that is comparable to the original data values. The correlation coefficient capitalizes on this relationship by scaling the covariance according to the standard deviations of both variables X and Y:

$$r_{XY} = \frac{cov_{XY}}{s_X s_Y}$$

By scaling the covariance according to the standard deviations of X and Y, the correlation now reveals the association between X and Y in *standard deviation units*. Thus, the association can be interpreted in standard deviation changes. Certain desirable properties emerge from this transformation. Most importantly, the scale of the correlation ranges between -1 (a perfect negative correlation) and $+1$ (a perfect positive correlation). From this, it follows that the strength of the correlation can be assessed without any other information being needed. Correlation values closer to $1(+ \; or \; -)$ indicate stronger associations. Correlation values close to 0.00 are weak associations.

For these reasons, the correlation coefficient is often presented as a way to demonstrate the association between any two variables of interest. As we introduce the various modeling strategies that follow in later chapters, you will notice that we refer to the correlation coefficient whenever it is necessary to relay information about the strength of a relationship. We will, however, retain an emphasis on the covariance because most of the modeling strategies we will discuss perform "better" if the covariance is analyzed rather than the correlation. Reasons for this will be discussed where appropriate. For now, it is only important that you realize the covariance and the correlation are kin—if you will—much like the variance and the standard deviation. They essentially tell us the same information. How they relay that information—and whether it is readily interpretable—is the only substantive difference between the two statistics.

2.3.5 The Regression Model

Over the past 50 years or so, the regression model has become one of the most popular and routinely employed analytic techniques in most areas of behavioral and social science. As you will see in later chapters, the regression model will form the foundation for several key analytic approaches in quantitative genetics. For example, the regression model is the basis for the DeFries and Fulker (1985) model that is introduced in Chapter 4. With this in mind, we offer only a brief introduction to the regression model here. We leave the more nuanced detail to later chapters that focus on unique uses and transformations of the regression model to handle data that are analyzed in quantitative genetics. Readers interested in a more thorough treatment of the basic regression model are encouraged to see any number of great resources including Gordon (2015) or Wooldridge (2013).

The basic bivariate—meaning only two variables are included in the analysis—regression model can be expressed algebraically as:

$$Y_i = \beta_0 + \beta_1(X_i) + \epsilon_i$$

where Y_i is the observed value of Y for case i; b_0 is the intercept value and represents the average value of Y when $X = 0$; b_1 is the regression parameter estimate for the relationship between X and Y (more on this below); X_i is the observed value on X for case i; and e_i is an error (or residual) term.

Omitting the error term from the model produces the following:

$$\hat{Y}_i = \beta_0 + \beta_1(X_i)$$

where $\hat{Y}_i$ is the predicted value of Y for case i.

The regression model is a basic statistical technique (but do not confuse this to mean that it is without complications; see generally Wooldridge [2013]) that is both well established and easy-to-use. And, as it turns out, the regression model has much in common with the covariance statistic discussed above. In fact, one way to solve for an estimate of β_1 is to first calculate the covariance of the two items and then scale the covariance by the variance in X:

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{cov_{XY}}{s_X^2}$$

Thus, as can be seen, the regression parameter estimate for the relationship between X and Y is nothing more than the covariance between X and Y scaled by the observed variance in X.

2.3.6 About Notation

As with the mean and the variance, there are many symbols one can use to denote the covariance, correlation, and regression parameter estimate. This text will adopt the convention of using *cov* when referring to the covariance. The standard symbol r is used to refer to the correlation, though you may have seen the Greek letter "rho" (ρ) in some occasions. The latter is typically reserved for "special" types of correlation that will be discussed in later sections but that do not drastically differ conceptually from the basic correlation statistic we have introduced above.

Finally, there are at least three ways of writing a regression parameter estimate that have become conventional in the statistics and econometrics literatures. The first is to use the upper-case greek letter beta, β. In some cases the upper-case β is reserved for scenarios when the regression parameter estimate is in standardized form. A second common symbol for the regression parameter estimate is the lower-case beta, b. Finally, the third common symbol for a regression parameter estimate is beta-hat, $\hat{\beta}$. We have elected to use the first symbol, β, throughout this text because it is easy to spot and typically recognized as a regression parameter. Note, however, that the third symbol introduced here, $\hat{\beta}$, is probably more accurate because the "hat" indicates that β is an estimate. Since, as we have discussed several times in this chapter, we will almost always rely on sample data, it is important to keep in mind that even the regression parameter estimates are just that, *estimates*.

2.4 Categorical Data & the Calculation of Proportions

One final point is important to note before we move into the chapters that follow. Specifically, we have introduced statistical equations that are most appropriate for continuous variables that have a distribution that is approximately normal. As a general rule of thumb, the normality assumption is robust to many types of violations. In other words, researchers need not stress too much if their data distribution does not "look" normal. Simulation research has shown that most estimators (e.g., the mean and the regression model) are robust to violations of the normality assumption as long as the sample size is moderate-to-large (a typical rule of thumb is that $n > 30$, but this value increases in proportion to the complexity of the model to be estimated) (Agresti & Franklin, 2013; Allison, 1999; Wooldridge, 2013).

There is, however, one scenario where the assumption of continuous data may cause problems with interpretation (note that the difficulties are rarely computational, rather the problems usually surround substantive interpretations of the results). That scenario involves the use of categorical data (also referred to as nominal, dichotomous, dummy coded, dummy variables, or binary). Categorical data are the lowest level of measurement, meaning they relay the least amount of information when compared to ordinal, interval, and ratio measurements. Categorical data are used to differentiate characteristics or categories, but they tell us nothing about the degree of differences that are observed. For example, typical cat-

egorical measures are sex, race, or treatment status. One might also consider a diagnosis a categorical indicator (e.g., coded 0 for cases with no diagnosis, 1 for cases with a diagnosis). Regardless of the substantive meaning of the categorical variable, it is important to keep in mind that the mean, variance, and covariance equations introduced above may perform poorly or they may provide results that are substantively difficult to interpret. This will be true of the most basic statistical applications (like those introduced in this chapter) and since those basic statistics are used as the building blocks for more complicated models (like those introduced in later chapters), the more complex models are equally likely to suffer from this shortcoming.

As a result, statisticians have developed tools that are tailored specifically to the use of categorical data. We will introduce these tools where appropriate in the chapters that follow. For now, however, we wish to quickly review the most appropriate way to handle categorical data when calculating the mean (now referred to as a proportion), the variance, and the covariance (referred to here as the tetrachoric correlation).

2.4.1 The Mean of Categorical Data: Proportions

The proportion, often denoted as a lower-case p, is the standard statistic for estimating the mean of a categorical level measure. The exact same equation used to calculate the mean can be used to calculate p. Indeed, there is nothing special to calculating p. The only difference from that which was introduced above is that the label on the left-hand side of the equation is different:

$$p = \frac{\sum_{i=1} X_i}{n}$$

2.4.2 The Variance of Categorical Data

Some would argue that it does not make sense to discuss the variance of a categorical variable since it can only take on a finite set of values, typically just 0 and 1. It is for this reason that copyeditors will often ask that the variance/standard deviation be omitted from a table of descriptive statistics if the variable of focus is categorical. Nonetheless, the variance of a categorical variable can be calculated and it is used when inferential statistics are desired (i.e., it is used in the numerator of the standard error equation [more on the standard error in later chapters]). The equation for the variance of a categorical variable is:

$$s^2 = p(1-p)$$

At first blush, the variance of a categorical variable appears to be calculated in a way that differs drastically from the equation outlined above for continuous data. Yet, it can be shown that this equation is identical to the standard variance equation if we simply "plug in" the right information and use n rather than $n-1$ in the denominator (the reason for the latter

is beyond the scope of this discussion, but note that the substantive impact is small). Recall that p is the mean. Assuming the categorical variable only takes on two values, 0 and 1, then the following proof demonstrates the equivalency of the two equations:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} = p(1-p)$$

The first step is to substitute the two possible values for a categorical variable, along with their frequencies into the first equation:

$$s^2 = \frac{n(1-p)(0-p)^2 + n(p)(1-p)^2}{n}$$

Above, we can see that $n(1-p)(0-p)^2$ has been substituted into the numerator of the standard variance equation. The logic is that $n(1-p)$ of the observations will be 0s, so multiplying this frequency count by the appropriate variance calculation will provide the variance contributed by the cases coded 0. The same was done for the cases coded 1: $n(p)(1-p)^2$. Notice that Σ has been removed from the above equation. The reason we were able to omit this term is that we will calculate the variance contributed by the 0s and then we will add that to the variance contributed by the 1s. Thus, we omit Σ because we explicitly show the summation in the equation.

Now, continue to reduce/simplify the equation:

$$s^2 = \frac{n(1-p)(0-p)^2}{n} + \frac{n(p)(1-p)^2}{n}$$

$$= (1-p)(0-p)^2 + p(1-p)^2$$

Now we need to deal with the right-hand side by expanding the square:

$$s^2 = (1-p)(0-p)^2 + p[(1-p)(1-p)]$$

Note that $(0-p)^2$ is just p^2 because 0 minus anything is just that value, negative. Squaring a negative gives a positive, so:

$$s^2 = (1-p)p^2 + p(1-2p+p^2)$$

Finally, we can continue to simplify:

$$\begin{aligned} s^2 &= p^2 - p^3 + p - 2p^2 + p^3 \\ &= p^2 + p - 2p^2 \\ &= p - p^2 \\ &= p(1-p) \end{aligned}$$

Chapter 3

The Foundation of Quantitative Genetics

This chapter introduces some of the most basic, yet most important, concepts necessary to understand the logic of quantitative genetic analysis. Much of the material provided in this chapter was drawn from the magisterial work of Falconer and Mackay (1989). Readers are encouraged to consult Falconer and Mackay (1989) or other authoritative texts on population and quantitative genetics such as the comprehensive work of Lynch and Walsh (1998) or Hamilton's (2009) introduction to population genetics. Posthuma et al. (2003) also offer an accessible overview in article length.

This chapter covers four main topics. The first concerns gene and genotype frequencies in the population (a hypothetical population at this point, but decades of population genetics analysis confirm the primary tenets that will be introduced). Gene and genotype frequencies can, in a general sense, be altered as a result of several forces including mutation, migration, and selection. We will not provide much detail on the ways in which gene and genotype frequencies can change over time, so interested readers are directed to Falconer and Mackay (1989). What is important for this text is the *result* of changes to gene and genotype frequencies. Enter the concept of equilibrium and the calculation of the Hardy-Weinberg equilibrium equation.

The second issue covered in this chapter is how genes and environments can influence the mean of a phenotype in the population. Recall that a phenotype is an umbrella term for any trait that is measurable (even if it is a categorical or nominal measure) in the population. Thus, we will demonstrate and discuss how genetic and environmental factors can lead to shifts in the mean of a phenotype in a population.

Third, this chapter will provide a discussion of how the variance of a phenotype can be affected by genetic and environmental factors. In short, we will demonstrate how variation in the aggregate mixing of genetic and environmental factors leads to more (or less) variation in a phenotype over time. This discussion will primarily use quantitative, continuous traits

as examples because the variance statistic is more obviously applicable in those scenarios. Yet, it is important to note that the same logic (though slightly different computations are necessary) applies when the phenotype is a categorical measure.

Finally, this chapter will close with a discussion of heritability. We will offer a conceptual definition of heritability, followed by a computationally based overview. As you will learn in later chapters, estimating heritability has become one of the main foci of modern quantitative and behavioral genetics. Scholars' interest in heritability dates back nearly a century, but as far as we can tell, the interest has not yet waned (Barnes et al., 2014). Somewhat surprisingly, however, heritability remains one of the most misunderstood and misapplied concepts by social scientists. Thus, the closing portion of this chapter offers a broader theoretical discussion of heritability, what it means, and where it fits in modern day behavioral science.

3.1 Gene & Genotype Frequencies in a Population

As was discussed in Chapter 1, genes come packaged on chromosomes and are defined as a contiguous string of base pairs along DNA molecules that work together to code for protein synthesis (Beaver, 2013; Snustad & Simmons, 2009). Although humans share much of their genetic make-up, many loci along the human genome vary from person-to-person. *How* that genetic variation enters the genome is beyond the scope of this text. See Falconer and Mackay (1989) or Hamilton (2009) for a detailed discussion of these issues. For now, suffice it to say that variation enters the human gene pool on a fairly regular basis. Most gene variants are silent mutations meaning they have no visible impact on the organism's structure or function. Yet, occasionally a gene variant will enter the population and it will confer an advantage for its carriers. In this situation, it might be expected that the advantageous gene will be passed along to the next generation with a non-zero probability. As more and more individuals inherit the gene variant, it will be detectable on a population wide scale and, as a result, it becomes a candidate for understanding the mean position or even the variance that is observed in the phenotype of focus.

Humans are diploid organisms, meaning we carry two copies of each gene (with the exception of those appearing on the sex chromosome in males). With this in mind, let us imagine that we have identified an autosomal (meaning it is located on one of the 22 autosomes and not on a sex chromosome) gene that is known to have a functional effect on levels of self-control; call the gene *LSC*.[1] Findings from contemporary quantitative and behavioral genetics reveals that most genetic influences are small, typically explaining less than 1% of the observed variance. So, let's say the scale for self-control ranges between 1 and 100. And the gene of focus is biallelic, meaning it comes in two forms: S (read: "big S") and s (read: "little s"). Biallelic genes like this can only come in one of three genotypes (It is important

[1]Conventionally, human genes are written in all capital letters and with italicized fonts. We adopt this convention in this text.

to note that the words "gene" and "genotype" are not synonymous. A gene is the unit of inheritance, meaning parents pass genes to their children. A genotype refers to the specific combination of genes a person(s) carries.):

$$\text{Genotypes} = \{SS, Ss, ss\}$$

Note that the heterozygote genotype can also be expressed as sS or Ss. Assuming S is the more common allele in the population, we can represent the frequency with which S occurs in the population as p. Here, p is a simple proportion of individuals in the population who carry at least *one* S allele. By definition, the frequency of the minor allele—the allele that occurs less often—can be calculated as:

$$q = 1 - p$$

and, therefore:

$$p + q = 1$$

We can represent the gene frequencies and the genotype frequencies in the mating population as:

	Alleles		Genotypes		
	S	s	SS	Ss	ss
Proportions	p	q	P	H	Q

where p and q represent the same concepts outlined above; P is the proportion of the population that is homozygous for the S allele; H is the proportion of the population that is heterozygous (Ss or sS); and Q is the proportion of the population that is homozygous for the s allele.

Given P, H, Q, and the knowledge that there are two "types" of heterozygotes (Ss or sS), it follows:

$$p = P + \frac{1}{2}H$$

$$q = Q + \frac{1}{2}H$$

Furthermore, we can represent the genotype frequencies in a 2×2 table like so:

		Maternal Gametes and their Proportions	
		$S(p)$	$s(q)$
Paternal Gametes	$S(p)$	$SS(p^2)$	$sS(pq)$
and their Proportions	$s(q)$	$Ss(pq)$	$ss(q^2)$

Which allows us to substitute p^2 for P, $2pq$ for H, and q^2 for Q. Then, the genotype for the progeny of the mating population can be expressed as:

	Progeny Genotypes		
	SS	Ss	ss
Proportions	p^2	$2pq$	q^2

And we can use these genotype frequencies to generate *expected* frequency values for the S allele:

$$\begin{aligned} f(S) &= p^2 + \frac{1}{2}(2pq) \\ &= p(p+q) \\ &= p(1) \\ &= p \end{aligned}$$

and, for the s allele:

$$\begin{aligned} f(s) &= q^2 + \frac{1}{2}(2pq) \\ &= q(p+q) \\ &= q(1) \\ &= q \end{aligned}$$

Putting it all together:

$$p + q = 1$$

so,

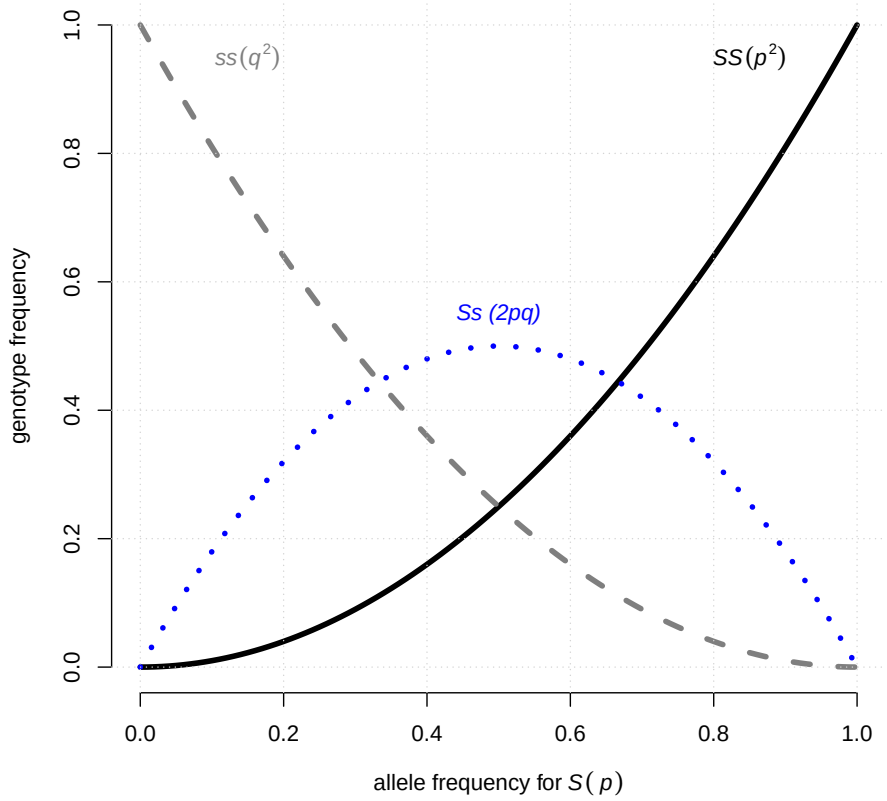
$$p^2 + 2pq + q^2 = 1$$

The purpose of this mathematical proof is to introduce the last equation from above as the Hardy-Weinberg Equilibrium (HWE) equation. HWE is one of the foundational concepts of quantitative and population genetics. Using the last equation, we can generate expected genotype frequencies knowing nothing more than the allele frequencies observed in the population (or vice versa). The relationship between genotype frequencies and allele frequency (for the allele represented as p) is demonstrated in Figure 3.1.

A quick glance at Figure 3.1 reveals several interesting points. First, as the allele frequency for S (denoted, as above, as p) increases toward 1.00, we see that the genotype frequency for SS homozygotes (i.e., p^2) increases toward its upper limit of 1.00 ($\lim_{p \to 1.00} f(p^2) = 1.00$). The opposite is true for the genotype frequency of ss homozygoes. As $p \to 1.00$, $q^2 \to 0.00$. Finally, the limit for heterozygotes is 0.50, meaning that the maximum number of heterozygotes in the population can never be larger than 50%. And this is only achieved when the frequency of both alleles (S and s) is equal to 0.50.

The HWE is especially important to researchers studying population and quantitative genetics because it provides a benchmark of expectations. These expected genotype frequencies (or expected allele frequencies) can be compared against the observed genotype (allele) frequencies in a population or sample to test the null hypothesis that flows from

Figure 3.1: Expected Genotype Frequencies as a Function of the S Allele Frequency p Based on the Hardy-Weinberg Equilibrium



HWE. Specifically, if a gene is in HWE, then we should not see any statistically significant differences from the expected values outlined above. Put differently, a population in HWE will not observe any changes in gene frequency from one generation to the next. One can even perform a simple chi-squared (X^2) test for independence to assess the relationship between the expected and observed values. If a statistically significant difference does emerge, then it suggests that the gene under observation has violated one of the assumptions of "normal" gene flow in a population. This can occur due to the allele(s) being selected *for* or *against* by natural selection, it can occur as a result of non-random mating, or any number of other scenarios (see Falconer & Mackay, 1989; Hamilton, 2009). The short of it is that any gene that is observed to reject the null model of HWE is likely to have been acted upon by one or more evolutionary and/or biological processes.

3.2 Phenotypic Values & Means in a Population

The preceding discussion introduced you to the way that genes and genotypes vary in the population. Yet, one critical link has not been considered. Specifically, it is not enough to say that a gene varies in the population (i.e., has more than one allele available for inheritance). As behavioral researchers, what we really want to know is whether (and how, but that question will be addressed in the closing portions of this text) genetic variation leads to observable differences in a phenotype among members of the population. When we consider this question, it quickly becomes apparent that two points must be addressed: 1) how genes affect the value of the phenotype at the individual-level; and 2) whether the gene affects the mean of the phenotype in the population.

For any individual in the population of interest, their phenotypic value (P)—meaning the score one receives on a measurement of a phenotype like height, weight, or IQ—can be parsed into two main components: a genetic component, G, and everything else, E, which will be referred to broadly as the environment but it need not be restricted to the social environment:

$$P = G + E$$

where P now refers to the phenotypic value of an individual and is distinct from the use of P from the above discussion of HWE. The present equation illustrates that a person's phenotypic value arises from genetic and environmental sources. As genetic factors play a larger role in determining the phenotypic value, the environment necessarily plays less of a role. Moreover, quantitative geneticists have argued that the phenotypic value is determined by the genetic value, plus whatever deviation is imparted by the environment (Falconer & Mackay, 1989). This is an important way of conceptualizing the phenotypic value because it represents a radical departure from the standard social science model that assumes the environment is the sole cause and that genetic factors are a nuisance term. Nonetheless, the overarching theoretical component being introduced here is that a phenotypic value results from the combination of genetic and environmental factors.

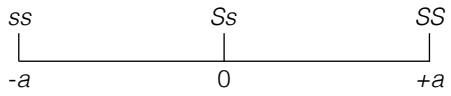
Regardless of whether we view genetic or environmental factors as the nuisance term, we must gain an understanding of *how* genes will affect the phenotypic value. This discussion is less a functional analysis (meaning molecular/biological function) and more a quantitative analysis revealing the logic of genetic effects for phenotypic outcomes. Yet, the connections to functional mechanisms such as neurotransmission are easy to see.

3.2.1 Single Gene Model

There are, as it turns out, several ways that a gene can influence the development of a phenotype. Perhaps the best way to relay this information is through the use of the diagram provided in Figure 3.2.

Figure 3.2 displays a spectrum of the degree to which the various genotypes might affect

Figure 3.2: Genotypic Effects of the Imaginary *LSC* Gene on Phenotypic Values Under a Strictly Additive Model



phenotypic values of self-control (sticking with the imaginary *LSC* gene discussed above). We will assume that SS homozygotes have higher levels of self-control, so we code this genotype as $+a$. Likewise, we assume ss homozygotes have the lowest levels of self-control (on average), so we code this genotype as $-a$. Finally, the heterozygote genotype (Ss *or* sS, but we will use Ss to represent both for simplicity) has been placed at the midline (represented as 0 for simplicity; this is akin to mean-centering the phenotype score), which is halfway between $-a$ and $+a$. Since heterozygotes fall on the midline, we are safe to assume there is no dominance observed for this trait. That is to say the imaginary *LSC* gene has as strictly additive influence. We will assume no dominance at the genotypic level for now in order to simplify some of the math that follows later. It is only important to point out that most genotypic influences are likely to involve some level of dominance deviation. We will return to this point later.

Pushing forward with this example, we can begin to build a model to describe how the three genotypes can influence the mean of the phenotype in the broader population. It is simply an exercise in calculating a weighted average. We must consider the size of the $-a$ effect, the degree to which heterozygotes exhibit a dominance deviation, and how much higher than the heterozygotes are the $+a$ genotypes.

Under a model where the phenotype of focus is affected solely by the single focal gene, we can generate an expected value for the phenotype using the following genotypic information:

Genotype	Genotype Frequency	Genotypic Value	Frequency*Value
SS	p^2	$+a$	p^2a
Ss	$2pq$	0	$2pq(0) = 0$
ss	q^2	$-a$	$-q^2a$

Summing across the three entries in the far right column provides an estimate of the phenotypic mean in the population:

$$\begin{aligned}\bar{P} &= p^2a - q^2a \\ &= a(p^2 - q^2) \\ &= a(p+q)(p-q) \\ &= a(1)(p-q) \\ &= a(p-q)\end{aligned}$$

As the proof shows, when a gene works additively (i.e., there is no dominance), the phenotypic mean $\bar{P}$ can be estimated as $a(p-q)$, where a is the simple difference between phenotypic scores between heterozygotes and either homozygote group [2] This equation should be intuitive when you approach the problem as a weighted average because $p-q$ will index how many more (proportionally speaking) SS homozygotes there are compared to the ss homozygotes. Once this simple difference is calculated, the relative increase in the phenotype that is associated with the SS homozygote is multiplied by the proportional difference between SS and ss genotypes to provide an estimate of the mean value to be expected in the population.[3]

One interesting observation that flows from this equation is that changes in the genotypic makeup of the population (whether it be due to mating patterns or selection effects) will result in changes to the phenotypic mean in the population. Think of it this way: if you were to adjust p or q in the population, the mean ($\bar{P}$) would be affected. This reveals that population means for phenotypes are directly tied to the genetic factors that underlie them!

3.2.2 Multiple Gene Model: Polygenics

To this point, our discussion of the phenotypic mean has assumed a relatively simple and straightforward connection between one gene with three genotypes and no dominance deviation, a point to which we will return later. Let us now consider the result of increasing beyond a single gene model to one where two genes are believed to influence the phenotype. As it turns out, with some simplifying assumptions (such as the assumption that the genes are not in linkage disequilibrium, that there is no dominance, and no epistasis), the logic remains the same as was outlined above. Rather than consider the role of one gene, we now *add* a second gene to the equation. The weighted average now becomes a weighted average of the proportion of the population who have all possible genotypes for gene A and for gene B (it is necessary to assume both are bi-allelic such that gene A has two alleles A and a and gene B has two alleles B and b.). There are now 16 possible combinations that fall into 9 categories after collapsing the symmetrical heterozygote combinations. Beyond extending the number of categories, the logic is unchanged:

		Gene A		
		AA	Aa or aA	aa
Gene B	BB	$AABB$	$AaBB$ or $aABB$	$aaBB$
	Bb or bB	$AABb$ or $AAbB$	$AaBb$ or $AabB$ $aABb$ or $aAbB$	$aaBb$ or $aabB$
	bb	$AAbb$	$Aabb$ or $aAbb$	$aabb$

[2]Note that this is only true under conditions like the one used in this example where there is no dominance deviation and the heterozygotes fall exactly at the midpoint.

[3]We have simplified this discussion by ignoring the point that $\bar{P}$ is actually the mean deviation—from the midline—so $\bar{P}$ must be added to the observed mean of the heterozygotes to get an estimate of the mean of the phenotype in the population.

Let us assume the trait under examination is criminality, where A and B each confer $+1$ increase in criminality and a and b confer a one point decrease in criminality. The table above can, therefore be written as deviations from the mean, which we will assume is 0 (i.e., assume we have mean-centered criminality) to simplify the math:

		Gene A		
		AA	Aa or aA	aa
Gene B	BB	$+4$	$+2$	0
	Bb or bB	$+2$	0	-2
	bb	0	-2	-4

Now, notice that folks who are heterozygous for A and B represent the midline of the phenotype and all others fall somewhere above or below. Attaching gene frequency information is all that would be necessary to generate the expected mean phenotypic value in the population. In short, we can begin to see how both the mean and variance in a phenotype is affected by genetic influence even when restricted to just one or two genes.

3.3 Variance of a Phenotype in a Population

One of the most compelling, aggravating, and incendiary questions one can ask is: "why aren't we all the same?" It turns out that there are very interesting evolutionary answers to these questions, but the easiest response is that we vary because we have different genes (and environments, but we will get to that in a minute). In fact, the preceding discussion of phenotypic values and phenotypic means might have alerted you to an obvious corollary: genotypic values drive the mean phenotypic value and they are also responsible for variation around the mean.

Let us define the variance in the phenotype P as V_P. Earlier, we noted that a phenotypic score (or the phenotype mean, $\bar{P}$) was defined as $P = G + E$. We can extend this equation to also account for V_P. Specifically, assuming there is no covariance between G and E, then:

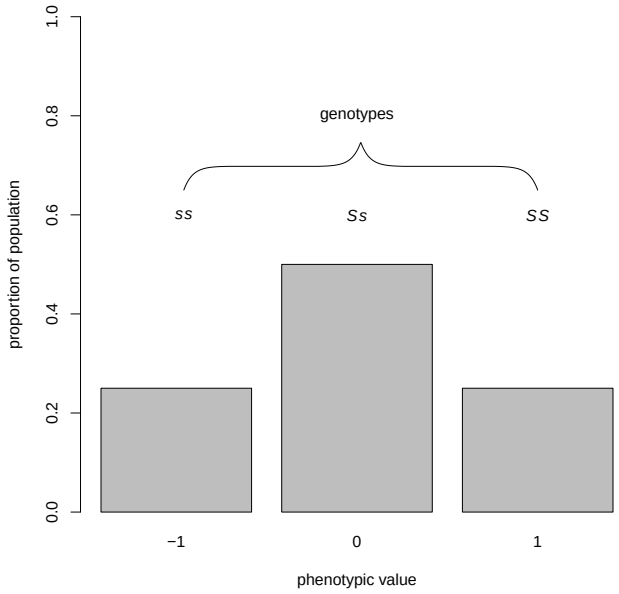
$$V_P = V_G + V_E$$

To understand how we can move from P to V_P (and similarly for $G \rightarrow V_G$ and $E \rightarrow V_E$), we will need to reconsider the single gene example from Figure 3.2. Recall there are three genotypes that one might carry (i.e., SS, Ss, and ss). If we conceive of this gene as impacting one's location on a spectrum of the phenotype (again, consistent with Figure 3.2), then it quickly becomes apparent that genetic influences affect the mean *and* the variance of the phenotype.

Consider the information presented in Figure 3.3, which is a bar chart revealing the proportion of individuals in the population who are expected to carry the three possible genotypes

when $p = 0.50$. We see that there are three possible groups that one may be placed in, each having its own average phenotypic value. In this respect, the distribution is simply the binomial distribution where the number of trials is equal to the number of alleles being considered (two alleles in the present case) and the probability of "success" is equal to p. As we learned from the HWE (see Figure 3.1), the distribution we expect is that the heterozygotes will account for 50% of the population and the homozygotes will each account for 25% of the population when $p = 0.50$. This was the setting used to generate the heuristic data presented in Figure 3.3.

Figure 3.3: Bar Chart Showing Proportion of Population With Different Phenotypic Values Due to a Single Bi-allelic Gene at a Single Loci



Having a visual depiction of how genotypic values can create variance in a phenotype, let us now consider it from a mathematical framework. As it turns out, it is does not require much to translate genotype data (i.e., the phenotypic values) into *variance*.

Sticking with the single gene model—and continuing to assume there is no dominance deviation—that was presented above, we can extend the table that was presented earlier to reveal how V_P can be computed directly from the genotypic information.

Genotype	Genotype Frequency	Genotypic Value	Mean Deviation
SS	p^2	$+a$	$2qa$
Ss	$2pq$	0	$a(q-p)$
ss	q^2	$-a$	$-2pa$

To see how this information translates into V_P, recall the variance equation that was reviewed

in chapter 2:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Which can be translated into a form that reflects the notation that we have introduced in this chapter:

$$s^2 = \sum f_i(X_i - \bar{P})^2$$

where f_i is the genotype frequency (i.e., p^2, $2pq$, or q^2) and $(X_i - \bar{P})^2$ is the squared mean deviation. Combining the genotype frequency with the mean deviation, it can be shown that the variance contributed by any genetic variant g can be expressed as:

$$\begin{aligned} V_g &= p^2[2qa]^2 + 2pq[a(q-p)]^2 + q^2[-2pa]^2 \\ &= 2pqa^2 \end{aligned}$$

Plugging in some heuristic values may help solidify this information. Imagine a gene where $p = q = 0.50$, $a = 1$, and $d = 0$. In this case, the above equation would result in $V_g = 2(0.50 * 0.50)1^2 = 0.50$. We can even confirm this using the traditional variance equation from above: $\sum f_i(X_i - \bar{P})^2 = [p^2(a-0) + 2pq(0-0) + q^2(-a-0)] = [0.50^2(1-0)^2 + 0.50(0-0^2) + 0.50^2(-1-0)^2] = 0.25 + 0 + 0.25 = 0.50$.

If one gene with two alleles can produce three different phenotypic scores—thereby also producing phenotypic variance—you might be wondering what happens when a phenotype is influenced by multiple genes. When this occurs, a condition know as polygenic variation emerges. Polygenic variation reveals that—all else being equal—the variance in a phenotype increases when it is influenced by more than one gene. Modern quantitative and biometrical genetics suggests that most of the human complex traits of interest to social scientists (e.g., intelligence, earnings, and criminality) are polygenic. Thus, unless otherwise noted, we will assume polygenic variation underlies phenotypic variance from this point forward.

Luckily, if we are willing to make certain assumptions—again, assuming no dominance, no epistasis (see below), and no linkage disequilibrium (see chapter 7)—then it is easy to extend the above equations to account for the influence of *multiple* genes. Specifically, we simply sum up the V_gs for each gene, such that the total genetic variance is calculated by summing across all genetic variants that influence V_P:

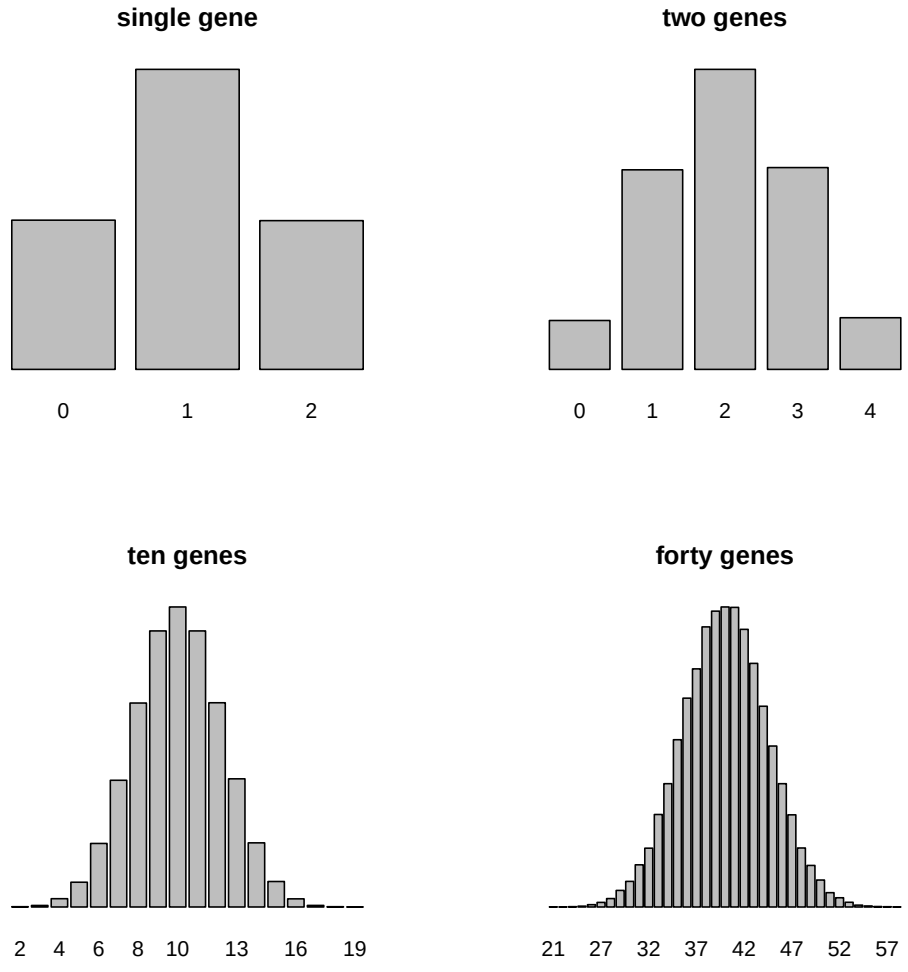
$$V_G = \sum_{g=1}^{G} 2p_g q_g a_g^2$$

where $g = 1$ tells us to start with the first genetic loci and calculate the variance it contributes to V_P. Then repeat the exercise across all loci G and end by summing up all components into V_G.

Such polygenic variation is displayed graphically in Figure 3.4, which has four panels. The distribution in the top-left panel of the figure is identical to the distribution that was presented in Figure 3.3. In other words, we see the expected distribution of phenotypic

scores as a result of a single bi-allelic gene. Moving to the panel in the top right of the figure reveals what happens to the distribution when *two* genes (both are assumed to be bi-allelic and genotype frequencies are equal for both genes) affect the phenotype with equal impact. As you can see, the distribution begins to take on a shape that resembles the normal distribution with an obvious modal value in the middle and symmetric tails to the sides. The bottom row in the figure reveals what happens when *ten* genes (bottom left) and *forty* genes (bottom right) (all bi-allelic with all the same assumptions as before) affect the same phenotype in equal amounts. Under these conditions, we very clearly see a normal distribution emerge. This means that we can begin to approach the study of phenotypic variance with the statistical tools that were provided in Chapter 2.[4]

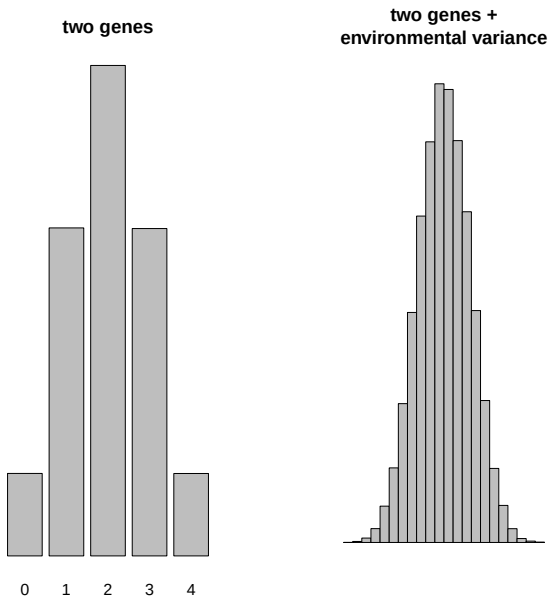
Figure 3.4: Bar Charts Showing the Distribution of Phenotypic Scores in the Population as a Function of Polygenic Variation (All Genes Assumed to be Bi-allelic)



[4]It is important that we tip our cap to the trailblazing work of R. A. Fisher here. The points we are making in this section—and, indeed, in this chapter—were all very cogently spelled out by Fisher in his seminal paper from 1918.

But recall that V_P results from *both* V_G and V_E. As before, we can conceive of the environmental variance as a deviation or a nuisance parameter that adds "noise" to the overall phenotypic variance. This is not to say that the environment is a stochastic, unpredictable, and non-influential part of the equation. To be sure, we will return to a full discussion of the role of the environment in explaining phenotypic variance later in this text. For now, however, it will simplify our argument if we conceive of the environment as a random variable (though, keep in mind that calling something a random variable is not the same as saying its influence is random). The two distributions presented in Figure 3.5 demonstrate this point. Note that the distribution on the left is the same as the one that was presented in the top-right panel of Figure 3.4. Specifically, this is the distribution of phenotypic scores as a result of two bi-allelic genes with equal influence, but the assumption in Figure 3.4 was that the environmental variance component (i.e., V_E) was zero, meaning V_P was assumed to equal V_G. The distribution on the right of Figure 3.5 reveals the same distribution when V_E is non-zero. This time, a random environmental component was added. As can be seen, there is more phenotypic variability observed in the population in the latter scenario. In essence, there is more "noise" to mask the genetic effect.

Figure 3.5: Phenotypic Values as a Result of Two Bi-Allelic Genes (First Panel) and Two Bi-Allelic Genes + Environmental Variance (Second Panel)



3.3.1 Key Focal Point

The remaining sections and chapters in this book will deal with the very issue highlighted in Figure 3.5. Specifically, quantitative and behavioral geneticists seek answers to questions

that surround the decomposition of variance into genetic and environmental components. Relating this to Figure 3.5 will reveal the challenging nature of the problem facing behavioral geneticists. Essentially, these researchers have access to data that looks like the right panel of Figure 3.5 (although rarely do the data display such a neat and obvious normal distribution). The problem they are faced with is trying to devise a strategy to separate the sources of variance into a genetic component, the left panel of Figure 3.5, and an environmental component. Of course, they don't have access to the left panel of Figure 3.5! So how do they do it? How can they possibly "slice up" a distribution into its constituent genetic (V_G) and environmental (V_E) components? That will be the focus of the remaining chapters of this text. Before we tell you *how* it is done, though, we must first introduce you to a few more concepts.

3.3.2 Heritability (h^2)

A researcher who is interested in studying the degree to which genetic factors influence phenotypic variance is really asking the following question: "what is the heritability of this phenotype?" As we will show in later chapters, estimating heritability is one of the primary aims of behavioral genetics. For this reason, the concept *heritability* will be one of the key organizing elements for this text.

So what is heritability; what does that term mean? A typical textbook-style definition might read something like this: heritability is the degree to which genetic factors explain variance in a phenotype. It is often symbolized as h^2.

Although that definition is technically accurate, it is a bit unsatisfying because it does little to communicate the underlying essence of what the concept is trying to tap into. In this regard, the concept of heritability is meant to communicate how much genetic factors influence an outcome. Eric Turkheimer, speaking on the David Pakman show (`https://youtu.be/rAKHmzrb6RA`), recently said heritability tells us "how much more similar we would all be if we were all [genetic] clones." That is a nice working definition because it resonates and one can immediately connect with it—if we were all genetic clones of one another, then traits that are highly heritable would be very similar from person-to-person. Take, for example, height, which has a fairly high heritability coefficient (probably somewhere between 0.60 and 0.80). This means, if we were all genetic clones, then we would expect very little variation in height—perhaps only about 20% of the variation we see in society now.

Estimating heritability is one of the aims of modern behavioral genetics; though behavioral geneticists certainly take their work well beyond the estimation of heritability. We will, in short order, provide you with a nauseating level of detail on how to generate heritability estimates in your own work. For now, we simply want to explore the concept of heritability. What does it mean? How, in a general sense, does it apply to quantitative genetics and to social science research?

Fewer scientific concepts have been misunderstood and maligned more often than heri-

tability. Although researchers have been interested in estimating heritability for centuries (going all the way back to Francis Galton, R. A. Fisher [yes, the famous statistician was perhaps even more famous for being a geneticist while he was alive], and Sewall Wright), many social scientists remain unsure what to make of heritability. They do not quite understand what it means to say that something is heritable. Moreover, scholars are prone to interpret heritability studies as a challenge to the environmental influences they have been trained to study. In some ways, perhaps this view is warranted. We will consider the ways in which heritability research impacts the standard social science model in later chapters. Yet, in our view, heritability research has much to offer the criminological discipline. Additionally, we do not see heritability research as being antithetical to environmentally focused research agendas. Rather, heritability research provides the tools necessary to study the environment, free of one of the largest confounding influences known to behavioral researchers: genes.

Before we can move into a detailed discussion of heritability, it is first necessary to reveal that the equation spelled out above, the one for phenotypic variance, is actually more complicated that it seems at first. Indeed, phenotypic variance (V_P) cannot be neatly divided into a genetic component (V_G) and an environmental component (V_E). Instead, there are at least five unique components that must be considered when explaining the variance of a phenotype because the genetic variance comes in three "types" while environmental variance can be broken into two unique forms:

$$V_G = V_A + V_D + V_I$$

and:

$$V_E = V_C + V_E$$

so:

$$V_P = V_A + V_D + V_I + V_C + V_E$$

where V_A is the additive genetic component; V_D is the dominance deviation genetic component; V_I is the epistatic genetic component; V_C is the shared environmental component; and V_E is the nonshared environmental component. We will discuss each of these variance components in turn.

V_A: Additive Genetic Factors

Additive genetic factors are those that operate like the examples that were given in the preceding sections of this chapter. When the heterozygotes are on the midline between homozygotes, the predicted phenotypic scores are said to be additive. There has been a lot of theoretical and empirical attention devoted to additive genetic effects, and with good reason. If one assumes additivity, then many other downstream concerns computations become more tractable. For example, we will discuss something known as "missing heritability" in chapter 8 when we cover recent advances in genome-wide association studies (GWAS). Assuming that most or all genetic variants act additively has important implications for one's interpretation of GWAS results that appear to explain only a small portion of the heritability observed from more traditional behavioral genetic results.

Given that there are three unique genetic components, you may have already anticipated that there is more than one way to calculate heritability. In fact, heritability estimates come in two "flavors": 1) narrow-sense heritability (h^2) and 2) broad-sense heritability, which we will differentiate from the former by using H^2. Broad-sense heritability will be introduced below, after dominance deviation and epistatic genetic factors have been introduced.

At the moment, our focus is on the simplest form of genetic influence, narrow-sense heritability. Narrow-sense heritability reveals the degree to which phenotypic variance V_P is explained by additive genetic factors V_A. In other words, narrow-sense heritability can be thought of as the "...maximum variance that can be explained by a linear combination of the allele counts..." (Zuk et al., 2012: 1194). Thus, when V_A is studied separate from the other two forms of genetic factors, we are said to be estimating the impact of narrow-sense heritability. Algebraically, narrow-sense heritability can be expressed as:

$$\text{narrow-sense } h^2 = \frac{V_A}{V_P}$$

Given an estimate of the degree to which *additive* genetic factors impact phenotypic variance, one can garner an estimate of narrow-sense heritability with relative ease. The equation above shows that narrow-sense heritability is simply a proportional estimate of the degree to which phenotypic variance is attributable to additive genetic variance. The trick, as you might have guessed, is getting an estimate of V_A. Starting in the next chapter, we will begin to discuss ways in which estimates of V_A (as well as the other components) can be generated using individual-level data such as family pedigrees, genetic relatives, or even information from genome-wide scans.

At this point, you may be wondering why anyone would want to estimate narrow-sense heritability. Why not prefer broad-sense heritability in all scenarios, that way you capture a "full" image of the genetic influenced. There are at least two reasons scholars are typically, though not always, interested in narrow-sense over broad-sense heritability. The first is a practical/mathematical issue. As we will see in the next chapter, it is often impractical to estimate broad-sense heritability with the methods typically used by quantitative geneticists. We will explain this point in more detail in the next chapter. The second reason scholars are often interested in narrow-sense heritability is that it gives us an indication of the degree to which a trait will *breed true* in the population. A trait that breeds true is one that shows an expected relationship across generations. Thus, narrow-sense h^2 gives us an idea of the degree to which parents and their offspring will show phenotypic similarities.

About "Additivity"

The previous section outlined what it means to say a gene acts "additively." Confusingly, we will also use the term "additive" when we discuss the way that genetic and environmental influences affect phenotypic development. As you noticed throughout this chapter, we have consistently set V_P on the left-hand side of an additive equation. Given the slightly different

usages of the term "additive", we will be careful to specify when we mean "additive" in terms of V_A and "additive" in the more common mathematical sense.

Until recently, many scholars felt comfortable assuming that most genetic influences act in an additive fashion, meaning narrow-sense h^2 was thought to be sufficient to understanding the genetic architecture of most complex traits (Crow, 2010). To be sure, Hill and colleagues (2008) performed many computations and simulations that appeared to support this assumption. But more recent developments by Zuk and colleagues (2012) question the veracity of this assumption and reveal, with impressive depth and clarity, how non-additive genetic influences may play an important role in phenotypic development. We will return to a discussion of Zuk and colleagues' findings in chapter 8. For now, suffice it to say that Zuk et al. were able to show that observational data are quite consistent with the idea that non-additive effects underlie a large portion of phenotypic variance.

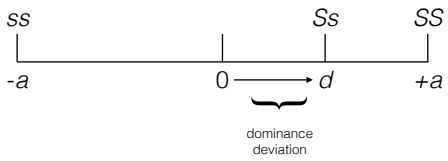
V_D: Dominance Deviation

Up to this point, our discussion has primarily focused on additive genetic influences. Even the earlier sections of this chapter relied on additive effects to demonstrate how genotypic values impact the phenotypic value, the phenotypic mean, and phenotypic variance. Yet, it turns out that nature has made things a little more difficult than we have let on! In fact, it is typically assumed that most genetic factors do *not* work in an additive fashion. But wait a second. We just said—in the section immediately above this one—that additive genetic effects account for the majority of the genetic variance. What is going on?

The short answer is that much of the human genome likely works in non-additive ways. Yet, when we combine the influence of all the genes that impact a phenotype and calculate the average phenotypic score, the non-additive effects tend to be "washed out" and, instead, start to resemble a simpler additive model. In other words, nature is more complicated than our models might suggest, but there are few reasons to suspect our models are systematically biased if we ignore this complexity when estimating h^2.

Because the dominance genetic component (V_D) is a non-additive component, you may have guessed that the impact will be non-linear. Recall the additive effect that was demonstrated in Figure 3.2. If the effect were non-additive and, instead, showed a dominance deviation, the figure might have looked more like that which is displayed in Figure 3.6. Note that Figure 3.6 looks a lot like Figure 3.2. The primary difference is that the heterozygotes are more similar to the SS homozygotes than would be expected under a strictly additive model. In other words, there is a dominance deviation (symbolized as d) because the heterozygotes do *not* fall at the midline (represented as 0) of the phenotypic values.

Figure 3.6: Genotypic Effects on Phenotypic Values When Dominance Deviation is Present



V_I: Epistatic Genetic Factors

Epistatic genetic factors (V_I) represent the third type of genetic influence on phenotypic variance. Epistasis refers to inter-gene interaction, meaning that the *effect* of a gene on a phenotype is contingent on the presence of a second (or more) gene also being present in the genome. Epistatic influences, therefore, are commonly referred to as gene X gene interactions ($G \times G$) and they can be considered consistent with the statistical meaning of the word "interaction." Epistasis is defined as, "the deviation from additive combination of the genotypic values" (Falconer & MacKay, 1996: 119) **CONFIRM QUOTE and PAGE number**. In other words, epistasis allows for situations where the presence of gene B changes the influence of gene A on the phenotype.

As was mentioned above when introducing additive genetic factors, there are two ways of calculating heritability. Narrow-sense heritability, as was shown, considers only the additive genetic component and calculates the proportion of phenotypic variance that is attributable to variance caused by additive genetic factors. Broad-sense heritability captures *all* genetic effects on the phenotype of interest, including those due to V_A ,V_D , and those due to V_I . Thus, broad-sense heritability can be defined mathematically as:

$$\text{broad-sense } H^2 = \frac{V_G}{V_P}$$

$$= \frac{V_A + V_D + V_I}{V_P}$$

Consistent with Zuk et al. (2012), we use a capital H to distinguish broad-sense H^2 from the more limited narrow-sense h^2 that was discussed above.

3.3.3 Environmental Influences

V_C: Shared Environment

Quantitative geneticists often start with the assumption that phenotypic variance is influenced by genotypic variance, plus an environmental deviation. The environmental deviation is thought to capture all non-genetic influences on phenotypic variance. This means that the environmental components need not be restricted to the types of environments that criminologists usually study such as parents, peers, and neighborhoods. Instead, the "environment" to a quantitative or behavioral geneticist would capture everything from the prenatal environment to exposure to toxins in drinking water.

The foundational assumption that phenotypic variance results from genetic factors and an environmental deviation has led many scholars to misinterpret quantitative and behavioral genetic research as being "gene-centric", with the environment being ignored or taking a backseat to the genetic influences. As you will see, however, nothing could be further from the truth. Quantitative and behavioral geneticists have spent much time conceptualizing and quantifying the influence of environmental factors on phenotypic variance. As a direct result of these efforts, an important discovery about the "types" of environmental influences emerged. Specifically, quantitative and behavioral geneticists soon began to realize that environmental factors could be classified as shared environmental influences and nonshared environmental influences. The nonshared environment is discussed below. The shared environment operates to make individuals more similar to one another. As you might imagine, shared environments can come in a variety of forms, ranging from a shared prenatal environment (such as for twins) to living through a major historical event like the terrorist attacks that struck Manhattan island on September 11, 2001.

V_E: Nonshared Environment

Finally, the last component affecting phenotypic variance is referred to as the nonshared environment. The nonshared environment captures all environmental (read: non-genetic) influences that make two people dissimilar. With this in mind, it is quite easy to understand why the nonshared environment is often conceptualized as capturing unique environmental effects that operate to make people different. Examples of nonshared environments are easy to envision, as long as we realize that any experience or exposure that is unique to one person will act as a nonshared environmental source of variance when that person is compared to someone else. Thus, nonshared environments might highlight that two people have different peer groups, that they lived in a different neighborhood, or even that stochastic events put one on a different life-course trajectory than the other (see generally, Sampson & Laub, 1993).

3.3.4 Gene-environment Interplay: Gene-environment Correlation (rGE)

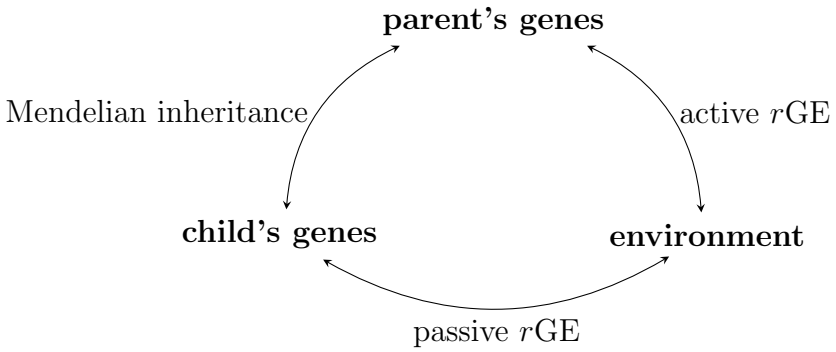
Now that the various sources of genetic and environmental influences have been outlined, it is important to discuss the ways in which genes and environments may combine, covary, and interact to affect phenotypic variance. There are two main types of gene-environment interplay that are important to keep in mind as you move through this text: 1) gene-environment correlation and 2) gene-environment interaction. The issues that underlie both types of gene-environment interplay is the realization that genetic influences do not work in a vacuum absent the influence of the environment. On the contrary, genetic factors influence phenotypic development alongside environmental sources of variance. Recall the two distributions that were presented in Figure 3.5. We noted that researchers will typically observe something like the distribution in the right panel of Figure 3.5, meaning the researcher will only observe phenotypic variance that includes genetic and environmental influences. The goal of the researcher is to "pull apart" the different sources of variance. But this assumes that genetic influences can be neatly cleaved from the environmental influences. In other words, up to this point, we have assumed that the genetic component is unique and separately estimable alongside the environmental component. But what happens if these two influences covary or if they interact? For example, what happens if the additive genetic factors that affect phenotypic development also make it more likely that one will be exposed to nonshared environmental influences that also impact phenotypic development? When this sort of situation occurs, we observe a phenomenon known as gene-environment correlation (rGE).

There are three types of rGE. The first is known as active rGE. Active rGE occurs when genetic factors influence a person's choice about environmental exposures. Keep in mind that the genetic influence is an indirect influence that most likely will be mediated by personality characteristics that are unique to each person. Nonetheless, active rGE can be thought of as the vehicle that drives a person's self-selection into one environment but not another. For example, imagine two people, person A and person B. Let us say, for the sake of demonstration, that person A's genotype makes him/her more likely to enjoy thrill-seeking activities. Person B, however, inherited a suite of genetic markers that make him/her more likely to experience anxiety at the very thought of engaging in a dangerous activity. We might expect, as a result of the genetic predispositions that affect personal preferences, that person A will be more likely than person B to ride motorcycles, to try skydiving, and to enjoy roller-coasters.

The second type of rGE is known as evocative rGE. Evocative rGE occurs when one's genetic predispositions affect the way in which the environment responds to that individual. In other words, genetic factors that drive one's personality may make it more (or less) likely that s/he will be treated in a certain way by others in his/her social group. Individuals who are funny, intelligent, and/or attractive tend to be well liked by others (Harris, 2009). Thus, the genetic factors that influence personality development (along with other phenotypes) will impact the way in which that person is received and treated by the environment. Put

differently, the genetic factors can evoke certain responses from the environment.

The third way in which genes and environments can covary is referred to as passive rGE. Passive rGE occurs when parents pass along genetic influences to their children that correlate with the environmental factors that are also passed along to their children. For example, parents who are athletic (a phenotype that is at least partially influenced by genetic factors) will tend to raise their children in an environment that is conducive to participation in athletic endeavors. Intelligent parents tend to surround their children with lots of books. Creative parents surround their kids with a variety of ways to stimulate their imaginations. And so on. What is interesting about the passive rGE, though, is that it requires active rGE to explain the causal pathway between the parents' genes, the child's genes, and the environmental exposures. Indeed, the causal pathway can be displayed graphically as it is in the diagram below. Notice that active rGE connects the parents' genetic influences to the environmental influences. Mendelian inheritance explains the correlation between the parents' genes and the child's genes. Finally, passive rGE explains the correlation between the child's genes and the environmental influences.



Although rGEs are important to consider from a theoretical standpoint, it is also imperative that they be included in any model that seeks to explain phenotypic variance. Indeed, anytime two sources of variances are summed, any overlap that they share (i.e., any covariance between the two) will also show up twice (because their overlap is present in both sources of variance) in the resulting value. When placed in the present context, this point means that our model of phenotypic variance must also account for any covariance(s) between the constituent items:

$$\begin{aligned} V_P &= V_G + V_E \\ &= V_A + V_D + V_I + V_C + V_E + 2(cov_{A,D}) + ... + 2(cov_{C,E}) \end{aligned}$$

As you might imagine, the equation for phenotypic variance can become unwieldy quite quickly. Fortunately, we can make certain assumptions based on logic and/or prior research to simplify the model. Specifically, as Barnes and colleagues (2014) noted, we can assume that most of the genetic factors do not covary, therefore striking any covariance statement that includes two genetic components. Also, we can safely assume that shared (V_C) and

nonshared (V_E) influences do not covary. Behavioral geneticists typically make certain assumptions that allow this equation to be simplified to include only additive genetic factors (V_A), shared environmental factors (V_C), and nonshared environmental factors (V_E) (see Barnes et al., 2014; and see the assumptions and limitations sections of the following chapters):

$$V_P = V_A + V_C + V_E$$

3.3.5 Gene-environment Interplay: Gene-environment Interaction ($G \times E$)

There are various ways the word "interact" can be used in a colloquial sense. Humans interact with one another during a conversation. A chemist looks for just the right level of chemical reactions and interactions to produce his latest drug. And it is often said that genes and environments interact to produce human behavior. Although the same term is used in various ways, only the last use of the term "interact" will be relevant for discussing gene-environment interactions ($G \times E$). $G \times E$ refers to a very specific relationship between genes and environments. It does not mean that genes and environments contribute equally to affect the phenotype. In fact, it may be the case that one or the other has a vanishingly small main effect on phenotypic variance. When quantitative and behavioral geneticists refer to $G \times E$, they are referring to a very specific relationship. Specifically, $G \times E$ captures the influence of an environment (gene) on the *effect* of the gene (environment) on the phenotype. In other words, $G \times E$ tells us that the *effect* of a gene (or an environment) may depend upon the presence (or level) of an environment (or a gene). The best way to demonstrate this point is with the visual example provided in Figure 3.7.

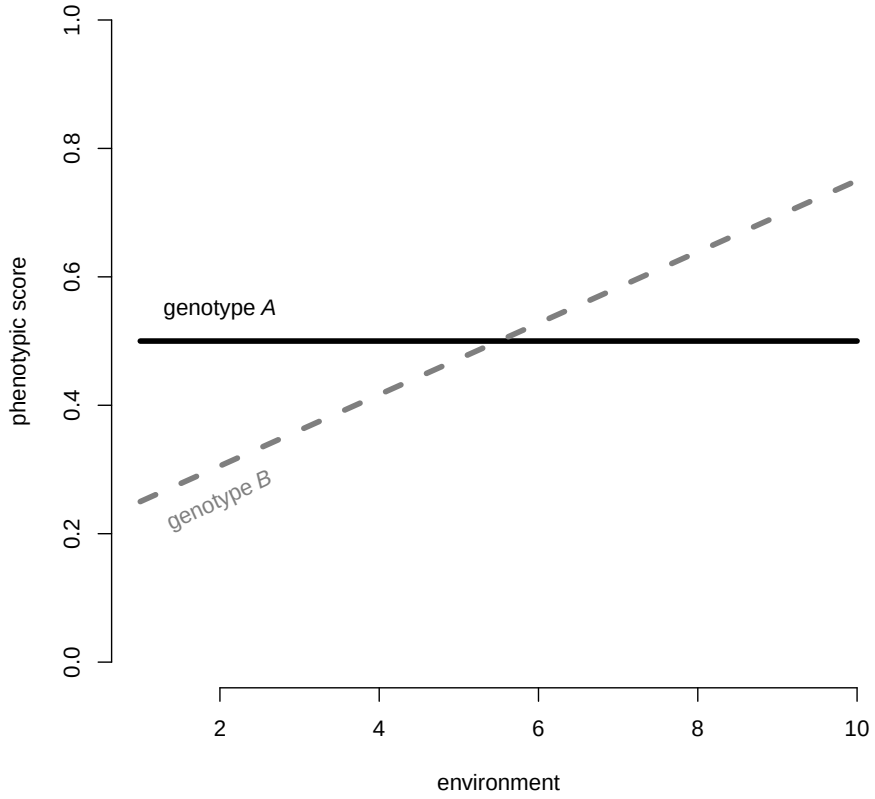
Two genotypes (A and B) are presented in Figure 3.7. These heuristic data reveal a crossover interaction where the environmental influence on the phenotype is inconsequential for those who carry genotype A. Yet, for those who carry genotype B, the environmental influence is strong and positive. Not all $G \times E$s will look like the one displayed in Figure 3.7. Indeed, rarely will the distinction be so dramatic. But, the important point to draw from this discussion is that $G \times E$s reveal that the impact of the environment (or the gene) on phenotypic variance is contingent on the presence of the gene (or the environment).

As with the discussion of rGE, the equation for the phenotypic variance must be expanded whenever $G \times E$s are present:

$$V_P = V_A + V_C + V_E + V_{A\times C} + V_{A\times E} + 2(cov_{A,A\times C}) + \cdots + 2(cov_{E,A\times E})$$

Whether $G \times E$s are prevalent enough to include in the phenotypic variance equation will be discussed in more detail in Chapter X. For now, we will assume that, on average, the net effect of all $G \times E$ is zero, meaning they can be safely omitted from the phenotypic variance equation. This assumption is, however, tentative.

Figure 3.7: A Graphical Display of Gene-Environment Interaction ($G \times E$)



3.4 Conclusion: A Working Model of P and V_P

We now have all the components necessary to understand phenotypic development. Recall that our original model began simply with $P = G + E$ when we are concerned with a *phenotypic score* and it was adjusted to $V_P = V_G + V_E$ when we are concerned with *phenotypic variance*.

But we now know—after having reviewed the information in this chapter—that the *actual* models must include or assume away many more elements. In fact, our fully saturated models look like:

$$\begin{aligned} P &= G + E + rGE + G \times E \\ &= A + D + I + C + E + rAD + \ldots + rCE + A \times D + \ldots + C \times E \end{aligned}$$

In short, one must alter the simple $P = G + E$ to account for arbitrary gene-gene and/or gene-environment interplay. To do so, but to keep the task manageable, we will allow the following to represent a generalized version of the phenotypic score equation (see Golan et al., 2014; Zuk et al., 2012):

$$P_i = \Psi(G_i, E_i)$$

where Ψ represents any arbitrary function between the genetic components G and the environmental components E for person i. This might include—but is not limited to—gene-gene interactions ($G \times G$), gene-environment interactions ($G \times E$), and gene-environment correlations (rGE).

Adjusting our equations to account for phenotypic variance (V_P) is also straightforward if we imagine P is a normally distributed trait where $P \sim N(0, V_P = 1)$. Here, the variance in P (i.e., V_P) can be thought of as arising from V_G (and, by extension, V_A, V_D, V_I, V_C, and V_E), V_E (and, by extension, V_C and V_E), and their interplay (i.e., rGE and $G \times E$). From this simple thought experiment, we can build on the classical principles of molecular genetics (e.g., Mendelian inheritance, dominance, and recessivity) to develop a mathematical model that moves us from phenotypic scores P to variance in P V_P:

$$\begin{aligned} V_P &= V_G + V_E + 2cov(V_G, V_E) + V_{G\times E} \\ &= V_A + V_D + V_I + V_C + V_E + 2cov(V_A, V_C) + \ldots + 2cov(V_I, V_E) + V_{A\times C} + \ldots + V_{C\times E} \\ &+ 2cov(V_A, V_{A\times C}) + \ldots + 2cov(V_E, V_{C\times E}) \end{aligned}$$

Of course, it is probably obvious that one cannot hope to ever estimate such a model. There are far too many parameters and there would scarcely be enough statistical power. Estimating such a model would become unmanageably complex. Thus, again, we can present a simplified model that still captures all the elements outlined above:

$$V_P = \Psi(V_G, V_E)$$

where, like before, Ψ represents an arbitrary function that can account for the genetic and environmental components, as well as any interactions/correlations that may exist between the individual elements.

We will seek to simplify matters by compartmentalizing and dealing with certain parts of the model(s) individually. This too requires some assumptions, so we will highlight the assumptions that are being made with all of the models that are presented in the following chapters. Typically, we will be forced to assume that certain parameters from the models above are zero, thus allowing us to estimate one part of the model assuming the other part(s) has minimal-to-no influence. Whether these assumptions are violated and the biasing effect that those violations may have will be the focal point of our discussion at the end of each model description.

The modeling chapters will consider/estimate different configurations of our fully saturated model. For example, the models discussed in chapter 5 will estimate the degree to which V_A, V_C, and V_E explain V_P. The methods in chapter 7 will attempt to unpack the A (or D or I) components of the phenotypic score model by looking for relationship between specific genes and P.

Chapter 4

How Do Genes Influence Human Behavior?

This chapter re-introduces many of the concepts that were covered in the previous chapter, but this time the focus is much more conceptual—in other words, the goal of this chapter is to give the reader a more intuitive feel for how genetic variation might impact human behavior. Toward this end, we will occasionally refer the reader to relevant sections from chapter 3. But it is not necessary to have read chapter 3 to follow the material presented here because in this chapter, we rely on logic and intuition to help guide us to the plausible mechanisms that link genetic variation to variation in human outcomes.

There tends to be a lot of confusion regarding the role of genes in the development of behavior among social scientists, particularly among social scientists lacking any formal training in genetics, biology, and neuropsychiatry. Much of this misperception, fortunately, can be ameliorated by learning some basic principles about genetics, human biology, and the mechanisms by which genetic variation is likely influencing phenotypic variation. For the most part, the confusion stems from an overly simplistic account of how genes influence behavior, wherein the path from gene to phenotype is seen as being straightforward, direct, and la*r*GEly deterministic. Such a view is completely unfounded because the nexus between a gene and a phenotype is highly complex, conditional, and probabilistic.

There is little doubt that as more research is published that examines the genotype-phenotype association, a clearer understanding of this association will eme*r*GE. There are some key findings that have eme*r*GEd from the literature that have been used as the foundation for different models and explanations highlighting the likely ways that genes influence behavior. Keep in mind that these are working models that may need to be adjusted in the future based on newer findings. What is more is that these models are phenotypic dependent and that multiple models may apply to the same phenotype. The models and explanations outlined below and in the rest of this chapter will help to add "flesh" to the statistical techniques introduced later in the book. They will also allow for research questions to be generated and for a helpful crutch to be available when interpreting the results of empirical

studies. In short, they provide a nice starting point for understanding the numbers produced from quantitative genetic analyses and for building theories of human behavior that incorporate genetic, biological, and environmental processes.

4.1 Monogenic, Polygenic, & Pleiotropic Effects

The link between a genotype and a phenotype is generally considered to be quite complex and often includes environmental moderators, feedback loops, and intermediary processes. Later on in this chapter (and book) we will discuss some of these complex processes in greater detail, but for now we want to focus on three relatively straightforward mechanisms that lead directly from genotype to phenotype.

4.1.1 Monogenic Effects

The first mechanism is known as a monogenic model (or sometimes referred to as OGOD [the acronym for one gene, one disorder]) and this model presupposes that each phenotype is caused by variation a single gene.

There are two main types of monogenic influences: dominant and recessive. Dominant monogenic effects occur when there is a single allele that, if inherited, causes the phenotype to occur. Recessive monogenic effects, in contrast, occur when two alleles must be inherited for the phenotype to emerGE. Inheriting only one allele for these recessive phenotypes will not cause the phenotype to develop and, in fact, inheriting only one allele for recessive monogenic effects is often advantageous as it can act as a protective factor against diseases. For example, sickle-cell anemia is a monogenic recessive disorder, but inheriting only one sickle-cell allele protects against malaria. With a monogenic effect, the allele (or alleles) is a necessary and sufficient condition for the phenotype to develop. If a person inherits the allele(s), then the phenotype will develop 100% of the time; if the allele(s) is not inherited, then that particular phenotype will never develop. This discussion aligns with the Single Gene Model discussion in section 3.2.1 of chapter 3.

There are thousands of human phenotypes that are caused by monogenic patterns of inheritance, including Huntington's disease, Fragile X syndrome, and Cystic Fibrosis. Although monogenic effects account for a substantial number of lethal disorders, when it comes to complex human phenotypes (e.g., personality traits and behaviors), monogenic influences are thought to not be all that important. Most scholars, in fact, dismiss the possibility that any behavior or personality trait could be the result of a monogenic influence. This viewpoint is not entirely correct as there have been isolated instances where monogenic effects on behaviors have been detected. Perhaps the most noteworthy example was discovered by Hans Brunner (Brunner et al., 1993). In the early 1990s, Brunner and his research team genotyped a Dutch kindred, wherein certain males displayed signs of cognitive impairments,

were characterized as having impulse control deficiencies, and engaged in serious acts of violence—collectively referred to as Brunner syndrome. Females in the family appeared to be immune to this constellation of behaviors. As a result, Brunner and his associates posited that these behaviors were caused by a mutation to a single gene located on the X chromosome.

Why did they hone in on the X chromosome? Because males possess only one X chromosome whereas females possess two X chromosomes. What that means for monogenic patterns of inheritance is that if males inherited a mutated allele for the gene (on the X chromosome), then they would not have an alternative allele that would be able to compensate for it. If, however, a female inherited a mutated allele for the gene (on the X chromosome), then they would have another copy of that gene on their other X chromosome. In this way, the "back-up" copy might compensate for the deleterious effects of the mutated allele. Males do not have back-up copies for X-linked genes and so if they inherited the mutated allele, they would ultimately develop Brunner syndrome.

Brunner's genetic analysis confirmed his suspicions: this syndrome was traced to a mutation on the monoamine oxidase A (MAOA) gene which is located on the X chromosome. MAOA codes for the production of the enzyme, monoamine oxidase A (MAOA), which is responsible for the degradation of neurotransmitters, such as serotonin, norepinephrine, and dopamine. It plays a critical role in the modulation of normal brain function. Not only is the effect of MAOA on neurobiological functioning borne out in genomic-imaging studies (more on this later in the chapter), but it is also highlighted by the fact that a class of antidepressants (known as monoamine oxidase inhibitors [MAOIs]) get their pharmacological properties by manipulating MAOA activity.

The mutation that Brunner discovered resulted in an MAOA gene that was unable to produce the MAOA enzyme. Males in this kindred with Brunner Syndrome, therefore, had inherited a mutated MAOA allele and thus did not produce any MAOA. Although Brunner syndrome is one example of a monogenic effect on antisocial behavior and maladaptive traits, this mutation has only been detected in only a handful of families (Palmer et al., 2016). This should not be all that surprising because most human phenotypes are much more complex than can be accounted for by a monogenic pattern of inheritance. Other genetic models of behavioral influence, however, are more in line with the way in which genes are thought to affect phenotypic variance. Perhaps the most noteworthy of these models is known as the polygenic model (see also section 3.2.2 in chapter 3).

4.1.2 Polygenic Effects

With a polygenic model, phenotypic variance is accounted for by multiple genes, usually thought to include hundreds or thousands of genes. Each of the genes works in a probabilistic fashion with each gene increasing (decreasing) the probability that a phenotype will occur or with each gene increasing (decreasing) the level of continuous phenotypes. Although the genes involved in a polygenic phenotype tend to have relatively small effects, cumulatively

they account for a la*r*GE percentage of phenotypic variance. To illustrate, consider a polygenic phenotype that is accounted for by 200 genetic polymorphisms. If the phenotype was entirely genetic, and if each gene had an equal influence on the variance, each gene would account for 0.5% of the variance. Clearly the effect of each gene is quite small, but when aggregated these genes account for 100% of the variance. Phenotypic variance is rarely 100% due to genes, meaning that environmental factors matter, too, and that, as a result, single genes likely have even smaller effects (because a la*r*GE number of genes is accounting for $< 100\%$ of the variance [see also section 3.3 in chapter 3]). The average effect of each gene will decrease as 1) the total amount of variance accounted for by genes decreases and/or 2) as the total number of genes involved in the polygenic model increases.

Most human behaviors and traits, particularly those studied in the social sciences, are thought to be created, in part, by polygenic influences. Take a look at genetic effects on height. In one study, more than 400 genes were found to influence height, meaning that on average each gene accounted for a small fraction total height (Wood et al., 2014). Other studies have revealed similar results for most other social science phenotypes, such as intelligence, criminality, and obesity. Keep in mind that with a polygenic model, genes are not necessary or sufficient conditions for a phenotype to eme*r*GE. Each gene simply works probabilistically and thus there is nothing deterministic or fatalistic when it comes to the link between genotypic variance and phenotypic variance under a polygenic model.

4.1.3 Pleiotropic Effects

The last model by which genes directly affect phenotypic variance is known as the pleiotropic model. This model is best understood by viewing it as the polygenic model turned on its head. Rather than each phenotype being caused by multiple genes, with the pleiotropic model each gene accounts for phenotypic variance on multiple phenotypes. The number of phenotypes that each gene accounts for varies across genes and phenotypes and the total amount of variance accounted for by each gene also varies across genes and phenotypes.

Pleiotropic effects are of particular interest to social scientists because they hold the potential to explain how and why certain phenotypes are almost always interrelated. Attention deficit hyperactivity disorder (ADHD) and criminal involvement, for instance, have been found to be consistently related across studies. The precise mechanisms that account for this association are not well understood. Explanations range from ADHD causing crime to ADHD and crime being nothing more than a spurious association caused by socioeconomic status. To date, there has not been a real serious effort to examine whether ADHD and crime covary because they share some of the same genetic underpinnings. This is not unique to the ADHD-crime connection; there is virtually no social science research examining whether pleiotropic effects could account for well-documented phenotypic associations. Findings from candidate gene studies provide some evidence that pleiotropic effects are likely to be widespread in the social sciences. For instance, alleles of a dopaminergic polymorphism (DAT1) have been found to be related to age at first arrest, alcohol consumption,

general delinquency, and sexual involvement (Beaver, 2016). It should also be noted that results from bivariate and multivariate genetics models also reveal substantial support for pleiotropic effects (see Chapter 6).

Although these three models are often pitted against each other, the polygenic and pleiotropic patterns of inheritance are not necessarily mutually exclusive. A single polymorphism, for instance, could very well account for a small percentage of variance in two behavioral phenotypes. And these two behavioral phenotypes could certainly be affected by hundreds of polymorphisms. Monogenic effects, in contrast, are certainly incompatible with a polygenic model and are most likely are incompatible with a pleiotropic model, too. It is unlikely (though possible) that a single allele from a single polymorphism could account for 100% of the variance in two (or more) phenotypes. Given that monogenic effects are quite rare for even a single behavioral phenotype, it is unlikely that a single allele would cause two phenotypes to emerGE (though it may account for multiple phenotypes of the same syndrome [e.g., mental abilities and violent behaviors that are part of Brunner syndrome]).

4.2 Gene-Environment Interplay

The aforementioned models outlining how genotype could cause phenotypic variation capture the direct effects between the gene and the outcome. Given that these three models focus exclusively on the direct route between genotype and phenotype, these models are unable to capture all of the potential complex pathways leading from genotype to phenotype, particularly those that involve environmental effects. Findings from research studies underscore that the environment is important and thus needs to be integrated into studies interested in estimating genetic influences. There are two key ways in which environments and genes combine together to produce phenotypic variance (a third way is epigenetics and will be discussed briefly in Chapter ??): gene-environment interaction ($G \times E$) and gene-environment correlation (rGE). $G \times E$s capture the process by which genetic effects interact with environmental effects to produce phenotypic variance. With $G \times E$s, the amount of phenotypic variance accounted for by the independent effects of genes and environments is greater than the sum of its parts (genetic [G] + environmental [E] effects does not capture all of the observed variation). Stated differently, this means that there is something about G and E that when they intersect, their influence increases exponentially.

What is often overlooked in discussions of $G \times E$s is how they can be used to specify the conditions under which genes or environments ultimately influence phenotypes. Most social science research, for instance, posits that certain environmental conditions cause variation in behaviors and even personality traits. For the most part, environments tend to have relatively small effects on phenotypes and numerous explanations have been employed for why this is the case. Perhaps it is due to measurement error, the inability to operationalize complex social processes, or even because environments are supposed to have very small effects. Another possibility that is typically overlooked is that genotype is responsible for conditioning responses to the environment. In this case, not all people react to and are

affected by the same environmental stimuli in the same way. Individualized responses are due, at least in part, by unique genotypes, such as when two people who experience the same environment and yet react to it in very different ways. Without taking into account G, the effect of E is attenuated to the point where it might not even reach conventional levels of statistical significance. If the dual effects of G and E were examined, however, the conditions under which E might exert an influence could be realized (e.g., E matters for people with certain genotypes, but not for people with other genotypes). Seen in this way, $G \times E$s have the potential to increase explanatory power and to add greater specificity to social science explanations of phenotypes.

The opposite is also true: $G \times E$s can also provide insight into why certain genes do not fully operate in the same way across all people. Some people might possess a specific allele and it confer no risk to developing a phenotype and others might have that same allele and it confers a great risk. Once again, without taking E into account, the effects of G will be attenuated because the effect of G might be governed, in part, by exposure to certain Es.

To date, there has been a considerable amount of research devoted to examining $G \times E$s, but perhaps the most well-known example of a $G \times E$ as it relates to behavior comes from a study published by Caspi and his colleagues. In this study, Caspi et al. (2002) were interested in examining why childhood maltreatment is related to an increase in antisocial behavior later in life, but why most maltreated children do not turn out to be antisocial—that is, why there is variation in how children are affected by maltreatment. This team of researchers reasoned that a polymorphism in the MAOA gene (the same gene that was involved in Brunner syndrome, but a different polymorphism) might account for why some males are susceptible to the adverse effects of maltreatment, but others are resilient against it. To test this possibility, they analyzed data from a sample of males from Dunedin and the results confirmed their hunches: the effect of maltreatment on antisocial phenotypes only occurred for males who possessed a particular MAOA allele; childhood maltreatment did not have an effect on antisocial behaviors for males who possessed a different MAOA allele. This effect of the MAOA $\times$ maltreatment interaction was actually quite strong as only 12% of the sample had the MAOA allele and were maltreated, but they accounted for 44% of all violent crime convictions.

$G \times E$s are typically detected by testing for statistical interactions between G and E (see Chapter 9). Recently, however, there has been a great deal of interest in trying to figure out the mechanisms that account for $G \times E$s. Two of the most popular explanations are the diathesis stress model and the differential susceptibility model. The most conventional way to explain $G \times E$s is through the logic of the diathesis-stress model. According to this model, different genotypes create different propensities for certain behaviors and traits. These different predispositions often are nothing more than resting potential, needing some type of trigger to set them in motion. Without the trigger, the potential will remain dormant. The diathesis-stress model purports that the trigger is some type of environmental stimuli. When the environmental stimuli is present, the genetic predisposition emerGEs; without the environmental stimuli present, the genetic effect remains dormant.

Up until a few years ago, the diathesis-stress model was the most popular model used to interpret and explain $G \times E$s. Recently, an alternative to the diathesis-stress model has been advanced by Jay Belsky with what he calls the differential susceptibility model (Belsky & Pluess, 2009). With this model, the belief is that genes do not necessarily create predispositions for certain phenotypes, but rather dictate how susceptible a person is to environmental effects. Instead of viewing alleles as creating a risk for some type of phenotype, these alleles are seen as measuring the amount of "plasticity" of each person; the more plastic that they are, the greater the likelihood that environments will affect them. As a result, persons who possess more plasticity alleles will be equally affected by all environments whether they have positive or negative influences. Alternatively, persons who possess relatively few plasticity alleles will be immune to all environmental effects, including ones that are positive and ones that are negative. Stated differently, the most plastic individuals will turn out the most positive when exposed to positive environments and they will also turn out the most negative when exposed to negative environments. That is why the differential susceptibility model is often referred to as having a "better-or-for-worse" effect on behaviors (they turn out the best in good environments and the worst in bad environments).

Ever since Belsky advanced the differential susceptibility explanation, there has been a great deal of interest in pitting diathesis stress against differential susceptibility and in developing newer techniques for doing so (e.g., Roisman et al., 2012). Research has revealed support in favor of both explanations. If these mixed findings continue to be replicated, then they would suggest that $G \times E$s are not exclusively due to diathesis stress and they are not exclusively due to differential susceptibility; rather, $G \times E$s are likely the result of differing combinations of these explanations (and perhaps others) that might vary across samples, phenotypes, and genes.

As with other aspects of genetic influences, there is confusion regarding the contributions and limitations of $G \times E$s. While $G \times E$s certainly have made some significant advancements in the understanding of human behavior, all too often there is a widespread belief that $G \times E$s are so pervasive and so ingrained in the etiology of human behavior that it makes it impossible to study G or E in isolation. In fact, some social scientists have gone so far as to say that since all human phenotypes are the result of $G \times E$s, dividing phenotypic variance into a genetic and environmental components is conceptually misguided.

To understand this criticism, consider an example with LeBron James. LeBron James is arguably one of the most accomplished professional basketball players in the history of the National Basketball Association (NBA). If we were to try to figure out why he is so talented, we could point to his environment, his hardwork, dedication, and practice routine. We could also point to his genetics, his hand-eye coordination, height, and muscular stature. But if we tried to divide his abilities into a genetic component (say, 50%) and an environmental component (say, 50%), we can see why this is a futile thought exercise. After all, if we removed the environment (e.g., no exposure to basketball, no practicing, etc.), then LeBron James would not be an NBA star. Likewise, if we removed the genetic gifts from LeBron James, then he would also not be an NBA star. For this reason, that is why some scholars argue that in order to understand any human phenotype—ranging from criminal involvement

to IQ to athletic talents—$G \times E$s must be modeled directly and estimating the effects of G and E separately should be abandoned.

At first glance, all of this seems to be commonsensical and focusing on $G \times E$s is the only logical way to study genetic (and, for that matter, environmental) effects. The problem with this approach is that in practice social scientists do not focus on the outcome of single individuals; rather, the interest is in explaining variance in a sample of people. What that necessarily means is that statistical techniques are designed to explain why people vary, not phenotypic scores for a single person (e.g., LeBron James). Applied to this example, the interest for social scientists would be in determining why some people excel at basketball and others do not. The focus is on between-person differences, not on the development of basketball skills for a single individual.

Behavioral geneticists have thoroughly addressed the issue that $G \times E$s are so omnipotent and so widespread that it is impossible to partition variance into separate genetic and environmental components using what Judith Rich Harris calls the "damned rectangle" example (Harris, 2006). More than a decade ago, Harris laid waste to the criticism of separating variance into G and E still exists, but it is worth highlighting once again. Consider trying to determine the area of a rectangle by estimating the percentage that is due to the length (l) of the rectangle and the percentage that is due to the width (w) of the rectangle. Clearly, this makes no sense as for an area to exist, the rectangle must have both a length and a width; remove one or the other, and the area ceases to exist. When the logic of this example is applied to phenotypic variance it highlights how all phenotypic variance (much like the area of a triangle) is due to $G \times E$s and it is a futile exercise to estimate G (length) and E (width) separately.

Recall that social scientists do not attempt to explain individual scores, but rather are interested in explaining variance. Referring back to the rectangle example, it certainly does not make sense to divide the area of a single rectangle into l and w. However, what if we included a sample of rectangles where they all had varying areas. The variation in area would be a function of differences in the length and width across rectangles; some might be relatively wider and some might be relatively taller. The point is that the variation in the area of the rectangles is due to variation in the length and width for each one. Now, we are interested in trying to explain variation in the area of rectangles by examining the l and w of each rectangle. When viewed in this way, it is easy to see that we could figure out the percentage of the variation in area that is due to variation in length and that is due to variation in width. This is akin to estimating phenotypic variance (not individual scores) by determining the percentage of the variance that is due to genetics and that percentage that is due to environments. When critics argue that there is no sense in separating genetic from environmental influences, they are confounding the difference between explaining unique traits for each person with explaining the variance that exists across an entire sample. The latter is what social scientists are interested in accomplishing and thus the criticism regarding the separation of genetic and environmental influences is entirely erroneous.

The second way in which genes and the environment work in unison is known as gene-

environment correlation (rGE) (Jaffee & Price, 2007). rGE describes the mechanisms by which genes and the environment are correlated with each other. Three types of rGEs have been identified: passive, active, and evocative. Passive rGE seeks to explain the close connection between rearing conditions and genotypic predispositions. To do so, it is important to understand that children receive both their genes and their environment from their parents. Given that their genes and their environment are tied to the same source (i.e., their parents), it is likely that the two will be correlated. To illustrate, suppose that a biological parent is depressed. Given that depression is heritable, their child will be at increased genetic risk for also developing depression. At the same time, the depressed parent is also at-risk for providing an environment that is linked to depression, such as being a cold, withdrawn, and uninvolved parent. The end result is that the child's genetic predisposition (for depression) is correlated with the environment that they are being reared (that is, an environment that is conducive to depression). Passive rGEs are thought to be the most salient early in life when the child is most likely to encounter environments provided by their parents. Later in life, when the importance of other socialization agents take over, and as they gain autonomy and can escape the constant surveillance of their parents, passive rGEs tend to wane in significance.

Although passive rGEs provide some insight into the genotype-environment correlation early in the life course, they have important methodological implications for studies of family and parental influences. A long line of social science research attempts to uncover the potential association between family and parental factors and the way in which they might be tied to a broad swath of child outcomes, ranging from educational achievement and employment success to criminal involvement and drug use. Virtually all research that examines the parent/family-child outcomes nexus fails to account for the potential of genetic confounding (Harris, 1998). This is a serious mistake as it leaves open the possibility that any detected association between parenting and family life and the development of the child could be the result of shared genetic influences (Pinker, 2002).

For instance, suppose a researcher is interested in determining whether serious physical abuse against a child (by their parents) increases the likelihood of the child becoming a violent criminal later in life. In all likelihood the findings would reveal that children who were abused are more likely to develop into criminals when compared to children who were not abused. Although most social scientists would interpret such a finding as evidence that is consistent with a causal explanation, there is another explanation that is equally consistent with the evidence: that parents who physically abuse their children are also passing along a genetic tendency to their children to be violent and aggressive. Without controlling for genetic influences, it is impossible to tease apart these two explanations. Anytime that a passive rGE could be involved in studies that are attempting to isolate the effects of specific environmental influences, it virtually requires some type of methodology and statistical technique that is able to remove any shared genetic variance between the phenotype of interest and the environment. Failure to do so will likely lead to upwardly biased parameter estimates for the environmental influence (Barnes et al., 2014) and incorrect interpretations of the causal influence of environmental effects (Harris, 1998).

The second type of rGE is known as an active rGE. Active rGE accounts for the correlation between genetic influences and environments by focusing on the fact that people construct their own lives by the power of their choice. In most contemporary societies, for instance, people are able to select virtually all of their environments, ranging from whom to marry and who to befriend to whether they want to have children and which neighborhood to reside. Given that the choice of one environment over another is governed, in part, by genetically influenced traits (e.g., personality, temperament, cognitive abilities, etc.), environmental variation will be accounted for, in part, by the genetic influences that operate via such traits.

Essentially active rGE is attempting to account for why there is environmental variation and why environmental variation is not randomly distributed. In the social sciences, this is often referred to as self-selection. Active rGE is in the same vein as self-selection except that that self-selection typically focuses on traits (e.g., self-control) and behaviors (e.g., violent criminal acts) that propel people into or away from certain environments whereas active rGE focuses on genetic influences that undergird these phenotypes. Active rGE and self-selection are not incompatible with each other; rather, active rGE examines genetic influences that are antecedent to any traits that might be involved in the self-selection of environments. Seen in this way, active rGEs represent a threat to social causation arguments if the genes that might account for self-selection into environments are not accounted for accurately. Once again, failing to include a research design capable of directly modeling active rGE could result in conclusions that are misleading.

The last type of rGE is known as an evocative rGE. Evocative rGE refers to the fact that people often elicit specific responses from their environment based on their personality traits or the way in which they behave. These traits and behaviors are genetically influenced and thus the response from the environment originated from genotype. For example, take a person who is extremely aggressive, a trait that is known to be quite heritable. Highly aggressive persons are likely to elicit negative reactions from their environment, such as losing their jobs, being suspended from school, and being involved in serious physical fights on a consistent basis. Variation in these environments (e.g., being fired from a job) are likely to be accounted for by genetic factors because the traits causing environmental variation (e.g., being aggressive) are under genetic influence.

Evocative rGE highlights the complexity of establishing the causal direction between environmental variation and phenotypic variation. In the social sciences, most theories and explanations are designed to identify potential causal environments that are antecedent to the outcomes of interest. In order to account for the appropriate temporal ordering, scientists typically employ some type of longitudinal data in an attempt to ensure that environmental conditions predate the outcome. Evocative rGE shows that if genotype is not fully modeled into the research design, then any conclusions regarding the causal impact of environments on phenotypes are untenable.

All three rGEs provide a more complete understanding of the close nexus between genes and environments. Of course, the question remains whether there is any evidence finding

support for rGEs. In order to address this possibility, it is first necessary to recognize that rGEs are frequently tested for by estimating a traditional twin-based model (see chapters 5 and 9), wherein the environment is used as the outcome measure (as opposed to a traditional phenotype). The variance in the environmental measure is then decomposed and if the heritability estimate is significantly greater than zero, then that parameter estimate would indicate that environmental variation is accounted for by genetic variation; that is, there is evidence in favor of rGE.

A relatively la$r$GE body of research has estimated the heritability of environmental variance to examine the merits of rGE. The findings from these studies have provided relatively consistent findings, indicating that most environments are under some level of genetic influence. Heritability estimates vary across studies and depend on the precise environment that is being examined, making it somewhat difficult to summarize the results with a single number. A la$r$GE review of the literature, however, concluded that the heritability of most environments ranged between 0.15 and 0.35 and that the total heritability for all environments was 0.27 (Kendler & Baker, 2007). These heritability estimates are not as la$r$GE as they are on human phenotypes, but they do highlight that virtually all environments are influenced, at least to some degree, by genetic variation.

An alternative to testing for rGEs via traditional methodologies that generate heritability estimates is by conducting candidate gene studies (see chapter 7) or even genome-wide-association studies (GWAS; see chapter 8). With these types of research designs, genetic polymorphisms are examined to determine whether the allelic combinations for them correlate with variation in environmental measures. Significant correlations between genetic polymorphisms and environmental variation provides empirical support for rGE. A number of studies have tested for rGEs in this way and the results have been decidedly mixed. Even so, there is some evidence providing support for rGEs. Studies, for instance, have revealed that genes of the dopaminergic system are related to a number of risky environments, including associating with delinquent peers. Relative to research on $G \times E$s, there has not been nearly as much research examining rGEs and the role that they play in the development of human phenotypes.

Although the available evidence does provide some support rGEs, for the most part these studies are unable to precisely test for different types of rGEs (though there are some studies that pit passive rGE vs. active/evocative rGE) or to examine the mechanism that is accounting for the rGE. All that most of the studies are able to identify is the presence or absence of an rGE. Recently, however, there have been a number of studies emerging that are using novel research designs and clever techniques to begin to delineate among the three different types of rGEs (Knafo & Jaffee, 2013). For now, what is important to realize is that the available research does consistently show that most environments are under some level of genetic influence and studies are now beginning to more closely examine the various types of rGEs that might be involved in the development of various phenotypes.

4.3 Endophenotypes, Neurobiology, & Genomic Imaging

The models described above provide a general basis for understanding some of the basic ways in which genes and gene-environment interplay are related to the development of human behaviors and traits. In many ways, however, these explanations do not fully capture the complexity of the connection between genotype and phenotype as they intimate that the link is relatively straightforward, direct, and short. In reality, the link between genotype and phenotype is infinitely more complex than these models are able to capture. The causal influence that genes have typically is indirect, can include feedback loops, and, includes an array of intermediary processes. Trying to map the entire linkage between genotype and phenotype is beyond the scope of this chapter (and book). The purpose of the remainder of this chapter is to provide a short introduction to some processes that are likely involved and that can be applied to a wide array of phenotypes studied in the social sciences. Specifically, we will focus on endophenotypes and the role of neurobiology and genomic imaging.

An endophenotype in its simplest terms can be of thought of as a phenotype that mediates the link between genotype and phenotype (Gottesman & Gould, 2003). More accurately, it is a trait that is posited to be closer to the gene of interest than the phenotype of interest, that is under genetic influence, that is more stable than the phenotype of interest, and that falls on the causal path between genotype and phenotype. Endophenotypes are particularly important for complex behavioral and trait outcomes as these outcomes are often difficult to measure and study because they are polygenic and because they are created from a multifactorial arrangement of genetic and environmental factors. Moreover, the effects of single genes can be difficult to detect as their effects may lie far upstream from the ultimate phenotype; however, these genes likely have much la*r*GEr effects on the endophenotypes that ultimately mediate the genotype-phenotype association. While endophenotypes ideally would be less genetically complex than the phenotype, this is not always the case nor does it have to be the case when searching for putative endophenotypes. Endophenotypes are also posited to be more stable and more reliably measured than the ultimate phenotype. What that necessarily means is that once an endophenotype is identified, it should be less malleable and more easily measured than complex outcomes, such as antisocial behaviors, which may ebb and flow across time and space.

Endophenotypes are widely variable and there is not a "one-size-fits-all" endophenotype that is involved in all genotype-phenotype connections. When it comes to the study of human behaviors and traits, most endophenotypes are thought to reside in the brain and are either involved in the functioning of the brain or in some aspect of brain structure. For instance, candidate endophenotypes could include structural brain abnormalities to the prefrontal cortex, glial cell abnormalities, or problems with spatial working memory. Candidate endophenotypes should be identified based on the gene being studied, the phenotype being studied, the biological mechanisms that might be involved, and the guidelines reviewed above and elsewhere (for a thorough overview see, Gottesman & Gould, 2003).

Take, for example, research examining the genetic architecture for schizophrenia. Studies have consistently shown that schizophrenia is heritable and there has been a movement to identify the specific polymorphisms that are involved in the development of this phenotype. To date, numerous genetic loci have been uncovered that might play a role in the causation of schizophrenia. To advance the study of the genetics to schizophrenia, and to more fully understand the causal pathway from genotype to schizophrenia, there has been a concerted effort to identify endophenotypes that are involved. The candidate endophenotypes have la*r*GEly been neurocognitive endophenotypes, such as attention, verbal and working memory, as well as facial processing. The results of some studies have found that certain genetic loci are related to some of these neurocognitive endophenotypes which, in turn, are related to the development of schizophrenia. Endophenotypes have been central to the study of schizophrenia and in identifying the etiological pathway to schizophrenia.

Using endophenotypes can be quite informative when trying to fully unpack the genotype-phenotype connection. Of course, it does add another layer of complexity to the genetic analysis because of the need to identify candidate endophenotypes and then examine whether the genetic loci related to the phenotype are also related to the endophenotype. Nonetheless, there is an emerging body of research that can be consulted to help with the identification of endophenotypes. This line of research is known as imaging genetics. Imaging genetics is designed to identify specific functional polymorphisms that are expressed in the brain in some capacity. Imaging genetics is thought to hold particular promise for identifying endophenotypes because 1) approximately 60-70% of all genes are expressed in the brain, 2) brain functioning and brain structure are shown to be highly heritable, with heritability estimates hovering around 60% or higher, and 3) brain structure and function are thought to mediate the association between genetic polymorphisms and most phenotypes. Social science research has been slow to directly examine the brain as an endophenotype, but rather examines the direct association between functional polymorphisms and phenotypes. In other fields of study, though, there has been a tremendous amount of effort and resources funneled into genetic imaging studies. To understand these studies and the findings that they produced, it is first necessary to understand at a basic level what is meant by structural neuroimaging and functional neuroimaging.

Structural neuroimaging studies are designed to map quantifiable features of particular regions of the brain, such as its size and volume. In this way, structural neuroimaging studies allow for the comparison of regions of the brain across people in terms of physical properties (e.g., the thalamus is la*r*GEr in Person A vs. Person B, the volume of the hippocampus is below the mean in this subject, etc.) and then this variation in brain structure is often examined to determine whether it corresponds to variation in phenotypes. For instance, it would be possible to test to see whether variation in the volume of the amygdala is related to anxiety or aggression. A number of brain-imaging techniques are used in structural neuroimaging studies, including computerized tomography (CT) scans, computed axial tomography (CAT) scans, and magnetic resonance imaging (MRI). All of these techniques are designed to identify variability in the quantifiable features of specific regions of the brain and do so through the use of X-rays (e.g., CT and CAT scans) or via radio waves and magnetic fields (e.g., MRIs). The key point to bear in mind is that structural neuroimag-

ing studies—regardless of the brain-imaging technique that is employed—are interested in mapping variation in specific brain structures.

Functional neuroimaging studies, in contrast, are designed to map the processing of information throughout the brain. In this way, it is possible to examine the activity level or general functioning of particular regions of the brain. The most commonly used techniques to assess brain functioning are functional magnetic resonance imaging (fMRI), positron emission tomography (PET), and single photon emission computed tomography (SPECT). The results from studies that use these techniques allow for researchers to examine the functional activity of the brain and then, in some studies, to determine whether variation in brain functioning corresponds to variation in specific phenotypes.

There have been literally tons of studies published using structural and functional imaging techniques. The results of these studies have produced a wealth of detail regarding variability in the structure and functioning of key regions of the brain thought to be implicated in human phenotypes, such as the amygdala, the hippocampus, the corpus callosum, and the cerebral cortex. Some studies have even tied variation in the structure and function of the brain to variation in behavioral phenotypes. For instance, studies have shown that the structure of certain regions of the brain, such as the corpus callosum, and the functioning of other regions of the brain, such as the amygdala, are related antisocial phenotypes, including psychopathy (Kiehl et al., 2001; Raine et al., 2003).

With an impressive amount of research providing detailed information about brain structure and brain functioning and how both relate to human phenotypes, another line of research began to develop to examine what causes variation in the structure and function of the brain. Brain science researchers focused their attention on the role of genetic factors. To do so, they employed the twin-based methodology (where they compared MZ twins to DZ twins) to estimate the heritability of the structure and function of the brain. The results of these studies have produced some staggering findings showing that more than half of the variance (on average) in brain structure and brain functioning is due to genetic factors (Thompson et al., 2001). These findings are highly robust and have been detected across a broad range of measurement techniques and studies (Peper et al., 2007).

Against this backdrop, studies started to examine the specific polymorphisms that might account for the heritability in brain structure and function by conducting what are called genomic-imaging studies. Genetic imaging can be thought of as a combination of molecular genetic association studies and functional/structural neuroimaging studies. Essentially, genetic imaging studies examine whether the alleles of certain genetic polymorphisms correspond to variation in brain structure and/or brain functioning. To the extent that they do, then scientists assume that the specific genetic polymorphisms are involved—at least in some capacity—in the creation of variance in neurobiology. Keep in mind that brain structure and/or brain functioning is typically viewed as an endophenotype that falls somewhere between the genetic polymorphism and the phenotype of interest.

There has been a lot of interest in genetic-imaging studies during the past decade and,

out of this interest, there has been a significant amount of research produced. These studies have identified a number of genetic polymorphisms thought to be involved in the genetic architecture of various human phenotypes. One of the more noteworthy findings is in regards to a polymorphism in MAOA, a polymorphism that has been implicated in aggression, violence, and antisocial traits. Researchers have long been interested in the possible intermediary mechanisms that link together MAOA and antisocial phenotypes. As a result, there have been a number of genomic-imaging studies that have examined the connection between this polymorphism and different aspects of brain structure and function. One study, for example, found that the alleles of MAOA are related to the activity level of the anterior cingulate (Fan et al., 2003) whereas other studies revealed that MAOA was associated with amygdala activation and functional coupling with ventromedial prefrontal cortex (Buckholtz et al., 2008) as well as reductions in limbic volume and hyperresponsiveness of the amygdala (Meyer-Lenderberg et al., 2006). Although the MAOA-antisocial link is one of the most replicated findings when it comes to single gene effects on phenotypes, the genomic-imaging results clearly show that this association is complex and involves multiple endophenotypes that include variation in brain structure and functioning.

Other studies have uncovered additional polymorphisms related to a wide range of activities, functions, and structures in the brain. To illustrate, alleles of a polymorphisms in the catechol-o-methyltransferase (COMT) gene have been shown to relate to prefrontal response activation, and a polymorphism in the apolipoprotein E (APOE) has been lined with the activity level of memory-related brain systems (as reviewed in Hariri & Weinberger, 2003). Most of the genetic-imaging research is being conducted by brain science researchers. As a result, there has not been much genomic-imaging research conducted that is applied to understanding the development of behaviors and outcomes that are of broad interest to social scientists. Much of the literature, for instance, applies to medicine and diseases, such as the onset of Alzheimer's disease or the development of schizophrenia.

More recently there has been a slight variant to the genomic-imaging studies, wherein rather than focusing on a single gene or a small handful of genes, entire genomes are scanned and then examined to see which (if any) of the genes are related to brain structure/function. These studies can be thought of as a hybrid between GWAS (see Chapter ??) and neuroimaging studies. Some of these studies, moreover, not only scan the entire genome, but also scan the entire brain (Medland et al., 2014). These types of studies employ a more data-driven approach, wherein no theoretical or biological rationale is employed to study specific genes or systems of genes as they relate to specific aspects of brain structure/function. Rather, these studies are more exploratory in nature and focus on what the data reveal. They can be quite informative as they provide insight into the genetic-brain linkage that might otherwise not be discovered. One study, for example, scanned more than one-half million SNPs and more than 140 measures of grey matter to determine whether there was an association between any of these genes and any of these neurobiological measures (Shen et al., 2010). The results revealed some significant associations between certain genetic markers and certain neuroimaging measures. Collectively, the results of genomic-imaging studies (whether they focus on one gene or the entire genome) can help elucidate the complex linkage between a gene and outcome by mapping the causal pathway that leads from gene to

endophenotype(s) to phenotypic outcome. The results to date have provided some important information regarding the development of phenotypic variance for certain phenotypes. However, the genomic-imaging literature remains in its infancy and much remains to be discovered about the nexus between genes and neurobiology and how they are involved in the etiology of most human phenotypes.

4.4 Conclusion

Social scientists are provided with very little formal training during the academic careers regarding genetics and biology (Cooper, Walsh, & Ellis, 2010). Consequently, they are not equipped with the knowledge to understand the basic principles for how genes may ultimately influence and affect certain phenotypes. This lack of knowledge likely fuels fears regarding genetics and also likely contributes to confusion regarding the link between genes and phenotypes. Without having been trained in basic human biology and genetics, social scientists also are not primed to think about the role that genes play in the development of phenotypes nor do they think about how genetic influences could be integrated into theoretical explanations or guide research questions in their area of specialization. Learning a bit about the link between genes and phenotypes can go a long way toward assuaging some outdated beliefs about genes and for showing how genes can augment and advance existing social science theories and explanations. The information reviewed in this chapter should help in this regard.

Understanding some of the basic ways in which genes may ultimately account for phenotypic variance is critical as the logic of the mathematical models presented throughout the rest of the book rests on these principles. Moreover, what is important to realize is that statistical analyses in isolation do not provide any insight into the mechanisms that ultimately are responsible for linking a gene to a phenotype. Statistical models must be built and research questions must be developed based on knowledge regarding how genes likely contribute to variance in the phenotype of interest. Moreover, conceptual information about genes and genetic influences can be quite useful when attempting to interpret the results provided by some type of quantitative genetic analysis. The rest of this book will provide the knowledge and statistical skills needed to test these models, integrate these models, and ultimately apply them to the phenotypes that are of interest to you.

Part II

Modeling Strategies

Chapter 5

Biometrical Model-fitting I: Univariate Models

Biometrical model-fitting has formed the backbone of quantitative genetics since its humble beginnings in Francis Galton's lab in the late 1800s (Mather & Jinks, 1982; Lynch & Walsh, 1998). Indeed, many of the statistical techniques that are used today by criminologists—even those who would not consider themselves biosocial scholars—were developed by statisticians who were interested in decomposing the variance in a phenotype into its constituent genetic and environmental components. This early work often focused on the correlation between relatives (e.g., mom and child) on quantitative traits like physical characteristics or personality measures. Francis Galton (1889) and Karl Pearson (1896) famously studied observable physical phenotypes like height, weight, cranial size, and wingspan in an attempt to identify the degree to which genetic influences affected the development of these observable characteristics.

During the first few decades of the 20^{th} century, Sewall Wright (1921a-c) formalized many of the techniques introduced by Francis Galton (1889) and R. A. Fisher (1918) into what is now referred to as path analysis. And thus, biometrical model-fitting offered a way to understand how Mendelian genetics can be used to explain traits that do not segregate neatly into categories like Mendel's pea plants did. In other words, quantitative genetics and techniques of biometrical model-fitting were developed as a way to adapt Mendelian genetic principles to more complex, quantitative traits. Scholars such as Wright, Fisher, and Galton showed that one could capitalize on the known levels of genetic relatedness between relatives to decompose the variance in a phenotype into genetic and environmental factors. Specifically, if mother and child share 50% of their genes (which, as we know from the principles of Mendelian inheritance is true), then any correlation between them can be used to determine the degree to which genetic influences underlie the trait of focus. The correlation itself is not directly interpretable as the genetic influence, though. As we will show below, one must utilize the genetic relatedness coefficient (which will be denoted as R) to scale the observed correlation to garner an approximation of the genetic influence (i.e.,

V_G) on the phenotypic variance (i.e., V_P).

The point to take away from this discussion is that scholars have long been interested in estimating the genetic and environmental influences on phenotypes. Statisticians, biologists, and geneticists (often scholars fit into more than one of these categories) have spent a considerable amount of time constructing statistical models that can estimate the genetic (which, as we revealed in the last chapter is referred to as h^2) and environmental (i.e., c^2 and e^2) components of a phenotype.

Biometrical model-fitting is an umbrella term that captures research methods aimed at decomposing the variance in a quantitative phenotype into genetic and environmental parts. This chapter and the next (Chapter 5) will introduce three general approaches to biometrical model-fitting: the univariate ACE model (this chapter), regression-based approaches (this chapter and Chapter 5), and bivariate and multivariate models (Chapter 5). In a general sense, these three approaches are related, but somewhat unique, strategies for accomplishing the same goal. The primary differences between them is that the univariate ACE model deals with one phenotype at a time. The bivariate and multivariate models (Chapter 5), as their names reveal, can handle two (bivariate) or more (multivariate) phenotypes at once. These models, furthermore, can even reveal the degree to which the correlation(s) between (among) the phenotypes is due to shared genetic/environmental overlap. This is an incredibly useful advantage of the bi-/multivariate approaches. We will spend a considerable amount of time discussing the utility of these models in the next chapter.

Finally, the regression-based approaches (discussed in this chapter and the next) will primarily be developed using the model proffered by DeFries and Fulker (1985). This easy-to-implement model (indeed, it can be used in much the same way as a "typical" OLS regression model) is extremely flexible and, as we will show, converges on parameter estimates that are reliable and comparable to those gleaned from the more complicated models discussed above. As you can imagine, then, the DF model is a popular tool and we recommend that scholars interested in learning biometrical model-fitting techniques begin with this model. We will provide a full mathematical treatment of the DF model, we will describe the parameters that are gleaned from the model, and we will provide several demonstrations in the last section of this chapter.

5.1 Conceptual Overview

Recall our working model for phenotypic variance that was derived at the end of chapter 3:

$$V_P = \Psi(V_G, V_E)$$

where the right-hand side can be expanded to more explicitly capture the interactions and correlations between the various components:

$$
\begin{aligned}
V_P &= V_G + V_E + 2cov(V_G, V_E) + V_{G\times E} \\
&= V_A + V_D + V_I + V_C + V_E + 2cov(V_A, V_C) + \ldots + 2cov(V_I, V_E) + V_{A\times C} + \ldots + V_{C\times E} \\
&+ 2cov(V_A, V_{A\times C}) + \ldots + 2cov(V_E, V_{C\times E})
\end{aligned}
$$

Recall also that one of the key goals of biometrical models is to estimate the degree to which the variance in P (i.e., V_P) can be attributed to variance in G (i.e., $V_G = V_A + V_D + V_I + V_C + V_E$) and to variance in E (i.e., $V_E = V_C + V_E$).

The biometrical modeling techniques discussed below will reveal *how* one can take data from a sample of individuals and estimate, for example, V_A and V_E in a latent way. The term *latent* here reveals that one need not have to observe all the genetic and environmental components individually (that will be the focus of chapters X and X). Instead, biometrical models will rely on information taken from pairs of siblings with known levels of genetic relatedness. Doing so will allow us to assume certain levels of genetic relatedness would have led to *phenotypic* covariances among siblings that should follow certain patterns when stratified by level of genetic relatedness. Thus, the primary goal of the techniques discussed here and in chapter 6 is to estimate V_A and to use it to generate a heritability (h^2) estimate.

5.2 Univariate Biometrical Models

We begin with the univariate biometrical model, which takes a single phenotype at a time and provides the user with estimates of h^2, c^2, and e^2. In order to garner estimates of these three parameters, the model must be estimated using a dataset that has pairs of cases with known genetic relationships (R).[1]

What do we mean by "pairs of cases with known genetic relationships"? While there are many suitable answers to that question, the most commonly used "pairs" are twins and other siblings. Imagine a pair of monozygotic (MZ) twins. MZ twins are colloquially known as "identical" twins because they often look, act, and share personalities that are nearly identical to one another. So, what exactly does it mean to say twins are "identical"? While it may seem overly reductionist, the simplest response is that they are identical at the genetic level. Indeed, MZ twins share nearly 100% of their DNA![2] Outside of *de novo* and random mutations that occur throughout the life course, MZ twins share *all* 3 billion of their As, Ts, Cs, and Gs (recall that A, T, C, and G are the four letters in the genetic alphabet [see Chapter 1]).

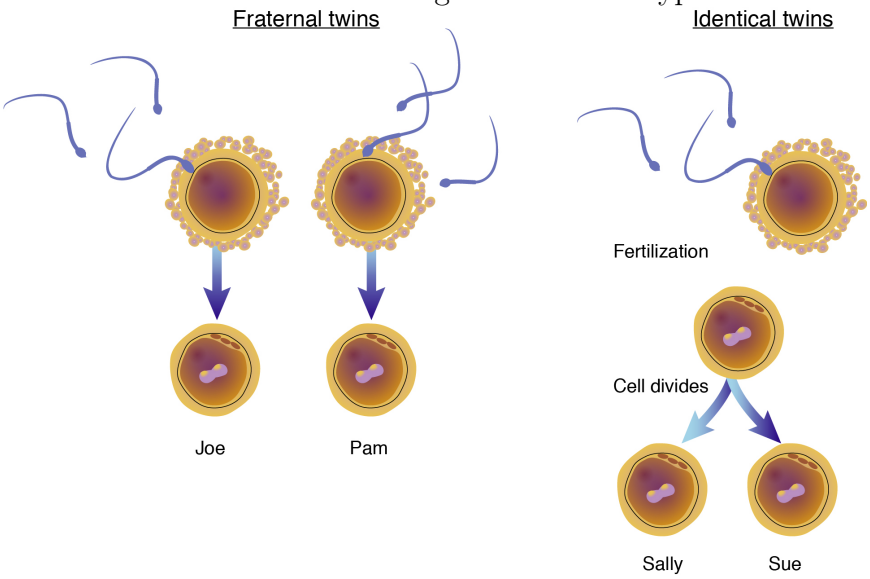
[1] It is extremely important to note that users are *not* restricted to dyadic, paired data. Indeed, one can use the ACE model on triadic data (i.e., groups of three) or data with larger groups if available. In our experience, though, dyadic data are far more common, so we will restrict our comments to data with genetically related pairs.

[2] We used the qualifier "nearly" here in light of recent research that shows MZ twins differ at *some* spots along the genome. For a discussion that is tailored to a general audience, see `http://www.huffingtonpost.com/2012/11/10/identical-twins-genes_n_2110016.html`

The reason MZ twins share so much genetic material is quite interesting. MZ twins begin life, at the moment of conception, as one zygote, which is the term for a fertilized egg. For reasons that remain unknown, however, MZ twins split into two independent zygotes during the first weeks of development. Those two independent zygotes, in turn, develop into two separate human beings who share 100% of their genetic material. In essence, it is accurate to think of MZ twins as being genetic clones; nature's clones!

MZ twins are not the only "type" of twin typically observed in nature. As you are no doubt aware, twins can sometimes be opposite sex (i.e., one male and one female). In these cases, the twins clearly do not share 100% of their genes because one has two X chromosomes (the female) while the other has an X and a Y chromosome (the male). These twins are discordant at many genetic loci—the term often used by geneticists to refer to a location in the human genome. So how do these twins come about? Where MZ twins begin as one sperm and one egg, the other type of twins, referred to as dizygotic (DZ) twins, begin as two sperm and two eggs. In essence DZ twins—often known colloquially as fraternal twins—are just like any "regular" (i.e., non-twin) siblings. They began life as two independent sperm and two independent eggs, each carrying their own unique mix of paternal genes and maternal genes. When the eggs are fertilized by the sperm, the genomes for the two (or more) siblings are *not* identical. These relationships are depicted graphically in Figure 5.1.

Figure 5.1: Two Types of Twins



Given that MZ twins arise from a single fertilized egg and DZ twins arise from two fertilized eggs, we can begin to build a biometric model based on the expected level of genetic overlap between twins. To be direct, knowing whether twins are MZ or DZ tells us something about the degree to which they share genetic material. As was noted above, MZ twins share 100% of their DNA and DZ twins share 50% (on average) of their DNA. Translating this into the

terms necessary to explain *variance* (see Chapter 2) in a phenotype: which is *the* goal of the univariate biometrical model—MZ twins share 100% of the additive variance (V_A see Chapter 3) in any phenotype and DZ twins share 50%. Dominance variance can also be inferred for twins: MZ twins share 100% of the dominance variance (V_D) while DZ twins share 25% of V_D for any phenotype.

While we have restricted our discussion thus far to twins, the level of genetic relatedness between *any* pair of genetically related individuals can be used to infer the level of genetic overlap for V_A and V_D. The table below provides these values for a range of different genetically related pairs that are commonly analyzed by quantitative geneticists.

	Proportion of Genetic Variation Shared (R)	
Pair Type	Additive (V_A)	Dominance (V_D)
Monozygotic Twins (MZ)	1.00	1.00
Dizygotic Twins (DZ)	0.50	0.25
Full-siblings (FS)	0.50	0.25
Parent-Offspring (PO)	0.50	0.00
Half-siblings (HS)	0.25	0.00

We now have all the information necessary to build a biometrical model to decompose the variance in a phenotype (V_P) into its constituent genetic (V_A, for the time being) and environmental (V_C and V_E) parts. Let us start with MZ twins. Imagine we have a dataset with $n = 100$ MZ twin *pairs*. It might look something like this:

Pair ID	Pair Type	V_A Shared (R)	phenotype$_1$	phenotype$_2$
1	MZ	1.00	5	5
2	MZ	1.00	7	8
3	MZ	1.00	3	1
⋮				
100	MZ	1.00	10	9

where the Pair ID simply indexes the number of twin *pairs* in the data; Pair Type identifies each of the twin pairs as an MZ twin; V_A Shared (R) reveals the same information that was displayed in the previous table, that MZ twins share 100% (shown here as a proportion, 1.00) of the additive genetic variation (V_A); phenotype$_1$ shows the score on the phenotype of focus for twin 1; and phenotype$_2$ is the score on the phenotype for twin 2.

Hold the information from the previous table in memory for just a moment while we switch gears. Recall that the primary goal of all the models presented in this chapter is to decompose variance in a phenotype (V_P) into genetic and environmental components. We now know that different types of genetically related pairs share different amounts of genetic variation (R). We also know that we can observe phenotype scores for individuals in a genetically related pair, just like was shown in the table immediately above. So how

can we translate these pieces of information into a biometrical model that will tell us the proportion of V_P that is due to V_A, V_C, and V_E? In order to do that, we need to have a model that explains why MZ twin pairs would covary on the phenotype of focus. Recall from our discussion in Chapter 2 that we can analyze the *covariance* between two variables and that value can be translated into a standardized coefficient known as a correlation (r). Looking at the dataset in the previous table, we see that phenotype$_1$ and phenotype$_2$ appear as two separate variables. Thus, one can calculate the correlation between phenotype$_1$ and phenotype$_2$ in exactly the same way as one calculates the correlation for any two "normal" variables x and y:

$$r_{xy} = \frac{cov_{xy}}{s_x s_y}$$

so:

$$r_{p_1 p_2} = \frac{cov_{p_1 p_2}}{s_{p1} s_{p2}}$$

where p_1 is the phenotypic score for twin 1 (phenotype$_1$ in the table above) and p_2 is the phenotypic score for twin 2 (phenotype$_2$ in the table above).

Now we need to identify all the sources of phenotypic variance that may lead MZ twin pairs to correlate on the phenotype. As it turns out—assuming, as we are here, that V_D is negligible—there are only two sources of phenotypic variance that could cause MZ twins to turn out similarly: V_A (additive genetic variation) and V_C (shared environmental influences).

Thus, any $r_{p_1 p_2}$ observed between two twins in an MZ twin pair—denoted as r_{MZ}—is the result of genetic influences $R{*}V_A$, which can be estimated by $R{*}h^2$ and shared environmental influences V_C, which are estimated by c^2. This leads to:

$$r_{MZ} = R * h^2 + c^2$$

Because $R = 1.00$ for MZ twins, the equation simplifies to:

$$\begin{aligned} r_{MZ} &= 1.00h^2 + c^2 \\ &= h^2 + c^2 \end{aligned}$$

Notice that the equation above has three unknowns: r_{MZ}, h^2, and c^2. Given the dataset from above, we can fill in one of the unknowns, r_{MZ}, by simply calculating the correlation between phenotype$_1$ and phenotype$_2$. But what about the other two unknowns, h^2 and c^2? Unfortunately, with the present information we have no way to fill in meaningful estimates for these items. In fact, these are the estimates that we are hoping to garner by observing r_{MZ}. Thus, we have run into a classic problem in mathematics: we have more unknowns than we have observable information. When this occurs, one cannot calculate an estimate of *either* unknown because there are an infinite number of solutions that would fit. Think about it this way: imagine $r_{MZ} = 0.50$, so:

$$0.50 = h^2 + c^2$$

What could you fill in for h^2 and c^2 in order to make the value on the left-hand side of the equation true? Hopefully you can see that there are an infinite number of possibilities. You

could fill in $h^2 = 0.50$ and $c^2 = 0.00$. Or it could be $h^2 = 0.00$ and $c^2 = 0.50$. Or anything in between. This problem is referred to as parameter indeterminacy (Keller & Coventry, 2005).

So how do we solve this problem? We must observe data from at least one other level of genetic relatedness. In other words, we need to add another "layer" of information to our dataset. This is most commonly solved by adding DZ twins to the study. Doing so might adjust our dataset to look like the following:

Pair ID	Pair Type	V_A Shared (R)	phenotype$_1$	phenotype$_2$
1	MZ	1.00	5	5
2	MZ	1.00	7	8
3	MZ	1.00	3	1
⋮				
100	MZ	1.00	10	9
101	DZ	0.50	4	9
102	DZ	0.50	6	5
103	DZ	0.50	10	2
⋮				
200	DZ	0.50	6	7

In much the same way that we parsed r_{MZ} into two components, we can do the same for r_{DZ}. Specifically:

$$\begin{aligned} r_{DZ} &= R * h^2 + c^2 \\ &= 0.50h^2 + c^2 \end{aligned}$$

But, just as before, we have three unknowns—r_{DZ}, h^2, and c^2—and one piece of information: r_{DZ}. You may be wondering, therefore, how adding DZ twins to the dataset solves our parameter indeterminacy problem. The solution is to *combine* the two equations—the equation for r_{MZ} and the equation for r_{DZ}—because they share two of the unknowns! For example, we could begin by solving for c^2 in the DZ equation:

$$c^2 = r_{DZ} - 0.50h^2$$

Realizing that *both* r_{MZ} and r_{DZ} share c^2 and h^2 immediately reveals the solution to garnering an estimate of h^2 by substituting our solution for c^2 in the DZ equation *into* the MZ equation:

$$\begin{aligned} r_{MZ} &= h^2 + c^2 \\ r_{MZ} &= h^2 + (r_{DZ} - 0.50h^2) \end{aligned}$$

Then, we simply solve for h^2:

$$\begin{aligned} 2r_{MZ} &= 2h^2 + 2r_{DZ} - h^2 \\ 2r_{MZ} &= h^2 + 2r_{DZ} \\ h^2 &= 2(r_{MZ} - r_{DZ}) \end{aligned}$$

We can, therefore, garner an estimate of h^2 simply by observing r_{MZ} and r_{DZ}![3]

And the same is true for c^2. The only difference is that we would solve the MZ equation for h^2 and then plug that value into the DZ equation. Doing so leads to the following:

$$\begin{aligned} r_{DZ} &= 0.50(h^2) + c^2 \\ r_{DZ} &= 0.50(r_{MZ} - c^2) + c^2 \\ r_{DZ} &= 0.50r_{MZ} - 0.50c^2 + c^2 \\ r_{DZ} &= 0.50(r_{MZ}) + 0.50c^2 \\ 2r_{DZ} &= r_{MZ} + c^2 \\ c^2 &= 2r_{DZ} - r_{MZ} \end{aligned}$$

By this point, you may be thinking, "what happened to e^2?". We've been telling you all along that phenotypic variance is the result of genetic (h^2) variation, shared environmental (c^2) variation, and nonshared environmental (e^2) variation. Yet, the above discussion focused on the first two and ignored e^2. This was done because we were focused on the *covariance* between twins: r_{MZ} and r_{DZ}. Only h^2 and c^2 can explain *covariance* between twins. The nonshared environmental (e^2) variation can, therefore, only explain why twins would differ. In this way, e^2 contributes to the phenotypic variance (V_P) but it does *not* contribute to the phenotypic *covariance*.

[3]At this point, it is important to note that we do *not* need to square the coefficient—nor do we square any of the correlation coefficients that went into the constituent equations—to get an estimate of the variance in the phenotype (i.e., V_P) that is due to the variance in additive genetic factors (V_A): $h^2 = \frac{V_A}{V_P}$. This may be counterintuitive to those who have studied econometrics/statistics and have long been taught to square a correlation coefficient to estimate the degree to which variance in X explains variance in Y (i.e., in most contexts, r^2 = variance explained). It is unnecessary—and, indeed, is incorrect—to square *these particular* coefficients in the context of a biometrical model. The simplest explanation for why this is so is that squaring a correlation coefficient provides an estimate of the degree to which variance in X explains variance in Y. In the biometrical model, we use correlations to express the *proportion of variance in X that is shared with variance in Y*. Although the pervious two statements may, at first, appear to express the same thing, upon careful inspection, we see that they are in fact different. The former—the squared correlation—deals with scenarios where one wishes to predict Y with known values of X. The latter—the case we are interested in when estimating a biometrical model—only wishes to compute an estimate of shared variance between the two. Thus, in the context of a biometrical model, phenotypic correlation coefficients—meaning, for example, r_{MZ} and r_{DZ}—are not squared when generating estimates for h^2, c^2, and e^2. At the risk of breeding even more confusion, though, we must point out that heritability, shared environment, and nonshared environment estimates *are* often squared during the actual estimation routine. Thus, those estimates are expressed as h^2, c^2, and e^2. This point will be explained with more detail below. For now, suffice to say that the squared values used to estimate h^2, c^2, and e^2 are necessary for simple computation purposes. They are distinct from the point we are trying to make above about r versus r^2.

In order to understand this distinction, it may help if we recall our discussion of the variance-covariance matrix from Chapter 2. A heuristic variance-covariance matrix is presented below for MZ twin pairs:

$$\begin{bmatrix} 1.00 & 0.500 \\ 0.500 & 1.00 \end{bmatrix}$$

Recall that the variance-covariance matrix has a very systematic structure: the values on the diagonal (in this case, the trait is standardized so both variances are 1.00) are the *variance* estimates for phenotype$_1$ and phenotype$_2$, respectively. The off-diagonal elements (in this case, 0.500) provide an estimate of the *covariance* between phenotype$_1$ and phenotype$_2$. We can, therefore, re-write the variance-covariance matrix like so:

$$\begin{bmatrix} s^2_{p_1} & cov_{p_2p_1} \\ cov_{p_1p_2} & s^2_{p_2} \end{bmatrix}$$

where the p_1 and p_2 index phenotype$_1$ and phenotype$_2$, respectively.

We can further expand the MZ variance-covariance matrix by filling in the information we now know about the variance and covariance of a phenotype and how our estimates of V_A, V_C, and V_E (i.e., h^2, c^2, and e^2) play a role:

$$MZs : \begin{bmatrix} h^2 + c^2 + e^2 & h^2 + c^2 \\ h^2 + c^2 & h^2 + c^2 + e^2 \end{bmatrix}$$

The corresponding variance-covariance matrix for DZs is:

$$DZs : \begin{bmatrix} h^2 + c^2 + e^2 & 0.50h^2 + c^2 \\ 0.50h^2 + c^2 & h^2 + c^2 + e^2 \end{bmatrix}$$

As you can see from the diagonal elements in both the MZ and the DZ matrix, the phenotypic variance values result from $h^2 + c^2 + e^2$. We have shown above how to garner estimates of h^2 and c^2 using the off-diagonal elements, the covariances/correlations. If one desires an estimate of e^2, the solution is simple (assuming the phenotype has been standardized to have a variance=1.00):

$$1 = h^2 + c^2 + e^2$$
$$e^2 = 1 - (h^2 + c^2)$$

5.3 The ACE Model

So how do we *actually* get estimates of h^2, c^2, and e^2 from the MZ and DZ variance-covariance matrices presented above? That process can take several forms, two of which will be discussed in this chapter: the ACE model and the DF model. The former—the ACE model—is discussed in this section. A discussion of the DF model follows in a subsequent section.

5.3.1 Conceptual Discussion

It is necessary to understand two points if we are to successfully relay the process of generating h^2, c^2, and e^2 estimates with the ACE model. The first point is that the variance-covariance matrices presented above reveal the *expected* values for the variance and the covariance. Anytime you are working in a statistics/quantitative framework and you see the word *expected*, you should automatically think "average." Thus, in this context, the expected values presented in the variance-covariance matrices tell us what we would see, on average, if we were to observe MZ and DZ twins in the population an infinite number of times with perfect replication. As you are no doubt aware, however, the reality of science does not allow us to observe nature (at least not human nature) in a way to conforms exactly to expectation.

This leads to the second point that must be kept in mind: *observed* variance-covariance matrices will *never* perfectly match their expected versions. The expected matrices presented above neatly cleaved all the estimates so that they were uniquely observed. Indeed, we saw that the covariance between MZ twins was a function of h^2 and c^2. While the covariance for DZ twins was a function of $0.50h^2$ and c^2. But as we noted earlier, we cannot directly observe these values. Instead, these are precisely the values we are trying to estimate. All we see are the *observed* variance and covariance values. For instance, we might see something like this if we were to analyze the heuristic twin dataset presented earlier:

$$MZs: \begin{bmatrix} 8 & 4 \\ 4 & 8 \end{bmatrix}$$

and:

$$DZs: \begin{bmatrix} 8 & 2 \\ 2 & 8 \end{bmatrix}$$

These represent our *observed* variance-covariance matrices and both are unstandardized so we see variances on the diagonals and covariances (not correlations) on the off-diagonals. Working backward and using the information outlined above (i.e., that we can garner an estimate of, say, h^2 by combining information from the MZ matrix with information from the DZ matrix), we can draw from these matrices to produce the estimates of interest: h^2, c^2, and e^2.

There are at least two ways this can be done. First, we could simply solve—by hand—for h^2, c^2, and e^2. Recall that the calculation for h^2 is:

$$h^2 = 2(r_{MZ} - r_{DZ})$$

But we have variances and covariances, not correlations. Recall, however, that the latter (correlations) are just standardized forms of the covariance. For our purposes here, the equation above—which uses correlations—is just a simplified form of the more general covariance

version:

$$h^2 = 2(r_{MZ} - r_{DZ})$$

$$= \frac{2(cov_{MZ} - covDZ)}{\sqrt{s^2_{pMZ}} * \sqrt{s^2_{pDZ}}}$$
$$= \frac{2(4-2)}{(8)}$$
$$= 0.50$$

We could then, similarly, solve for c^2 and e^2.

This may seem an appealing strategy. Indeed, it is simple and straightforward. It requires no great skill, nor does it even require one of those pesky statistical programs! But, there are some obvious limitations to solving for h^2, c^2, and e^2 by hand. Three stand most prominent. First, humans make mistakes. Even the simplest calculations are prone to an errant keystroke on the calculator. Second, the variance-covariance matrices presented here had equal variances for phenotype$_1$ and phenotype$_2$ *and* for MZ and DZ twins. In other words, the variance of the phenotype was exactly the same no matter who or what type of twin we observed. The "real world" is never quite so clean, so the simple hand calculation above would need to be adjusted to take into account unequal variances. Third, and perhaps even more important, there is no obvious way to calculate tests of statistical inference such as p-values and confidence intervals. Moreover, how would you ever be able to determine whether your calculations are a "good" fit if you only used the hand calculations. Short answer: you couldn't.

This leads to our second approach for calculating estimates of h^2, c^2, and e^2. As you have almost certainly anticipated, the solution to the problems with the hand calculation approach is to let a computer do all the heavy lifting. Computers are prone to fewer computation errors, they are a bit (understatement) faster than humans, and they can overcome our statistical inference limitation. But, in order to understand how a computer would go about computing estimates of h^2, c^2, and e^2, it is best if we think of the process as if a human were going to approach it. This will help to dissolve some of the mystique that surrounds the univariate ACE model.

Let's return to our consideration of *observed* matrices, but this time let's make a few adjustments so they look more like real data:

$$MZs : \begin{bmatrix} 7.2 & 4.5 \\ 4.5 & 6.9 \end{bmatrix}$$

and:

$$DZs : \begin{bmatrix} 6.8 & 2.3 \\ 2.3 & 7.1 \end{bmatrix}$$

In order to garner estimates of h^2, c^2, and e^2, a computer would approach things in much the same way that we did with the hand calculations above. Yet, rather than solve for

one set of *expected* matrices, a computer could carryout these calculations thousands of times, all the while trying to find the set of values that does the best job of reproducing the *observed* matrices. This process we have just described is known as maximum likelihood estimation (MLE). In short, the computer will fill in a set of values for h^2, c^2, and e^2, compute *expected* variance/covariance values, and then compute the probability (i.e., the likelihood) that those values are the best-fitting values. Next, the computer will "jiggle" those estimates a little (adding a little here, subtracting a little there) and then it would recompute the probability. This process will be repeated until a set of values *maximizes* the *likelihood* that the parameters of focus are the correct parameters.

Think of it this way. Imagine all the possible combinations of h^2, c^2, and e^2 were laid out onto an infinite plane. This plane—think of it like a blanket that goes on forever—represents the parameter space. Also imagine that there is *one* combination of values that reflects the correct solution. This set of values will show up for us as the solution that maximizes the likelihood, the latter of which will be represented by the y-axis in a plot. Using iterative sampling (which is what we described above when we noted the computer would jiggle the values) the computer no longer needs to traverse the infinite parameter space. Instead, the computer can simply compute the probabilities each time a solution is reached. When the computer sees the probabilities increasing, it knows that it must be approaching an optimal (i.e., maximum) solution. Once the probabilities begin to decrease, the computer will back up one solution. In this way, the computer "finds" the most optimal solution without the need to have observed all the infinite parameter space (and, for that reason, integral calculus solutions for the maxima are avoided as well).

This process can be expressed graphically as in Figure 5.2. As the only shows, there are many possible solutions that can be fit. But, there is only one solution that will maximize the likelihood (i.e., the solution that sits right at the top of the peak in the parameter space). Regardless of the angle from which the computer approaches the maximum, it will always converge on the same set of parameters because there is only one set of values for the optimal solution.

Let's work through an overly simplified example that can be carried out as a thought experiment. This will help clarify any points of lingering confusion. Imagine you observe the matrices that were presented above. What combination of estimates for h^2, c^2, and e^2 would you fill in as a first guess? Perhaps you might go with $h^2 = 0.40$, $c^2 = 0.20$, and $e^2 = 0.40$ because these values conform closely to the values expected by the three laws of behavior genetics (Turkheimer, 2000). So, let's go with these values for a second. When we do, we get the following *expected* matrices:

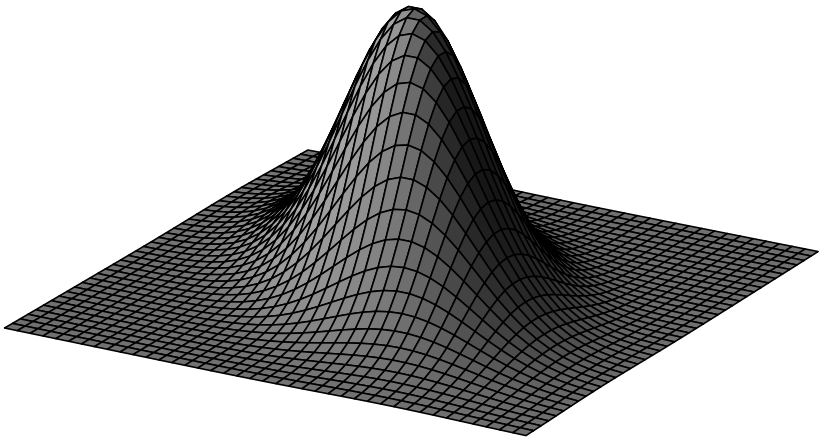
$$MZs: \begin{bmatrix} 0.40 + 0.20 + 0.40 & 0.40 + 0.20 \\ 0.40 + 0.20 & 0.40 + 0.20 + 0.40 \end{bmatrix}$$

and:

$$DZs: \begin{bmatrix} 0.40 + 0.20 + 0.40 & 0.50(0.40) + 0.20 \\ 0.50(0.40) + 0.20 & 0.40 + 0.20 + 0.40 \end{bmatrix}$$

But recall we were working with variances and covariances. Each estimate in the matrices above represents the *proportion* of the variance explained by that factor. Thus, to produce

Figure 5.2: Parameter Space with Maximum Shown



an *expected* variance-covariance matrix we will need to multiply each value by the observed variance. Let's make a simplifying assumption that the average variance across both sets of twins is 7. This produces:

$$MZs : \begin{bmatrix} (7)0.40 + (7)0.20 + (7)0.40 & (7)0.40 + (7)0.20 \\ (7)0.40 + (7)0.20 & (7)0.40 + (7)0.20 + (7)0.40 \end{bmatrix}$$

$$= \begin{bmatrix} 2.8 + 1.4 + 2.8 & 2.8 + 1.4 \\ 2.8 + 1.4 & 2.8 + 1.4 + 2.8 \end{bmatrix}$$

$$= \begin{bmatrix} 7 & 4.2 \\ 4.2 & 7 \end{bmatrix}$$

and:

$$DZs : \begin{bmatrix} (7)0.40 + (7)0.20 + (7)0.40 & [(7)0.50(0.40)] + (7)0.20 \\ [(7)0.50(0.40)] + (7)0.20 & (7)0.40 + (7)0.20 + (7)0.40 \end{bmatrix}$$

$$= \begin{bmatrix} 2.8 + 1.4 + 2.8 & 1.4 + 1.4 \\ 1.4 + 1.4 & 2.8 + 1.4 + 2.8 \end{bmatrix}$$

$$= \begin{bmatrix} 7 & 2.8 \\ 2.8 & 7 \end{bmatrix}$$

Thus, we have the following *observed* matrices (from above):

$$MZs: \begin{bmatrix} 7.2 & 4.5 \\ 4.5 & 6.9 \end{bmatrix}$$

$$DZs: \begin{bmatrix} 6.8 & 2.3 \\ 2.3 & 7.1 \end{bmatrix}$$

and these are our *expected* matrices based on our model where $h^2 = 0.40$, $c^2 = 0.20$, and $e^2 = 0.40$:

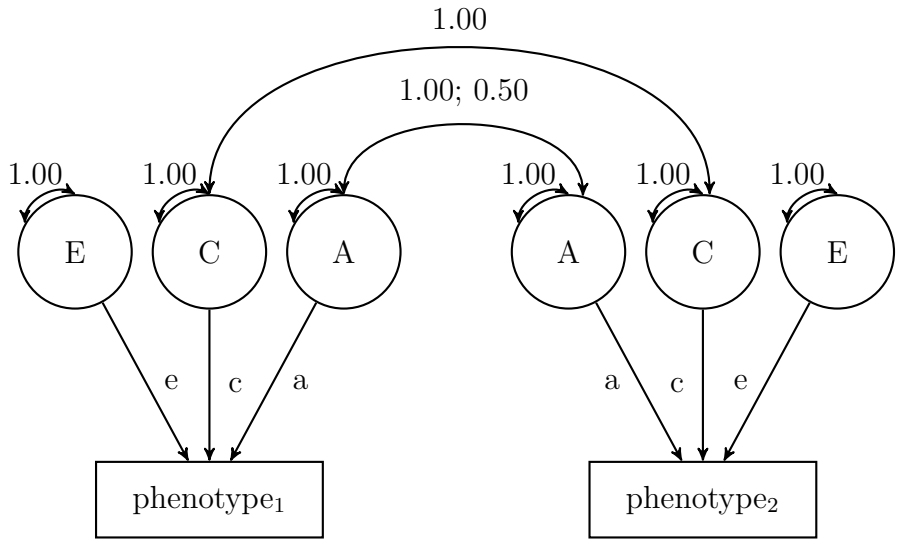
$$MZs: \begin{bmatrix} 7 & 4.2 \\ 4.2 & 7 \end{bmatrix}$$

$$DZs: \begin{bmatrix} 7 & 2.8 \\ 2.8 & 7 \end{bmatrix}$$

Comparing the *expected* matrices to the *observed* matrices reveals that our first guess was not correct: we overestimated some elements of the matrices and underestimated others. With this information, though, we could refine our first guess to get a little closer to the *observed* matrices. Perhaps we tweak things a little and go with the following for a second guess: $h^2 = 0.45$, $c^2 = 0.15$, and $e^2 = 0.20$. As before, we could then compare the *expected* matrices that emerge from these guesses with the *observed* matrices. We would, no doubt, be incorrect again. But hopefully this time our error would have reduced by a little bit. We could then follow this iterative process of guessing and comparing until we reached either 1) a perfect solution (theoretically possible but practically unlikely) or 2) we reached a solution that could not be improved upon (i.e., no other set of estimates would further reduce the difference between the *observed* and *expected* matrices).

This is precisely the process a computer follows whenever it is fitting a biometrical model like the ACE model. Of course, the computer is much faster and can, thus, perform thousands of computations in the blink of an eye. Furthermore, the computer can calculate a chi-squared (X^2) test each time to determine whether the differences between the *observed* and *expected* matrices is statistically significant. Humans could easily do the latter, but that would add several minutes more of computation *per* guess.

Now that we have a working understanding of *how* the ACE model "finds" parameter estimates, let's look at a visual depiction of the model so we can consider a few final points before demonstrating how to compute it. The univariate ACE model is depicted in the diagram below.



We have previously outlined all the elements necessary to understand *how* the ACE model "finds" parameter estimates for h^2, c^2, and e^2. Thus, the visual diagram above is not a necessary component. Rather, it is a visual aid that will help some understand the estimation routine. Notice that there are a few unknown parameters in the model: a, c, and e. There are two observed indicators: phenotype_1 and phenotype_2. A, C, and E appear as latent terms and, thus, are depicted as circles. As is standard with any structural equation modeling program, we are required to either fix the variance of the latent factors to 1.00 or we must fix one of the path loadings to 1.00. The path loadings (a, c, and e) are the values of interest, so it is commonplace to set the variances to 1.00 for each of the latent balloons for A, C, and E. This is depicted as the double-headed arrow on each latent balloon.

Using the maximum likelihood routine outlined above, a computer would fill in "guesses" for a, c, and e (which appear on the single-headed arrows leading *from* the latent balloons capturing A, C, and E *to* the observed phenotypes). Notice that A, C, E, a, c, and e all appear twice; one set for each phenotype. Yet, notice that none of these elements were given subscripts. This was intentional because it reveals that they are *fixed* for both phenotypes. Indeed, we only need one set of estimates for a, c, and e to garner estimates for h^2, c^2, and e^2.

Recall that covariance between MZ twins is h^2+c^2. Relying on standard path tracing rules that state we can move along arrows in the diagram—all the while collecting the parameters we encounter—reveals the exact same solution. Specifically, begin with phenotype_1. Move up the "a" path, across the double-headed arrow (1.00), and down the "a" path for phenotype_2. This results in: $a * 1.00 * a = a^2$. The same process can be followed for the "c" paths, resulting in: $c * 1.00 * c = c^2$. Thus, the covariance between phenotype_1 and phenotype_2 for MZ twins results from $a^2 + c^2$. The same is true for DZ twins, with one exception: when we follow the "a" paths we collect the 0.50 rather than the 1.00. Thus, we end up with the following for the covariance for DZ twins: $0.50a^2 + c^2$. Hopefully you can already see the parallels between these equations and the ones outlined above for h^2 and c^2.

The total variance for either phenotype can also be computed by following path tracing rules. As we pointed out above, each of the latent factors has a variance of 1.00. Thus, when computing the total variance for, say, phenotype$_1$, you would start at the phenotype, move up the "a" path, collect the 1.00 variance, and then move back down the *same* "a" path. This results in: a^2. The same is true for the "c" and "e" paths. Thus, the total variance for *both* phenotype$_1$ and phenotype$_2$ $= a^2 + c^2 + e^2$.

If we were to fill in values for a, c, and e, we could easily compute estimates for the variance of the phenotype$_1$, phenotype$_2$, the DZ $cov_{p_1p_2}$, the MZ $cov_{p_1p_2}$, h^2, c^2, and e^2. Realizing that the total (expected) variance for the phenotypes is equal to $a^2 + c^2 + e^2$ provides several obvious solutions:

$$h^2 = a^2/(a^2 + c^2 + e^2)$$
$$c^2 = c^2/(a^2 + c^2 + e^2)$$
$$e^2 = e^2/(a^2 + c^2 + e^2)$$

5.3.2 Demonstration

Installing OpenMx in R

One of the most popular analytical techniques in all of quantitative genetics is the univariate ACE model that was depicted above. As was noted, the univariate ACE model is used to decompose variance in one phenotype at a time. This model has been fully described elsewhere (Neale & Maes, 2004; Plomin et al., 2013). The previous section offered an overview of the inner workings of the model from a conceptual angle and from a mathematical perspective. This section will demonstrate how to estimate the ACE model using simulated data so that readers can reproduce, alter, and re-estimate everything we present. All codes for the data simulations and for the estimation of the ACE model will be provided in the text. The codes presented in this chapter and in the next chapter are written in `R`, which is a freely available statistical programming package (see `https://cran.r-project.org`). It is important that you note the typeface `R` will be used to refer to the statistical program. We have already introduced a different typeface—*R*—to refer to the proportion of genetic overlap between relatives. *R* and `R` will be used in a very different capacities, so it is imperative that you recognize and understand the differences.

In order to use `R` to calculate estimates for the ACE model (and the bi-/multivariate models discussed in the next chapter), readers will need to download and install the OpenMx package (`https://openmx.ssri.psu.edu//`). If you already have installed `R`, then you can easily install the latest version of OpenMx by typing the following into the `R` console: `source('https://openmx.ssri.psu.edu/software/getOpenMx.R')`.

OpenMx is a freely available package run in `R` for estimating structural equation models. It was developed by behavioral geneticists for the purposes of estimating the biometrical

models (in addition to others) discussed in this text. It is important to note, though, that R and OpenMx are not the only packages available to researchers. On the contrary, the models presented in this text can be estimated in many statistical packages such as Python, Stata, SAS, and Mplus. Indeed, Mplus has a suite of model scripts tailored specifically to quantitative geneticists (https://www.statmodel.com/geneticstopic.shtml). We have opted to present the models in R for several reasons. R is freely available, it is extremely flexible, and it has a nice support community online (just Google "how to do _____ in R" and you will see what we mean). Finally, R has become the go-to package for many quantitative geneticists, so learning it and its programming language (actually, R *is* a programming language) will afford many benefits beyond those provided in this text.

Minimal Working Example (MWE)

Throughout this text, we will provide demonstrations and working examples of the statistical methods that are introduced. Each example will be presented with simulated data so that you can instantly replicate our results, you can make your own changes, and you can begin to learn/work with the methods introduced in the text. This is our first demonstration in the text, so it will require a little bit of introduction and background.

Our examples will be written in R code. Here is a minimal working example (MWE) of R code for you to begin familiarizing yourself with the structure of the language:

```
# anything that follows the # symbol is a comment and will not be processed by R

# let's set the working directory so all files go to the right place
setwd("/Users/JC/Box Sync/Manuscripts/Book_--_QuantitativeGenetics/_ch4")

# we're going to use the 'xtable' package in this example.  You may need to install it first:
install.packages("xtable")

# R requires that you "load" packages that will be used during your session.  We'll use 'xtable' to generate neatly formatted tables that are then placed in the text of the book.
library(xtable)

# clear the workspace
remove(list=ls())

# set seed so results are reproducible
set.seed(1)

# simulate data for variable x
x<-rnorm(100)

# simulate data for variable y, build in an association with x
y<-x*.2+rnorm(100)

# calculate the covariance between x and y
cov<-cov(x,y)

# calculate the correlation between x and y
cor<-cor(x,y)
xtable(cbind(cov, cor))

# estimate a regression model of y on x
```

```
32 lm<-lm(y~x)
33 xtable(lm)
```

As you can see, there are 33 lines of code in this MWE. Yet, roughly half of those lines of code are not code at all. Instead, they are comments that are intended to help guide you (and us!) through the script. You will always be able to identify a comment in R code because the line will begin with the pound symbol, #, as in the very first line of code in the MWE. Any line of code that begins with # will not be processed by R. Rather, it will simply be printed back to the user in any output. So you can place notes to yourself in your R code, just be sure to put # in front. This is also a good way to "comment out" broken lines of code that you are trying to work on and need to save for later.

Moving through the MWE, we see that the first line of "actual" code begins on line 4. Here, we set the working directory so that R knows where to look for and to place any files referenced during our session. The next line of code (line 7), installs a package that will be used in the MWE. The package is then, on line 10, loaded into the library for our current session. In short, you must tell R when you plan to use a command that is not pre-loaded in the base R environment. Base R has most things you will need, but certain specialized commands must be called on explicitly. This helps keep R running fast by not loading a bunch of stuff you will never use.

Next, on line 13, we clear out the memory so we can begin our current session with a clean workspace. Line 16 is where the example really takes off. Here, we set the seed so that any (pseudo)random numbers produced from here on out can be *re*produced perfectly. You can think of it like this: imagine R has a vast storage facility somewhere on your hard drive. That storage facility houses a bunch of (pseudo)random numbers. We could ask R to pull a random number at random. Let's say we do that and R spits out 5. We could then do this again. This time we might get 12. We could repeat this exercise *ad infinitum*. But what if, for some reason, you needed to perform the same exercise, but it would help if you received the same sequence of numbers? For instance, imagine you wanted to generate a dataset with random numbers but you would like to share you code with a friend so that s/he could end up with the same dataset. That scenario can be made possible if we set the seed before drawing any numbers. Setting the seed, to return to our storage facility example, is like telling R which shelf to start on when pulling the random numbers. Thus, setting the seed allows us to draw (pseudo)random numbers in a way that can be reproduced later without any alterations![4]

After setting the seed, we then ask R to generate a new object—which is R speak for something you create and store during your session. We name that object x by defining it

[4]The world of pseudo-random number generation is actually quite interesting. It turns out that even computers are incapable of producing *truly* random numbers (http://www.rand.org/pubs/monograph_reports/MR1418/index2.html). Over the long haul, when a computer is asked to produce a series of random numbers, scientists have discovered that the computer will begin to fail, meaning it will "favor" certain sequences of numbers more than others. This would be imperceptible to the human eye, but statistical tests (such as a X^2) confirm that computers eventually breakdown when asked to produce random sequences. Nonetheless, the pseudo-random numbers produced by R will be more than sufficient for our purposes.

with R speak: <-. If R were a person, it would read line 19 as: "create a new object, call it 'x' and place in that object 100 numbers drawn from a random normal distribution."

On line 22, we again ask R to create a new object, y, that has 100 observed values drawn from a random normal distribution. This time, however, we build in a correlation with the object x from above. This, in effect, will create a new variable that is correlated with a variable we have already created. Line 25 then asks R to calculate the covariance between x and y and place it in its own object, called cov. Next, line 28 asks R to calculate the correlation and place it in a new object called cor. Line 29 uses the xtable command to produce a table including the covariance and the correlation between x and y in a format that can be rendered in LaTeX, the program we have used to write this text.

Here is the table that resulted from the call to xtable on line 29:

	cov	cor
1	0.16	0.18

Of course, we can always edit the table, adding any custom formatting we might like. For instance, were this table going to be used for anything other than this demonstration, we would have omitted the "1", changed the column headings, and given the table a title.

Returning to the MWE, we see that there are only two lines of code left. Line 32 asks R to estimate a linear regression model, which R calls lm. The linear regression model is specified by listing the dependent variable (here, y), followed by the tilde "∼" and then any independent variables (here, x). The results from the linear regression model were stored in the object we named lm and then, those results were transformed into a table that could be rendered in LaTeX using the xtable command:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0377	0.0970	-0.39	0.6984
x	0.1989	0.1077	1.85	0.0678

Substantive Example: Simulating a Twin Data File

Let us now move into a working example of the ACE model. In order to use a reproducible example, we must first simulate a twin dataset. The chunk of R code shown below will accomplish that task[5]:

```
1 setwd("/Users/JC/Box Sync/Manuscripts/Book_--_QuantitativeGenetics/_ch4")
2 remove(list=ls())
```

[5]It is important that we acknowledge this code was adapted from example code provided by Hermine Maes and Benjamin Neale at the International Workshop on Statistical Genetic Methods for Human Complex Traits (March 2014), which is hosted by the Institute for Behavioral Genetics at the University of Colorado, Boulder

```
3
4  #install.packages(c("MASS","psych"))
5  library(MASS)
6  library(psych)
7
8  set.seed(2016)
9
10 asq<-0.50
11 csq<-0.25
12 esq<-0.25
13 nmz<-200
14 ndz<-500
15
16 # define covariance matrices for simulation
17 mzcov<-matrix(c(asq+csq+esq, asq+csq,
18                 asq+csq,     asq+csq+esq),2,2)
19 dzcov<-matrix(c(asq+csq+esq, 0.5*asq+csq,
20                 0.5*asq+csq, asq+csq+esq),2,2)
21
22 # simulate data using the mvrnorm command
23 # MZs first
24 mzData<-data.frame(mvrnorm(nmz,mu=c(0,0),Sigma=mzcov),rep(1,nmz))
25 colnames(mzData)<-c("p1","p2","R")
26 describe(mzData)
27 cor(mzData$p1,mzData$p2)
28
29 # DZs
30 dzData<-data.frame(mvrnorm(ndz,mu=c(0,0),Sigma=dzcov),rep(0.5,ndz))
31 colnames(dzData)<-c("p1","p2","R")
32 describe(dzData)
33 cor(dzData$p1,dzData$p2)
```

We will use two auxiliary packages during this session, so we must load them in lines 5 and 6 (uncomment line 4 if you need to install either/both packages). The `MASS` package has a great function for generating random data based on a variance-covariance matrix; perfect for creating a simulated twin file. The `psych` package has a nice function for summarizing the data, so we'll use it.

As was described above, we set the seed in line 8. Lines 10-14 store values that will be used later. Anything that is stored in an `R` object can later be recalled. So, here, we store values for `asq` (which is our way of using `R` code to represent a^2), `csq` (c^2), `esq` (e^2), `nmz` (total number of MZ twin pairs), and `ndz` (total number of DZ twin pairs).

Next, lines 17-20 set up the variance-covariance matrices. They are specified just as we saw above, where:

$$MZs: \begin{bmatrix} a^2+c^2+e^2 & a^2+c^2 \\ a^2+c^2 & a^2+c^2+e^2 \end{bmatrix}$$

becomes this in `R`:

$$\begin{bmatrix} \texttt{asq+csq+esq,} & \texttt{asq+csq,} \\ \texttt{asq+csq,} & \texttt{asq+csq+esq} \end{bmatrix}$$

And for DZs:

$$\begin{bmatrix} a^2+c^2+e^2 & 0.50a^2+c^2 \\ 0.50a^2+c^2 & a^2+c^2+e^2 \end{bmatrix}$$

$$\downarrow$$

```
[asq+csq+esq,  0.5*asq+csq,]
[0.5*asq+csq,   asq+csq+esq ]
```

Notice that the structure of the code in R is such that we define the four elements of each matrix, then we tell R to store those elements as a 2x2 matrix (this is what R "sees" when it interprets the 2,2 at the end of both the mzcov statement and at the end of the dzcov statement).

Now that we have given R all the information it will need to generate the twin data, we ask it to simulate data for the MZ twins using the mvnorm command (which is part of the MASS package). The MZ data are then placed into a data frame (using the data.frame command) and are then placed into a new object called mzData. This process is carried out in line 24 and the same sequence of events is enacted on line 30 for the DZ data. Note that lines 24 and 30 are practically identical. The main distinguishing factors are that line 24 asks mvnorm to simulate data using the mzcov matrix that was set above. Line 30, the DZ data, uses the dzcov matrix.

After creating the data, we ask R to produce basic summary statistics with the describe command (lines 26 an 32), which is part of the psych package we loaded earlier. Then, we ask for the phenotypic correlation between twin 1 and twin 2 (lines 27 and 33).

Substantive Example: Fitting the ACE Model

ACE model fitting is typically carried out in a series of steps that begin with a fully saturated model, followed by a series of trimmed models. The fully saturated model is one that fits the maximum number of parameters to the data. If you are familiar with SEM, you will recognize this as a just-identified model. The trimmed models are nested (i.e., over-identified) versions of the saturated model, where one or more parameters are dropped (i.e., set to 0.00) and the model re-estimated. Goodness of fit statistics guide the user to the best-fitting model.

To fit ACE models in R, you will need to install and load the OpenMX package (see the beginning of this section for installation instructions). The R script for estimating an ACE model is presented below. As you can see, there a few lines of code involved in estimating the ACE model. As with the MWE, we will walk you through each one so that you can begin working with these models immediately.

```
1  setwd("/Users/JC/Box Sync/Manuscripts/Book_--_QuantitativeGenetics/_ch4")
2  library(OpenMx)
3  library(xtable)
4
5  # set up data file for OpenMX
6  pNames<-"p"
7  nv<-1
8  selVars<-paste(pNames,c(rep(1,nv),rep(2,nv)),sep="")
9
10 # begin by creating matrices a, c, and e to store a, c, and e path coefficients
11 pathA<-mxMatrix(type="Lower",nrow=nv,ncol=nv,free=TRUE,values=.6,label="a11",name="a" )
12 pathC<-mxMatrix(type="Lower",nrow=nv,ncol=nv,free=TRUE,values=.6,label="c11",name="c" )
13 pathE<-mxMatrix(type="Lower",nrow=nv,ncol=nv,free=TRUE,values=.6,label="e11",name="e" )
```

```
14
15 # matrices A, C, and E compute variance components
16 covA<-mxAlgebra(expression=a %*% t(a),name="A")
17 covC<-mxAlgebra(expression=c %*% t(c),name="C")
18 covE<-mxAlgebra(expression=e %*% t(e),name="E")
19
20 # algebra to compute total variances and standard deviations (diagonal only)
21 covP<-mxAlgebra(expression=A+C+E,name="V")
22 matI<-mxMatrix(type="Iden",nrow=nv,ncol=nv,name="I")
23 iSD<-mxAlgebra(expression=solve(sqrt(I*V)),name="iSD")
24
25 # matrix & Algebra for expected means vector
26 meanG<-mxMatrix(type="Full",nrow=1,ncol=nv,free=TRUE,values=0,label="mean",name="Mean" )
27 expMean<-mxAlgebra(expression=cbind(Mean,Mean),name="expMean")
28
29 # algebra for expected variance/covariance matrix in MZ & DZ twins
30 expCovMZ<-mxAlgebra(expression=rbind(cbind(V,A+C),
31                                      cbind(A+C,V)),name="expCovMZ")
32 expCovDZ<-mxAlgebra(expression=rbind(cbind(V,0.5%x%A+C),
33                                      cbind(0.5%x%A+C,V)),name="expCovDZ")
34
35 # register the data with OpenMx
36 dataMZ<-mxData(observed=mzData,type="raw")
37 dataDZ<-mxData(observed=dzData,type="raw")
38 objMZ<-mxFIMLObjective(covariance="expCovMZ",means="expMean",dimnames=selVars)
39 objDZ<-mxFIMLObjective(covariance="expCovDZ",means="expMean",dimnames=selVars)
40
41 # set up model parameters
42 pars<-list(pathA,pathC,pathE,covA,covC,covE,covP,matI,iSD,meanG)
43 modelMZ<-mxModel("MZ",pars,expMean,expCovMZ,dataMZ,objMZ)
44 modelDZ<-mxModel("DZ",pars,expMean,expCovDZ,dataDZ,objDZ)
45 minus2ll<-mxAlgebra(expression=MZ.objective+DZ.objective,name="minus2loglikelihood")
46 obj<-mxAlgebraObjective("minus2loglikelihood")
47 ACE_Model<-mxModel("twinACE",pars,modelMZ,modelDZ,minus2ll,obj)
48
49 # estimate the model!
50 ACE_Fit<-mxRun(ACE_Model,intervals=T)
51
52 # Generate Table of Parameter Estimates using mxEval
53 varComponentsACE<-round(mxEval(cbind(A/V,C/V,E/V),ACE_Fit),4)
54 rownames(varComponentsACE)<-'varComponents'
55 colnames(varComponentsACE)<-c('a^2','c^2','e^2')
56
57 # % variance explained
58 xtable(varComponentsACE)
```

The first line of code specific to the ACE model (line 6) sets up a few parameters that will be used repeatedly when fitting the ACE model. It is easier to ask `R` to remember them up front. Then, when we need the information later, we can just ask `R` to recall the object. This is particularly useful when we have a repeated element that we then need to change. Setting up our script this way allows us to make one change that is reflected throughout the script. The first line of code names the phenotype we are analyzing. The information entered here will only be used for "filing" purposes and will not affect the estimation routine. Thus, we have chosen to name the phenotype `p` for simplicity and in keeping with our general discussion up to this point. The next two lines of code enter more information that will be used throughout the code. For instance, the code `nv<-1` creates a new object named `nv` and we place the value `1` inside. Here, `nv` is short for "number of variables." Since we are dealing with univariate models in this chapter, we will use `1`. In the next chapter, you'll notice that `nv<-2`.

The next chunk of code (lines 11-13) create three new objects that are specified as matrices and they are labeled `pathA`, `pathC`, and `pathE`. Each of these objects is created so that it will store the path coefficients for a, c, and e, that are estimated when the ACE model is run. Each of the three lines of code are set up identically (except for the names of the object, the labels, and the name that we specify at the end of each line of code; the latter two are required by OpenMX).

The next three lines of code (lines 16-18) instruct OpenMx on how to calculate the variance components of interest: A, C, and E. There is no information that should need to be changed here, so we will move on to the next three lines (lines 21-23). Here, we instruct OpenMx to set up a few more objects that will assist in the computation and that will hold information that is used during the computation. Again, there is not any information here that will need to be changed, so we will move to the next two lines: line 26 and 27. Here, we create a matrix that will hold the estimate for the mean of the phenotypes (line 26). The next line (27) instructs OpenMx on how to compute those means. Neither of these lines requires editing.

The next four lines of code (lines 30-33) instruct OpenMx on how to compute the expected covariances for MZ twins (lines 30-31) and DZ twins (lines 32-33). Notice the `0.5%x%A` in the DZ statement. This tells OpenMx to compute a covariance that only includes half of A for DZs, as we discussed earlier when setting up the conceptual background for the biometrical model.

The next four lines of code (lines 36-39) register the data (separately for MZ [line 36] and DZ [line 37] twins) so that observed variance-covariance matrices can be computed from the `raw` data. `OpenMx` works nicely when the data have been parsed into the two groups of focus. Thus, why we never combined the MZ and DZ data files into a single data file.

Then, we create two objects to hold the expected variance-covariance matrices (lines 38 and 39). These lines of code require no editing. Below that, lines 42 through 47 pull everything together and format the information in a way that OpenMx can understand. As with the previous chunk of code, these lines require no editing.

We are now ready to estimate the ACE model! Line 50 of the code will estimate the model and store the results in an object we call `ACE_fit`. Below that, we format the information in `ACE_fit` so that it will be interpretable to a human (lines 53-55). Finally, we ask `R` to show us the output, which reflects the proportion of variance in the phenotype explained by a^2 (i.e., h^2), c^2, and e^2. The values that resulted from this script are presented in the table below.

	a^2	c^2	e^2
varComponents	0.60	0.14	0.26

The table reveals that the ACE model has converged on estimates of $h^2 = 0.60$, $c^2 = 0.14$, and $e^2 = 0.26$. In other words, 60% of the variance in the simulated phenotype was explained

by h^2. The shared environment (c^2) accounted for 14% and the nonshared environment (e^2) accounted for the remaining 26%.

From Saturated to Best-Fitting Model

As we noted earlier, ACE model fitting often follows a series of steps from a fully saturated model to a trimmed "best-fitting" model. The full model was fit in the previous section.[6] Here, we'll quickly show how you can fit trimmed models and how you can determine which is the best-fitting model using model fit statistics.

The `R` script below will fit a trimmed version of the model presented above. Specifically, we will drop the c^2 parameter since it was the smallest parameter. Then, we will compare model fit between the ACE model (above) and the AE model estimated here.

```
1  # AE model
2  AE_Model<-mxModel(ACE_Fit,name="AE")
3  AE_Model<-omxSetParameters(AE_Model,label="c11",free=FALSE,values=0)
4
5  # drop c at 0
6  AE_Fit<-mxRun(AE_Model)
7
8  # compare trimmed AE model with fully saturated ACE model estimated earlier
9  aceAE<-mxCompare(ACE_Fit,AE_Fit)
10 xtable(aceAE)
11
12 # Generate Table of Parameter Estimates using mxEval
13 varComponentsAE<-round(mxEval(cbind(A/V,C/V,E/V),AE_Fit),4)
14 rownames(varComponentsAE)<-'varComponents'
15 colnames(varComponentsAE)<-c('a^2','c^2','e^2')
16 varComponentsAE
17 estimates<-rbind(varComponentsACE, varComponentsAE)
18 xtable(estimates)
```

Since we already set up the ACE model, we have very few "new" things that must be done to estimate the AE model. In fact, we only have three lines of code that are necessary to tell `OpenMx` to drop C and re-calculate the parameter estimates. As you see in line 2, we ask `OpenMx` to recall the `ACE_Fit` that was estimated earlier. Line 3 then instructs `OpenMx` to set the `c11` path to `value=0`. This will fix any path that draws on `c11` to 0.00, thus giving us an AE model. Line 6 re-estimates the model. Line 9 compares the model fit statistics between the ACE model (listed first, this is important because it is the model to which we want to compare the AE model; always list the comparison model first) and the AE model. The results from this comparison are listed in the table below.

	base	comparison	ep	minus2LL	df	AIC	diffLL	diffdf	p
1	twinACE		4	3675.79	1396.00	883.79			
2	twinACE	AE	3	3679.10	1397.00	885.10	3.31	1.00	0.07

[6] We refrain from referring to the previously fit model as a "fully saturated model" because it constrained the means and variances to be equal across twin type. Technically speaking, a fully saturated model would allow *all* parameters, including the means and variances, to vary across twin types.

The table provides several important pieces of information. The first row reveals the model fit statistics for the ACE model. The second row reveals the model fit statistics for the AE model, along with comparisons between the ACE model and the AE model (the last three columns). As we can see from the table, the X^2 comparison between the ACE model and the AE model ($X^2 = 3.31$) was not statistically significant, though it was close ($p = 0.07$). In a borderline case such as this one, one is justified in selecting either the ACE or the AE model as the best-fitting model.

The parameter estimates gleaned from the AE model are collected, stored, and formatted in `varComponentsAE` in lines 13-15. Line 17 combines parameter estimates from the ACE model with those gleaned from the AE model. The estimates from both models are presented in the table below (from line 18). The ACE model results appear in the first row and the AE model estimates are in the second row.

	a^2	c^2	e^2
1	0.60	0.14	0.26
2	0.75	0.00	0.25

5.4 The Regression-based DeFries-Fulker (DF) Model

The ACE model discussed above utilizes structural equation modeling (SEM) to decompose variance in the phenotype of focus. Its flexibility to integrate control variables, to be extended to the bi-/multivariate case (discussed in the next chapter), and its ability to seamlessly estimate h^2, c^2, and e^2 have made the ACE model one of the most popular approaches to decomposing variance. It is not the only approach available, though. On the contrary, a popular alternative is referred to as the DeFries-Fulker (DF) model after the scholars who proposed the approach in the 1980s (DeFries & Fulker, 1985).

The DF model is a regression-based approach, meaning it does not require one to utilize SEM. This can be seen as both an advantage of the DF model and a disadvantage. Advantages of using a regression-based model are that the estimation routine is simpler (we mean this practically, the codes to estimate the DF model are a fraction of the length necessary to estimate the ACE model; computationally, they're about the same), model-fitting routines are more straight-forward (typically, the user only needs to estimate three models), and it can be estimated in any statistics package that has regression capabilities (i.e., one does not need special software like is necessary for the ACE model). Of course, though, these advantages are counterbalanced by certain disadvantages of the DF model compared to the ACE model. For example, the DF model does not directly estimate e^2. We will discuss this limitation below.

Nonetheless, the DF model is one of the most concise and easy-to-use biometrical models available. Given that criminologists are exposed to the ordinary least squares (OLS) model as part of their graduate training, we anticipate that the DF model will be the favored ap-

proach to integrating quantitative genetics into criminology. Indeed, the estimation routine is intuitive, with only a few steps the user must be sure to follow. We outline the estimation routine, along with a mathematical treatment of the DF model, in the next section.

5.4.1 Conceptual Discussion

The DF model can be expressed as a simple regression equation with four observed variables and four parameter estimates:

$$Y_i = \beta_0 + \beta_1(X_{1i}) + \beta_2(X_{2i}) + \beta_3(X_{3i}) + \epsilon_i$$

As you can see, the equation above is nothing more than a multiple OLS model (see generally, Wooldridge [2014]). Translating the standard regression model into one that will decompose variance in a phenotype requires that one have access to a genetically informative dataset like the one presented below (which is the same dataset that was presented earlier when discussing the ACE model):

Pair ID	Pair Type	V_A Shared (R)	phenotype$_1$	phenotype$_2$
1	MZ	1.00	5	5
2	MZ	1.00	7	8
3	MZ	1.00	3	1
⋮				
100	MZ	1.00	10	9
101	DZ	0.50	4	9
102	DZ	0.50	6	5
103	DZ	0.50	10	2
⋮				
200	DZ	0.50	6	7

In essence, if you wish to decompose the variance in a phenotype into genetic and environmental sources, then you will need information from pairs of individuals with known genetic relationships (R) *and* you will need at least *two* levels of genetic relationships (R). We have the first requirement covered in the dataset above by knowing the proportion of V_A shared (i.e., R) for each twin pair. We have the second requirement covered by including MZ *and* DZ twins in the same dataset. Note, however, that it is not necessary to analyze MZ and DZ twins. One is free to choose any level of genetic relatedness (e.g., full siblings, half-siblings, parent-offspring). MZ and DZ twins are the most common, however, because they simplify some (though not all) of the assumptions that will be outlined at the end of this chapter. Analyzing different types of pairs (e.g., parent-offspring) sometimes requires additional assumptions and/or violates other assumptions in known ways.

There are three steps to estimating the DF model on a dataset like the one above. The first step is to generate an observed variable that will be used to estimate h^2. In order to

understand what this might look like, it is important to take a step back and contemplate what a regression parameter estimate (i.e., the βs) is actually telling us. From the regression equation presented above, β_0 is known as the Y-intercept and it provides an estimate of the expected value of Y when all of the X variables are set to 0.00 (i.e., Y-intercept $= \mathbb{E}[Y|X_1 = 0, ..., X_k = 0]$). All of the other βs in the regression equation tell us how much the expected value of Y changes for a one-unit change in the focal variable X. For instance, $\beta_1 = \frac{\Delta Y}{\Delta X_1}$. Let us imagine that Y represents phenotype$_1$ (i.e., the phenotypic score for twin 1) and X_1 represents phenotype$_2$ (i.e., the phenotypic score for twin 2). Let us also imagine that $\beta_1 = 0.25$. One could interpret β_1 by noting that a one-unit increase in the phenotype for twin 2 predicts a 0.25 increase in the phenotype for twin 1. In short, β_1 in this example provides information about the *covariance* between phenotype$_1$ and phenotype$_2$. Recall from our discussion of the ACE model that the covariance plays a central role in decomposing variance into h^2, c^2, and e^2. The same is true for the DF model.

As we saw during our discussion of the ACE model, it is necessary to estimate the covariance between phenotype$_1$ and phenotype$_2$ across a minimum of two levels of genetic relatedness in order to generate estimates of h^2, c^2, and e^2 (see our discussion of parameter indeterminacy above). The same is true in the DF model. Thus why we included MZ and DZ twins in the heuristic dataset.

As you have probably deduced by now, we are going to need to devise a way to estimate the covariance between phenotype$_1$ and phenotype$_2$ for MZ twins and separately for DZ twins. In the context of the DF model, we need a way to estimate b_1 once for MZ twins (i.e., $\beta_{1.MZ}$) and again for DZ twins (i.e., $\beta_{1.DZ}$). As you are no doubt aware, what we have just described is known as a statistical interaction. In short, we need to find a way to statistically interact b_1 with a variable that taps into the level of genetic overlap shared by the pairs of focus. We already have all the necessary information in the dataset from above; recall that R indexes the level of genetic overlap (1.00 for MZ twins and 0.50 for DZ twins). Thus, we have an obvious solution: generate a variable that will tap into the multiplicative interaction between β_1 and R. This is easily done by simply creating a new variable, call it "*inter*", that is $R*$phenotype$_2$. Include this new variable on the right-hand side of the regression model, along with the two constituent terms, and the OLS model from above becomes:

$$phen_{ij} = \beta_0 + \beta_1(phen_{i'j}) + \beta_2(R_j) + \beta_3(inter_{i'j})$$

where i indexes the individual respondent (i.e., $i = \{1, 2\}$), i' is the co-twin for the focal respondent i, j indexes the twin pair to which individual i and individual i' belong, $phen_{ij}$ is phenotypic score for the focal twin i from pair j, $phen_{i'j}$ is the phenotypic score for the co-twin i' from pair j, R_j is the level of genetic overlap shared by the twins from pair j ($R_j = 1.00$ when the pair is MZ and $R_j = 0.50$ when the pair is DZ) from pair j, and $inter_{i'j}$ is the multiplicative interaction between $phen_{i'j}$ and R_j.

The regression model presented above is consistent with the DF model that was originally proposed by DeFries and Fulker in 1985. As any student of econometrics/statistics will no doubt recognize, however, there are a few concerns that can easily be addressed if this model is augmented slightly. Two concerns stand out. First, multicollinearity between $inter_{i'j}$ and

its constituent terms $phen_{i'j}$ and R_j can be reduced if one "centers" the constituent variables prior to generating $inter_{i'j}$. Second, the substantive meaning of the constituent term R_j is not obvious. More directly, it is not clear how we should interpret β_2. What does it *mean*? As it turns out, β_2 captures the impact of sharing more genetic relatedness on the expected value of $phen_{ij}$. If we are willing to assume that the mean level of $phen_{ij}$ does not vary across levels of genetic relatedness (something that is often observed in the descriptive phase of any study), then we can safely omit R_j from the equation, simplifying things to (Rodgers & Kohler, 2005):

$$phen_{ij} = \beta_0 + \beta_1(phen_{i'j} - \bar{phen}_{i'}) + \beta_2(inter_{i;j})$$

The above equation now provides estimates of c^2 and h^2. Specifically, β_1 provides an estimate of c^2 and b_2 provides an estimate of h^2. As with any "normal" OLS model, inferential tests can be carried out because standard errors for all of the parameter estimates will be calculated. Thus, one can compute confidence intervals and perform tests of statistical significance just like with the ACE model.

You may be wondering *why* β_1 estimates c^2 and b_2 estimates h^2. The logic is straightforward and intuitive. Let us start with h^2. Recall from above that the *inter* variable is a multiplicative interaction between R and $phen_{i'j}$. The logic of this measure was outlined above and, hopefully, you can see how *inter* approximates the information we glean from the ACE model by fitting a variance-covariance matrix separately for MZs and DZs. Recall also that the covariance between $phen_1$ and $phen_2$ is explained by h^2 and c^2. If *inter* estimates h^2, then the *only* remaining source of covariance between $phen_1$ and $phen_2$ is a product of c^2. Thus, the remaining covariance between $phen_1$ and $phen_2$ *must* be captured by β_1.

At this point you may also be wondering a few additional things. First, you might be asking yourself how we garner an estimate of e^2 from the DF model. Recall from our discussion of the ACE model that only h^2 and c^2 contribute to the covariance between pairs of genetically related individuals. e^2 contributes to phenotypic variance, but not the covariance. Because the DF model unpacks the *covariance* between phenotype_1 and phenotype_2, there is no direct way to estimate e^2. One can, however, derive a point estimate for e^2 (but standard errors, confidence intervals, and tests for statistical significance cannot be calculated): $e^2 = 1 - h^2 + c^2$.

The second point you may be wondering about is how we conduct the model-fitting steps that were outlined for the ACE model. The model-fitting procedure for the DF model is considerably shorter than for the ACE model because the former does not directly model variances, only covariances. Thus, we need not fit a fully saturated model and sequentially work through more parsimonious models like with the ACE model. That does not mean we automatically accept the results gleaned from the "full" DF model. On the contrary, we *can* fit a version of the fully saturated model and then trim it until we reach a more parsimonious solution. First, fit the fully saturated model:

$$phen_{ij} = \beta_0 + \beta_1(phen_{i'j} - \bar{phen}_{i'}) + \beta_2(R) + \beta_3(inter_{i'j})$$

If β_2 is not statistically significant (as is often the case), fit the more parsimonious model

outlined above:

$$phen_{ij} = \beta_0 + \beta_1(phen_{i'j} - \bar{phen}_{i'}) + \beta_2(inter_{i'j})$$

Observe the statistical significance of β_1 and β_2. If either is *not* statistically significant, drop it from the model. If *both* are statistically significant, terminate model-fitting because the most parsimonious solution has been reached. If *neither* are statistically significant, we suggest omitting $phen_{i'j} - \bar{phen}_{i'}$ and re-estimating the model (something similar to the AE model). We suggest omitting $phen_{i'j} - \bar{phen}_{i'}$ rather than *inter* in light of the three laws of behavioral genetics which state that c^2 is often negligible (Turkheimer, 2000). One can perform F tests to determine the best-fitting solution. This is similar to the X^2 test discussed above for the ACE model.

Two final issues should be considered before we move to the demonstration of the DF model. First, the user should always consider whether the data are "single-entered" or "double-entered". The distinction between the two is easy to understand. Single-entered data look like the heuristic data we have presented up to this point. Each row in the dataset represents a set of genetically related pairs. The columns index the information for twin 1 (i.e., $phen_1$) and twin 2 (i.e., $phen_2$). Estimating the DF model on single-entered data requires no further adjustments (as is true for the ACE model).

Double-entered data "look" a little different than single-entered data. An example of a double-entered dataset is presented below.

Pair ID	Twin ID	Pair Type	V_A Shared (R)	$phen_1$	$phen_2$
1	1	MZ	1.00	7	8
1	2	MZ	1.00	8	7
2	1	MZ	1.00	3	1
2	2	MZ	1.00	1	3
⋮					
100	1	MZ	1.00	10	9
100	2	MZ	1.00	9	10
101	1	DZ	0.50	4	9
101	2	DZ	0.50	9	4
102	1	DZ	0.50	6	5
102	2	DZ	0.50	5	6
⋮					
200	1	DZ	0.50	6	7
200	2	DZ	0.50	7	6

As you can see, the dataset has been transformed so that each twin pair now occupies *two* rows of the datafile: one row for twin 1 and one row for twin 2. The twins are now identified by their own unique identifier, Twin ID, which takes on two values: 1 for twin 1 and 2 for twin 2. In short, a double-entered dataset is akin to a longitudinal dataset that has been formatted in "long" structure. The single-entered datasets we have been dealing with up to

this point are similar to a longitudinal dataset that has been formatted in "wide" structure. As with longitudinal data, the formatting of the data does not pose any estimation problems as long as the user records the formatting choice with his/her estimation program. For instance, Stata allows the user to "set" whether the data are long or wide. When estimating the DF model, it is not technically necessary that one "set" the formatting structure with the estimation program. Instead, it is only important that you, the user, know how your data are formatted so certain corrections can be made.

The primary concern here is that double-entered data (i.e., long format) "trick" the estimation package into thinking you have twice as many cases as you actually do. Think of it like this, when the data are single-entered, your statistical package will recognize j rows, where j is the number of twin pairs, and it will use this value in the denominator when calculating statistics such as the standard error. If you were, instead, to feed in a double-entered dataset, all of the parameters for the DF model would be identical because you've simply duplicated the single-entered data file, but *now* your statistical program thinks you have doubled the sample size because each j will appear twice. In essence, your statistical package will assume $j * 2$ is your effective sample size when it should really be j.

There are two ways to handle this problem. First, and most obvious, is to reformat a double-entered data file into a single-entered data file. This can easily be done by randomly dropping one twin from each twin pair (e.g., drop if Twin ID is equal to 1 [assuming, of course, that the Twin ID values are randomly assigned]). The second solution—which is the one we often use—is to correct the standard errors from the DF model to account for the clustering of cases within families j. This is similar to the approach utilized by developmental researchers who need to correct standard errors for the clustering of data points within cases (e.g., i is clustered within j). Standard error corrections are available in all statistical packages. Be sure to correct for "clustering" if you choose this route.

Why $h^2 \neq R^2$

Readers who are familiar with the OLS model will, almost certainly, also have a working understanding of what R^2 provides. Briefly, R^2 tells us the proportion of the variance in the outcome variable that has been explained by the independent variables. Let Y stand for the outcome variable and X stand for the lone independent variable. If we were to estimate the association between X and Y using an OLS equation, it would take the following form:

$$Y_i = \beta_0 + \beta_1(X_i)$$

From there, R^2 is typically calculated as:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

where $\hat{Y}_i$ is the predicted value of Y for person i that is gleaned from the regression equation (i.e., by observing the values for β_0 and β_1 and then plugging in the observed value of X for person i), $\bar{Y}$ is the mean of Y, and Y_i is the observed value of Y for person i.

Here, R^2 is interpreted as the degree to which variation in Y is explained by variation in X. When estimating the DF model, then, one might expect that $h^2 = R^2$. Or, at a minimum, we might expect $h^2 \approx R^2$. But, as it turns out, this is not the case. In fact, h^2 and R^2 are often quite different and will only come close to coinciding in certain types of "boundary condition" cases.[7]

Thus, one might ask, "why is it that $h^2 \neq R^2$?" Part of the answer is that they are simply different metrics that tell us different things. But, in many ways, this is an unsatisfying—and, if we are being honest, a lazy answer—because R^2 provides an estimate of the proportion of variation in Y that is explained by knowing X. And we told you in chapter 3 that $h^2 = \frac{V_A}{V_P}$, which, in words, is the proportion of the variance in P that is explained by the variance in additive genetic factors. Thus, it would seem that R^2 and h^2 should be linked in some way. They are are, in fact, linked, but the point is that the link is not a 1:1 relationship. Rather, R^2 is a metric that is calculated based on the properties and the "fit" of the DF model. h^2 tells us how much of the variance in the phenotype is explained by genetic factors, so they *are* capturing different, albeit related, things.

Another way of looking at the discrepancy between h^2 and R^2 is to, once again, distinguish between a correlation coefficient r and its squared cousin r^2. We noted earlier in this chapter (see footnote X) that r^2, and in the present context R^2, deals with the degree to which X *predicts* Y. But in the DF model, we are not concerned about *predicting* twin 1's score on P based on twin 2's score on P. Rather, we simply want to know how much of the variance in P is attributable to h^2. We garner an estimate of h^2 by observing different types of twins who have known levels of genetic overlap. Thus, we are less concerned about the degree to which twin 1's score is a good predictor of twin 2's score, and instead we are using the correlation r between the two scores to provide the information necessary to compute h^2.

5.4.2 Demonstration

As with our discussion of the ACE model, we will briefly reveal how one can estimate the DF model by first simulating a twin dataset and second calculating the DF model to retrieve parameter estimates for h^2 and c^2 (recall that e^2 is not directly estimated in the DF model).

We will rely on the same simulated twin dataset from above. As a reminder, here are the codes in `R` to simulate the file:

[7]For example, imagine you estimate a DF model on a data file that contains MZ and DZ twins only (further, let us simply and assume there are only two twins in each twin pair and that there are no missing data points). In this case, h^2 will *approach* $\sqrt{R^2}$ as the ratio of MZ pairs to DZ pairs approaches infinity.

```
setwd("/Users/JC/Box Sync/Manuscripts/Book_--_QuantitativeGenetics/_ch4")
remove(list=ls())

#install.packages(c("MASS","psych"))
library(MASS)
library(psych)

set.seed(2016)

asq<-0.50
csq<-0.25
esq<-0.25
nmz<-200
ndz<-500

# define covariance matrices for simulation
mzcov<-matrix(c(asq+csq+esq, asq+csq,
                asq+csq,     asq+csq+esq),2,2)
dzcov<-matrix(c(asq+csq+esq, 0.5*asq+csq,
                0.5*asq+csq, asq+csq+esq),2,2)

# simulate data using the mvrnorm command
# MZs first
mzData<-data.frame(mvrnorm(nmz,mu=c(0,0),Sigma=mzcov),rep(1,nmz))
colnames(mzData)<-c("p1","p2","R")
describe(mzData)
cor(mzData$p1,mzData$p2)

# DZs
dzData<-data.frame(mvrnorm(ndz,mu=c(0,0),Sigma=dzcov),rep(0.5,ndz))
colnames(dzData)<-c("p1","p2","R")
describe(dzData)
cor(dzData$p1,dzData$p2)
```

Now, let's see how to estimate the DF model using the simulated data from above. As we have noted throughout this chapter, the DF model can be thought of as a "typical" OLS model. Thus, we can rely on `R`'s built-in linear model command (i.e., `lm` in `R`). The code below shows how to estimate the DF model in the order described above.

```
1  library(xtable)
2
3  #combine MZs and DZs into one dataset
4  data<-rbind(mzData,dzData)
5  head(data)
6  head(mzData)
7  tail(data)
8  tail(dzData)
9  describe(data)
10
11 # mean center p2
12 data[,4]<-data$p2-mean(data$p2)
13 colnames(data)[4]<-"c_p2"
14 head(data)
15
16 # create "inter"
17 data[,5]<-data$c_p2*data$R
18 colnames(data)[5]<-"inter"
19 head(data)
20
21 # estimate full saturated DF model
22 df1<-lm(p1~c_p2+R+inter,data=data)
23 summary(df1)
24 xtable(df1)
25
26 # estimate DF model: drop R
27 df2<-lm(p1~c_p2+inter,data=data)
28 summary(df2)
29 xtable(df2)
30
31 # estimate DF model: drop R & c2
32 df3<-lm(p1~inter,data=data)
33 summary(df3)
34 xtable(df3)
```

Our first order of business is to combine the MZ and the DZ files into a single dataset that can be used to estimate the DF model. This is achieved on line 4 where we create a new object named `data` by relying on the base `R` command `rbind` which will combine the two objects `mzData` and `dzData` by stacking the former on top of the latter; just as we would see with a "real" twin data file. Lines 5-9 ask `R` to show us the data file in various ways: `head` shows the top 10 rows, `tail` shows the bottom 10 rows.

Next, we must create the mean-centered version of $phen_2$ and the *inter* variable. These are done in lines 12-19.

The table for the first DF model estimated on line 22 (i.e., the fully saturated DF model) appears below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0971	0.0950	-1.02	0.3074
c_p2	0.1631	0.0958	1.70	0.0891
R	0.0705	0.1395	0.51	0.6136
inter	0.5686	0.1362	4.17	0.0000

As you can see, all four parameters discussed above are estimated. Note that above we

referred to them generally as β_0, β_1, β_2, and β_3. R provides the parameter estimates for each of the variables by listing the name of the variable. So, in the table above, β_0 = Intercept, β_1 = c_p2, $\beta_2 = R$, and β_3 = inter.

Let's now turn to the substantive interpretations that can be gleaned from the above table. Rarely will you need to interpret the Intercept value (i.e., β_0), so we will skip that estimate. Recall that the parameter estimate for c_p2 (i.e., β_1) provides an estimate of c^2. Recall also that we simulated this data file so that c^2 would equal 0.25 on average (i.e., if we were to repeat the simulation an infinite number of times, the average value for all the β_1s would be 0.25). In this particular data file, the estimate for c^2 was 0.1631, meaning that 16% of the variance in the phenotype of focus was explained by shared environmental (c^2) influences. The difference between the observed value of 0.1631 and the specified value of 0.25 is due to the random error injected into the simulations. Note that the estimate for c^2 was not statistically significant at the conventional $\alpha = 0.05$ level, but that it was approaching significance since the p-value shown in the last column was 0.0891.

Moving to the estimate for R, we see that the value is substantively small (0.0705) and it is not statistically significant ($P > 0.05$). Given the latter observation of a non-significant t-value, we will drop R from the next model. For now, turn your attention to the coefficient estimate for *inter*, which as we noted above provides an estimate of h^2. Note that the estimated value was 0.5686—indicating that approximately 57% of the variance in the phenotype was explained by additive genetic factors (h^2). The parameter was statistically significant ($P < 0.05$).

Dropping R from the DF model results in (lines 27-29):

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0518	0.0315	-1.64	0.1006
c_p2	0.1627	0.0957	1.70	0.0896
inter	0.5695	0.1361	4.18	0.0000

Here, we see that the estimate for c^2 was relatively unaffected by the omission of R. Similarly, the estimate for h^2 was relatively unaffected. One point to note is that the estimate for c^2 remained non-significant ($P > 0.05$), so one could reasonably drop this parameter from the model based on the information provided by the model-fitting routine. Doing so results in (lines 32-34):

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0526	0.0315	-1.67	0.0958
inter	0.7877	0.0454	17.34	0.0000

Here, we see that the estimate for h^2 was increased, which makes sense given that it is the only parameter now included to explain the covariance between $phen_1$ and $phen_2$. Yet, keep

in mind that the "true" parameter is 0.50 so the present scenario reveals that dropping c^2 from the model will have biased h^2 upward. Such was the case for the present example, but this may not necessarily be borne out in every situation.

Because the ACE model is the more common approach, we thought it would be prudent to show that the DF model performs well in terms of estimating c^2 and h^2. The only way to truly do this is to estimate the above models a large number of times, each time simulating a "new" dataset. This will allow us to parse out the random noise that is injected into each unique simulation, so that we can see how the DF model performs *on average*. This is the real test of how well the DF model performs.

Thus, to do so, we simulated 10,000 datasets (`R` code for this simulation is included in the appendix). Each of them were simulated using the format outlined above. We then estimated the trimmed DF model that included a parameter for c^2 and h^2. The average estimates for both parameters are listed in the table below. As you can see, the average estimate (with rounding at the fourth decimal place) for c^2 was 0.25 and the average estimate for h^2 was 0.50. In short, the DF model will, on average, retrieve the correct values for h^2 and c^2.

	Estimates from 10,000 simulations
p2	0.25
inter	0.50

5.5 Assumptions & Limitations

Before concluding this chapter it is important to outline the various limitations and assumptions that underlie the univariate biometrical models discussed above. Keep these limitations and assumptions in mind as you work through the next few chapters as well. We will periodically return to these points when we discuss more advanced models. In some cases, the more advanced models overcome the issues outlined below. In other cases, the advanced models require these assumptions plus additional ones.

5.5.1 The Equal Environments Assumption (EEA)

The equal environments assumption (EEA) can be stated succinctly as the assumption that *genetic factors are solely responsible for the increased similarity between MZ twins relative to DZ twins*. The EEA allows one to solve for h^2, c^2, and e^2 using the methods above by assuming V_C (i.e., the shared environment) carries a similar influence across all sibling pairings (e.g., MZ and DZ). In an attempt to assess the current empirical reality of the EEA, Barnes and colleagues (2014) performed an exhaustive search of the literature bearing directly on the EEA. Their systematic review revealed that violations of the EEA had been empirically tested for more than 1,200 environments and violations were detected in only

9% them. There are a few studies that estimated the impact of unequal environments on h^2 estimates, with the average effect being an upward bias of about 0.012 (or about 1%) in the h^2 estimate. The review by Barnes et al. (2014) was quite revealing and convincingly showed that the EEA likely has little-to-no influence on heritability estimates.

This conclusion is backed by the use of an ingenious methodological design: misclassified twin samples. More specifically, in samples where zygosity is determined via responses to self-report questionnaires tapping confusability, misclassifications can occur where MZ twins are initially classified as DZ twins and vice versa. Once genotyping tests are conducted, however, these classifications can be corrected. A number of studies have drawn on this unique situation to employ a more robust test of the EEA (e.g., Conley et al., 2013). This situation presents an ideal way to test whether the EEA is violated and whether such violations result in meaningful changes in estimates of h^2 and c^2. Assuming that unequal environments result in biased h^2 estimates, DZ twins which are mistaken as MZ twins should more closely resemble one another across the phenotypes of interest relative to correctly identified DZ twins. Similarly, if critics of twin studies are correct, then MZ twins incorrectly classified as DZ twins should be less similar to one another across the examined phenotypes relative to correctly classified MZ twins. The results do not follow this pattern. Instead, twin similarity is best explained by pair type (i.e., MZ or DZ) rather than misclassification status.

Nonetheless, if the EEA fails, then estimates from the ACE model and the DF model will be biased, with h^2 systematically *over*estimated and c^2 systematically *under*estimated. The logic behind this conclusion is rather simple: if certain types of siblings/twins receive more V_C than other types of siblings/twins, then those siblings will be more similar to one another as a result of having greater levels/impacts of V_C. This becomes problematic because critics often argue that MZ twins will receive greater levels of V_C relative to DZ twins because they tend to look more similar to one another than DZ twins. Directly related to these observations, critics of twin research have correctly pointed out that MZ twins tend to have more environments in common relative to DZ twins (Loehlin & Nichols, 1976).

Given these concerns, it is understandable why the EEA has garnered so much attention. Yet, despite the apparently damaging effects of EEA violations, Barnes and colleagues (2014) revealed that it is a robust assumption (Carey, 2003); one that can be violated but that does *not* lead to large biases in parameter estimates. When Barnes et al. (2014) simulated EEA violations, their models revealed that realistic levels of violation biased parameter estimates no more than by about 5 percentage points. Thus, it appears that Carey's (2003, p. 301) take on the EEA was correct:

> Taken together, all these lines of evidence suggest that the equal environments assumption meets the definition of a robust assumption. A *robust assumption* is one that might actually be violated, but the effect of violating the assumption is so small that the estimates and substantive conclusions are not altered. For example, Newtonian physics is incorrect, but one can use Newtonian principles to build a bridge or design a skyscraper. In these situations, the assumptions of Newtonian physics are robust even though they are technically wrong.

5.5.2 Random Mating

The assumption of random mating is required to fit all biometrical models because this assumption defines the R in the variance-covariance matrix for all pair types except for MZ twins. Take, for example, the covariance between DZ twins: $0.50h^2 + c^2$. The assumption of random mating defines the 0.50 portion of the equation. When two humans reproduce, germline cells formed through meiosis (which is the process of genetic mixing for sexual reproduction) fuse and form the zygote that will eventually develop into an independent and genetically unique human (Carey, 2003). As a consequence of meiosis and fertilization, a quasi-random 50% of the genes from each parent (i.e., a random 50% maternally and a random 50% paternally) are combined to create the offspring which will not be genetically identical to either parent at all genetic loci. Thus, one may assume that any offspring produced by two humans will be 50% similar to their mother and 50% similar to their father at the distinguishing loci. Based on this logic, full siblings and DZ twins are 50% similar, on average, within the distinguishing regions of the genome.

There is an impressive body of research regarding mate similarity (see Barnes et al., 2014). This literature is so consistent that scholars have now taken to concluding that "In human populations, assortative mating is almost universally positive, with similarities between partners for quantitative phenotypes[1?6], common disease risk[1,3,7-10], behaviour[6,11], social factors[12-14] and personality[4,5,11]" (Robinson et al., 2017:1). In other words, the assumption of random mating is unlikely to hold except for the rarest of phenotypes.

Findings from a diverse line of scholarship, moreover, suggest humans select mates who display similar levels of antisocial and aggressive behavior (Boutwell et al., 2012; Capaldi et al., 2008; Rowe & Farrington, 1997). Each of these analyses reports a positive and statistically significant correlation between mates for antisocial outcomes. In short, empirical evidence shows that sexual partners do not mate randomly, and thus the the assumption of random mating is likely consistently violated in biosocial criminological research.

The process of meiosis ensures that, on average, full siblings and DZ twins will share 50% of their distinguishing genotype if mating is random. If mating is not random, then the 50% figure may be an underestimate which would lead to underestimates of h^2 in the biometrical models. Substituting hypothetical values and solving for h^2 clearly indicates that a trait which is completely influenced by V_A will produce a h^2 estimate that is below 1.00 due to a violation of this assumption. If the assumption of random mating is violated, then the 0.50 value in the DZ correlation matrix will be too low, producing an overall estimate of h^2 that is below 1.00 because the correlation for MZ twins will be 1.00 but the correlation for DZ twins will be above the expected 0.50; the value will reflect the amount of genetic correlation that is actually present. When this occurs in practice, the biometrical models attribute any portion of the DZ correlation that is above 0.50 to c^2. Thus, violation of the random mating assumption leads to inflated estimates of c^2 and deflated estimates of h^2.

5.5.3 No $G \times E$ & No rGE

Researchers using the classic twin design often assume no covariance (i.e., rGE) and no interactions (i.e., $G \times E$) between genetic and environmental influences. By assuming no covariance between genetic and environmental influences, researchers can simplify the biometrical model because the $2cov(x, y)$ statements are not necessary. This assumption may be necessary to identify mathematically the traditional biometrical model, although previous studies have revealed that $G \times E$ and rGE can be addressed with extended latent variable modeling strategies (Purcell, 2002). For the sake of brevity, suffice it to say that researchers often assume $G \times E$s and rGEs are not present, allowing for a simplified equation.

This assumption may be problematic given the ever-growing base of research into both sources of gene-environment interplay (Dick, 2011; Kendler & Baker, 2007). A violation of this particular assumption could potentially result in overestimated or underestimated parameters. In general, $G \times E$s that occur between additive genetic effects (i.e., V_A) and shared environmental factors (i.e., V_C) result in overestimated h^2 and underestimated c^2. Heritability is underestimated and e^2 is overestimated, however, when the $G \times E$ is between V_A and V_E. When rGEs underlie the covariances, any correlation between V_A and V_C is attributed to the e^2 and any correlation between V_A and V_E is attributed to h^2 (Purcell, 2002).

Moreover, Del Giudice (2016) recently showed that biometrical models may "hide" $G \times E$s that are consistent with theories of human plasticity (see differential susceptibility theory [Belsky, 1997; 2005] and biological sensitivity to context theory [Boyce and Ellis, 2005; Ellis and Boyce, 2008]). These theories—and, more broadly, research into $G \times E$—are growing in popularity and there is now a lot of evidence to suggest they should be given closer attention. That is to say, the assumption of no $G \times E$ is likely to be one that should be relaxed in many circumstances, meaning researchers will need to estimate models that allow for $G \times E$ to have a non-zero effect on V_P. How to do this is in a biometrical model is beyond the scope of this text, so we point the reader to Purcell's (2002) discussion. How to allow for $G \times E$ in molecular genetic studies will be the focus of chapter ??.

5.5.4 No Dominance & No Epistasis

The V_D parameter captures dominance deviations (see Chapter 3). Dominance involves the interaction of alleles at the same loci (i.e., the same genotype) in a Mendelian inheritance framework. If one assumes genetic dominance does not influence the trait under investigation, then the total variance equation simplifies to the ACE model. If V_D matters but is ignored it may upwardly bias estimates of V_A and V_C.

Epistasis (V_I) is the second type of non-additive genetic effect that may influence variance in a trait. Epistasis occurs when two genes at separate loci interact to influence the phenotype. This is only applicable for traits that are influenced by multiple genotypes (i.e.,

polygenic traits), which is likely true of all complex human traits, especially those related to antisocial behavior. To the extent that epistatic effects influence trait variance, the V_I parameter will have a nonzero effect on the total variance.

Traditionally, the classic twin design was primarily concerned with partitioning phenotypic variance into h^2, c^2, and e^2. Over time, advances in quantitative behavior genetics have made it possible to model V_D and even V_I on phenotypic variation. Yet, several genome-wide complex trait analyses, which do not rely on samples of kinship pairs, have found similar heritability estimates to twin studies (see Kendler, 2013). The strongest evidence against concerns over violating the assumption of no epistatic (or no dominant) effects comes from a study conducted by Hill, Goddard, and Visscher (2008) that assessed the presence of nonadditive genetic effects on more than 80 quantitative traits. The results revealed that more than half of the total genetic variation was attributable to V_A regardless of the amount of dominant or epistatic genetic influence on the individual loci (Hill, Goddard, & Visscher, 2008).

5.5.5 Generalizability

Finally, concerns over the generalizability of biometrical models has been raised by critics. Quantitative geneticists have often assumed that the information gleaned from datasets of twins will generalize to the broader non-twin (referred to as singletons) population. This assumption can be called into question by arguing that twins represent a unique group, and for that reason, their experiences may not represent the larger singleton population. Consistent with the argument, scholars have shown that twins tend to differ in systematic ways from the singleton population on a range of developmental outcomes at least throughout the first few years of life. Does this mean the findings from twin studies are not representative of the broader population for most outcomes? Not necessarily. As Barnes and Boutwell (2013) recently demonstrated, the subpopulation of twins available in the Add Health data was statistically representative of the broader and nationally representative sample of singletons on most measures of personality development, behavioral tendencies, and social outcomes. Importantly, Barnes and Boutwell analyzed antisocial behaviors and reported no consistent difference between twins and singletons in these behaviors or in the factors that predicted involvement in antisocial behavior.

Chapter 6

Biometrical Model-fitting II: Bivariate Models

6.1 Conceptual Overview

The basic univariate twin model was discussed in Chapter 5 and highlighted the mathematics to decomposing variance in a single phenotype into three components: a heritability component, a shared environmental component, and a nonshared environmental component. While the univariate model is widely used and provides important information regarding the variance of a single variable, it is limited in that it can only be used for research questions that pertain to one variable at a time. Much research, however, is concerned with the interrelationships among two or more variables, and in these research scenarios, the univariate model would be inappropriate.

Fortunately, the univariate model can be extended to a bivariate model or even a multivariate model which allows for the analysis of two or more variables in a biometric model. As the name implies, rather than estimating variance components for a single phenotype, with bivariate models, the focus shifts to estimating variance components for the covariance between two (or more) phenotypes. In doing so, bivariate models estimate the extent to which the same genetic and environmental effects that account for variance in one variable are also accounting for the variance in another variable—that is, the degree to which there is genetic/environmental overlap between two or more variables. These models are particularly germane to phenotypes that consistently co-occur or that are consistently found to correlate with each other.

Bivariate models can address research questions regarding whether the covariance between two measures is accounted for by genetic and environmental influences. Or, the two variables of interest can be the same variable measured at different developmental time periods. For instance, it is possible to use a bivariate model to examine the stability in a phenotype over

time and to determine the extent to which that stability is due to common genetic and/or common environmental influences. Seen in this way, the bivariate model is robust and can be used to address a wide range of questions that focus on the covariance among two or more variables.

Keep in mind that while the mathematics underlying the bivariate model might appear to be a bit more complex and cumbersome, the logic underlying these models is the same as the more simple univariate twin model. That is to say, these models will also rely on MZ and DZ twin correlations to estimate genetic and environmental effects.

Whereas the basic univariate twin model produces phenotypic variance estimates by focusing on MZ and DZ cross-twin (intraclass) correlations, the bivariate twin model compares cross-twin, cross-trait correlations for MZ and DZ twins. Recall that cross-twin correlations refer to the phenotypic correlation between two twins from the same twin pair (i.e., the score for one variable for one twin is correlated with the score for that same variable with their co-twin). Cross-twin, cross-trait correlations are used when examining two variables, such as the potential association between antisocial behavior and depression. In this case, the score on one variable (e.g., antisocial behavior) for one twin is correlated with the score on the other variable (e.g., depression) for their co-twin. These cross-twin, cross-trait correlations are estimated separately for MZ and DZ twins and then, as will be discussed in greater detail later, are compared to estimate the heritability, shared environmental influence, and nonshared environmental influence that accounts for the phenotypic correlation between the two variables.

In the sections that follow, we will provide a conceptual and mathematical overview of the bivariate model. After that, we will simulate a twin dataset and we will then demonstrate how to fit a basic bivariate model. We will then discuss how to extend the analysis we have demonstrated to a multivariate context (i.e., 3+ variables).

But before we move forward it is important to note that we consider this chapter a companion to Chapter 5. Thus, we will carry forward much of the language and terms that were developed in that chapter. In the interest of parsimony, whenever concepts are re-used—or, for example, if an assumption that was previously introduced also applies here—we will simply refer the reader to the appropriate section of the preceding chapter. That will allow our discussion to remain streamlined, with minimal redundancy.

6.1.1 Bivariate Models

Let us begin by imagining we have a sample of $n = 100$ monozygotic (MZ) twins and $n = 100$ dizygotic (DZ) twins. Each twin is assessed for two phenotypes, X and Y. Because we have two twins in each twin pair, we will need to adopt the subscripting logic of using "1" to indicate twin 1's scores on a specific phenotype and "2" to indicate twin 2's scores on that phenotype. Thus, we will use X_1 to reflect twin 1's score on phenotype X, X_2 will reflect twin 2's score on phenotype X, Y_1 will be twin 1's score on phenotype Y, and Y_2 will indicate

twin 2's score on phenotype Y. Recall MZ twins share 100% of their DNA, so we can say that our relatedness value, R, is 1.00 for all MZ pairs. DZ twins only share, on average, 50% of their DNA, so we can use $R = 0.50$ for DZs. Such a sample might produce a dataset like the one below:

Pair ID	Pair Type	V_A Shared (R)	X_1	X_2	Y_1	Y_2
1	MZ	1.00	5	5	2	3
2	MZ	1.00	7	8	1	7
3	MZ	1.00	3	1	4	5
⋮						
198	DZ	0.50	7	4	5	10
199	DZ	0.50	3	6	1	9
200	DZ	0.50	10	9	2	8

How might we analyze data that look like this? It is best if we approach it in three stages. First, it is important to understand the *phenotypic* association between X and Y. Let us denote that association using the covariance (e.g., $cov_{X,Y}$ or, equivalently, $cov_{Y,X}$). Because we have two twins in each twin pair, there will be two covariances we can use to assess this relationship–cov_{X_1,Y_1} and cov_{X_2,Y_2}. It is typical to assume these covariances will be equivalent, meaning there is no reason to expect the phenotypic association will be stronger for twin 1 than for twin 2 on average. This is a weak assumption—meaning it is likely to hold—when twins have been arbitrarily (i.e., randomly) assigned to appear as twin 1 or twin 2.

The second stage is to assess the cross-twin covariance *within* phenotype (cov_{X_1,X_2} and cov_{Y_1,Y_2}). This will help us to get a sense of the degree to which genetic (h^2), shared environmental (c^2), and nonshared environmental (e^2) influences affect each phenotype. As was discussed in chapter 5, one can get a rough estimate of the portion of variance explained by these various factors by observing the cross-twin covariance and performing the simple hand calculations (or by estimating the more extensive ACE model or DF model) that were presented in chapter 5.

Finally, the third stage is to assess the cross-twin, cross-trait covariance. To what degree does twin 1's score on X covary with twin 2's score on Y? In other words, what is the value for cov_{X_1,Y_2} (or, we assume equivalently, cov_{X_2,Y_1}). And to what degree does cov_{X_1,Y_2} vary according to twin type?

The logic here aligns with the logic used in chapter 5: if MZ twins are more similar to one another compared to DZ twins, then we will use that as evidence that genetic influences factor into the variance—or covariance—being analyzed. Thus, we will be looking for differential patterns of cross-twin, cross-trait covariance by twin type. This will allow us to estimate the degree to which genetic and environmental factors affect the covariance between the two traits.

We can garner information for all three stages by estimating a cross-twin, cross-trait covariance matrix. Let us first establish a convention for presenting this information. We will use a variance-covariance matrix like we did in chapter 5. But this time we will need to present information for both phenotypes. This will cause the matrix to grow in size, so to keep this exercise tractable, we will only present the information from the lower triangle of the matrix. Note, however, that we have omitted the upper triangle only for simplicity. One can still perfectly represent the upper triangle by observing the information in the lower triangle—they are mirror images of one another.

$$
\begin{array}{cc}
 & \begin{array}{cccc} \text{Twin 1:} & & \text{Twin 2:} & \end{array} \\
 & \begin{array}{cccc} X_1 & Y_1 & X_2 & Y_2 \end{array} \\
\begin{array}{c} \text{Twin 1:} \\ \\ \text{Twin 2:} \\ \end{array} &
\begin{array}{c} X_1 \\ Y_1 \\ X_2 \\ Y_2 \end{array}
\begin{bmatrix} V_{X_1} & & & \\ cov_{Y_1,X_1} & V_{Y_1} & & \\ cov_{X_2,X_1} & cov_{X_2,Y_1} & V_{X_2} & \\ cov_{Y_2,X_1} & cov_{Y_2,Y_1} & cov_{Y_2,X_2} & V_{Y_2} \end{bmatrix}
\end{array}
$$

Notice that the covariance matrix presents information for both phenotypes X and Y for twin 1 and for twin 2. On the diagonal, we see the variance for each phenotype, by twin. For example, the first entry—V_{X_1}—is the variance for phenotype X for twins labeled twin 1. As we have noted before, it is best to randomly assign twin labels. Doing so will allow us to simplify our biometrical models. In the case of the above covariance matrix, if we have randomly assigned twins to appear as twin 1 and twin 2, then we can simplify by assuming V_{X_1} is equivalent to V_{X_2}, which appears as the third entry along the diagonal. Similarly, for the phenotype Y, we can see the variance is represented by V_{Y_1}, and we assume $V_{Y_1} = V_{Y_2}$.

Looking down the first column of the matrix, we see the next entry in the covariance matrix (after V_{X_1}) is cov_{Y_1,X_1}. As was noted above, this represents the phenotypic covariance. This estimate is similar to the covariance you might estimate in a non-genetically informed study—what we have repeatedly referred to as a standard social science methodology. In other words, this provides us with an estimate of the degree to which X and Y covary. And if twins have been randomly ordered, we can assume cov_{Y_1,X_1} is equivalent to cov_{Y_2,X_2}, which appears as the last covariance entry in the bottom right of the matrix.

Staying in the first column of the covariance matrix, the third entry reflects the cross-twin covariance for phenotype X (i.e., cov_{X_2,X_1}). This information could be used to decompose the variance in phenotype X into h^2, c^2, and e^2 as we did in chapter 5. Similarly, note that the cross-twin covariance for phenotype Y appears in the bottom of the second column of the matrix (i.e., cov_{Y_2,Y_1}).

Finally, the last entry in the first column shows us the cross-twin, cross-trait covariance (i.e., cov_{Y_2,X_1}). We assume this value is equivalent to the cross-twin, cross-trait covariance that appears in the second column (i.e., cov_{X_2,Y_1}). It is this covariance that will be used to estimate the degree to which the genetic and environmental factors that impact X also impact Y. Or, put a different way, we can rely on cov_{Y_2,X_1} (or cov_{X_2,Y_1}) to: a) estimate the

genetic correlation between X and Y, r_A; to estimate the shared environmental correlation between X and Y, r_C; and the correlation between the nonshared environmental effects that impact both X and Y, r_E.

In order to gain a different perspective on this, let us represent the bivariate model in a diagram (see Figure 6.1). As with the ACE model diagram revealed in chapter 5, we will use A, C, and E to represent additive genetic influences, shared environmental influences, and nonshared environmental influences, respectively. In many respects, the bivariate ACE model is similar to the univariate version from chapter 5. But two key differences are important to note.

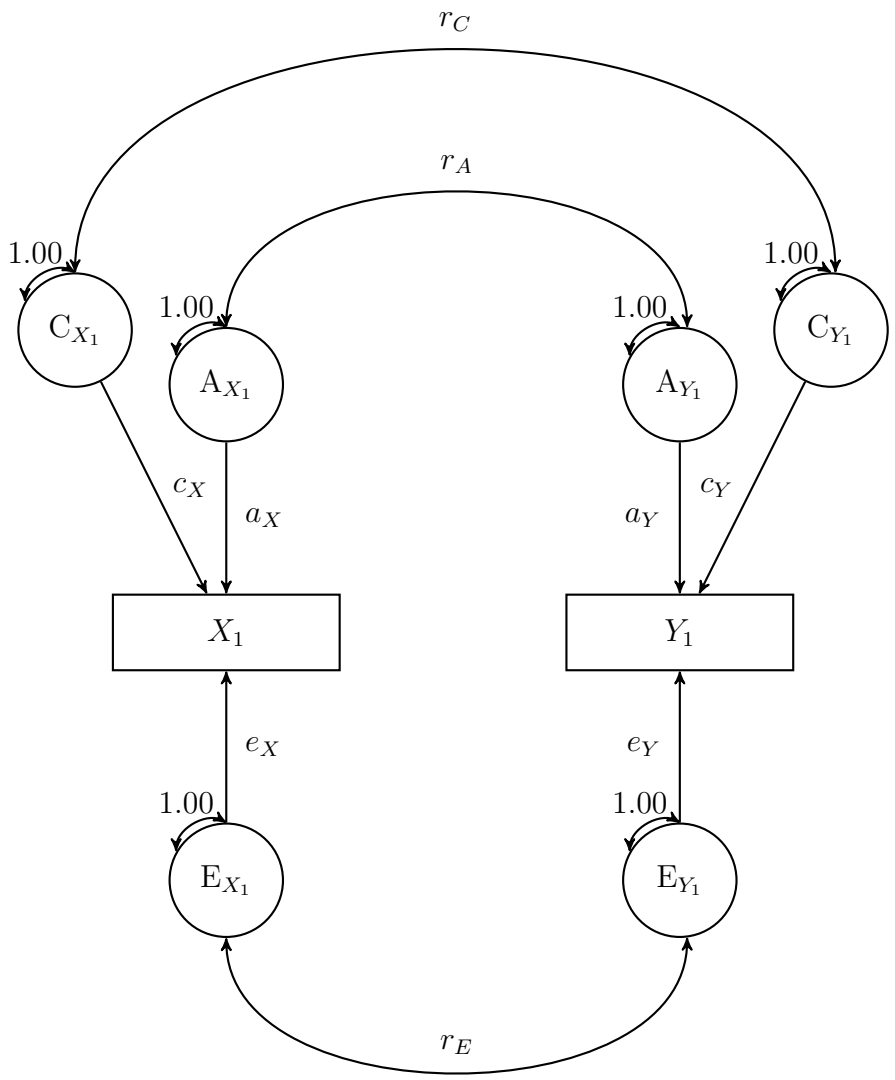
First, notice that instead of analyzing the same phenotype for twin 1 and twin 2 (as we did with the univariate ACE model in chapter 5), the bivariate version analyzes one phenotype for twin 1 and a different phenotype for twin 2. In this way, we estimate the degree to which A, C, and E impact *both* phenotypes. Which is to say the bivariate ACE model estimates the degree to which genetic influences impact X, the degree to which genetic influences impact Y, and the degree to which the genetic influences that impact X are shared with Y. Similar points extend to the shared environmental influences captured by C and the nonshared environmental influences captured by E.

The second key difference between the bivariate ACE model presented here and the univariate ACE model from chapter 5 is that the bivariate version shown in Figure 6.1 includes three new paths: r_A, r_C, and r_E. These represent the correlation between the A, C, and E factors across the two phenotypes X and Y. Think of it this way: if there are genetic influences on X, then the path from A_{X_1} to X_1 will be significant. Now, imagine a portion of the genetic influences on X also influence Y (recall our discussion of pleiotropy from chapter 5). This *correlation* between the genetic influences that impact X and those that impact Y will show up in r_A. Also note that r_A will be positive if the genetic influences operate in the same direction on both phenotypes. If, on average, the alleles that raise scores on X also raise scores on Y, then r_A will be positive. Conversely, r_A will be negative if, on average, alleles that raise scores on X work to lower scores on Y.

Similarly, the correlation between shared environmental influences (r_C) and the correlation between nonshared environmental influences (r_E) can be positive or negative depending on their substantive influences on X and Y. It is not difficult to imagine that certain environmental influences like living in poverty might positively impact (meaning it will raise scores) on some phenotypes like mental distress and simultaneously lower scores on other phenotypes like positive affect. In this way, environmental influences can drive r_C and r_E to be negative or positive.

These correlations are not just necessary for estimation of the other parts of the equation, it is important to recognize they have substantive meaning that can provide insight into the shared etiology (i.e., comorbidity) of phenotypes. Imagine, for example, trait X and trait Y tend to be comorbid in the population, such that a person who presents with X is at heightened risk of developing Y. For this to be the case, either $|r_A > 0.00|$, $|r_C > 0.00|$, or

Figure 6.1: Bivariate ACE Model for Twin 1



$|r_E > 0.00|$. In other words, if X and Y covary, then it must be the case that one of these correlations is non-zero.

We can then extend this point to make another one: if an unbiased phenotypic correlation (r_P) is non-zero, then it must be due to either a correlation at the genetic or environmental level (or a combination of both). And, as such, if we are able to estimate the phenotypic correlation and we have estimates of r_A, r_C, and r_E, then we can compute an estimate of the degree to which r_P is driven by r_A, r_C, and/or r_E by computing:

proportion of r_P due to common genetic factors:

$$\frac{\sqrt{a_X^2} \times r_A \times \sqrt{a_Y^2}}{r_P}$$

proportion of r_P due to common shared environmental factors:

$$\frac{\sqrt{c_X^2} \times r_C \times \sqrt{c_Y^2}}{r_P}$$

proportion of r_P due to common nonshared environmental factors:

$$\frac{\sqrt{e_X^2} \times r_E \times \sqrt{e_Y^2}}{r_P}$$

So how do we estimate values for all the parameters shown in Figure 6.1? Just like with the univariate ACE model from chapter 5, we will rely on a maximum likelihood algorithm to find the most optimal solution. We will also rely on the same package in `R`—the `Open Mx` package—that was used in chapter 5.

Mechanically, the bivariate ACE model operates similarly to the univariate ACE model from chapter 5. The software will create an observed variance-covariance matrix for MZ twins and another for DZ twins (and so on for all types of pairs observed in the data). It will then compute estimates for each of the paths and then generate a predicted variance-covariance matrix. The predicted variance-covariance matrices are compared to the observed matrices and the process is repeated until the divergence between the observed and predicted matrices has reached a minimum and cannot be improved with further iteration.

6.1.2 Multivariate Models

Before we turn to a demonstration of the bivariate model, let us briefly consider how one could adjust the bivariate ACE model to include more than two phenotypes. It is easy to think of reasons why one might want to extend an analysis to three or more traits: most human complex traits are correlated with many factors. Thus, if your research focus suggests

several phenotypes are correlated with one another and you are interested in whether those correlations are due to common genetic and environmental factors then you may consider estimating a multivariate ACE model. To do so, one only needs to expand the variance-covariance matrix presented earlier to include a third phenotype. Similarly, the diagram presented above is easily expanded to include a third trait.

In the interest of keeping our discussion tractable, we will only demonstrate the bivariate model here. Note, however, that the logic and the analytical approach extends easily to multivariate scenarios. One of the only things that changes when one moves from the bivariate to the multivariate study is the variance-covariance matrices becomes progressively larger as a function of the number of phenotypes under study. But this is not really a problem given the state of modern computing power. The only real challenge this poses is that statistical power can drop as more phenotypes are analyzed. And, of course, the researcher's ability to explain the pattern of findings can be strained if there are too many phenotypes to grasp conceptually.

6.2 Demonstration

NOTE: The demonstration portion of this chapter is still under construction.

As with the demonstration from chapter ??, we must begin by simulating a twin data set with two phenotypes. To keep things tractable, we will simulate a twin data file that has MZ twins and DZ twins only—no other types of pairs will be considered.

To give the demonstration some context, let us imagine we are performing a research project on the covariation between two psychological traits: conduct disorder and depressive symptomatology. Let us assume these traits are normally distributed in the population and in our sample data. The variables (i.e., our observed representations of these two concepts) will be coded so that higher values reflect higher values of each trait. Without any more context, it is easy to anticipate the direction of the phenotypic correlation between these two traits. To be direct, we expect conduct disorder scores will be positively correlated with scores on a depressive symptomatology scale (Caspi and Moffitt, 2018).

To begin, let us first simulate a dataset that has two continuous variables that are positively correlated:

```
1  setwd("/Users/jcbarnes/Box Sync/manuscripts/book_--_QuantitativeGenetics/_ch6")
2  remove(list=ls())
3
4  #install.packages(c("MASS","psych"))
5  library(MASS)
6  library(psych)
7
8  set.seed(2217)
9
10 nmz<-500
11 ndz<-500
12
```

```
13 # define covariance matrices for simulation
14 mzcov<-matrix(c(1.00, 0.29, 0.79, 0.50,
15                 0.29, 1.00, 0.49, 0.59,
16                 0.79, 0.49, 1.00, 0.29,
17                 0.50, 0.59, 0.29, 1.00),4,4)
18
19 dzcov<-matrix(c(1.00, 0.31, 0.39, 0.24,
20                 0.31, 1.00, 0.25, 0.43,
21                 0.39, 0.25, 1.00, 0.31,
22                 0.24, 0.43, 0.31, 1.00),4,4)
23
24
25
26 # simulate data using the mvrnorm command
27 # MZs first
28 mzData<-data.frame(mvrnorm(nmz,mu=c(0,0,0,0),Sigma=mzcov),rep(1.00,nmz))
29 colnames(mzData)<-c("p1_twin1","p2_twin1","p1_twin2","p2_twin2","R")
30 describe(mzData)
31 cov(mzData[,1:4])
32 cor(mzData[,1:4])
33 colMeans(mzData)
34
35 # DZs
36 dzData<-data.frame(mvrnorm(ndz,mu=c(0,0,0,0),Sigma=dzcov),rep(0.50,ndz))
37 colnames(dzData)<-c("p1_twin1","p2_twin1","p1_twin2","p2_twin2","R")
38 describe(dzData)
39 cov(dzData[,1:4])
40 cor(dzData[,1:4])
41 colMeans(dzData)
```

Although there are several approaches that can be taken to simulating data for a bivariate analysis like this, we will approach the issue by building correlation matrices for MZ twins and then for DZ twins. We begin by setting a seed (see our earlier discussion in chapter ?? if you would like more information about what exactly a "seed" represents). After that, we indicate that we would like to simulate $n = 500$ MZ twins and $n = 500$ DZ twins.

After that, we specify the variances and covariances that will be used to inform the simulation. Notice that we set the variances (the values on the diagonal in both matrices) to be 1.00. This is not strictly necessary, but rather a convenience. It allows us to interpret the off-diagonal values as correlations.

As can be seen, the correlation matrices will be 4×4, meaning they will each have four rows and four columns, one row and column for each observed variable. Even though we only have two phenotypes, we will observe both phenotypes twice: once for twin 1 and again for twin 2. And this will be true in both the MZ subsample and the DZ subsample.

Lines ?? through ?? of the simulation file reveal the structure of the MZ matrix and lines ?? through ?? reveal the structure of the DZ matrix.

The remaining lines of code in the simulation file simply as R to create the MZ data file (line ??), name the variables in that file (line ??), and provide some basic summary and bivariate statistics for the variables that were simulated (lines ??). The corresponding commands for the DZ file are shown on lines ?? through ??. We can confirm that the simulation was successfully executed when we compare the correlation matrix that is observed in the simulated data (line ??) to the "expected" correlation matrix that we specified on line

?? to create the simulation. As you will see if you replicate this example, the observed correlations are quite close to the expected ones, so we can move on to the next stage of the demonstration: estimation of Cholesky decomposition model.

6.3 Assumptions & Limitations

The assumptions and limitations inherent to bivariate (and multivariate) models overlap considerably with those discussed in the previous chapter on univariate models. This is not too surprising given that, as we discussed above, one can think of the bivariate model as two univariate models that are being estimated simultaneously. To avoid needless repetition, we invite the reader to see our discussion of the assumptions and limitations in chapter ??.

The only additional element that needs to be considered here are assumptions and limitations surrounding the the genetic and environmental influences that overlap between the two (or more) phenotypes being studied. Recall from above that the overlap between phenotype A and phenotype B is modeled as a correlation (or covariance). Correlations do not assume a direction of influence, meaning the models we have discussed in this chapter do not necessarily require the user to have worked out whether the phenotypes cause one another, whether one causes the other, or whether they are correlated due to other mechanistic pathways (e.g., they are both outcomes of a shared causal process).

In general, this is not a problem for the models we have discussed. But, it does limit the utility of the results unless the researcher is willing to make assumptions about why there is genetic/environmental overlap. Those assumptions, of course, will have consequences for how results of bivariate and multivariate models are interpreted. If, for example, the researcher assumes the correlation between the phenotypes is due to phenotype A causes phenotype B, then it would be reasonable to assume that the reason there is a genetic correlation is due to the fact that genetic factors influence the development of A, which then go on to influence the manifestation of B. A different conclusion would be reached if the causal direction were assumed to be reversed or if it was thought that the two phenotypes were byproducts of a shared causal system.

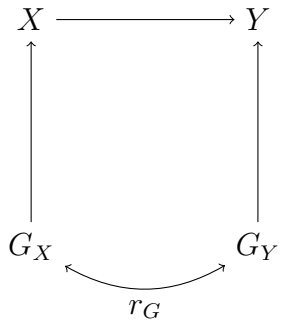
The point we intend to make here is that the results from bivariate (and multivariate) models can be informative, but enriched explanations can only be gained if the researcher has reason to believe the etiological pathway works in a specific direction. Otherwise, the results should be cautiously interpreted as evidence of a shared pathway, with the direction of influence remaining unknown.

6.4 Conclusion

This and the previous chapter provided an overview of some of the most popular methods in the behavioral geneticists' toolkit. There are, of course, many other methods that can be applied to twin, sibling, or family data to estimate the relative contributions of heritability and environmental influences on a trait or traits. Our goal here was to introduce you to some of the most commonly applied methods, so that you would have a foundation to build on if you encounter different strategies in your own work or in the published literature.

Before we close this chapter, we would like to direct your attention to an issue that we have yet to consider in any detail—the issue of genetic confounding. The term *confounding* is widely understood in scientific circles to describe a situation where one variable, X, appears to share a causal relationship with another variable, Y, when before making adjustment for a third variable C. After adjusting for C, however, the original association between X and Y disappears (or is reduced substantively). When this occurs, the original relationship between X and Y is said to have been "confounded by variation in C."

Concerns over *genetic* confounding can be tied to the present discussion if we consider that most human outcomes have a genetic component (Polderman et al., 2015). What this means is that, for any two variables X and Y, it is likely that both are to some degree influenced by genetic factors, such that:



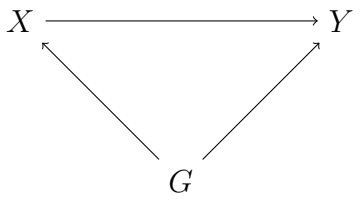
The question becomes, to what degree do the genetic factors that influence X (i.e., G_X) and those that influence Y (i.e., G_Y) correlate (i.e., what is the value for r_G)?

This is an important question, because the answer will dictate the degree to which the relationship between X and Y will be confounded by shared genetic factors. If $r_G = 0.00$, then there is no risk for genetic confounding. But, as r_G increases, genetic confounding becomes more of a concern.

Pondering this issue raises one final question: what exactly does r_G represent? In the statistical sense, r_G simply tells us the degree to which genetic influences that affect X also have an impact on Y. It does not tell us which genetic factors are shared and it does not tell us the degree to which any given genetic factor has a shared influence. In other words,

imagine there are 10 genes that influence X and 5 of those are shared with Y. It is not immediately clear what r_G would be in this scenario. To be sure, we would need more information to sort it out. For example, we would need information about how much of an influence those shared genes have on both X and Y.

Additionally, to make sense of r_G, we would need to know something about how the genetic influences actually impact X and Y. If it works like this:



then we might feel comfortable concluding that the relationship between X and Y is confounded.

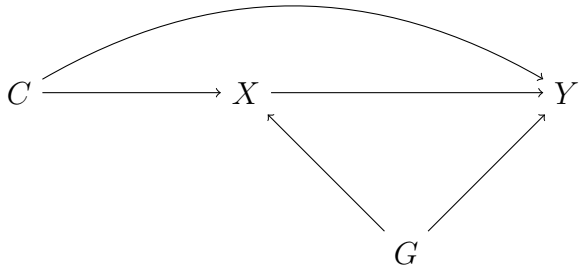
But, if it works like this:

then the role of r_G becomes more complicated. Indeed, it is entirely possible that a genetic correlation (r_G) between two variables could reflect the fact that genes impact the development of X, which goes on to affect Y. Only with a thorough understanding of the mechanisms that link genetic influences, X, and Y will scholars be able to differentiate between these possibilities in practice.

6.5 Appendix: Estimating the Degree of Genetic Confounding

The material presented in this appendix expands on our above discussion of genetic confounding to demonstrate a novel method that was developed to help scholars estimate the degree of genetic confounding that might be present between any two arbitrarily chosen variables. This material is also available in the supplemental material for Barbaro et al. (2017), which can be found at: `https://ars.els-cdn.com/content/image/1-s2.0-S1090513816303671-mmc1.pdf`.

Imagine a researcher has a dataset with information on an outcome Y, a key independent variable X, and a host of covariates C, but s/he does not have a way to control for genetic factors G, as in the diagram below:



This sort of scenario presents itself in many—if not most—behavioral science studies. When this happens, scholars are forced to either: 1) abandon viable ideas for fear of producing biased parameter estimates; 2) expend more resources to collect additional data (e.g., identify and interview MZ twins) so that a genetically sensitive design can be used; or 3) publish potentially biased parameter estimates. The purpose of this discussion is to present a novel alternative.

Specifically, the new tool developed here blends a well-established equation for estimating the degree to which a phenotypic correlation r_p is driven by genetic correlation r_g with modern statistical simulation methods. By combining these two elements, one is able to simulate the degree to which an observed correlation may be sensitive to uncontrolled genetic influences. The degree to which r_p is sensitive to genetic influences will be referred to as h^2_{cov}. In the context of the above diagram, the tool developed here will allow one to estimate the degree to which the $X \rightarrow Y$ association (i.e., r_p) is sensitive to the inclusion of G.

In order to understand the estimation routine, it is first necessary to introduce the various pieces of information that must be supplied by the user. Then, the equation that sits at the center of the estimation routine—the equation for h^2_{cov}—will be introduced.

Necessary Information: r_p, h^2_X, h^2_Y, & r_g

The estimation routine is carried out in several steps. The first step is to estimate the phenotypic correlation between X and Y, referred to as r_p. This step can be carried out using any statistical analysis package and, it is worth pointing out, partial correlations can be used when available. In other words, there is no requirement that the unconditional correlation between X and Y be preferred over a partial correlation that has already accounted for other measured covariates C.

The second step is to arrive at an estimate for the heritability of X, h^2_X. Recognizing that this value is not directly estimable—because if it were, one of several other methods would be preferable to the present approach—the researcher is encouraged to consult the available behavioral genetic literature that has bearing on the heritability of the phenotype of focus (see Polderman et al. [2015] and/or the accompanying webpage: `http://match.ctglab.nl/#/home`).

The same is true for the heritability of Y, h^2_Y. While it is not necessary that the user be an expert in behavioral genetics, the utility of this novel tool is contingent upon the user inputting heritability estimates that are both meaningful and realistic.

We now have three pieces of information necessary for estimating h^2_{cov}, but in order to garner an estimate of h^2_{cov}, we will also need an estimate of the genetic correlation r_g between X and Y. In essence, r_g provides an estimate of the degree to which the genetic factors that affect X also impact Y. Thus, r_g is simply an estimate of the correlation between the genetic factors that influence the phenotypes—X and Y—of interest.

Building a Distribution of h^2_{cov} Estimates

Researchers interested in unpacking the covariance between X and Y often rely on one of several bivariate biometrical models (Loehlin, 1996). What is unique about the bivariate biometrical model is that the covariance between X and Y can be decomposed into a heritability component that we will refer to as h^2_{cov}. This value represents the proportion of the phenotypic correlation r_p that is due to a shared genetic overlap between X and Y.

One can calculate h^2_{cov} as:

$$h^2_{cov} = \frac{\sqrt{h^2_X} * r_g * \sqrt{h^2_Y}}{r_p}$$

where: $\sqrt{h^2_X}$ is the square root of h^2_X; $\sqrt{h^2_Y}$ is the square root of h^2_Y; r_g is the genetic correlation between X and Y; and r_p is the phenotypic correlation between X and Y. Conceptually, the equation provides an estimate of the proportion of r_p that is due to shared genetic influences between X and Y. For this reason, the equation for h^2_{cov} is the centerpiece of the new estimation tool developed here.

One might be tempted to simply solve for h^2_{cov} using estimates that come to mind for h^2_X, h^2_Y, r_p, and r_g. Indeed, one can easily calculate the proportion of the phenotypic correlation that is due to genetic factors (i.e., h^2_{cov}) knowing nothing more than these four values. One key point, however, would be overlooked. Specifically, the inaccuracy of the estimates for h^2_X, h^2_Y, r_p, and r_g are ignored if one solves the equation with just one set of values. Of course, the very foundation of statistical analysis rests on the assumption of random error, meaning that any estimate we receive from this equation is likely to be too high or too low.

Fortunately, drawing on certain principles and techniques that have become commonplace in Bayesian analysis can help solve this problem. The mechanics of modern Bayesian statistical analysis is one of "brute force" sampling and simulation (Gelman et al., 2014; Gill, 2013; Jackman, 2000). Recognizing that integrating over the posterior distributions of interest—even for very simple problems—is often too complicated to calculate with closed form integral calculus, contemporary Bayesian statisticians have adopted Markov chain Monte Carlo (MCMC) routines of simulation and sampling as their primary workhorse for generating

estimates of the posterior distribution. The logic is straightforward: if you cannot directly calculate a solution to a problem, use MCMC to simulate and estimate the problem a large number of times and create a distribution of posterior estimates.

Thus, the estimation tool developed herein will allow for the uncertainty of the estimates provided by the user to be taken into account when calculating the posterior distribution of h^2_{cov} estimates. This is done by solving the above equation k times, each time including a slightly different configuration of values for the heritability estimates and for r_p. The value k will be supplied by the user—it is recommended that k be set to a large value (e.g., $k = 10,000$) in order to ensure adequate coverage of the parameter space. Rather than force one to calculate the equation for h^2_{cov} for all the possible combinations of heritability estimates and r_p estimates—a procedure that would necessarily ignore the probability distributions of the various statistics—the approach developed herein allows one to randomly sample values from a distribution of heritability estimates and from a distribution of r_p estimates. But first, the user must construct said probability distributions. The beta distribution makes this task tractable.

The Beta Distribution

The beta distribution is appropriate for building a probability distribution of prior estimates for r_p and the heritability estimates because it is bounded at 0 and 1, but can take on any real value between those two integers. The beta distribution is a well-defined univariate distribution that has a direct relationship with the normal distribution and has a probability density function of (Gelman et al., 2014: 578; Leemis, 1986: 146):

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} * x^{a-1}(1-x)^{b-1}$$

where both a and b are greater than 0 and can be thought of as shape parameters that affect the form and location of the distribution along the support region. The three values of interest—the expected value [$\mathbb{E}(x)$], the mode [$mode(x)$], and the variance [$var(x)$]—are calculated as:

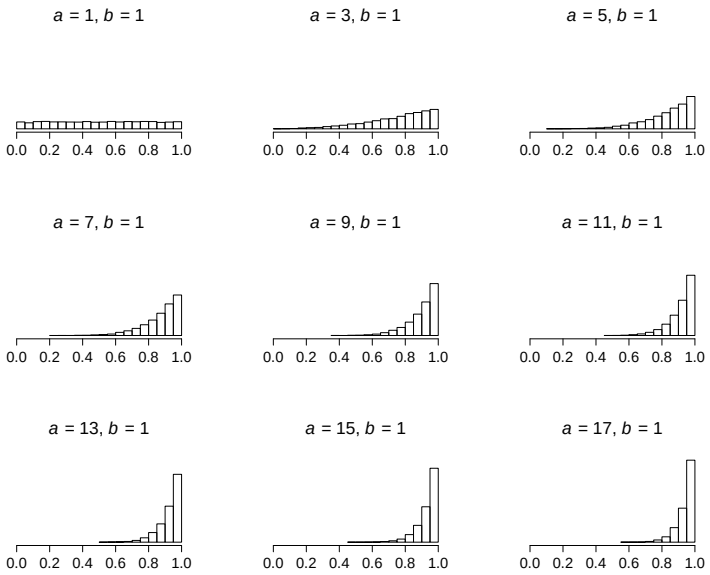
$$\begin{aligned} \mathbb{E}(x) &= \frac{a}{a+b} \\ mode(x) &= \frac{a-1}{(a-1)+(b-1)} \\ var(x) &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}$$

Thus, the shape parameters can be used to adjust the balance point (i.e., the mean or expected value) of the distribution, the modal value, and the dispersion (i.e., variance)

around the expected value. Generally, a can be thought of as the right shape parameter meaning that larger values for a, relative to b, will place more density in the right portion of the support region. The opposite is true for the shape parameter b, which is the left shape parameter. Taken together, this means that the user can adjust the beta distribution to load more density for the heritability estimate distribution and/or the r_p estimate distribution in the right side of the support region if a is increased relative to b and the opposite effect is achieved if b is increased relative to a.

These points are demonstrated graphically in Figure 6.2 and Figure 6.3. Note also that the user can set the beta distribution to reflect his/her level of confidence in the estimates by setting the shape parameters to higher or lower values. Higher values for the shape parameters will load more density in increasingly smaller regions of the distribution, meaning the variance approaches its lower limit as a and b approach ∞.

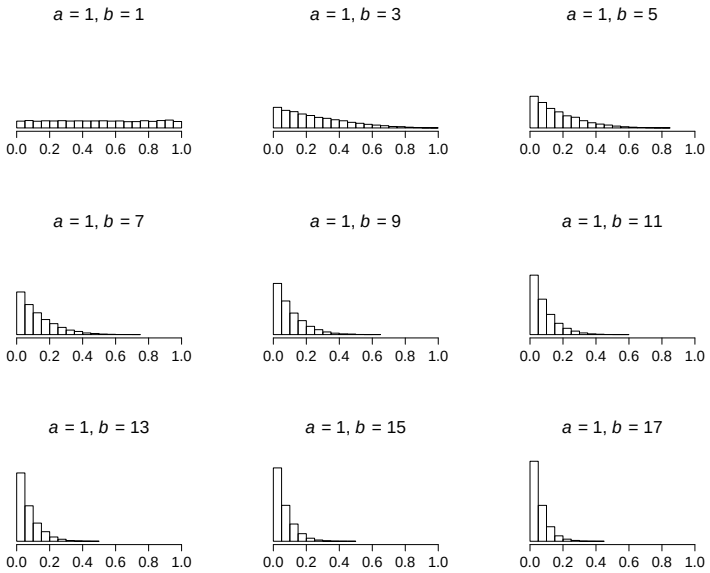
Figure 6.2: The Effect of Changing a



Several other useful features of the beta distribution are worth pointing out. Imagine a scenario where the researcher is unsure what the heritability estimate(s) and/or the r_p estimate should be. In this case, the researcher would benefit from relying on something similar to the Bayesian diffuse/uninformative prior. This can be achieved by setting both shape parameters to equal 1 (i.e., $a = 1$ and $b = 1$). The panel in the top-left of Figure 6.2 and Figure 6.3 reveals the beta distribution is uniform under this condition.

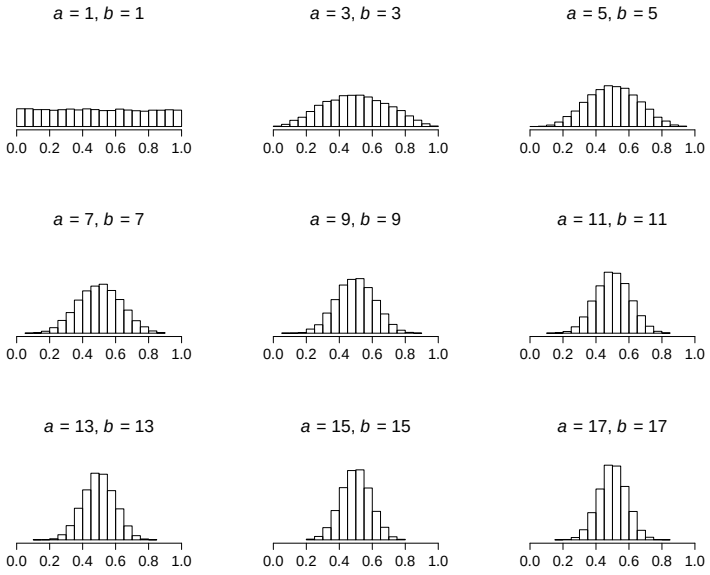
Imagine another case where the researcher believes the heritability estimate(s) and/or the r_p estimate is approximately 0.50. This is an especially important value for the former (i.e., heritability estimates) because much of the behavioral genetic literature converges on heritability estimates that are approximately 0.50 (Polderman et al., 2015). These types of estimates can be modeled with the beta distribution by simultaneously increasing both

Figure 6.3: The Effect of Changing b



shape parameters in equal magnitude (i.e., $a = b$). This relationship is revealed graphically in Figure 6.4.

Figure 6.4: $a = b$



Recognizing that the true population parameter is an unknown that is only estimated in any given study, the beta distribution will capture the uncertainty in the estimates by building a range of values that will be fed through the equation for h^2_{cov} k times. In the end, a posterior distribution of estimates for h^2_{cov}—the degree to which r_p may be biased due to

uncontrolled genetic influences—is retrieved.

Recommendations for Estimating a Distribution of h^2_{cov}

The above sections introduced a novel estimation tool that can be used by any researcher who is concerned that the relationship between two variables X and Y might be inflated due to uncontrolled genetic factors. All of the codes—in `R`—necessary to carry out the estimation routine have been posted to the following GitHub page: `https://github.com/jcbarnescrim`. Thus, access to genetically sensitive data is no longer necessary to estimate to extent to which a phenotypic correlation is sensitive to omitted genetic factors.

A brief summary of the estimation procedure is outlined here:

1. The researcher observes (whether from a novel data analysis or from the available literature) a relationship between two variables X and Y. The relationship should be measured in the form of a correlation coefficient (r_p), but note that a partial regression coefficient (i.e., an estimate that already accounts for other known confounders) can also be used as long as the value has been standardized.
    - Form a distribution of r_p values using the beta distribution. The expected value (or the mode if the distribution is skewed) of the beta distribution should be set to equal the observed correlation coefficient.
    - The shape parameters, a and b, are used to construct the desired beta distribution.
2. The researcher specifies the heritability estimate for X (h^2_X). This information should be based on the available behavioral genetic literature. Scholars are encouraged to see Polderman et al. (2015) for heritability estimates.
    - Form a distribution of h^2_X values using the beta distribution. The expected value (or the mode if the distribution is skewed) of the beta distribution should be set to equal h^2_X.
    - The shape parameters, a and b, are used to construct the desired beta distribution.
3. The researcher specifies the heritability estimate for Y (h^2_Y). This information should be based on the available behavioral genetic literature. Scholars are encouraged to see Polderman et al. (2015) for heritability estimates.
    - Form a distribution of h^2_Y values using the beta distribution. The expected value (or the mode if the distribution is skewed) of the beta distribution should be set to equal h^2_Y.
    - The shape parameters, a and b, are used to construct the desired beta distribution.

4. The researcher specifies the genetic correlation between X and Y (r_g). This information may not always be available. In cases where r_g is unknown, the researcher is encouraged to try a range of potential values.

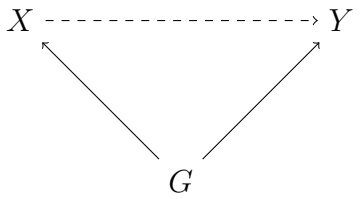
5. Enter the information from steps 1 through 4 into the program code located at (`https://github.com/jcbarnescrim`) and generate a posterior distribution of h^2_{cov} estimates. This distribution of estimates is calculated by feeding randomly drawn values from the above distributions through the equation for h^2_{cov} k times.

    - k is set by the user and should be a large value (e.g., 10,000) to ensure adequate coverage of the parameter space for the posterior distribution of h^2_{cov} estimates.

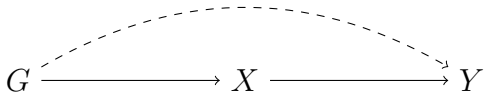
Conclusions

While there are many ways researchers could use this tool, the most obvious is to estimate the sensitivity of a parameter estimate of the association between X and Y (i.e., r_p). Rather than simply speculating about the degree to which a relationship might be confounded, a probability distribution of values can now be formed by carrying out the five simple steps outlined above.

But, it is important to caution researchers from blindly estimating a distribution of h^2_{cov} values. In fact, the distribution of h^2_{cov} values is only meaningful if genetic factors G serve as confounding influences. Confounding influences are those that are antecedent to X and Y and have a causal effect on variance in the two measures:



It is important to note, however, that confounding variables are statistically indistinguishable from mediator variables. This may complicate the interpretation of the results gleaned from the proposed tool if the true relationship is:



where X mediates the influence of G on Y. This chain of causation is quite different from that which is expected from a confounded relationship. Estimates of h^2_{cov} mean something

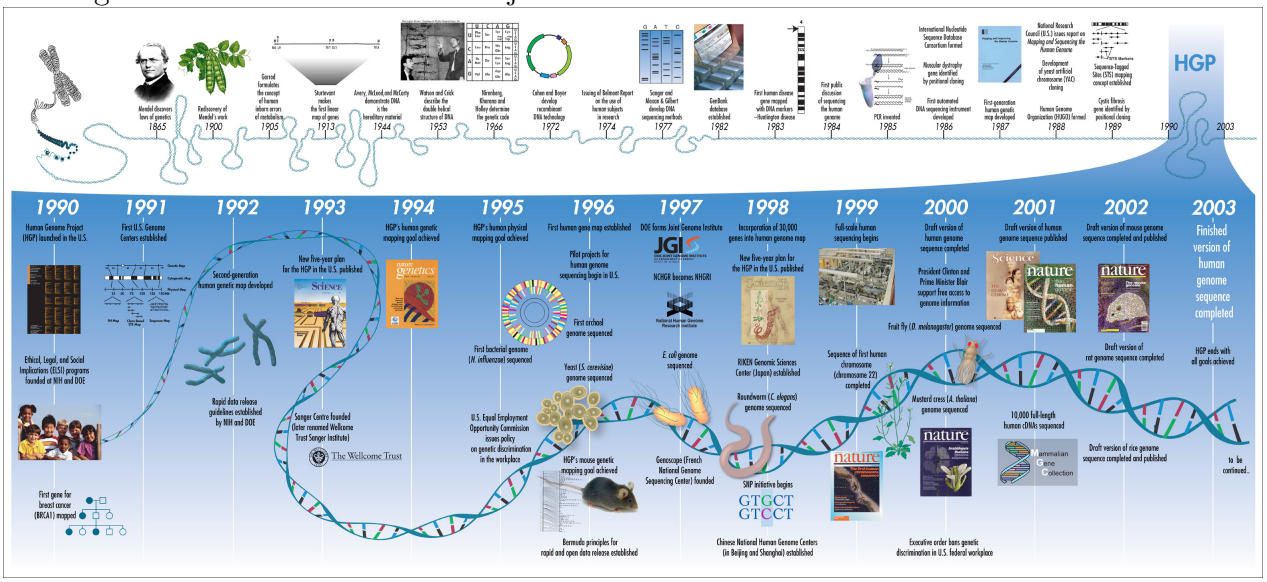
different in this case, so researchers must rely on theory, empirical evidence, and logical deduction to determine whether the genetic correlation (i.e., r_g) between X and Y is due to genetic confounding or something else.

Chapter 7

Candidate Gene Studies

From its humble beginnings in a monestary in the 1800s (Mendel, 1866), genetics research has gone from a topic of focus for farmers and those with an interest in animal husbandry to something that touches nearly every corner of the modern scientific enterprise. Indeed, one would be hard pressed to identify an area of study that has grown faster than genetics and genomics. From pea plants, to complex human traits like intelligence and schizophrenia, scientists have never known more about how all those As, Ts (or Us), Cs, and Gs, translate into phenotypic variation (see Figure 7.1, which is available here: `https://www.genome.gov/imagegallery/#nanogallery/58md7m602nv/72157672130483806/26964377742`)

Figure 7.1: Human Genome Project Timeline: From Mendel to Modern Genomics



In one sense, the latter statement is quite *un*remarkable. Of course we have never known more than we do now! Putting this point aside, though, the overarching idea is that scientists

have made important strides in understanding how the genome affects everything about what it means to be human.

The techniques discussed in the previous chapters help illuminate the progression of knowledge in behavioral genetics. In fact, one can—roughly speaking—view this text as an indicator of the temporal progression of behavioral and quantitative genetics research. As scientists unlocked the basic building blocks of DNA and genetic transmission, the latent variable decomposition models discussed in chapter 5 and chapter 6 were developed. As genetic technology progressed at an accelerating rate, scientists began to realize that it was no longer necessary to estimate genetic effects in a latent fashion (think: h^2). Rather, they began to imagine the possibilities of examining specific locations of the genome and linking those genetic sequences with outcomes of interest.

It takes no stretch of the imagination to realize that these technologies developed first and foremost on animal models (CITES). But as the technologies were refined—and ethical/safety issues were addressed—these methods became common place in human research. The present chapter is focused on the modern application of genetic research methods to human phenotypes. Our emphasis on modern applications will streamline the discussion by avoiding some of the early techniques such as linkage analysis. Readers interested in a more thorough treatment of linkage association studies—and others we omit—are encouraged to see Plomin et al. (2013).

This chapter and the next are companion chapters and, thus, are divided into two broad sections. The first section—covered in the present chapter—will address candidate gene studies. A candidate gene is a broad term applied to any study that selects one (or a few) gene(s) of interest and analyzes it to see whether it correlates with the human phenotype of interest. Candidate gene research exploded on the landscape of human behavioral science research after two highly influential papers were published in the prestigious journal, *Science* (Caspi et al., 2001; Caspi et al., 2002). It did not take long, though, for scholars to realize that candidate gene research—as a general approach—suffered from several limitations. Recognition of these issues, which will be discussed in detail below, led scientists to further refine their methods and their approach to the study of human complex traits.

Refinements were made to the methods that had become popular in the early 2000s, but it quickly became apparent that if the issues with candidate gene research were to be fully addressed, a new approach would be necessary. Luckily, around the same time, commercial technology was being developed that would afford scientists the opportunity to scan a person's *entire genome* with just a little bit of DNA being extracted from saliva, blood, or buccal cells. It is no exaggeration to say that this technology was a game changer. Indeed, research into human complex traits has not been the same since.

The next chapter—chapter 8—will cover this new technology and, more specifically, how it is being applied to the study of human complex traits. This new approach is often referred to as genome-wide association (GWA) research. We will discuss the application of GWA and the closely related technique known as genome-wide complex trait analysis (GCTA) in the

next chapter. For now, we turn our attention to candidate gene studies.

7.1 Conceptual Overview

Candidate gene studies sprang into popularity in the 1990s and 2000s after the first candidate gene for a common human disorder was discovered (Goate et al., 1991). When placed in historical context, it is not hard to see why. The turn of the century was marked by increasing demands for advanced technologies, especially in industrialized societies like the United States. The personal computer was the great liberator of knowledge. Within just a few years, the computer became something that was accessible to the majority of American households. This meant that most Americans had access to a machine that could answer any question they might have with the stroke of just a few keys. An "ask Jeeves" search could provide you with the information you requested, as long as you asked nicely (i.e., used the correct search terms)![1] Knowledge was no longer isolated to the privileged.

At the same time—and probably as a consequence of the technological revolution—the United States undertook one of the most ambitious projects in the history of science. That project had one goal: map the *human* genome. Other animal species had been mapped, of course, so the technology was in place. The costs, however, were exorbitant and the task, time consuming. Undeterred, though, two groups of scientists set out to do the unthinkable and identify all of the genes in the human body.[2]

Looking back now, it is quite funny to think about some of the early missteps and misconceptions that emerged during this time. For instance, genetic researchers had previously mapped the genome of the cat and the common mouse. These animals were known to have roughly 20,000 genes. Humans, of course, are far more complex and sophisticated than the cat and/or the mouse. So, it was reasoned, humans might have a million genes or more. That might have even been considered a conservative estimate. This number started to come down as soon as the early findings began to emerge. Indeed, early estimates suggested the number might be closer to half a million. Then it was estimated that humans would have 100,000 genes. Then 50,000. Finally, a startling reality began to set in. Humans do not really have any more genes than the common house cat! And to make matters worse, many of our genes are not even "human" genes in the sense that they are unique to our species. In other words, we share a large proportion of our genes with the cat and even the mouse (Collins, XXXX; Ridley, XXXX). We are not all that unique from other mammalian species (Wilson, XXXX).

Although these results may seem discouraging from an existential standpoint, they encouraged scientists to ponder many important questions. For example, if humans share much of

[1] Today, the "ask Jeeves" search would be a Google search.

[2] For those who are interested, the two groups were 1) the Human Genome Project, which was a (large) group of publicly funded scientists and 2) a privately funded venture led by Craig Venter. Interested readers are encouraged to see Mukherjee (2016) for more information.

their genetic material with the common house cat, then how can we be so different? There were at least two obvious—but contradictory—answers. The first answer would suggest that humans are largely "immune" to genetic influence. The logic would go something like this: humans are a product of the environment and, therefore, are under very little genetic influence.[3] Thus, it does not matter that we share much of our genetic material with other mammals; genes are simply a means to an end. The second argument—the counterargument, in a sense—would note that genetic variance is *very* important for human outcomes, but the way in which the genes are used, when/where they are expressed, and how flexible they are must vary as a result of environmental inputs.

Clearly, of course, the second argument is the closer representation of reality. One need only consult the results from the most recent behavioral genetics study to find evidence that genes influence variance in human outcomes (see, for example, the massive meta-analysis by Polderman et al., 2015). But these studies only show us that genes matter. They do little—if anything—to highlight *which* genes might matter and under which circumstances they might matter. In order to address these questions, one must utilize a type of research design that can capitalize on information about specific *genotypes* and their relation to the phenotype of interest. Put differently, if one wishes to understand which genes play a role in human variance, one will need to actually gather information from the genomes of individual humans in much the same way that other information is gathered: draw a sample from the population and estimate the relationship between the variables of interest.

Thinking through a hypothetical scenario will facilitate a better understanding. Imagine you read the latest twin study and the findings suggest that political affiliation is highly heritable, perhaps $h^2 \approx 0.80$. This gets you thinking about the individual genes that may underlie this h^2 estimate. Recall from chapter 3 (and some of the discussion in chapters 5 and 6) that a phenotypic value P for person i is expected to be caused by his/her genetic (G) and environmental (E) inputs. Thus:

$$P_i = G_i + E_i$$

Starting from this baseline equation (and assuming many factors like gene-environment interactions are either not operating or can be safely omitted for the time being), though, we know that each of the components on the right-hand side act like "global" values that sum over the collective influence of the individual genetic and environmental inputs. In other words, we might re-express the above equation using an expected value notation (i.e., $\mathbb{E}$):

$$\mathbb{E}(P_i|G = g_i, E = e_i) = \theta_0 + \theta_G(g_i) + \theta_E(e_i)$$

The above can be thought of as a general model for the expected value of phenotype P given one's known genotypic values g and environmental values e. One could use this

[3]We should point out that no one would ever make this exact argument. It is an extreme case that is being used merely as a talking point to reveal a broader perspective.

model to inform estimation of $\mathbb{E}(P)$ in a dataset using any type of modeling strategy that is appropriate for estimating the impact of G on P. This equation could also be re-expressed to make the same points, but to illuminate a few additional details:

$$\begin{aligned}\mathbb{E}(P_i) &= \theta_0 + \theta_{g_1}(g_{1i}) + \ldots + \theta_{g_k}(g_{ki}) + \theta_{e_1}(e_{1i}) + \ldots + \theta_{e_m}(e_{mi}) \\ &= \theta_0 + \sum_{k=1}^{K} \theta_{gk}(g_{ki}) + \sum_{m=1}^{M} \theta_{em}(e_{mi})\end{aligned}$$

The second equation states that we can predict individual i's phenotypic value if we know the collective influence of all the genes G that might play a role in the etiology of P. And, likewise for all the environmental influences E that underlie P's etiology. If we had this information, we could simply sum over the collective influence of the various genetic influences G (i.e., we sum across all the θ_{gk} estimates that might impact P, given the individual's specific genetic endowment g_i). This point is expressed as $\sum_{k=1}^{K} \theta_{gk}(g_{ki})$ in the equation. And similarly, $\sum_{m=1}^{M} \theta_{em}(e_{mi})$ reflects the impact of all the environmental inputs, given person i's observed environmental values.

Several obvious problems prevent scholars from ever estimating an equation like the one outlined above. First and foremost, we never have an individual's information for all the possible genetic variants that may play a role in the etiology of P. Some techniques that are moving in that direction are discussed in chapter 8. But, suffice it to say for now, it is not yet possible to fully specify the $\sum_{k=1}^{K} \theta_{gk}(g_{ki})$ component of the equation. And, at the risk of pointing out the obvious, we will never have all the information necessary to specify the environmental component of the equation. This is all to say that behavioral scientists will never be able to perfectly specify the various factors that influence the development of P. Instead, we look to specify models that offer a good representation of the underlying reality, all the while knowing that the model is incomplete and, perhaps, incorrect. A popular phrase highlights the point we are trying to make. Specifically, Box and Draper (XXXX) noted quite pointedly that, "all models are wrong, but some are useful."

With that point in mind, the question now becomes, how might scholars begin to specify *elements* of $\sum_{k=1}^{K} \theta_{gk}(g_{ki})$ and $\sum_{m=1}^{M} \theta_{em}(e_{mi})$ in an effort to "piece together" the various factors that might play a role in the etiology of P? Candidate gene studies represent one of the ways that researchers have attempted to address this question. Specifically, candidate gene studies aim to unpack the $\sum_{k=1}^{K} \theta_{gk}(g_{ki})$ portion of the equation, one gene at a time.

The candidate gene study is, therefore, relatively straightforward: the researcher hypothesizes that some known genetic variant g_k affects the phenotype of interest Y (we now switch from using P to represent the phenotype to allowing Y to represent the phenotype so that

it will conform more closely with a typical regression-based approach). From there, the researcher estimates the statistical association between g_k and Y using the typical method(s) that prevail in modern applied statistical analysis (e.g., with regression estimation). For instance, if Y is a continuous trait with an approximately normal distribution, then the association between g_k and Y can be estimated by ordinary least squares (OLS) regression:

$$\mathbb{E}(Y_i|G = g_i) = \beta_0 + \beta_1(g_{ki})$$

If Y is a categorical variable—say, a dichotomous indicator of the presence of some trait—then, the logistic regression model may be estimated (Long, 1997):

$$\log\Big[\frac{\mathbb{P}(Y_i = 1|G = g_i)}{1 - \mathbb{P}(Y_i = 1|G = g_i)}\Big] = \pi_0 + \pi_1(g_{ki})$$

Although it is not necessarily conventional, we introduce $\mathbb{P}$ as the symbol for probability in order to differentiate it from the other various uses of "P" throughout this text. $\mathbb{P}(Y_i = 1)$ can, therefore, be read as "the probability that Y_i is equal to 1." With this in place, it it easy to see that the left-hand side of the logistic regression equation represents the (natural) log of the odds of $Y_i = 1$. Those log odds are fit as a linear function of the right-hand side variables, which for this example only considers information from the genotype g_k.

But logic tells you that just about any human phenotype is a complex trait, meaning it will be (highly) unlikely to result from a single genetic loci g_k. In fact, modern genetics research reveals that only a few human outcomes are governed by a single gene. This finding is so well-established that Vink and Boomsma (2002) brought it up in their review of gene finding strategies more than 15 years ago. Specifically, Vink and Boomsma (2002: 63) tell us, "The ideal candidate gene has been shown to be functional: it influences the concentration of the (iso)form of a protein, its functionality or efficiency, or perhaps most importantly, its responsiveness to environmental factors triggering the expression of the gene. The problem with a candidate gene approach for most complex traits is the potentially huge proportion of genes, which can serve as candidates."

The complexity of human outcomes and what that necessarily means for candidate gene research recently led Christopher Chabris (2016: 304) and his colleagues to extend Eric Turkheimer's (2000) three laws of behavior genetics to account for a fourth law: "A typical human behavioral trait is associated with very many genetic variants, each of which accounts for a very small percentage of the behavioral variability."

The takeaway is obvious: the modern scientist ignores the complexity embedded within the study of how the genome affects variance in behavioral phenotypes at his/her own peril. Rare will be the human trait that is governed by a single gene in isolation. So rare in fact, that we are willing to make a prediction: you will *never* find one! This raises important questions about the utility of studying one gene at a time, as many candidate gene studies

have done in the past. We will revisit this point in the next chapter in much more detail. As we have noted throughout this chapter already, the present discussion is meant as a companion to the discussion that follows in chapter 8. Thus, we will, for now, put aside the potential limitations that arise whenever a scientist aims to unpack $\sum_{k=1}^{K} \theta_{gk}(g_{ki})$ by studying one gene (or even a handful of carefully selected genes) at a time.

Returning focus to the practice of candidate gene research, the point we have tried to make clear so far is that conducting a candidate gene study with only a single gene of focus does not require much—in terms of applied statistical tools—beyond the techniques typical in social science research. There are, however, several practical and theoretical issues that must be considered when developing a candidate gene study. Based on the "recipe" provided by Dick and colleagues (2014), we discuss two issues that should be considered when one plans to conduct a candidate gene study. Specifically, when conducing a candidate gene study, the researcher must 1) develop a hypothesis about the underlying causal pathway between the gene(s) and the phenotype(s) and 2) determine how s/he will operationalize the genetic information.

7.1.1 The Causal Pathway Between a Gene and a Phenotype

During the design phase of a candidate gene study, the researcher must consider several important issues. Chief among these considerations is the choice of the candidate gene of focus. Specifically, how does one go about identifying a candidate gene to study? As noted by Dick and colleagues (2014), it is no longer acceptable to analyze the "usual" suspects (e.g., $DRD2$, $DRD4$, $MAOA$, $5HTT$, and $COMT$) simply due to their availability in a dataset. Thus, one must take extra steps to justify the analysis of any specific candidate gene.

There are, broadly speaking, at least two approaches to justifying a candidate gene study. The first is to note that the study is exploratory and the candidate gene of focus may or may not be causally linked to the phenotype under consideration. If this route is taken, it is essential that the researcher disclose several other important pieces of information. For instance, it is necessary to report how many *other* genes were available for analysis and why this gene was chosen over the others. If the gene of focus was chosen because it was found to have a relationship with the phenotype, then it is imperative that the researcher also report how many other tests were performed. This information is critical so that the resulting p-value can be adjusted to account for multiple testing bias. We will discuss multiple testing bias in more detail in the Assumptions & Limitations portion of this chapter.

The second route one can take to justify a candidate gene study (i.e., the selection of gene A over gene B) is to produce an *a priori* hypothesis about the relationship to be analyzed. Behavioral geneticists now recommend applying a sort of "prior probability check" to candidate gene research. The prior probability is a probabilistic statement that tries to represent the researcher's (or the research literature's) best guess of the probability that the relation-

ship being tested actually exists. In other words, a prior probability is a way to quantify the level of confidence the researcher has about the relationship *prior to conducting the study.* Prior probabilities—or "priors" for short—are a staple feature of modern day Bayesian statistical analysis (Gelman et al., 2014; Gill, 2013). In essence, the prior probability can be used to adjust observed empirical estimates to the researcher's initial/prior beliefs.

The reason this is important for candidate gene research is that scholars have argued that much of the available candidate gene literature suffers from having a low prior probability (Duncan and Keller, 2011; Munafo, 2009). In practice, research that proliferates—as candidate gene research has—despite low prior probabilities is more likely to build a body of evidence that has a relatively high rate of false-positive findings. In this context, a "positive" result is one that appears to support the researcher's hypothesis. Thus, false-positive findings are those that appear to support the researcher's hypothesis, but in fact the underlying reality does *not* support the hypothesis. In other words, false-positive findings are errors and when an area of study is built on findings that have low prior probabilities, errors will abound (Ioannidis, 2005).

On the one hand, this concern reflects a central principle of the scientific enterprise. Scientists develop hypotheses based on the best available evidence. They test those hypotheses using error-prone (but the best available) tools. Using error-prone tools means that a portion of any evidence-base is likely to be incorrect. So, one might argue that low prior probabilities driving up false-positive results reflects the natural evolution of an area of scientific inquiry. Science, one might argue, is self-correcting and will eventually weed out the errors in favor of the truth.

On the other hand, the false-positive problem can be interpreted as a major threat to the advancement of research into human complex traits. Publishing an erroneous result may have little impact in, say sociology, but in the modern genetics literature the consequences can be quite drastic. One could argue that findings from single studies are sometimes used as the justification for expensive interventions, trials, or targets for drugs. If that single study turned out to be wrong, then countless resources would have been wasted, to say nothing about the potential harms that could be done.

Juxtaposing these two positions reveals the true importance of taking care when conducting a candidate gene study. While we have purposely drawn attention to extreme arguments, it is critical that scholars considering taking up a candidate gene study understand the full consequences and the challenges that lie ahead. Thus, candidate genes should not be chosen lightly. One should use the best available evidence to target specific loci that are reasonably likely to detect an association.

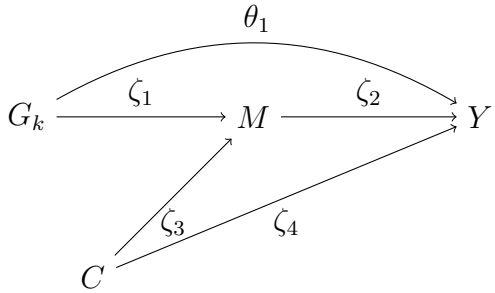
With this in mind, the obvious next question is how one should go about targeting specific loci that have an acceptable prior probability. Dick and colleagues (2014: 49) offer the following guidance, "...methods of gene selection that have a greater likelihood of producing meaningful [candidate gene-environment interaction] results include focusing on candidates suggested by well-powered GWAS or meta-analyses, or by model organisms work, ideally

with replication." In other words, given the recent explosion of GWAS research (see chapter 8), one can lean on the findings from those studies to identify potential candidate genes for study. Alternatively, one could turn attention to candidate genes that have shown promise in animal models or that have been linked to brain functioning thought to play a role in the development of the phenotype.

Once a candidate gene is identified, the onus is on the researcher to justify why the candidate gene is believed to be related to the phenotype. Of course, it is assumed that the relationship is a causal one, with the causal influence flowing from the gene to the phenotype: $G \rightarrow P$ (but see the discussion below considering whether genes are causal agents or attributes). Assuming $G \rightarrow P$, the next obvious question concerns *how* the gene is expected to causally affect the phenotype. In other words, the biological substrates that underlie the $G \rightarrow P$ relationship should be described—to the extent that it is possible—and used to help build the hypotheses for a candidate gene study.

Recall our hypothetical scenario from above, where the outcome of interest was political affiliation. Let y_i stand for the political affiliation for person i, such that we can estimate the probability that $y_i = \textit{Republican}$ (i.e., $\mathbb{P}(y_i = \textit{Republican})$) with sample data drawn from the population of focus. We might set out to estimate the conditional probability that $y_i = \textit{Republican}$ given person i's genotype at the loci of interest g: $\mathbb{P}(y_i = \textit{Republican}|g_i)$.

One would want to justify the choice of G_k as the genotype to be studied by noting that it had emerged as genome-wide significant in a recent GWAS (see chapter 8) or that research into traits thought to be linked with political affiliation given reason to suspect a relationship. Any causal pathway that is surmised is, of course, theoretical. It reflects an open empirical question that is just as vulnerable to falsifiability as any other scientific hypothesis. Moreover, it is important to realize that the causal pathway connecting any gene to a phenotype is not direct. Instead, there is will be a *large* black box of unknown intervening factors (e.g., neurological mediators). Thus, many candidate gene studies seek to estimate θ from the figure below, ignoring all of the other factors that play a role in the etiology of the phenotype (i.e., the ζs):



As one might imagine, ignoring that many relationships could prove problematic for any conclusions drawn about $\hat{\theta}_1$.

7.1.2 Operationalizing a Gene

Single Gene

Let us imagine you identified a gene to study, call it *PA*. Moreover, imagine you have reason to believe the gene of focus causally impacts political affiliation. With your hypothesis in hand, you develop a study to estimate the correlation between *PA* and political affiliation. To do so, you draw a random sample of size $n = 1,000$ from the population (let us say that the population is students at the local university). Each of the participants is asked a series of survey questions that is intended to tap a range of outcomes you think are relevant to political affiliation (i.e., potential confounders). One question asks the participants to indicate the political party with which they most identify. For simplicity, let us assume you code this item as a nominal variable where Democrat and Republican are the only options.

Finally, you ask participants to submit a saliva sample so that you may genotype them for *PA*. You collect the saliva samples and send them off to the lab for genotyping. You receive back information about each participants genotype at the *PA* loci. Recall that humans are diploid organisms, so each participant will carry two copies of *PA*. You can think of this as a data file where each participant (the rows) has two variables: *PA1* revealing their allele type on the first instance of *PA* (perhaps it is the maternally inherited *PA*) and *PA2* revealing their allele type on the second instance of *PA*. For simplicity and the sake of this demonstration, let us assume that *PA* comes in one of two forms (i.e., it is bi-allelic in the population): the minor allele, which you will code as 1 because your hypothesis suggests it is the "risk" allele, and the reference allele (the term *reference* allele is often used to refer to the allele that occurs more frequently in the population), which you code as 0. Thus, your genotype data file would have two columns of information, all coded as 0s and 1s. If the minor allele frequency (MAF) in the population was MAF=0.35, then you would expect to see roughly 35 1s for every 100 cases in your data file.

At this point, you have a coding decision to make. In essence, you will not want to leave the genetic information separated into two separate variables. Rather, you will want to combine that information in some meaningful way. There are at least four options from which you can choose. The first is to code the gene co-dominantly by summing the values observed in *PA1* and *PA2*. This would result in a single new variable that could take on three discrete values: 0, 1, and 2. This option makes the assumption that the gene does *not* deviate from additivity in the traditional quantitative genetics sense. And by drawing inference from Hardy-Weinberg equilibrium (see page X), you know what sort of group proportions to expect. Specifically, you can predict *a priori* that q^2 will be coded 0 as homozygotes for the reference allele, that $2pq$ will be heterozygotes, and p^2 will be homozygote for the "risk" allele (the minor allele). In this hypothetical scenario, we should expect the following proportions: $0.65^2 = 0.4225$ will be coded 0, $2(0.35 * 0.65) = 0.455$ will be coded 1, and $0.35^2 = 0.1225$ will be coded 2.

A second alternative would be to assume the gene operates recessively and, therefore,

your primary interest is whether participants inherit two copies of the gene. Only those cases would be expected to present with the phenotype, which in this case would mean only those cases would be expected to deviate from the expected value of political affiliation gleaned from the population. Again, assuming MAF=0.35 would suggest that you would observe $0.35^2 = 0.1225$ (recall from Hardy-Weinberg equilibrium) cases that inherit two copies of the "risk" allele. All others ($1 - q^2 = 1 - 0.1225 = 0.8775$) will be coded 0.

The third option is to treat the gene as if there is dominance deviation in the relationship between genotypic and phenotypic score. Recall from chapter 3 that dominance deviation occurs when heterozygotes deviate from the midline (see page X). If the heterozygotes resemble the homozygotes who inherited two copies of the "risk" allele, then we might choose to code *PA1* and *PA2* such that any participant who inherits *at least one copy of the "risk" allele'* is coded 1 and homozygotes who inherit no "risk" allele are coded 0. If MAF=0.35, then you would expect roughly $2(pq) + q^2 = 2(0.35 * 0.65) + 0.35^2 = 0.5775$ of the sample to receive a 1 when *PA* is coded as if the underlying genetic pathway shows dominance.

A fourth coding option is to remain agnostic about the way in which *PA* exerts its influence (assuming it has any influence at all) on the phenotype. This can be done if one treats each of the three possible combinations of *PA1* and *PA2* as distinct groups (i.e., 0 "risk" alleles, 1 "risk" alleles, 2 "risk" alleles). This is easily implemented in a regression model by treating the groups as dummy variables, where one of the three groups acts as a comparison for the other two. In this way, the researcher can identify empirically the functional relationship between *PA* and the phenotype of focus.

With these different coding strategies in mind, let us imagine the first 10 cases from your dataset look like this[4]:

caseID	polAff	c	PA1	PA2	PA
1	1	0.82	0	0	0
2	1	0.45	1	0	1
3	0	0.68	0	0	0
4	0	-0.12	0	1	1
5	0	-1.49	0	1	1
6	1	-1.23	1	0	1
7	0	0.31	0	0	0
8	1	-1.20	0	1	1
9	1	1.09	0	0	0
10	1	2.42	1	1	2

where caseID is the identifier variable and is not substantively important; polAff is the political affiliation variable, where 0 = *Democrat* and 1 = *Republican*; c is a covariate that we would want to include as a control variable; PA1 reveals the maternally inherited allele information; PA2 is the paternally inherited allele information; PA is the variable revealing

[4] `R` code for generating a similar dataset is provided in the Demonstration portion of this chapter

the number of "risk" alleles carried by each participant (the co-dominant coding mentioned above); PArec is the recessive coding scheme discussed above (**STILL NEED**); and PAdom is the dominance coding discussed above (**STILL NEED**). If we want to treat *PA* as dummy variables (the fourth option discussed above), then we can have `R` take care of it when we estimate our regression model by relying on the `factor` command (more on this later).

Multiple Genes

It is common for a candidate gene study to include more than one candidate gene. It is easy to understand why: the fourth law of behavior genetics tells us that any single gene will explain only a small portion of the variance in a phenotype. Thus, one can improve predictive power—and, then, perhaps publication chances—by including more than one gene. The benefits of this strategy are somewhat obvious: increased explanatory power is (almost) always a good thing. Yet, there are several complicating factors that must be kept in mind when one decides to expand beyond a single gene model. One of those complicating factors concerns operationalization. As with the discussion immediately above, the researcher must determine how *all* of the genes will be coded, a decision that should be based on the known biological pathway the gene takes to affect the phenotype. This task becomes more complicated, in an exponential fashion, when the study includes more than one gene.

The reason the task of operationalization is more complicated when one adds more genes is simple: the genes may have non-linear effects and they may interact in their effect on the phenotype. Thus, one is not justified in choosing to code all variables in one way just for simplicity/readability. Moreover, it may be misleading (at best) or lead to biased parameter estimates (at worst) if one ignores the added complexity of the relationship between two or more genes. Given these concerns, we are reluctant to offer specific recommendations on how to handle a research project that includes more than one gene. Instead, best-practices would dictate that all decisions be made transparent and that they be guided by the most up-to-date evidence. Overall, a general "best practices" approach is probably one that follows the above noted guidelines for operationalizing a gene, but also considers the added complexity mentioned here. Theory and available evidence should be used to guide decision making when determining whether to combine the genes into a polygenic risk score or to allow the genes to have their own independent (and/or perhaps interactive) effects (see, generally, Keller [2014] for model-fitting issues).

7.2 Demonstration

As with the preceding chapters of this text, we offer a brief demonstration of a candidate gene study here using simulated data. Our goal is to give you an inside look at some of the decision-making points that arise when conducting a candidate gene study. We will make several assumptions and restrictions that will serve to simplify the task of demonstrating a

candidate gene study.

First, we will restrict our simulated study to a single gene. We will consider polygenic models in chapter 8 when we discuss GWAS and GCTA, so it will streamline the present discussion if we overlook this complicating factor for now. Second, we will assume that the gene has a co-dominant effect on the phenotype. We will demonstrate how this assumption can be relaxed so that it is not strictly necessary to assume such effects. A third assumption that will be necessary to specify our estimation equation is that there is no gene-environment interplay between the focal gene and any omitted environmental factors. We will cover gene-environment interplay in a later chapter (chapter 10), so we leave that discussion for that chapter. Relatedly, a fourth assumption we will make is that the focal gene does not interact with any other genes, that there are no epigenetic factors that would systematically influence the biological pathway under investigation, and that the genotype is not confounded due to population stratification. We will consider the latter point—the role of population stratification—in the Assumptions & Limitations portion of this chapter. Finally, let us assume that the gene of focus is autosomal and bi-allelic, meaning every case will have one of two alleles at each loci, meaning there are three possible categories that will be observed.

With these assumptions in hand, let us imagine that you have access to the following dataset :

caseID	polAff	PA1	PA2	PA
1	1	1	0	1
2	0	0	0	0
3	1	1	0	1
4	1	1	0	1
5	1	1	0	1
6	0	1	1	2
7	1	0	1	1
8	1	1	0	1
9	1	0	0	0
10	0	0	0	0

These data were generated with the following code in R:

```
# load required packages
library(xtable)

# clear the workspace
remove(list=ls())

# set the seed for reproducibility
set.seed(2217)

# set sample size
n<-1000

# a dummy variable, which will act as PA1
PA1<-rbinom(n,1,.35)

```

```
16 # a dummy variable, which will act as PA2
17 PA2<-rbinom(n,1,.35)
18
19 # combine PA1 and PA2 into PA; assume additivity
20 PA<-PA1+PA2
21
22 # linear combination of the right-hand side variables; log odds
23 lo<-0.25+0.15*PA
24
25 # calculate probabilities by transforming lo into p
26 pr<-exp(lo)/(1+exp(lo))
27
28 # generate 0,1 outcomes based on the probabilities generated above
29 polAff<-rbinom(n,1,pr)
30
31 # combine everything into a dataframe
32 df<-data.frame(polAff=polAff,PA1=PA1,PA2=PA2,PA=PA)
33
34 # view the first 10 cases
35 xtable(head(df,10))
```

As before, the first few lines of code simply set up the `R` environment for the analysis we will perform. Starting on line 11, we begin to define the parameters of our simulation. Line 11 provides the sample size that will be used for our simulated study. Lines 14 and 17 simulate the two alleles for the *PA* gene, *PA1* and *PA2*. These lines of code utilize `R`s built-in random data generator for the binomial family of distributions. We specify that a random binomial distribution be referenced for each of the `n` cases (thus, far, we have specified $n = 1,000$), where the number of trials is set to 1 (that's the second part of the `rbinom` command) and probability of a "success"—which will be coded as a 1—is 0.35. The 0.35 can be thought of as the MAF.

Line 20 combines the two alleles into a single observed variable, which now corresponds to the number of minor alleles (we can think of these as the "risk alleles") carried by each participant. The effect of this line of code can be seen in the 10-case preview of the data that was displayed earlier.

Starting on line 23, we begin to build up the information that will be necessary to simulate the phenotype, *polAff*. Because we will estimate the impact of the *PA* gene *on polAff*, our estimation model takes the following form:

$$\text{polAff}_i = \theta_0 + \theta_1(PA_i)$$

Thus, in order to simulate data for *polAff*, we must "work backwards" by simulating the right-hand side variables first. Then, we can specify a model to generate observed scores on *polAff*. To begin, we will specify *polAff* as a categorical outcome with two groups: *Democrat* $= 0$ and *Republican* $= 1$. This will require us to estimate the parameters from the model from above using a logistic regression equation:

$$\log\Big[\frac{\mathbb{P}(\text{polAff}_i = 1|PA = pa_i)}{1 - \mathbb{P}(\text{polAff}_i = 1|PA = pa_i)}\Big] = \pi_0 + \pi_1(PA_i)$$

Notice that the logistic regression model fits the right-hand side variables to the log-odds of the probability that *polAff*= 1. Thus, Line 23 specifies the log odds (we name the object "lo") to be a linear combination of a constant term (randomly drawn from a normal distribution with mean=0 and standard deviation=1) and a weighted value of the observed *PA* gene. Using these logged odds, the probability that *polAff*= 1 is calculated in Line 26. Finally, these probabilities are used to generate the *polAff* variable by using them as the probability value in the `rbinom` statement in line 29. All of the variables we have simulated are combined into a data frame in line 32 and we view the first 10 cases of the data frame (in LaTeX format) in line 35. This final line of code was used to generate the data preview that was displayed above.

Now let us turn to the results of the logistic regression model. To run a logistic regression on these data we feed `R` the following code: `logit<-glm(polAff PA,data=df,family="binomial")`. Doing so produces the following results:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.1401	0.0911	1.54	0.1240
PA	0.2995	0.0967	3.10	0.0020

The first line of the table shows the coefficient estimate for the intercept, which here represents the average log odds of a case having a 1 on the outcome (*polAff*) if *PA* is equal to 0. We can think of the intercept as providing the baseline level of "risk" of observing a 1 on the outcome when the right-hand side variables are set to 0. We can even use the intercept value to compute the probability of having a 1 on the outcome for cases that have a 0 on the *PA* genotype by carrying out the following:

$$\mathbb{P}(\text{polAff}_i = 1 | PA = 0) = \frac{e^{\pi_0}}{1 + e^{\pi_0}} = \frac{e^{0.1401}}{1 + e^{0.1401}} = 0.5349$$

which indicates that about 53 percent of those with a 0 on the *PA* gene had a 1 on the *polAff* variable.

Turning now to the coefficient estimate for the *PA* gene, we see that the log odds are expected to increase by 0.2995 for every one unit increase in the *PA* gene. This indicates that the gene had a *positive* effect on the *polAff* outcome, meaning that individuals who inherited one of the "risk" alleles on the *PA* gene are expected to have a higher probability of having a 1 on the *polAff* variable compared to those did not inherit any "risk" alleles. But the log odds are somewhat difficult to interpret—it is rare that someone would think about risks and likelihoods in log odds units. Thus, it is common place to transform log odds values into some that is easier to interpret. It turns out that a relatively straightforward transformation will do the trick: exponentiating the log odds produces an odds ratio:

$$e^{\pi} = \text{odds ratio}$$

Carrying out this transformation on the coefficient estimate for the *PA* gene yields $e^{0.2995} = 1.3492$, which indicates that a one unit increase in the *PA* gene increases the odds of having

a 1 on the outcome (*polAff*) by approximately 35%. Another way of interpreting the odds ratio is to note that the odds of having a 1 on the outcome for those who have a 1 on the *PA* gene are 1.35 times the odds for those who have a 0 on the *PA* gene. This is all to say that the *PA* gene is estimated to have an impact on the *polAff* variable in these sample data.

We could even use these data to compute the probability of having a 1 on the *polAff* variable as a function of *PA* genotype. Doing so simply requires us to compute the probabilities for all three groups of cases (i.e., those with *PA*=0, *PA*=1, and *PA*=2):

$$\mathbb{P}(\text{polAff}_i = 1|PA = 0) = \frac{e^{\pi_0}}{1 + e^{\pi_0}} = \frac{e^{0.1401}}{1 + e^{0.1401}} = 0.5349$$
$$\mathbb{P}(\text{polAff}_i = 1|PA = 1) = \frac{e^{\pi_0+\pi_1}}{1 + e^{\pi_0+\pi_1}} = \frac{e^{0.1401+0.2995}}{1 + e^{0.1401+0.2995}} = 0.6082$$
$$\mathbb{P}(\text{polAff}_i = 1|PA = 2) = \frac{e^{\pi_0+2\times\pi_1}}{1 + e^{\pi_0+2\times\pi_1}} = \frac{e^{0.1401+2\times0.2995}}{1 + e^{0.1401+0.2\times2995}} = 0.6768$$

The first value is exactly equivalent to the one computed earlier when we discussed the substantive interpretation of the intercept. The other two show that the *PA* genotype was estimated to increase the probability of having a 1 on *polAff*.

7.3 Assumptions & Limitations

The demonstrations provided above required that we model the effect of the *PA* gene on political affiliation with a logistic regression model. This was not intended to give the impression that *every* candidate gene study will utilize the logistic regression model. On the contrary, the choice of modeling strategy will be unique to each individual study. In fact, it is not even strictly necessary that one use a regression-based approach. It is possible to use a genetic variable much like any other "typical" social science variable. It could, for instance, provide information used to construct a propensity score for a propensity score matching analysis. Or perhaps candidate gene data could be used to predict the intercept and/or slope of a developmental trajectory. The point is this: the number of unique ways one can capitalize on candidate gene data is practically limitless.

There are, however, several important concerns that naturally arise when one considers the limitless plane of possibilities that lies ahead. Readers are encouraged to give careful consideration to the assumptions and limitations outlined below. Each of the issues outlined here will serve to temper some of the enthusiasm that surrounds candidate gene research in the social sciences. At the very least, we hope that readers will consider the full range of concerns that currently surround candidate gene research because failure to do so could result in a higher-than-expected rate of false-positive findings, meaning you are more likely to reach an incorrect conclusion if you do not follow the current best-practices. With these points in mind, we must alert you to the fact that the issues outlined below should not be interpreted as an exhaustive list of possible problems that might arise. Given that the

possibilities for candidate gene research are seemingly endless, it would naive for us to give the impression that we have (or even *could*) cover all the potential concerns that might arise. We have attempted to outline the issues that have stood most prominently in the literature and/or have arisen in our own experience. For this reason, we feel the discussion outlined below offers a thorough—if incomplete—review of some of the more important topics facing modern scholars.

7.3.1 Sources of Bias

One of the over-arching goals of this chapter—and indeed the entire book—is to illuminate the fact that quantitative genetics research is similar in many ways to the standard social science approach. Key differences, of course, arise and are important to be aware of. Yet, at the end of the day, the quantitative geneticist has to worry about many of the same issues that concern every social scientist. Chief among these concerns is the issue of bias in the estimation of the impact of a gene on a phenotype.

Applied statistical analysis of any form[5] requires the researcher to produce two values of interest: 1) a parameter estimate (let β stand for a general parameter estimate for this discussion) and 2) a standard error (SE).

Any researcher who has carried out a statistical analysis knows the anxiety that arises from the thought of having estimated a biased β or a biased SE. The stressful reality is that there are *many* ways in which one could arrive at a biased β or SE. In order to understand how this could occur, it is first important that we briefly introduce one of the most important principles of the applied statistics enterprise: the law of large numbers (LLN).

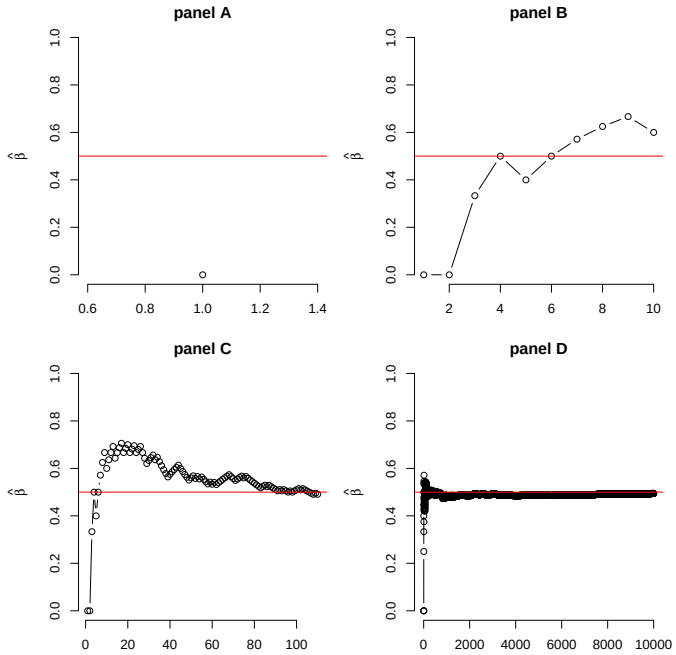
The LLN is one of the foundational assumptions upon which *all* applied statistical analyses rests. Luckily, the LLN can be understood intuitively without the need for a formal mathematical treatment. Essentially, the LLN reveals that over time, with repeated sampling from a distribution, our estimates of the value of interest will converge to the population parameter. Although any single estimate may (will!) be off by a little this way or a little that way, in the aggregate, our estimates will converge to the population parameter of interest.

This point is displayed graphically in Figure 7.2. The figure contains four panels, each of which represents the results of a trial (as in panel A) or several trials (as in Panels B, C, and D). Imagine that each trial attempts to estimate the population parameter for β, which is represented in each figure as the red horizontal line. No single trial provides an estimate that could be considered "accurate". But as we collect more and more information from the trials we see that the distribution of estimates (i.e., the $\hat{\beta}$s) begins to converge to the population parameter β.

This reveals several interesting and useful properties of the LLN. In essence, the LLN

[5]With the exception, perhaps, of certain Bayesian techniques.

Figure 7.2: Law of Large Numbers (LLN)



allows for error in our predictions! The only catch is that those errors must be *random* errors. As long as our errors are random, the LLN will take over as we add more information to the overall distribution of estimates. Translating this to candidate gene research, the LLN reveals that any *one* candidate gene study is unlikely to provide an exact estimate of the underlying parameter of interest. As more and more evidence accumulates, though, the aggregate literature *can* provide a robust/stable estimate. This helps to illuminate the value of the meta-analysis, a point to which we will return in the next chapter on GWAS.

A second property that emerges from the LLN—one that is not very obvious from our demonstration in Figure 7.2 (due primarily to the way in which we plotted the information)—is that researchers can estimate the degree to which their estimates *might* fluctuate due to the random errors inherent in their designs. The estimate of the degree to which estimates of β might fluctuate is referred to as the standard error (SE) and it is gleaned from a property of the LLN, which is known as the central limit theorem. Briefly, the SE is a researchers best estimate of the standard deviation of the sampling distribution of estimates that would be observed under random error. The SE can, in turn, be used to test hypotheses about the probability one would observe an estimate of a certain magnitude if the true value of the parameter (i.e., β) were equal to some value provided by the user. We will return to this point momentarily.

With this discussion in mind, it is important that we consider several sources of bias. Each of the sources of bias that are discussed below will impact our estimate of the β and/or the *SE* gleaned from a candidate gene study. Thus, the below is intended to serve as a general discussion of the sources of bias that may impact candidate gene studies. Any one study

will suffer from these sources of bias to varying degrees; some will be robust to all or most of these biases while others will be plagued by several or all of them simultaneously.

7.3.2 Biased Hypothesis Tests: Multiple Testing Bias

One of the most prevalent sources of bias in the candidate gene literature stems from something called multiple testing bias (MTB). To be direct: MTB results when a researcher estimates more than one association between a gene and a phenotype(s) but s/he fails to correct the p-value calculated for any one hypothesis test.

The hypothesis test is typically conducted by carrying out the following calculation for the *test statistic*:

$$\text{test statistic} = \frac{\hat{\beta} - H_0}{SE}$$

where $\hat{\beta}$ is the estimate gleaned from a statistical model; SE is the corresponding standard error; and H_0 represents the null hypothesis value, which is usually set to 0.00, so the equation often simplifies to:

$$\text{test statistic} = \frac{\hat{\beta}}{SE}$$

Once the researcher calculates a test statistic, that value is then compared against a table of critical values (in practice, a computer would carry this out automatically, but it is instructive to think about what the computer is actually doing). If the test statistic is larger than the critical value representing the 5^{th} percentile of the range of estimates expected under the null (i.e., H_0), then the observed estimate (i.e., $\hat{\beta}$) is said to be *statistically significant*. Here, statistical significance indicates that the probability of observing an estimate as large or larger than $\hat{\beta}$ is less than 0.05. This value—the probability of observing $\hat{\beta}$ as large or larger than the one that was observed—is known as the p-value. All else being equal, researchers want to see p-value ≤ 0.05. When this occurs, the chances of publication increase (see the discussion of publication bias below).

There is one sticking point, though, that often occurs with candidate gene research. The p-value that one calculates is a valid estimate of the probability of interest for any *one* test. When one performs more than one test, the p-value must be adjusted to account for the number of tests that were calculated.

Think of it like this. Imagine a carnival game where you are given the chance to draw a marble from an urn. Before you draw, the carnival worker tells you the urn is filled with with 100 marbles and 5 of them are red. The remaining 95 are black. If you draw a red

marble, you win a prize (probably an over-priced and equally over-stuffed animal). You pay \$10 for a single draw from the urn. On that occasion, your probability of drawing a red marble matches the proportion of red marbles in the urn: $\mathbb{P}(\text{red} = 0.05)$.

But imagine you find a loophole in the game's instructions. You notice that there is no explicit statement about the *number* of times you can draw from the urn. Having pointed this out to the carnival attendant, you proceed to draw 10 marbles in succession. Let us assume, though, that you are a good citizen and you draw the marbels *with* replacement, meaning each time you draw you return that marble to the urn before drawing again. Under this scenario, what is your probability of drawing *at least one* red marble? It is easy to see that you have substantially improved your chances of winning by drawing $k = 10$ marbles rather than $k = 1$. More specifically, your probability of drawing at least one red marble is now calculated as a function of the probability that you drew *all* black marbles in k attempts:

$$\begin{aligned}
\mathbb{P}(\text{any red}) &= 1 - (1 - \mathbb{P}(\text{red}))^k \\
&= 1 - 0.95^{10} \\
&= 1 - 0.599 \\
&= 0.401
\end{aligned}$$

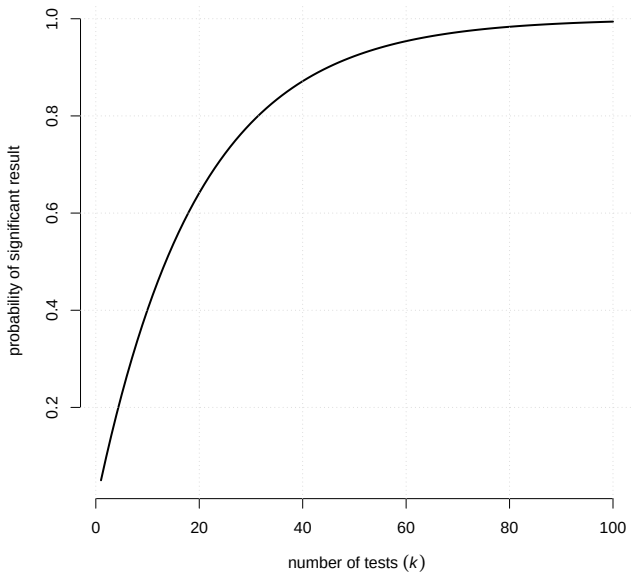
As the proof shows, the probability of drawing *at least one* red marble in $k = 10$ tries is nearly 0.401; a *substantial* difference from the original value of 0.05!

Figure 7.3 reveals the full extent of MTB as k increases from 1 to 100. It is unnecessary to allow k to increase beyond 100 because, as is clearly displayed in the figure, the function begins to approach its limit of 1.00 around $k = 100$.

As was probably obvious from the beginning of this hypothetical carnival game scenario, $\mathbb{P}(\text{red})$ represents the p-value for any given candidate gene study, k represents the number of analyses that a researcher conducts (e.g., the number of regression models that were run) and the corrected values (i.e., the values represented by the function in Figure 7.3) are the *actual* probabilities of observing at least one statistically significant result simply by chance. It is instructive to recall that a p-value of 0.05 means that 1 out of 20 tests will result in a statistically significant finding *even if none of the estimates are substantively different from the null value*. The takeaway is clear: the more tests you run, the more likely you are to find a statistically significant result by chance alone. Moreover, as the number of tests approaches 100, the probability of observing *at least one* statistically significant result by chance approaches 1.00 ($\lim_{k \to 100} \mathbb{P}(\text{at least one false-positive}) = 1.00$).

This represents a very significant problem because it reveals how easy it is to find a statistically significant result that is $p \leq 0.05$ even if the null hypothesis should not be rejected. In other words, MTB increases one's *overall* chances of stumbling on a false-positive. Left uncorrected, this problem could lead to a biased literature base that is riddled with false-positive results.

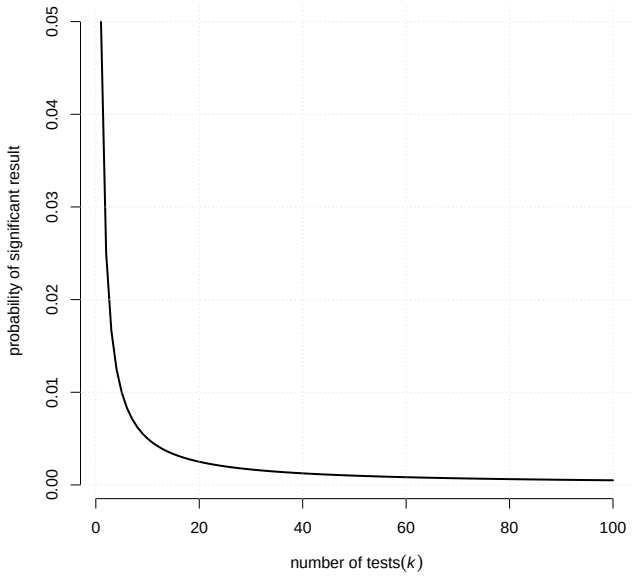
Figure 7.3: Multiple Testing Bias



Luckily, though, there is a straightforward way to correct the p-value for MTB. The correction is known as the Bonferroni method and it is carried out by dividing the p-value by k, such that the corrected corrected p-value $= \frac{p\text{-value}}{k}$. As k increases, the Bonferroni method forces one to search for smaller and smaller p-values. Those Bonferroni corrected p-values are plotted in Figure 7.4).

This method is sometimes criticized, though, because it is a conservative correction. True, it does help alleviate concerns over Type I error. Note, however, that even without the Bonferroni correction, the Type I error rate remains constant at the α level that was specified by the user (typically 0.05, which is the value we have assumed for this discussion). To reveal this point, we simulated a dataset with $k = 1,000$. The first 800 data points $(k_1, k_2, \ldots, k_{800})$ were drawn from a random normal distribution with mean $= 0$ and standard deviation $= 1$. The remaining 200 data points $(k_{801}, k_{802}, \ldots, k_{1,000})$ were simulated from a random distribution with mean $= 2$ and standard deviation $= 1$. Imagine these values represent test statistics from individual t-tests. Thus, any value larger/smaller than $\pm$ 1.96 will be statistically significant at the $\alpha = 0.05$ level. We should expect to see a Type I error rate of ≈ 0.05 prior to a Bonferroni correction. The Type I error rate will drop *after* we apply the Bonferroni correction of $\frac{\alpha}{k} = \frac{0.05}{1000} = 0.00005$. This is, in fact, exactly what we see from the table of results listed below.

Figure 7.4: Bonferroni Corrected p-values



	Type I Error Rate	Type II Error Rate
First 800 Cases:		
No Bonferroni Correction	$\frac{43}{800} = 0.054$	-
Bonferroni Correction	$\frac{0}{800} = 0$	-
Last 200 Cases:		
No Bonferroni Correction	-	$\frac{75}{200} = 0.375$
Bonferroni Correction	-	$\frac{196}{200} = 0.98$

Prior to any correction, the first 800 cases had a Type I error rate of roughly 0.05 (i.e., the α level we specified). After a Bonferroni correction, the Type I error rate dropped to 0.00. Though this latter result should not be expected in practice—recall these data were simulated for this exercise—it is instructive to see how well the Bonferroni does at correcting for Type I errors. Note, however, the premium we pay in terms of Type II errors. These rates are listed in the bottom two rows of the table. As can be seen, prior to the Bonferroni correction, we experience a Type II error rate of approximately 37.5%. This means that our tests failed to reject the null 37.5% when the null should have actually been rejected. Thus, our tests are *conservative*, which is almost always a good thing. But note the steep "penalty" that is applied when the Bonferroni correction is used. In this case—see the last row—our Type II error rate has jumped to 98%! This reveals that, under the conditions specified in our simulated scenario (where the test statistics were drawn from a distribution centered

around 2 and a standard deviation of 1.00, which is *just* beyond the 1.96 threshold), we were only able to reject the null hypothesis in 4 out of 200 cases. The message should be clear: applying a Bonferroni correction will decrease Type I error (which, by almost any account is the error rate that is most concerning), but it will also increase Type II error. When the latter is of concern, scholars should consider alternatives to the Bonferroni correction (for example, the XXX or the XXX).

7.3.3 Biased Parameter Estimate (β): Low Pre-study Odds, Reporting Bias, & Publication Bias

Many of the behavioral sciences are currently embroiled in a "mid-life" crisis type scenario. The issue can be summarized succinctly: a large portion of the evidence base is probably incorrect (Ioannidis, 2005). Relying on simple mathematical proofs, John Ioannidis (2005) showed that the probability that any given study reporting a positive (meaning statistically significant) result is correct is vanishingly small. So small, in fact, that Ioannidis titled his paper, "Why Most Published Research Findings are False"! Although the issue is layered, there are three primary concerns driving the "crisis of confidence" that has entrapped psychology and other social science disciplines (Pashler and Wagenmakers, 2012): 1) low pre-study odds; 2) reporting bias; and 3) publication bias. We will address each of these issues briefly in turn.

We will let Ioannidis (2005: 0696) explain the pre-study odds:

> Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated.

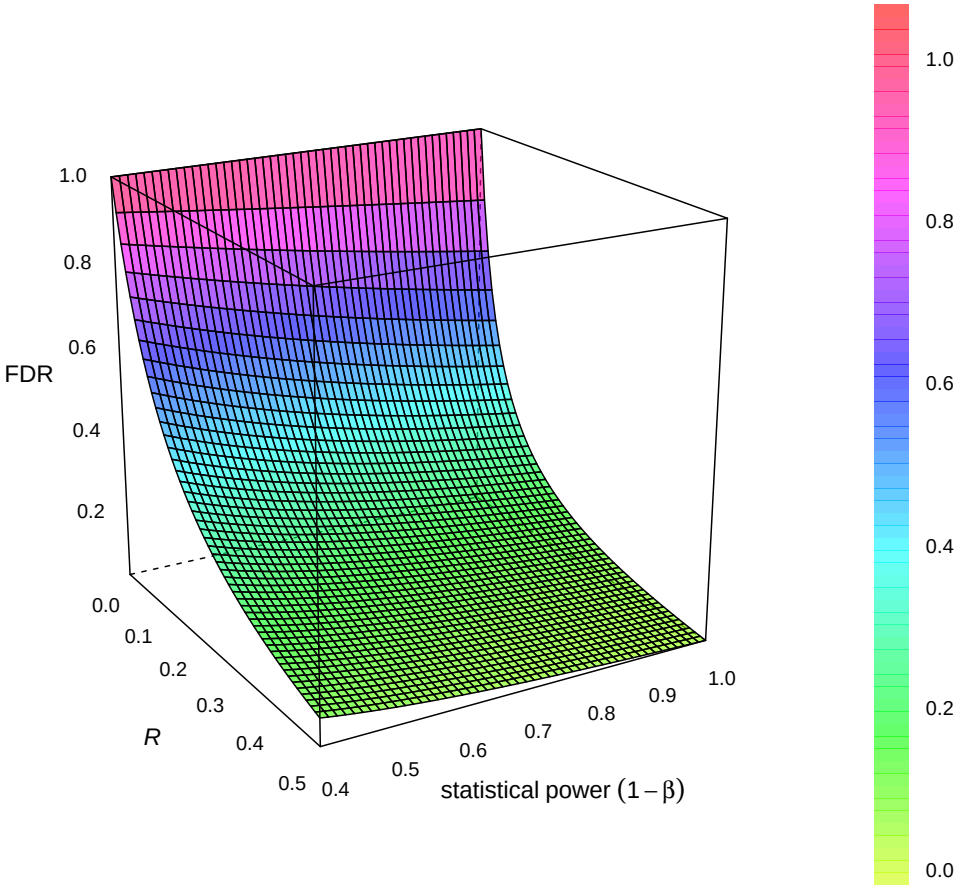
So, if we define R as Ioannidis explains, a relatively "young" discipline like psychology or even criminology might be expected to have a small ratio of true relationships to no relationships. Perhaps, for example, in criminology, R might be somewhere between 1:5 or even 1:100. Regardless of the *actual* value for R, though, the point is clear: the less definitive the literature is in an area, the lower will be the prevailing R.

A low R by itself is not a bad thing. On the contrary, a low R indicates that there is still much to be learned and that there are still many question to be asked, answered, and augmented. So, in a way, one might consider a low R a beacon of light indicating where more attention is needed.

A low R *is*, however, a cause for concern when one places it into the context of false-positive findings and the false-discovery rate (FDR), which ultimately end up leading to the "winner's curse." Briefly, the FDR is an estimate of the degree to which a body of literature

reports "positive" finding (meaning a hypothesis is supported with a statistically significant finding) when, in fact, the relationship does *not* exist. Intuition reveals that R would play a role in the FDR, along with other practical elements embedded in the practical workings of social science research (e.g., statistical power $(1 - \beta$; see Cohen, 1996). The relationship between R, statistical power $(1 - \beta)$ and the FDR is plotted in Figure 7.5).

Figure 7.5: The False-Discovery Rate (FDR)



As shown in the figure, the FDR increases as R decreases (meaning there are fewer "true relationships" relative to "no relationships") and the FDR increases as statistical power $(1-\beta)$ decreases. We will take more time to explain the latter (statistical power) in the next chapter (chapter 8). For now, it will suffice to provide a glimpse into the inner workings of statistical power $(1 - \beta)$. Statistical power $(1 - \beta)$ is primarily driven by two factors: a) effect size (ES) and b) sample size (n). As ES and/or n increase, so too does statistical power $(1 - \beta)$. As statistical power $(1 - \beta)$ increases, the FDR decreases. This can be seen quite clearly in Figure 7.5).

The end result is that any given candidate gene study—and especially if the study is the first to report an association between the gene and phenotype(s) of focus—is more likely to

suffer from the "winner's curse" than it is to report a "true relationship." The "winner's curse" is a phenomenon that explains how scientists who are the first to "find" a relationship are more likely to report upwardly biased ESs and downwardly biased SEs, in no part due to malfeasance. Rather, the issues highlighted by R and statistical power $(1 - \beta)$ underlie the "winner's curse" phenomenon.

The takeaway message should be obvious by now: research areas that have low R, low statistical power $(1 - \beta)$, or both run a high risk of stumbling on false-positive results. And, when one considers all the various ways that reporting and publication bias can influence the findings that eventually make it into print (e.g., authors often estimate a large number of statistical models but only report the ones that supported the hypothesis being tested because statistically significant results are more likely to be published compared to non-statistically significant results), it is easy to make sense of Ioannidis's (2005) claim.

Connecting this to candidate gene research, there are many reasons to believe that candidate gene studies suffer from a vanishingly small R *and* statistical power $(1 - \beta)$ levels that are quite low (often below 0.50) (Duncan and Keller, 2011). Thus, it is safe to conclude that candidate gene studies must contend with a large FDR. This means that when it comes to candidate gene studies, a great many published findings are more likely to be wrong (i.e., a false-positive) than to be right (i.e., a true positive). The solution to these problems is threefold: 1) target *known* candidate genes that have a well-understood causal pathway between the genotype and the phenotype; 2) given the option, prefer larger samples sizes; and 3) correct all p-values for the number of tests that were conducted (and report the number of tests that were conducted).

7.3.4 Biased Parameter Estimate (β): Linkage Disequilibrium

Linkage disequilibrium (LD) occurs when two or more genetic loci (e.g., alleles in two different genes or two different SNPs within the same gene; see the next chapter for more detail) tend to be inherited together at a rate that exceeds chance. Our discussion thus far has assumed that genes are transmitted in a "random" fashion, such that the probability you will inherit gene A is independent of the probability you will inherit gene B. But, as Lynch and Walsh (1998: 94) noted, "when loci are physically linked on the same chromosome, a statistical dependence can exist between the genes incorporated into gametes [the products of inheritance that will go on to form the progeny genotype]." This statement reveals the possibility that genes on the same chromosome will, in some cases, be inherited as a group, thereby violating Mendel's law of independent assortment (Plomin et al., 2013).

Providing a hypothetical example will help to clarify. Imagine you hypothesize that the *DAT1* gene—which is biallelic with a 9 repeat (9R) allele and a 10 repeat (10R) allele—is causally linked to risky behavior (e.g., Gou et al, 2010).[6] To test this hypothesis, you perform

[6]It is important to note that the linkage disequilibrium (LD) effects discussed here are purely hypothetical. In short, we reference the *DAT1* genotype only to make the example more tractable and in no way do we

an analysis much like the ones we demonstrated in the previous section. Results indicate that participants who inherited two copies of the 9R allele evince lower levels of risky behavior compared to those who inherit at least one 10R allele. These results, therefore, suggest the *DAT1* genotype may be linked to the etiology of risky behavior.

But what if that were not the whole story? Instead, imagine there was another gene on chromosome 5 (where *DAT1* is located) that was a close neighbor to *DAT1*. Let us call that gene *NEIGH*. Now, imagine *NEIGH* is a polymorphism with two alleles, N1 and N2. Finally, imagine N1 is the *actual* protective factor for risky behavior, but N1 just so happens to be inherited with the 9R allele from *DAT1* more often than we would expect by chance. Put a different way, allele frequencies between *NEIGH* and *DAT1* are *not* randomly distributed and are instead correlated; meaning a person who inherits N1 on *NEIGH* is more likely to inherit the 9R on *DAT1*. Under these conditions, it is easy to see that if *NEIGH* and *DAT1* are in LD, then any results produced using *DAT1* may be misleading. It may *look* like the 9R for *DAT1* is a protective factor for risky behavior, but it is instead acting as a proxy for the N1 allele on *NEIGH*.

Although LD is a known concern that can mislead researchers carrying out candidate gene studies, according to Carey (2003: 191), it is "...not a fatal problem with association design because refinements of research design can sometimes overcome the problem. In addition, studies of the physiology of the proteins themselves can assist in distinguishing which of two alleles in disequilibrium is the real culprit." In short, if you anticipate that the genotype you are studying is in LD, then it is prudent to consider the probability that it is the *other* gene that is acting as the causal agent. If available, consult the proteomics literature to help differentiate the two genotypes in question. Doing so will allow you to build a strong(er) case for the hypothesized relationship.

7.3.5 Biased Parameter Estimate (β): Population Stratification

Population stratification (PS) is an issue that, while not unique to the candidate gene literature, should be a central concern for any scholar planning to undertake a new study with an admixed population. More detail about PS—and several potential ways to address it—will be covered in the next chapter because this issue poses a major threat to the interpretation of results from GWA studies. Thus, we will save much of the discussion for chapter 8. Here, we only wish to introduce to basic elements of PS and offer a few thoughts on how to correct parameter estimates for the relationship between a gene and Y in a candidate gene study.

PS is, essentially, an issue of confounding that might arise due to the way in which a population (or a sample) is selected. If the population under study is of mixed genetic ancestry and the phenotype (Y) of focus varies across the groups represented, then one is more likely to find a relationship between a gene (G) and Y purely by chance if certain corrective actions are not taken. Imagine, for instance that your candidate gene varies

mean to imply that findings from studies using *DAT1* are spurious owing to LD.

between two groups (e.g., Eastern and Western Europeans). Any phenotype that also varies between these two groups is likely to show a relationship with the candidate gene, *even if the candidate gene has no influence—direct or indirect—on the phenotype.*

The famous "chopstick gene" example comes to mind (Hamer and Sirota, 2000). It is silly to think that there would be a gene that predicts chopstick use. But one could easily find such a gene if one were to analyze chopstick use among a sample that consisted of native Europeans and native Asians. In this example, *any* allele that was more prevalent in one population or the other would almost certainly share an association with chopstick use. Even if the polymorphism were a silent (i.e., substantively meaningless) mutation.

There are several ways that one can guard against PS as a confounding influence. Perhaps most obvious, the researcher should insert control variables for the j groups represented in the sampled population. This could be done simply by including dummy variables for $j-1$ groups that are represented in the sample. Another solution is to estimate a principal components analysis to try and capture the group dynamics that underlie the sample. Yet another alternative is to estimate the relationship of interest *within* the j groups. In other words, the researcher can estimate a within-groups design rather than a naïve between-subjects analysis. If the relationship between the genotype and the phenotype only appears in the latter (the naïve between-subjects analysis), then there is reason to believe the results are being driven by PS. Although it is impossible to offer a general recommendation about how to best handle PS, the most important point is that it be addressed in some way in any candidate gene study.

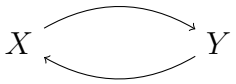
7.4 Appendix: Mendelian Randomization

An innovative analytical technique has emerged out of the candidate gene literature (and the GWA literature discussed in the next chapter). The technique, known as Mendelian Randomization (MR), capitalizes on Mendel's law of independent assortment and combines it with the logic of instrumental variables (IV) analysis that has long been a popular tool among economists (see, generally, Wooldridge, 2010). Although page space precludes a full treatment of MR here, we do wish to provide a general introduction to the conceptual and mathematical properties of the approach. Those interested in a directed discussion of MR are encouraged to see Smith and Hemani (2014), VanderWeele et al. (2014), and VanderWeele (2015). For those generally interested in IV, see Wooldridge (2010).

First, the conceptual framework. As noted above, MR capitalizes on Mendel's law of independent assortment, which tells us that most genetic loci are inherited independent of one another (but see our discussion of LD above). Moreover, genetic inheritance occurs independently of the environment. Thus, some have suggested MR can be considered an approximation of a randomized controlled trial (RCT). In an RCT, participants are randomly assigned to one of (at least) two conditions. In nature, humans are assigned genetic variants largely independently of their environment. Thus, Smith and Hemani (2014: R2) argue

that, "As in RCTs, groups defined by genotype will experience an on-average difference in exposure to trait A, whilst not differing with respect to confounding factors. Thus, a by-genotype analysis is equivalent to an intention-to-treat analysis in a RCT, in which individuals are analysed according to the group they were randomized into, independent of whether they complied to the treatment regimen or not."

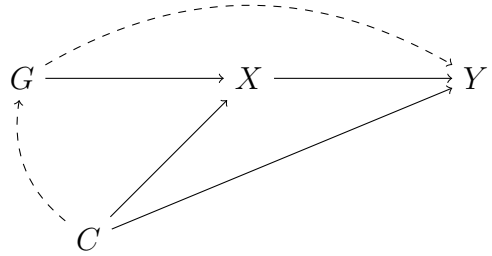
Now, let us think through the mathematical foundation of IV analysis by imagining that G has been linked with a phenotype X by previous research. Moreover, imagine you are interested in studying the impact of X on Y, but endogeneity concerns have precluded any definitive research in the area. For example, perhaps there is reason to suspect that a portion of the correlation between X and Y is due to Y causing X, but that another portion of the correlation is due to X causing Y. In other words, you have reason to think X and Y are wrapped up in a reciprocal feedback loop, like the following:



In a situation like this, any estimate of the association (e.g., r or β) between X and Y will reflect *both* arrows. In other words, one must utilize something other than the "normal" tools if an estimate of the $X \rightarrow Y$ relationship is the path of interest.

One of the best ways to estimate the $X \rightarrow Y$ relationship is to utilize IV analysis. In essence, IV analysis affords the researcher the opportunity to analyze a specific portion of the covariance between X and Y by specifying an instrumental variable(s) for the endogenous variable. The endogenous variable is the one we want to use as a predictor, but we suspect it is endogenous with the outcome. Here, we want to estimate the $X \rightarrow Y$ relationship, but we fear X may be endogenous to Y. Thus, we must specify an IV for X.

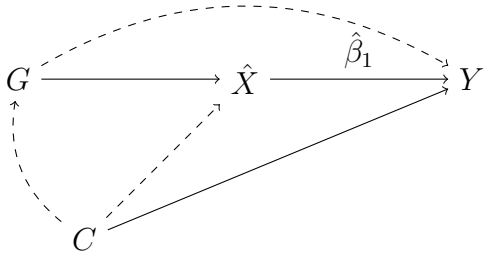
IVs can be thought of as an exogenous variable that has an indirect relationship on Y through X. Specifically, a "good" IV is one that predicts Y, but *only* because it predicts X (assumption IV1, which is sometimes referred to as the exclusion restriction). Further, an IV must be uncorrelated with U (assumption IV2), the error term for the equation predicting Y. In other words, IV2 states that there is no unmeasured confounding between the IV and Y. As shown in the figure below, a "good" IV will have no direct effect on Y (i.e., IV2, represented in the figure by a dashed line) and it will be uncorrelated with U (i.e., IV2).



it is often assumed that if IV1 and IV2 hold, we can estimate the $X \rightarrow Y$ relationship by allowing G to act as an instrument for X, such that the effect of X on Y can be estimated as:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1(\hat{X}_i) + \epsilon_i$$

These relationships can be expressed graphically as:



Notice that $\hat{x_i}$ is being used to predict y_i, not x_i. The notation is critically important because it tells us that only the *predicted* variance of x is being used to estimate $\hat{\beta}_1$. Thus, the importance of choosing an IV that is associated with X. If $Cov(G, X) = 0$, then the plim$(\hat{\beta}_1) = 0$ rather than plim$(\hat{\beta}_1) = \beta_1$

Under the logic of MR, scholars have begun to employ candidate genes as IVs. But, in keeping with the theme of this chapter, it is important that we caution readers about the potential for bias in an IV analysis. The two assumptions—IV1 and IV2—listed above are often discussed in MR analyses, but there is reason to believe that they are often not fully appreciated and that there is more nuance to be considered. As VanderWeele et al. (2014) and VanderWeele (2015) have noted, there are at least *six* "strong" (meaning, they are unlikely to hold) assumptions one must make when conducting an MR analysis. Specifically, when conducting MR analysis, one must assume that 1) X fully captures the phenotype thought to mediate the association between G and Y; 2) X is not time-varying or that if it is time-varying, the time ordering has been correctly specified; 3) there is no $G \times E$ interaction or if there is $G \times E$ interaction, it has been correctly specified; 4) there is no measurement error in X; 5) there is no unaccounted for reverse causation between X and Y; and 6) there is no linkage disequilibrium between the focal G and another variant on the same chromosome.

Along these lines, Pierce and colleagues's (2011) simulations of a range of conditions that are likely to prevail in an MR analysis are instructive. In general, their findings, along with VanderWeele's (2015) assessment, should invoke an air of caution when it comes to the applicability of MR to social science outcomes. Moreovoer, recall the point made above that plim$(\hat{\beta}_1) = 0$ if $Cov(G, X) = 0$. Put more directly, an IV estimator will be biased if G is not associated with X, or if it is only weakly associated with X. This is a critically important

point to keep in mind given that most of the candidate gene research in the social sciences has converged to show that any *one* G is likely to have a (extremely) small effect on the phenotype of interest. That is to say that any gene found to influence a complex human trait is likely to be a "weak instrument." That said, it is fitting to close this chapter by reiterating the newly minted fourth law of behavior genetics (Chabris et al., 2016: 304): "A typical human behavioral trait is associated with very many genetic variants, each of which accounts for a very small percentage of the behavioral variability."

7.5 Conclusion

As you have no doubt noticed from reading this chapter, we encourage much caution when conducting (or when contemplating whether to conduct) a candidate gene study. This should not be misinterpreted as a general resentment or devaluation of candidate gene research. Quite the contrary, we hold the candidate gene literature in high regard. We have conducted candidate gene studies and our work has—in a broad sense—been shaped by many findings that have come out of this line of inquiry. Yet, as with any area of academic interest, the approaches to behavioral genetic research have quickly evolved. One of the implications of this evolution is that candidate gene studies have fallen out of favor (Dick et al., 2014). The reasons for this can be traced to many of the issues that were outlined in this chapter. Indeed, there are many trapdoors in this area of study. But that does not mean candidate gene studies are "dead" or that they should be abandoned. Instead, we hope that the discussion in this chapter will encourage readers to conduct their own candidate gene studies, but that they will do so using the most up-to-date techniques, that they will consider each of the numerous confounding factors, and that they will be more transparent in their decision making. The candidate gene literature has plenty of "noise." Thus, we hope to encourage novel contributions that will better highlight the "signals."

Chapter 8

Genome-wide Association Studies (GWAS) & Extensions

Given nothing more than the sheer number of concerns with the candidate gene approach to behavioral genetics research, one is justified in wondering "what should I do?" In other words, it is reasonable to feel overwhelmed when contemplating the best way to handle the limitations and how to address the assumptions necessary to go "gene hunting" using the approach laid out in the previous chapter. These very concerns, it turns out, motivated behavioral geneticists to find an alternative way to study the genetics of human complex traits. That alternative is known as a genome-wide association study (GWAS).

As you will see in this chapter, GWAS capitalizes on modern computing power and modern genomics technology to avoid the trapdoors that come with the candidate gene study. Recall from the last chapter that one of the major limitations of the candidate gene study is that the researcher is forced to assume that the gene chosen for analysis is *the* causal variant or that it is in linkage disequilibrium (LD) with the causal variant. The candidate gene study has repeatedly shown that any given candidate gene is likely to explain only a fraction of a percentage of the variance in the phenotype P. Thus, candidate gene studies tell us almost nothing about the landscape of genetic effects for a human complex trait. This is to say that studying one (or two, or even three) genes at a time is error-prone at best and it is a slow process for trying to uncover the genetic variants that might lie beneath the heritability blanket.

On that last point—the heritability blanket—it is important that we connect the material in this chapter with the central equations that have formed the backbone of this text. Recall that we have traced everything in behavioral genetics back the basic equation:

$$P = G + E$$

But recall also that we (see page ??) noted early on that this basic linear model obscured many of the known complexities that researchers are interested in studying. For example, the above equation "hides" gene-environment interplay (i.e., gene-environment interaction

[$G \times E$, see chapter ??] and gene-environment correlation [rGE, see chapter ??]). We noted that this issue can be addressed by generalizing the above equation to account for any arbitrary gene-environment interplay like so:

$$P = \Psi(G, E)$$

where Ψ (psi) serves as a general function that can be thought of as capturing all the complexities (i.e., interactions and correlations) that may exist between the Gs and the Es in the etiology of P. Some of those potential complexities are covered in the next two chapters where $G \times E$ is addressed in chapter ?? and rGE is addressed in chapter ??. So, for now, we will set those issues aside. What is important to notice about Ψ is that it also captures any arbitrary combination of the components that go into G or the components that go into E. This chapter will consider the way in which the factors that make up the G component are identified and combined. A later chapter will take a similar approach to the factors that make up the E component (see section ?? of chapter ??).

Before we move to the conceptual overview, it is important to point out that the strategy for this chapter will depart from all of the others in Part II of this book. Specifically, this chapter will not include a demonstration section. The reason for this omission is simple: GWAS research strategies are too complicated and involve far too many technical details that require careful attention to fit into a single chapter. Any attempt to water down the approach so that it could be fit into a single chapter would paint too distorted of a picture and could, quite frankly, lead to more confusion about the method. Moreover, with the rate at which technology is being developed any text that did attempt to demonstrate GWAS would almost certainly be out of date before it even went to print.

Readers who desire an overview of the practical research process underlying GWAS are encouraged, of course, to read this chapter but to also seek out the latest information by reading up-to-date research literature on the topic. Over the past five years or so, the tools and techniques used by behavioral geneticists who carry out a GWAS project have changed dramatically. For example, as will be outlined below, some of the first GWAS studies on human behavior were published in the 2000s and early 2010s. Within just a few years, several problems were highlighted and the strategies used by GWAS researchers were quickly updated to accommodate and to handle the concerns. In recognition of this fast-paced growth, we have opted to focus exclusively on the conceptual issues that surround GWAS. If a "how to" guide is what you are really after, we encourage you to consider attending the International Workshop on Statistical Genetic Methods for Complex Human Traits that is held bi-annually in Boulder, CO (the webpage for the 2017 workshop can be found here: `https://ibg.colorado.edu/dokuwiki/doku.php?id=workshop:2017:start`). Also, for those looking for a user-friendly introduction to the GWAS analysis in R, we encourage you to visit the tutorial provided by the Open Resources in Statistical Genomics available here: `http://www.stat-gen.org/tut/tut_intro.html`.

8.1 Conceptual Overview

8.1.1 The Logic of GWAS

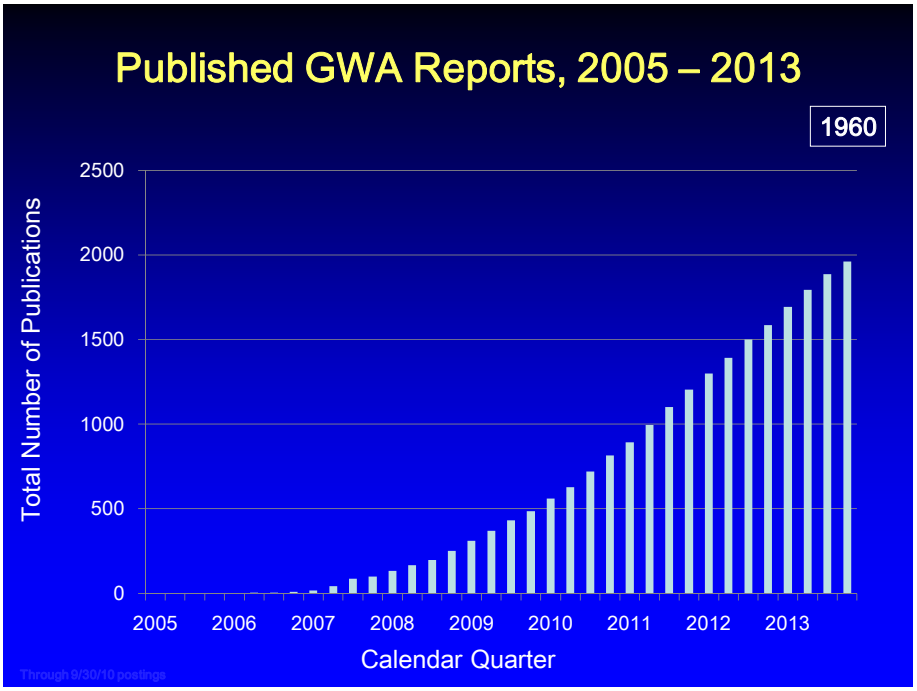
Modern "Gene finding" technologies have completely revolutionized behavioral genetics. As Ku and colleagues noted back in 2010 (p. 412):

> Genome-wide association study is a comprehensive and biologically agnostic approach to searching for unknown disease variants, and as demonstrated in more than 450 [now in the tens of thousands!] studies, this strategy has been very successful in identifying new genetic loci for various human complex traits. Most of the genes and loci that have been identified are not previously thought to be associated with their respective diseases.[122-125] More importantly, the GWAS findings have also provided new insights into the molecular pathways of complex diseases even when most of the disease causative variants remain to be discerned from the neighboring correlated markers. For example, the three new genes that have been linked to Crohn's disease: *IL23R*, *ATG16L1* and *IRGM* have highlighted the importance of interleukin-23 receptor and autophagy pathways underlying the pathophysiology of this chronic inflammatory bowel disease.[126,127] Notably, GWAS have been making some significant advances in our understanding and knowledge of the genetic basis of human complex diseases compared with the pre-GWAS approaches (that is, the candidate gene association and linkage studies).

That 450 studies cited by Ku and colleagues has exploded to well into the thousands now (as our insertion exclaimed). To be sure, the National Institute of Health (Welter et al., 2014) has kept a running catalog of GWAS studies since the early 2000s. Thousands of new studies are added to the catalog every year (see `https://www.genome.gov/26525384/catalog-of-published-genomewide-association-studies/`.

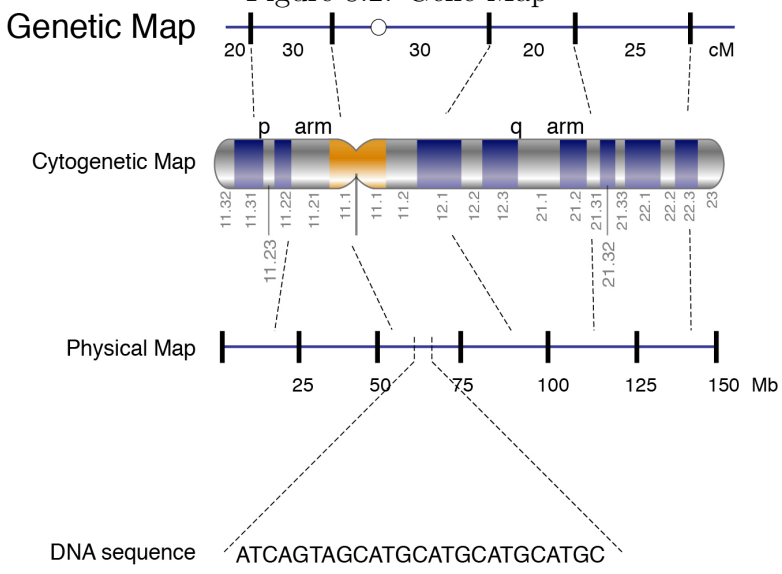
Now, assuming your convinced that GWAS is an impressive or—at a minimum, a popular—tool, the next question to be dealt with is "how does it work?". To answer that question, it is helpful if we start by imagining we are looking at a single chromosome. If we could pluck one chromosome from a human cell, amplify it so we could see it on the page of a book, it might look something like Figure 8.2, which is a "map" of chromosome 11 (Figure 8.2 is available at `https://www.genome.gov/dmd/index.cfm?node=Photos/Graphics`). Imagine it were possible to start on the p arm (i.e., the left-hand side of the chromosome in Figure 8.2) of chromosome 11 and work your way down to the end of the q arm. Imagine also that every time you encountered a single nucleotide polymorphism (SNP), you documented it. For example, take a look at the enlarged DNA sequence at the bottom of the figure. Notice how it begins with AT<u>C</u>AGT. But imagine some respondents in your data presented with the following sequence: AT<u>T</u>AGT. Notice how these two sequences differ at the third base where the first sequence has a C and the second sequence has a T. This is what we

Figure 8.1: Gene Map



mean when we refer to SNPs. When SNPs are identified, they can be used as predictors of P to see if individuals carrying one variant have a heightened risk of developing P compared to those carrying an alternative variant.

Figure 8.2: Gene Map



But, rather than stopping with the one SNP identified above on chromosome 11. Imagine you did this for *every* SNP that could be identified across *all* human chromosomes. Further, imagine that each time you found a SNP, you looked to see whether carries of that variant had a differential risk for P compared to others who carried an alternative variant. If you were to do this, you would be carrying out a GWAS.

This leads to an obvious question: how many SNPs are there? Like most questions addressed in this book, the answer to this one is not straightforward. But we can provide some insight by looking to the latest technology to see how many SNPs are known and, therefore, how many SNPs behavioral geneticists typically search when they want to compute the G portion of the $P = \Psi(G, E)$ equation.

It is estimated that there are approximately 10 million SNPs in the human genome (see `https://ghr.nlm.nih.gov/primer/genomicresearch/snp`). This number was derived by carrying out a simple calculation. In essence, geneticists have shown that a SNP occurs, on average, about every 300 nucleotides in the human genome. This means that if you were to line up all the As, Ts, Cs, and Gs, in the human genome and start from the beginning and read all the way to the end. You would encounter a SNP—a location where the genome you are reading would differ from another, reference, genome—every 300 nucleotides on average. It is critical that you not misinterpret this point. The "SNP every 300 nucleotides" prediction does *not* mean that you are guaranteed to see a SNP at regular intervals in the human genome anymore than you expect the roulette wheel at the casino to hit red every other spin. Instead, you *expect* a SNP about every 300 nucleotides, meaning this is your prediction. Predictions, of course, are not always correct so we know that SNPs may not occur at regular intervals, but on average this is a "good enough" description of the rate of SNPs in the human genome.

Now, taking the "SNP every 300 nucleotides" prediction and combining it with the knowledge that the human genome consists of roughly 3 *billion* (i.e., 3 with 9 zeros or 3×10^9) nucleotides, we can garner a rough estimate that there are approximately 10 million SNPs in the human genome: $3,000,000,000/300 = 10,000,000$. That's a lot SNPs to deal with!

The human brain is, of course, not equipped with the necessary hardware to analyze 10 million SNPs. Nor would it be possible to even keep track of all that data if it were not for the advances in modern computing that we have experienced over the past few decades. The point we are making here is that GWA data is "big data" in every sense of the term.

Think of it this way, there are 10 million SNPs in the human genome—the vast majority of them are going to have no effect on human development, meaning they are silent mutations—but some of them *will* impact the phenotype of focus. But there are 10 million SNPs to sort through. That's a big haystack.

So how could scientists possibly hope to sort through all this information to find the needle, the genetic "signal"? Oh, and to further complicate matters, it is worth reminding you about the fourth law of behavioral genetics that we discussed in chapter ?? (Chabris et al., 2016): most genetic effects are small.

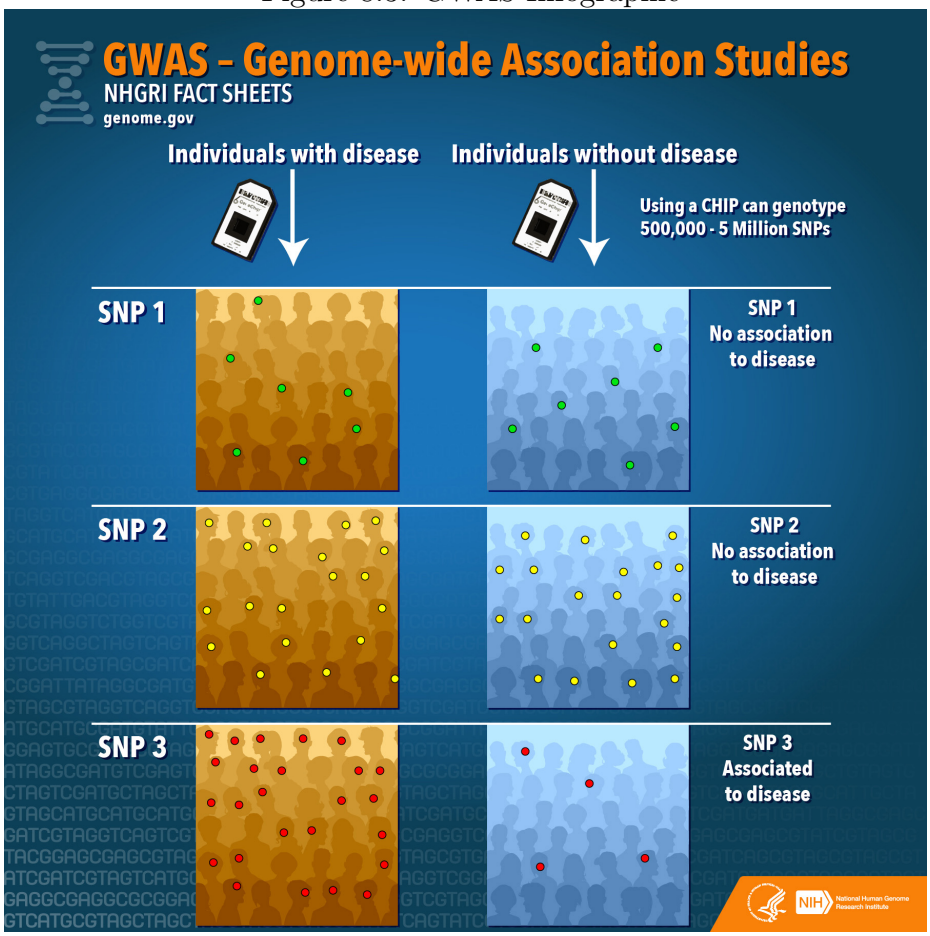
To being to understand how scientists sort through all the "noise" in search of weak "signals", consider the information provided in Figure 8.3, which is available at `https://www.genome.gov/20019523/`. This figure shows how modern GWAS is carried out—with, of course, some important details omitted! The figure contains two "types" of people: 1) those who are affected by some disease and 2) those who are not. The individuals affected by the disease appear on the left-hand side of the figure. Individuals are *not* affected by the disease are presented on the right. Three SNPs have are shown in the figure. The first SNP—creatively labeled SNP 1—appears in the top row of the figure. We see that there is no association between SNP 1 and the disease. We see this because the figure tells us, but we can also see it if we consider the dots in the figure space. The dots represent the *presence* of the "problem" SNP. Imagine, for instance that SNP 1 had an A allele and a G allele. Imagine further that the G allele was thought to be associated with the disease of focus. If this were borne out in the data, then we should see the G allele more often in the left column (i.e., among those who are affected by the disease) than in the right column. The presence of the G allele is represented as a dot in the figure. If you were to count up the number of dots that appear on the left-hand side you would see that there are seven folks with the G allele. How does this compare to those on the right? There are six healthy (meaning they do not suffer from the disease of focus) who carry the G allele. Although the numbers are not exactly the same, it is hopefully apparent that such a small difference in the presence of the G allele across disease status is not enough to reject the null hypothesis of no difference between the groups. In other words, despite the fact that those affected by the disease are more likely to carry the G allele, the difference in the proportion of cases with the G allele is not enough to warrant any firm conclusion about the link between the G allele and the disease outcome.

A similar story emerges when we look at the second row of Figure 8.3. In essence, the second row also shows that the SNP of focus—here, SNP 2—is not associated with the disease trait because those who are affected by the disease carry the focal allele at approximately the same rate as those who are not affected by the disease. In other words, variants of SNP 2 have no discernible (i.e., statistical) link to the risk of developing the disease.

A different pattern is seen in the bottom panel of Figure 8.3. Notice that here, those who developed the disease appear to carry a certain variant of SNP 3 more often than individuals who did not develop the disease. This is clearly represented by the sheer number of dots that are shown on either side of the figure. It is easy to see that the "problem allele" for SNP 3 is much more common among those with the disease than it is among those who did not develop the disease. All else being equal, this suggests that the disease of focus *may* be linked to SNP 3.

Although Figure 8.3 is an overly simplified demonstration, the logic and the approach it reveals are actually quite telling. Notice that the logic used to show that SNP 3 is predictive of the disease trait was straightforward: if one allele shows up more often among the "disease" group, then it might be related to the disease. In this way, you can think of GWAS as simply looking for an association between two binary variables: 1) a SNP with two categories and 2) a disease trait with two outcomes. Thus, it may help if you consider GWAS one SNP and

Figure 8.3: GWAS Infographic



one disease at a time. In this way, it is easy to see that GWAS simply computes a 2×2 table much like the one below:

	disease not present	disease present
allele 1	p_{00}	p_{01}
allele 2	p_{10}	p_{11}

Imagine that $p_{01} > p_{11}$ and that this difference was statistically significant. Such evidence would suggest that allele 1 had an association with the disease trait of focus. But scientists, of course, would never just rely on a simple observation of proportions like this. Rather, they would prefer to calculate whether the proportions shown in the table above depart from what we might expect under random error. In other words, it is crucial to carry out a statistical significance test to determine whether the proportions are statistically significantly different from one another.

Recall that whenever one has data that can be displayed in a simple table like the one above, the chi (say: "ki") squared test (i.e., X^2) is appropriate. The X^2 value can be computed and compared to the χ^2 distribution of values. If the observed value is greater

than the critical value gleaned from the χ^2 distribution (i.e., if X^2¿critical χ^2) with df = columns $-$ 1 $\times$ rows $-$ 1, then the focal SNP could be said to be a statistically significant predictor of the phenotype.

At its core, GWAS is a relatively straightforward technique. But as you have no doubt already considered, the practical implementation of GWAS is a bit more complicated than we have described here. There are, indeed, a number of critically important issues that must be considered and addressed before the results of a GWAS can be interpreted in any meaningful sense. We will briefly touch on some of the most important issues that must be considered in the following subsections. Again, though, it is important that you realize that this chapter does not cover *all* the issues that underlie modern GWAS. We hope to give readers a working knowledge and the vocabulary necessary to interpret GWAS findings. Those who wish to perform GWAS on their own data are encouraged to consult GWAS experts and/or consider attending the annual workshop hosted by the Institute of Behavioral Genetics at the University of Colorado, Boulder (here, again, is the link to the 2017 workshop: `https://ibg.colorado.edu/dokuwiki/doku.php?id=workshop:2017:start`)

8.1.2 Correcting P-values: Genome-wide Statistical Significance

This simple thought experiment from above raises several important points. One that might be standing out in your mind concerns the sheer number of statistical tests that are estimated in any GWAS. Recall that modern GWAS analyzes millions of SNPs. It does so by computing X^2 statistics for the relationship between the focal SNP and the phenotype. Recall from our discussion in chapter ?? the issue of multiple testing bias (MTB; see pages ??). The issue, simply put, is that P-values calculated for statistical tests assume the association of focus is the only association that was analyzed. In the context of GWAS, to show an association between any given SNP k and the phenotype, the P-value calculated would only be interpretable as the probability that one would see an association that large (or larger) by chance alone *if the focal association were the only one tested.* The last part of that sentence—the italicized part—can be thought of as the independence assumption. It is typically overlooked (in most cases, safely) in every day applied statistics usage. Hopefully, though, it is obvious to you that this point cannot be overlooked in GWAS. Indeed, GWAS represents a *major* violation of the independence assumption.

Given that GWAS violates the independence assumption on such a large scale—one million X^2 tests is a massive departure from the independence assumption!—something must be done. That "something", as it turns out, is a rather simple, yet elegant, solution.

The solution is to *correct* the P-values for the MTB by carrying out a Bonferroni correction. Recall from chapter ?? that the Bonferroni correction simply divides the desired P-value by the number of tests k that were performed: α/k. If GWAS typically performs approximately one million tests and the P-value desired is $\alpha = 0.05$ for *each* test, then the correction would suggest the *actual* P-value threshold for any given test should be set at $\alpha/k = 0.05/1,000,000 = 0.00000005 = 5 \times 10^{-8}$. This, it turns out is the P-value that most

scholars use to differentiate "significant" SNPs from those that are not. The term coined for this value is "genome-wide statistical significance." Thus, if you happen to see a GWAS that reports to have found one or more genome-wide statistically significant SNPs, you know that that simply means the relationship between the SNP and the phenotype had a P-value that was lower than 5×10^{-8}.

8.1.3 What Do Genome-wide Significant SNPs Tell Us?

What does it mean to say that the SNP is a genome-wide statistically significant predictor of the phenotype? Of course, we have covered what that means in a *statistical* sense int the previous subsection. But what does it mean on a *substantive* level?

It could mean several things. It could mean that the focal SNP is a causal variant of the phenotype P. But, given the sheer volume of SNPs in the human genome, when one finds a significant X^2 value, it is unlikely—probabilistically speaking—that it will be a causal variant. Instead, it is more likely that the statistically significant SNP is in linkage disequilibrium (LD) with the *actual* causal variant. We covered LD in more detail in chapter ??. Recall that LD occurs when two or more genetic loci (e.g., alleles in two different genes or two different SNPs within the same gene) tend to be inherited together at a rate that exceeds what we would expect by random chance.

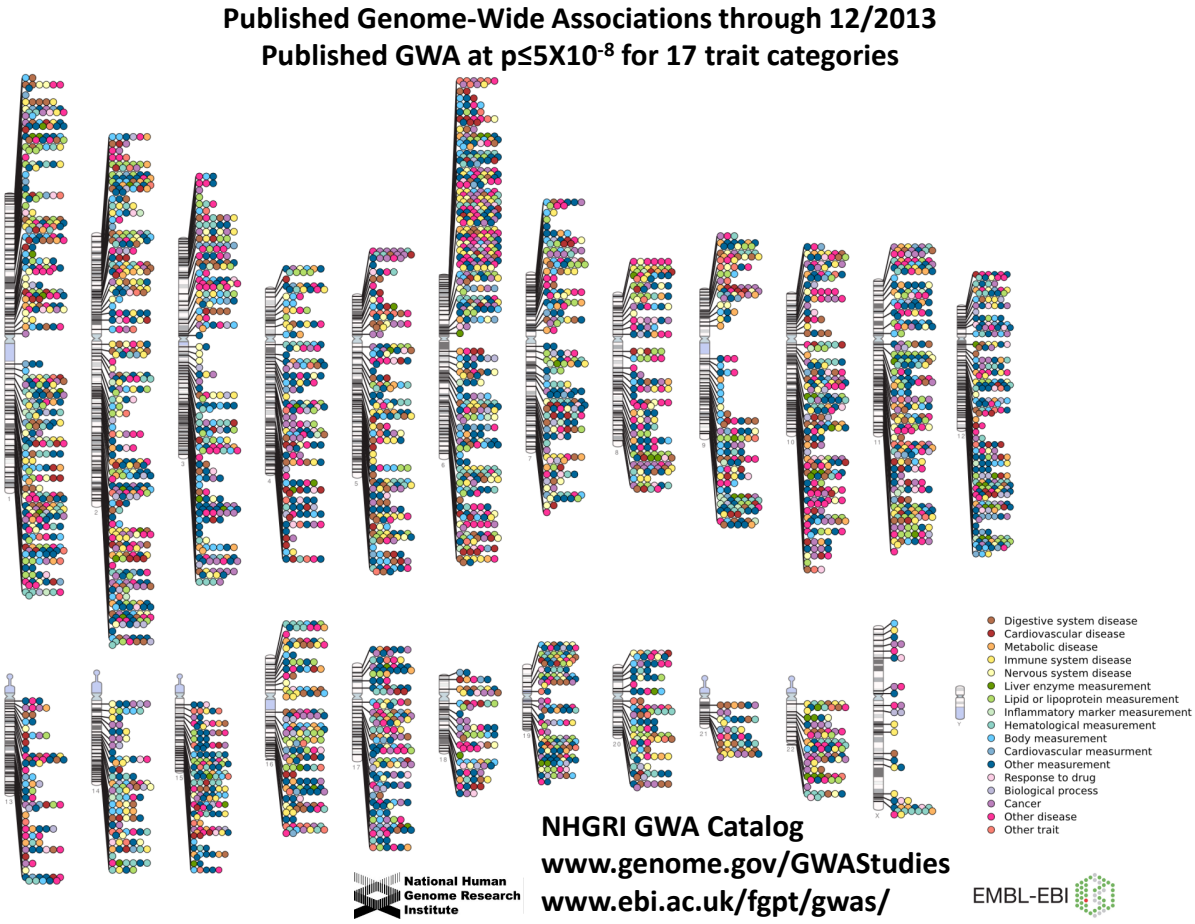
A hypothetical scenario will help to clarify how this could impact GWAS results. Imagine you perform a GWAS and find that a certain SNP emerges as genome-wide statistically significant. The next obvious task would be to determine whether the SNP was a functional variant or if it instead served as a proxy for another variant that was the actual cause. The latter scenario would represent a SNP that was in LD. If, for instance, the SNP that emerged as genome-wide statistically significant was in LD with a functional genetic variant—it is important to note that the actual functional variant need not be another SNP, it *could* be a different type of genetic variation such as a VNTR—then we would want to search the genetic code that immediately surrounded the focal SNP to see whether there was a gene nearby. This is often displayed in GWAS with a regional association plot like the one displayed in Figure ??.

As you can see, the regional association plot in Figure ?? shows that there was one genome-wide statistically significant SNP (rs??????). The nearest neighboring genes—meaning those that are most likely to be in LD with rs??????—are ???? and ????. These represent the best and most likely candidates to be the true causal variant.

That gives you an idea of the "how to" behind the GWAS procedure. With all the background information, you may have been wondering what we have learned from GWAS. In other words, it is important to pause briefly from the methodological discussions and consider what genome-wide statistically significant SNPs have told us to date. Have there been any meaningful discoveries to emerge from GWAS? In short, the answer is a resounding "yes!"

Take a look at Figure 8.4. What you see is a map of sorts. You probably already identified the 22 autosomes and the sex chromosomes on the map. But what are all those little dots you ask? Each dot represents a SNP that has been linked to a different phenotype in GWAS. Most of the phenotypes indexed on this plot represent disease traits or medical outcomes. Nonetheless, you can grasp a sense of the sheer volume of information that has been gained by GWAS in just a few short years.

Figure 8.4: Genetic Markers Identified by GWAS



But recall that as of the writing of this text, it is possible to genotype an individual (or a sample of individuals) for well over one million SNPs (of the 10 million that probably exist). Take a moment to think about how remarkable that is. For around $100 (weblink to commercial SNP chip????), geneticists can now scan your genome for 1,000,000 known SNPs. SNPs that are believed to be related to all sorts of complex outcomes ranging from educational attainment (Rietveld et al., 2013) to head circumference (CITE??) to schizophrenia (CITE??).

The number 1,000,000 should also give you pause, though, by revealing a point we have repeated several times throughout this text. When it comes to human complex traits, *there*

is no one gene for that outcome. Put differently, the GWAS revolution has made clear the importance of remembering the new fourth law of behavioral genetics: "A typical human behavioral trait is associated with very many genetic variants, each of which accounts for a very small percentage of the behavioral variability" (Chabris et al., 2016: 304).

In other words, when behavioral geneticists find a SNP to be associated with some complex human phenotype, you can bet—without knowing any details—that the effect is small and quite variable. Or, as Conley and Fletcher (2017: 5) put it: "...the 'small effects' aspect of this paradigm has called for ever-larger data sets to find the needles in the genomic haystack because the needles are now thought to be much smaller than originally suspected."

This point serves as a nice segue into the next subsection, which considers the role of statistical power in GWAS.

8.1.4 Statistical Power $(1 - \beta)$ & the False-discovery Rate (FDR)

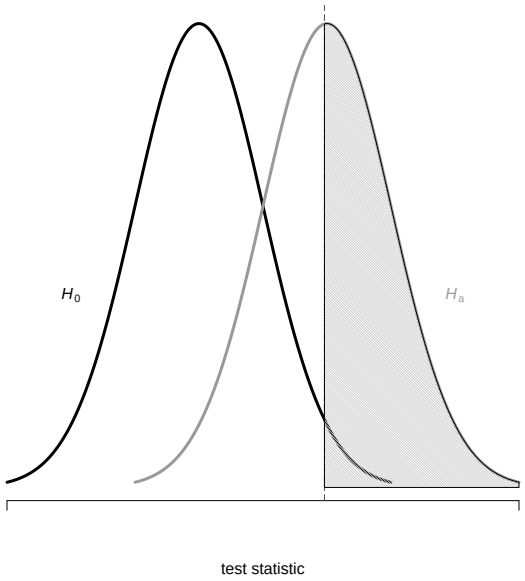
Given the P-value correction that is required to qualify a SNP as being genome-wide statistically significant ($P < 5 \times 10^{-8}$), it should be obvious that statistical power $(1 - \beta)$ will be an important concern in any application of GWAS. This raises several important issues. First, what is statistical power $(1-\beta)$ and how should we be thinking about it in the context of GWAS? Second, what are the consequences of having "low" statistical power? Finally, how might *under*-powered GWAS mislead researchers and why would that be bothersome? We will address each of these questions in turn.

First, let us consider what statistical power *is*. In order to understand statistical power, it is helpful to begin by discussing the three factors that go into it (Cohen, 1988; Sham and Purcell, 2014): 1) the α-level specified by the researcher, 2) the effect size of the relationship in question, and 3) the sample size (n) of the study. The role of these three ingredients in statistical power is easy to grasp when one considers the sampling distribution of an estimate (or, alternatively, one can think of the sampling distribution of a test statistic like the X^2 statistic discussed earlier or the t-statistic that is often calculated in regression models). The α-level is a threshold that differentiates a statistically significant effect from one that is not statistically significant. We have already discussed above the point that the GWAS α-level is 5×10^{-8}.

The central limit theorem tells us that estimates will always include error, meaning we will never observe the "true" effect in the population (i.e., the population parameter of interest). Thus, the estimate will have its own sampling distribution, known as the non-centrality parameter distribution. We can think of the non-centrality parameter distribution as a normal distribution like the one presented in grey on the right side of Figure **??**.

Larger effect sizes—all else equal—push the sampling distribution for the non-centrality parameter further away from the null hypothesis (H_0) distribution, which is the distribution displayed in black on the left-hand side of Figure 8.5. Smaller effect sizes allow these two

Figure 8.5: Probability Distributions for Test Statistics & Statistical Power



distributions to overlap more so than larger effect sizes. This is important because the degree to which the two distributions overlap directly affects the statistical power of the study because statistical power is calculated as the proportion of the non-centrality parameter distribution that lies beyond the α-level threshold in the H_0 distribution.

The α-level threshold is displayed in Figure 8.5 as the vertical line. Note that the α-level is set in the H_0 distribution. In other words, the α is the location in the H_0 distribution that we have arbitrarily chosen as the demarcation line for a statistically significant finding. In a very real sense, then, the α-level is the point at which we are willing to say that the observed effect may not have come from the H_0 distribution because it is an unlikely result if the H_0 value were actually the true population value.

The last ingredient in statistical power is the sample size (n) for the study. All else being equal, studies with larger n produce sampling distributions (both the sampling distribution for H_0 and the sampling distribution for the non-centrality parameter) with less variation, meaning the sampling distributions will have greater precision. This is reflected most often by standard errors, which are estimates of the standard deviation of the sampling distribution. Standard errors decrease as n increases. As n grows, the sampling distribution for the H_0 and the non-centrality parameter will be narrower with a much more obvious peak around the true parameter in the latter. The reverse is also true; as n drops, the sampling distributions for both will have greater variation.

It should be obvious at this point that parameter estimates vary and sometimes they will stray quite drastically from the true population that they are intended to represent. When an estimate strays far enough away from the H_0 value (which is almost always set to 0 such

that $H_0 = 0$ in most cases) that it passes an the α-level threshold, researchers claim the estimate is statistically significant (Fisher, 1925[1973]); which is to say that the researcher observes evidence to suggest that the parameter estimate did not come from the H_0 sampling distribution.

The shaded region of Figure 8.5 denotes the amount of density that lies beyond the threshold α-level and, therefore, the density in the non-centrality parameter distribution that lies beyond the α-level. That density is referred to as statistical power and it is symbolized as $1 - \beta$.

This may all seem a bit esoteric and difficult to follow. If that is the case, it may help if you imagine the two curves in Figure **??** are real entities and that they are sitting on the table in front of you. We can "see" what happens to the distributions as the non-centrality parameter (i.e., the true effect size) increases. When the true effect size increases, the distribution on the right slides even further to the right. As it does, the shaded region increases because the α-level threshold remains fixed in the H_0 distribution. Thus, the statistical power for your test increases because a larger portion of the non-centrality parameter distribution lies beyond the α-level threshold.

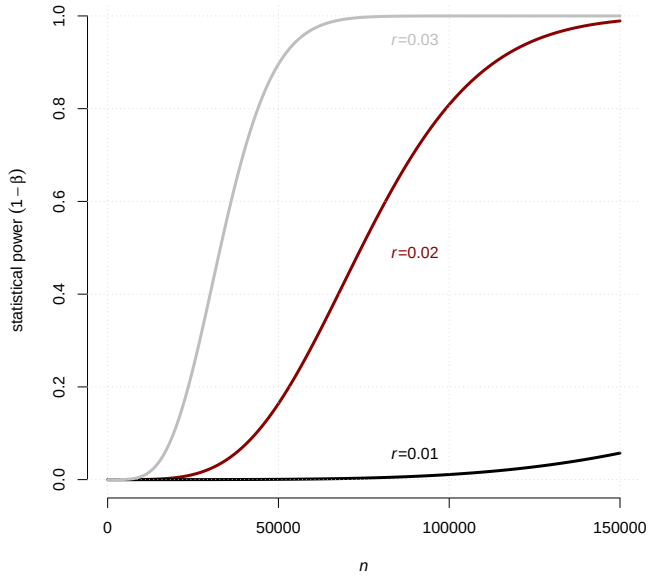
A similar "visual" can be painted to represent the effect of increasing n. Again imagine the distributions are sitting on the table in front of you. But this time, imagine they are made out of Play-Doh. Increasing n would be like taking each distribution and squeezing its sides. What would happen? It would become narrower and narrower until it was eventually just a flattened piece of material. This is, in effect, what happens to the distributions shown in Figure 8.5 as n approaches infinity. Note, however, that the center of the distributions stays in place while this happens. Thus, you can easily see that if you were to "squeeze" the sides of the two distributions, you would be forcing more density into the central regions of their distributions. This would increase statistical power because you would be decreasing the amount of overlap the two distributions share and, as as result, you would be pushing the α-level threshold out of the non-central parameter distribution.

The opposite occurs as n drops. Indeed, as n decreases, its like you push down on the top of your Play-Doh distributions. This smashes them until they are flat and have run together in an ugly brown mess.

The takeaway from this discussion is this: given that GWAS effect sizes are expected to be small (Chabris et al., 2016), it is imperative that GWAS be conducted with large n. Large n is the only way to ensure there is adequate statistical power to detect small effect sizes. But just how large are we talking? Really large. In the 10s or 100s of thousands large depending on the expected effect size. Perhaps the best way to get a feel for the sample sizes needed is to perform a statistical power test. Presented in Figure 8.6 are three statistical power curves for three different effect sizes. The effect size scale used here is the standard Pearson product-moment correlation coefficient (r) (see chapter ?? for a review). As can be seen, for effect sizes of $r = 0.01$, it will take extremely large n to bring statistical power into a region that is even distinguishable from 0.00. The prospects are brighter for effect sizes of

$r = 0.02$ and $r = 0.03$. As shown, statistical power can top the typical threshold of 0.80 for $n \approx 50,000$ if $r = 0.03$. Larger n is necessary to hit this threshold if $r = 0.02$.

Figure 8.6: Statistical Power Curves for Different Effect Sizes (r) as n Increases



The second point raised at the beginning of this subsection concerned the consequences of having "low" statistical power in GWAS. Typically, under-powered research designs are understood to increase Type II error. Type II error is the not-so-informative name given to the probability that one will fail to reject the null hypothesis when in fact the null should be rejected. In other words, Type II error when an effect actually is greater than zero, but the scientist "overlooks" it because there is not enough statistical power for the observed effect to reach statistical significance. This is, indeed, a problem and one that any researcher would want to mitigate if possible. The most obvious way to limit Type II error is to increase statistical power by increasing the n study. Of course, statistical power increases when an effect size increases, but hopefully it is obvious why we are not advocating for one to increase his/her effect size in order to limit Type II error. The effect size is observed in a study. The scientist cannot control it; although s/he *can* control *which* effects are explored, so in that sense s/he can control the effect size to some extent.

Although the problem of non-trivial Type II error rates is concerning, there is another concern that stems from using under-powered GWAS and this one is far more problematic. This point addresses our third issue that was introduced above. Specifically, we will show that under-powered research designs can increase the probability that a statistically significant finding is a false-positive. In other words, not only will under-powered GWAS *miss* important signals, it will also make it more likely that the statistically significant findings that emerge are errors.

Recall the false-discovery rate (FDR) that was discussed in chapter ??. While we introduced the concept in that chapter, what we did not show you was the computation for the FDR. The equation for calculating the FDR was not critical to our earlier discussion, so we omitted it there. But now, it is important that you see it because you will recognize the central role that statistical power plays in that calculation:

$$FDR = \frac{\alpha}{\alpha + (1 - \beta)}$$

Given the emphasis we have placed on statistical power $(1 - \beta)$ in the past few pages, it is likely that your eye jumped right to the $1 - \beta$ in the denominator of the FDR equation. Because statistical power appears in the denominator of the FDR equation, we can easily see that, all else being equal, the FDR will *increase* as statistical power *decreases*. In words, GWAS is more likely to "find" a false-positive association between a SNP and the phenotype whenever statistical power is low.[1]

8.1.5 Type M & Type S Error

When statistical power is low and effect sizes are expected to be small—two obvious features of GWAS—the possibility is raised that one will make a Type M (for "magnitude") or Type S (for "sign") error (Gelman and Carlin, 2014). A Type M error occurs when a researcher performs an analysis and finds a statistically significant result that is overstated in terms of its effect size. In other words, a Type M error is one in which the researcher's estimate of the effect size is larger than the actual effect size in the population. It is easy to see how this might happen if we consider the probability distributions that were shown earlier in Figure 8.5. Recall that the distribution on the right—the non-centrality parameter distribution—represents the distribution of estimates we would see if the effect size were equal to the mean of that distribution. The shaded portion of the plot shows the density of the non-centrality parameter distribution that sits to the right of the α-level threshold. This reveals that any estimate that falls in the shaded portion of the plot would be marked as statistically significant.

Now consider the range of estimates that would emerge as statistically significant. For Figure 8.5, that would only happen for estimates that were roughly equal to or *larger than* the actual effect size. Again, note that the *actual* effect size is represented as the mean of the non-centrality parameter distribution. Thus, you can see from Figure 8.5 that it is necessary to over-state the effect size in order to find a statistically significant result.

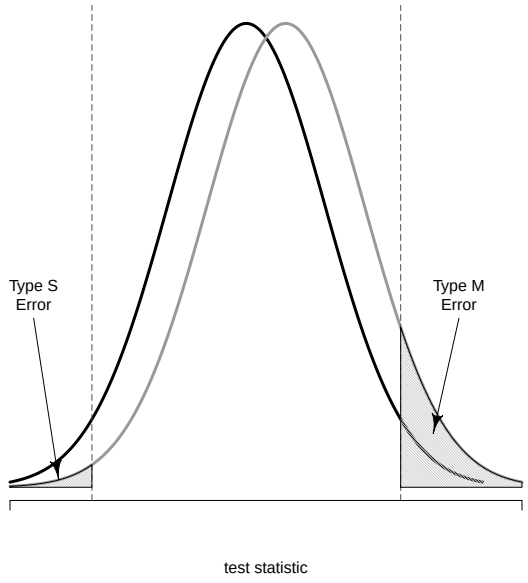
But note that the conditions shown in Figure 8.5 do not capture all scenarios. In fact, the scenario captured in the figure represents a reasonably well powered analysis (statistical power is approximately 0.50 in the figure). So what happens when statistical power is lower? Well, the short answer is that the situation worsens. You can visualize this if you once again

[1]Note that for simplicity we omitted the "prior" (Duncan and Keller, 2011) or what Ioannidis (2005) referred to as R from the FDR equation.

imagine the non-centrality parameter distribution slides to the left as statistical power weakens. As this happens, the shaded region shrinks to the right-tail of the distribution, revealing that one is required to find more-and-more extreme values if a statistically significant result is going to emerge.

A related problem known as Type S error (Gelman and Carlin, 2014) can also emerge when statistical power is low. In order to visualize Type S error, it will be helpful to consider Figure 8.7. This figure is similar to Figure 8.5 except that in this scenario, the statistical test has far less statistical power so more of the non-centrality parameter distribution overlaps the H_0 distribution. Another feature that is unique to Figure 8.7 is that the α-level threshold appears on both ends of the H_0 distribution. This occurs whenever the researcher performs a two-tailed statistical significance test. One typically uses a two-tailed test whenever there is ambiguity about the direction of the expected relationship between the two variables involved in the test. In the context of GWAS, this would represent just about any scenario where the researcher was unsure of whether the minor allele for the SNP would increase or decrease the probability of the phenotype occurring. In other words, unless the researcher had reason to suspect the direction of association between a SNP and the phenotype would be positive/negative, then s/he would most likely conduct a two-tailed test for the direction of association.

Figure 8.7: Type M & S Error



Although the two-tailed test is appropriate, it can prove problematic when statistical power is low because it raises the possibility that one will find a statistically significant association that is in the wrong direction! To see how this is possible, consult Figure 8.7. Notice how a portion of the non-centrality parameter distribution extends beyond the α-level threshold on the left-hand side of the figure? Recall that the H_0 distribution will be centered at zero. Thus,

the left-hand side of the figure represents *negative* associations. But the figure shows that the true association is *positive* because the center of the non-centrality parameter distribution sits to the right (albeit, just barely) of the H_0 distribution. Yet, because statistical power is low, there is a small portion of the non-centrality parameter distribution that extends beyond the α-level threshold on the negative side of the H_0 distribution. In this way, it is possible that one can find a statistically significant association between a SNP and a phenotype but conclude that the association is of the wrong sign (i.e., direction).

Type M and Type S error are errors that one can make in GWAS. The probability that any given statistically significant result represents a Type M or Type S error will depend on the statistical power of the GWAS analysis, which will itself depend on the sample size n and the true effect size that defines the relationship between the SNP and the phenotype. Thus, it is impossible to provide a general conclusion about the possibility that these errors occur in GWAS. Instead, Type M and Type S error must be considered for specific associations of interest, one at a time.

In the end, the point to realize is that under-powered GWAS can be problematic not only because it can miss signals (Type II error) but it can also lead one to overstate the estimated effect size for statistically significant relationships and it can even raise the possibility that one estimates the association to be in the wrong direction.

8.1.6 "Missing" h^2 problem

If you had told a geneticist from the early 2000s that nearly 20 years into the future, folks would still be debating and trying to estimate heritability (h^2) coefficients, they might have justifiably called you crazy. But, as it turns out, that is exactly where we are today. The GWAS technique (and its off-shoots like GCTA) have been used to estimate the degree to which the genetic variants "found" in a GWAS are enough to explain the h^2 estimates gleaned from earlier (or even contemporary) twin studies. Early attempts seemed to show a similar pattern: the h^2 gleaned from GWAS was smaller—often *much* smaller—than the h^2 gleaned from twin studies. In other words, $V_{known} < V_A$. The gap between V_{known} and V_A has come to be known as the "missing h^2" and can be computed simply as:

$$V_{missing} = 1 - V_{explained}$$

As it turns out, the proportion of V_P that remains yet unexplained is substantial for many (if not most) human complex traits (but see Hill and colleagues 2018 for evidence that some of the "missing" h^2 has been found for intelligence). This seemingly leaves researchers with two explanations: 1) the h^2 estimates gleaned from twin studies were wrong (inflated); or 2) the GWAS is simply overlooking a lot of genetic variants that impact the phenotype. But more recent developments suggest four interlocking explanations for the "missing" h^2.

These four explanations were covered in three papers published in the *Proceedings of*

the National Academy of Sciences (Golan et al., 2014; Zuk et al., 2012; Zuk et al., 2014). Briefly, these authors revealed that the "missing" h^2 may be partially attributable to: 1) unaccounted for gene-environment ($G \times E$) and gene-gene ($G \times G$) interactions in GWAS (Zuk et al., 2012); 2) rare variants that are missed by GWAS (Zuk et al., 2014); 3) common variants that are missed by GWAS due to false-negative findings (Golan et al., 2014); and 4) SNPs may not appropriately "tag" all the causal genetic variants.

On the first possibility, it is likely that there are many $G \times E$ that simply go overlooked in behavioral genetics research. Some have even coined this "Plomin's pradox" (Wachs and Plomin, 1991). Specifically, Bakermans-Kranenburg and van Ijzendoorn (2015: 392), when discussing "Plomin's paradox" noted that "...gene-environment interactions may be omnipresent (as the raw material for evolutionary variation and selection) (Rutter, 2006) but appear difficult to find and to replicate."

On the third possibility—that a portion of the "missing" h^2 could come from common variants that are simply false-negative findings in available GWAS, Golan et al. (2014) astutely point out that many variants will not reach genome-wide significance because of low effect sizes or low minor allele frequencies. We know this must be the case because, as Golan et al. (2014: E5274) note, "...the number of loci identified has been continuing to grow with sample size."

Finally, the fourth explanation for "missing" h^2 recognizes the point that modern GWA technologies typically only tag SNPs. While this is—on the surface—not a bad thing, it is important to recall that genetic variation comes in many forms. SNPs are but *one* of several ways that genetic loci can vary in the human body (Ku et al., 2010). There are, for example, copy number variations (CNVs), inversions, and deletions. The extent to which each of these types of genetic variants are captured by GWAS is not fully known. This is not to say that geneticists have ignored these other types of variants or that no one has thought about whether SNPs are appropriate proxies for them, but rather, we still do not know how well common genetic variants like variable number of tandem repeats (VNTRs) will be accounted for by GWAS. Thus, it is possible that a portion of the "missing" h^2 problem can be attributed to the fact that GWAS simply "overlooks" a portion (perhaps only a small portion) of the true causal variants. To demonstrate, we rely on the points made by Ku and colleagues (2014: 412):

> The genetic architecture of complex diseases remains elusive; it is unclear how much each type of genetic variation contributes to inherited risk and the relative proportion of rare versus common variants. If non-SNP genetic variants or rarer SNPs constitute most of the genetic component of complex diseases, then GWAS using the current genotyping arrays would be likely to miss them, simply because they are not covered directly by the genotyping arrays. How much they can be tagged through LD by the markers on the arrays still needs further investigation. Regardless, it is important to continue investigating other genetic variations to discover additional disease associated variants to explain the heritability.

It is now easy to recognize that each of these four explanations likely work in tandem with the others to affect the h^2—and, therefore, the "missing" h^2—for any given phenotype. When focused on rare disease traits, it may turn out that the rare variant concern largely explains the "missing" h^2. When focused on continuous variation in a common outcome (e.g., intelligence), it will likely turn out that the "missing" h^2 will be attributable to overlooked interactions and to overlooked common variants. Nonetheless, the current take on the "missing" h^2 problem is *not* that it implicates either the early twin studies or the current GWAS approach as being wrong. Rather, the discrepancy between the approaches (i.e., the "missing" h^2) reflects some methodological differences that will, in time, be identified and will point to where adjustments can be made to one or the other so that researchers can continue to obtain ever more accurate estimates about the degree to which genetic factors influence human phenotypes.

8.1.7 Population Stratification

We have already discussed population stratification (PS) in a previous chapter (see pages ??). To briefly review, PS is an issue that could confound an association between any given SNP and the phenotype. If the population under study is of mixed genetic ancestry and the phenotype varies across the groups represented, then one is more likely to find a relationship between a SNP and the phenotype purely by chance if certain corrective actions are not taken. The example we used in chapter ?? still makes the point. Imagine, you find a statistically significant SNP emerges from a GWAS but the SNP is known to vary in frequency between the ancestry groups in your sample. In such a case, the focal SNP could have no causal influence—direct or indirect—on the phenotype, but it could still emerge as a statistically significant genetic variant in GWAS if the genetic frequency across ancestry groups was not taken into account.

PS is typically addressed in GWAS by conducting a principal component analysis to capture the group variation observed in the sample. Then, these principal components are included as control variables in the GWAS.

8.1.8 Quality Control & Reference Panels

All GWAS analyses are preceded by a rigorous set of quality control (QC) tests. As any scholar who uses statistical analysis will tell you, the data cleaning and analysis preparation stage of a study almost always takes longer than the actual analysis. And, somewhat disappointingly, these are the very details that are almost always skipped over when it comes to writing up the results for publication. The same is true when it comes to GWAS—indeed, a the QC process can often take weeks or even months. And it typically involves dozens of scholars, with various expertise in both data analytics and genetics. This is one of the reasons that behavior geneticists have begun to embrace co-authorship so warmly—you may have noticed that the list of authors on genetics studies is growing. Indeed, it takes a village

to prepare genome-wide data for analysis.

Because we will not demonstrate a GWAS analysis in this chapter, we will simply direct the reader to sources that cover QC in more detail. What is most important to know is that GWAS data often do not come to the researcher "clean" and ready to use (unless one is analyzing data that have already been subjected to QC, like, for instance, the Add Health genomic data [Highland et al., 2018]). Rather, an extensive process of cleaning and review must take place before the researcher can take genome data that have been collected from a SNP chip and actually perform a GWAS (or GCTA, a polygenic score analysis, or anything else discussed below). The QC process is intended to reduce errors in order to minimize the chances of a false-discovery at the data analysis phase. For more detailed and formal discussion, we recommend Anderson et al. (2010) and Turner et al. (2012).

A second point that is somewhat related to the QC process—in that it cuts to the heart of the issue of whether genome-wide data are valid and reliable for making inferences—concerns the use of reference panels. Reference panels have increasingly become the topic of discussion among the genomics community. To explain what a reference panel is, it might help to begin with an example. To give some context, imagine for moment that you are sent a genome data file—perhaps it's your genome that you've downloaded from 23andMe. The ASCII file you download is filled with As, Ts, Cs, and Gs. Indeed, there are millions of them. What do they all mean? How can you determine where, in your genome, any specific A or T is located? And, even more overwhelmingly, how are you going to figure out which genomic loci vary in the population? In order to find answers to these types of questions, you must rely on a reference panel to tell you what the "typical" genome looks like. Think of it this way: everything is relative. If Every member of a population has a G at a specific loci, then that genetic marker cannot explain variation in some phenotype. But, if 20% of the population has a G and 80% has a T, then perhaps the G/T ratio can explain a portion of the variation in the phenotype of focus.

This should immediately spark questions in your mind like, "where does the reference panel come from?" Reference panels are like big catalogues of previously collected genomic information. The 1000 Genomes Project was one of the first such projects, running from 2008 to 2015. As noted on their website (`http://www.internationalgenome.org/about`), "The goal of the 1000 Genomes Project was to find most genetic variants with frequencies of at least 1% in the populations studied."

So you can think of it this way: genomics researchers collect data from participants they would like to study. In order to make sense of all that genomic information, they compare it to a reference panel (like the 1000 Genomes Project) to determine which loci vary and by how much. Thus, by looking at the reference panel, the researcher could identify that G as the minor allele (i.e., the allele that appears less often in the population) and s/he would also know that the other allele is T. S/he could then use this information to help make sense of any GWAS output that might reveal this specific loci as being correlated with the phenotype.

Although reference panels are a necessary piece for performing genomic analyses, they are not without their own limitations. One primary limitation—which turns out to be more of a practical limitation than a theoretical one—is that they are only informative for the ancestries that are represented in the panel. Put a different way: if a reference panel is like a catalogue, then its usefulness can be judged by its inclusiveness. If the catalogue is limited to one type of product (or, in this case, one ancestral group), then it may not be useful for researchers who wish to study other groups.

This, as it turns out, has become an important issue in genomics research because most major reference panels (i.e., those with the largest number of genomes in the panel) are heavily skewed toward humans of European ancestry. Thus, it can be difficult, if not impossible, to study genomic risk factors for phenotypes among, say, individuals of African ancestry. And, as you can imagine, this makes it challenging to study disparities in things like health outcomes that may have genetic origins. Some scholars have called for this to be addressed, arguing that the lack of diversity in reference panels could itself be considered yet another type of disparity that negatively impacts racial and ethnic minorities (West et al., 2017).

8.2 Extensions to GWAS

8.2.1 Estimating h^2 with Genome-wide Complex Trait Analysis (GCTA)

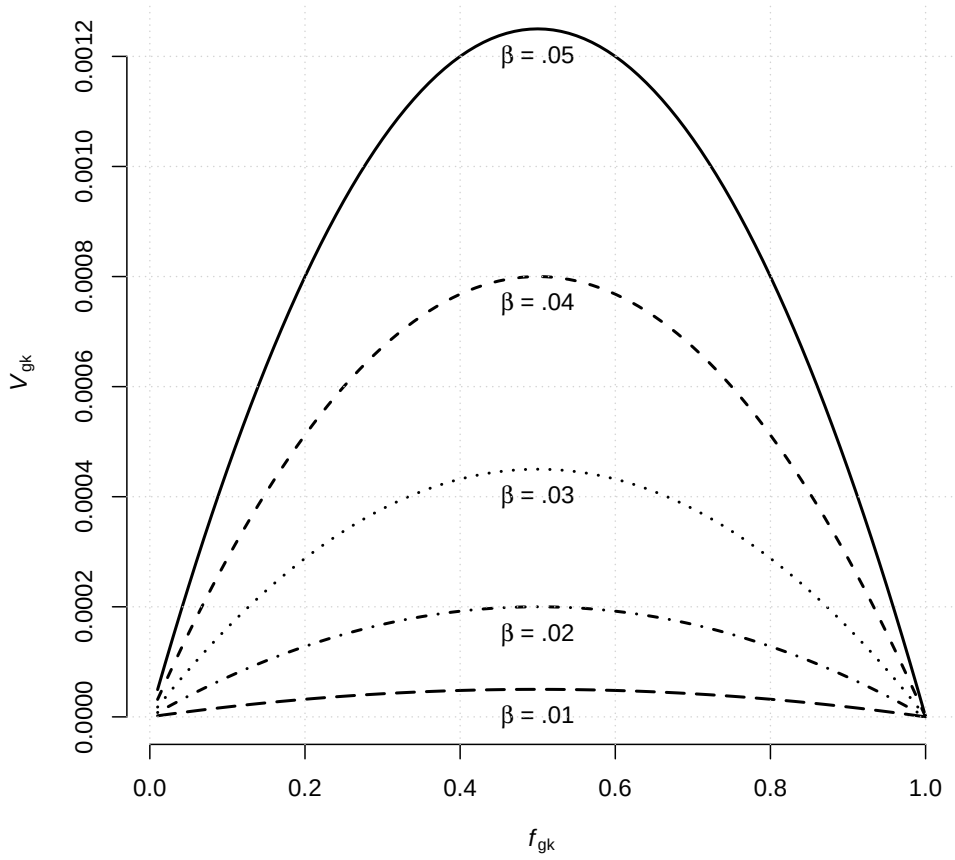
Although GWAS is ostensibly a "gene finding" method, it has also proven useful for estimating h^2. More specifically, h^2 that have already been established using phenotypic correlations in the population (e.g., h^2 estimates gleaned from a twin study [see chapter 5]) can be used as a benchmark to then analyze how much of the h^2 has been accounted for due to the genetic variants identified by GWAS. More directly, how "well" the GWAS did at finding influential genetic variants can be assessed in part by computing the phenotypic variance V_P accounted for by the identified variance and comparing it to the h^2 estimates established in the literature. Assuming additivity (see chapter 3), one can compute the variance V_{gk} accounted for by the kth genetic variant g_k as (see Falconer and Mackay, 1996:126-127; Zuk et al., 2012:1194):

$$V_{gk} = 2f_{gk}(1 - f_{gk})\beta_k^2$$

where f_gk is the frequency of the genetic variant. Recall from chapter 3 that the frequency of a focal allele can be defined as p and $1 - f = q$ (specifically, see page ??). Finally, β_k is the standardized parameter estimate gleaned from the regression of P on g_k. Two points about the variance of P that is attributable to g_k are obvious from this equation: 1) the degree to which g_k explains variance in P (i.e., V_{gk}) is a function of g_ks frequency f_{gk}; and 2) V_{gk} varies as a function of g_ks strength of association with P. These relationships are shown graphically in Figure 8.8.

The point here is that one can estimate the portion of the h^2 that is explained by the set

Figure 8.8: V_{gk} as a Function of f_{gk} and β



of known variants S (i.e., those identified by GWAS) by simply summing over the variance that is attributable to all g_k in S:

$$V_{known} = \sum_{k \in S} V_{gk} = \sum_{k \in S} 2 f_{gk}(1 - f_{gk}) \beta_k^2$$

Once V_{known} is calculated, it is easy to see that a simple ratio of V_{known} to h^2 from a twin study. Recall that the latter—the h^2 from a twin study—will almost always capture narrow-sense h^2, which is equivalent to V_A. Thus, the ratio of focus is almost always:

$$V_{explained} = \frac{V_{known}}{V_A}$$

8.2.2 Polygenic Scores

Imagine you have a rich set of data on a birth cohort that you have followed for 40 years of development. In early stages of data collection, you gathered information on all sorts of environmental exposures, prenatal conditions, parenting practices, and neighborhood factors. As the cohort aged into early childhood, you measured motor development, cognitive ability, and social skills. Once the cohort reached adolescence, you gathered information about the respondents' involvement in delinquency, their involvement in sporting activities, and their educational aspirations. Upon entering adulthood, you measured whether they had been married, how they raised their children, and their occupational prestige. In other words, you have access to an invaluable data source.

But imagine also that this sample is not very large by GWAS standards. For example, imagine that the number of respondents at any given wave of data collection was around $n = 1,000$. As you saw above, statistical power is one of the most prominent concerns for modern GWAS due to the level of P-value corrections that are necessary. What this means for your rich data source is that a GWAS would be unlikely to provide any useful information. Any statistically significant relationship that were to emerge would almost certainly be a severe *over*-estimate of the true effect size (see, generally, **gelman2014beyond**).

Yet, you know that nearly every phenotype of interest in your data is likely to have *both* a genetic component and an environmental component. In other words, if you were to ignore the genetic side of development then any associations reported in studies coming out of these data run the risk of being confounded (see chapter ?? and chapter ??). So, what do you do?

One option is to carry out a GWAS with your data, ignoring the warnings from statisticians and behavioral geneticists. We, of course, do not advise that you follow this route. But, we also would not advise you to ignore the genetic side of the equation. As an alternative, we would suggest you consider using polygenic scores as a way to account for the genetic influence on phenotypic development. The idea for polygenic scores has been around for quite some time but it was not successfully used to predict a complex trait until the late 2000s (Evans et al., 2009; Purcell et al., 2009), but they have only recently been developed out of GWAS data. The idea, though, is quite simple. An example may be the best way to explain.

Imagine you want to study the development of, say, aggression over the life course. Again, you note that your sample of $n = 1,000$ is too small to carry out a GWAS. But you *can* calculate polygenic scores based on previously conducted GWAS. As it so happens, there are a few GWAS analyses that have utilized a measure of aggression (or a related variant of antisocial behavior) that could be used to build a polygenic score in your data. For instance, Pappa and colleagues (2016) conducted a GWAS on childhood aggression and reported that a SNP on chromosome 2 that was nearly genome-wide significant ($P = 5.30 \times 10^{-8}$). In addition to this SNP, these authors also provided their summary data showing the estimated relationship between *all* 2.5 *million* SNPs that were analyzed. Relying on this summary data, you could construct a polygenic score by relying on the Pappa et al. (2016) data as a

reference panel.

In order to actually capitalize on this information, you would first need to genotype all of the participants in your sample using the latest SNP chip technology. Once you had the genotype information, it would be as simple as matching up the SNPs observed in your data to the SNP relationships estimated in the Pappa et al. (2016) data.

For instance, let us take the SNP on chromosome 2 that was nearly genome-wide significant from the Pappa et al. (2016) study. The SNP was rs11126630 and the two alleles observed were T and C. The results of the analysis were actually quite clear: it appears the T allele lowered the risk of aggressive behavior. The estimated effect size was approximately -0.03. You can think of this effect size as a "weight" that can be applied to a simple linear prediction equation. If you wanted to predict someone's outcome on P and you had data on all the SNPs that had been studied and their respective estimate association with P, then you could calculate a polygenic risk score for P for each individual based on their unique combination of SNPs.

This was precisely the strategy that Daniel Belsky and his colleagues(2016) followed for their study of the SNPs associated with educational attainment. The phenotype educational attainment has become one of the most studied complex traits in the GWAS community. It is not that GWAS researchers *really* want to know which genes affect educational attainment, but rather it is an issue of convenience. Because GWAS requires such huge sample sizes it is often necessary for researchers to combine their data to have any hope of identifying SNPs that are genome-wide significant. In order to combine data, though, it is necessary to have collected information that is comparable in different datasets. Well, as it turns out, one of the most frequently asked questions is: "how far did you go in school?" Because school completion data is available in just about every data source that collects individual-level measures, it is possible to combine various different surveys together and begin to estimate GWAS on sample sizes large enough to provide meaningful estimates. Just such a GWAS was estimated in 2013 on 100,000 individuals (Rietveld et al., 2013)!

Capitalizing on the opportunity, Belsky et al. (2016) relied on the summary statistics from the Rietveld study and created a polygenic score for educational attainment. In order to create the polygenic score, the authors "...counted the number of education-associated alleles (0, 1, or 2) and multiplied this count by the effect size estimate in the original GWAS [Rietveld et al., 2013]" (Belsky et al., 2016: 958-59).

Think of it this way, the polygenic score is simply:

$$\text{polygenic score}_i = \beta_1(SNP_{1i}) + \beta_2(SNP_{2i}) + \cdots + \beta_k(SNP_{ki})$$

$$\text{polygenic score}_i = \sum_{j=1}^{J} \beta_j(SNP_{ji})$$

where β_1 represents the estimated effect size of SNP_1 on the phenotype; β_2 represents the estimated effect size of SNP_2 on the phenotype; β_k represents the estimated effect size of SNP_k on the phenotype; SNP_{1i} represents the observed number of "outcome-associated

alleles" person i has for the first SNP of focus; SNP_{2i} represents the observed number of "outcome-associated alleles" person i has for the second SNP of focus; and SNP_{ki} represents the observed number of "outcome-associated alleles" person i has for the k^{th} SNP of focus.

Imagine, for simplicity, that there were only three SNPs analyzed in a GWAS and that their respective effect sizes were $\beta_1 = 0.03$, $\beta_2 = 0.01$, and $\beta_3 = 0.015$. Imagine also that a randomly drawn case from your data had the following observed SNP counts: $\text{SNP}_1 = 2$, $\text{SNP}_2 = 1$ and $\text{SNP}_3 = 0$. You would then use these effect sizes in combination with the observed SNP counts to calculate a polygenic score as:

$$\text{polygenic score} = 0.03(2) + 0.01(1) + 0.015(0)$$

where the polygenic score would obviously sum to 0.07. But rather than doing this for only three SNPs for one person, modern polygenic scoring does this for *millions* of SNPs for *every* member of a data set.

Returning to the Belsky et al. (2016) analysis, these authors used the polygenic score for educational attainment to predict a range of outcomes including socioeconomic indicators, social mobility, and reading aptitude relative to peers. The findings showed that individuals with higher polygenic scores—which is an indicator of higher levels of educational attainment—were positively associated with socioeconomic status, social mobility, and reading ability at an early age.

8.2.3 Linkage Disequilibrium Score Regression

One of the newest and most exciting techniques is known as linkage disequilibrium score regression (LD score regression). The LD score regression method was introduced by Bulik-Sullivan and colleagues (**bulik2015ld bulik2015atlas**). We direct readers interested in learning the technique and using it with their own data to the LD Hub webpage hosted at the Broad Institute (`http://ldsc.broadinstitute.org/ldhub/`). Users can upload their own data to the LD Hub database and receive estimates from the LD score regression analysis within seconds. Readers interested in estimating their own LD score regression models are encouraged to visit Bulik-Sullivan's GitHub page here: `https://github.com/bulik/ldsc`. Much like the rest of this chapter, we seek to introduce the concepts and the logic of the LD score regression method here.

The LD score regression technique capitalizes on the concept of LD itself. Specifically, as we noted above, LD occurs when two or more genetic variants tend to be inherited as a package more often than would be expected by chance. When this occurs, we can the effect sizes of SNPs in LD with one another and combine them (weighting them according to their observed level of LD) to generate an estimate of the degree to which the phenotype is influenced by genetic variants. In other words, similar to the goal of GCTA, LD score regression can take the SNPs observed in GWAS and use their individual effect sizes to calculate an estimate of the degree to which those SNPs explain the total observed variation in the phenotype.

But calculating SNP heritability is only one of the features users get from the LD score regression technique. There are three other important features worth noting here. First, the LD score regression method allows the user to calculate SNP heritability for more than one phenotype at a time. Given the ability to observe more than one phenotype at a time, the user is also afforded the opportunity to estimate the genetic correlation r_g between those two phenotypes. This is an extremely important feature because it opens up many new opportunities for researchers to study the developmental etiology of phenotypes in combination with one another.

Second, the LD score regression method has been shown to provide a more convenient and perhaps more appropriate way for controlling certain confounding influences that emerge in GWAS. More details are provided in **bulik2015ld**

Third, and perhaps most importantly, the LD score regression method does not require the user to analyze original GWAS data. On the contrary, the user is able to estimate the LD score regression method is estimated on the *summary* data from GWAS output. This is an incredible development because GWAS summary data are typically publicly available and open to download freely on the Internet. Two databases are, in fact, already up and running. The first is the LD Hub database we referenced earlier: `http://ldsc.broadinstitute.org/ldhub/`. Users can find summary data for—as of the writing of this text—nearly 220 unique phenotypes! The second database housing GWAS summary statistics (there is some overlap between the two web sites) is hosted by the Psychiatric Genomics Consortium at the University of North Carolina, Chapel Hill: `http://www.med.unc.edu/pgc/results-and-downloads`.

8.3 Conclusion

Genome-wide data is fast becoming available to social scientists. The implications of such data and the findings that stem from it have only recently begun to be discussed and, thus, we anticipate that the moral, ethical, and philosophical implications of genome-wide research will be a source of debate for years to come. In anticipation of such debate, we close this chapter with a brief consideration of perhaps the most frequently raised concern: what happens if genome data can be used to predict human behavior?

Although we believe it is highly unlikely that genome-wide information will ever provide enough predictive power to be considered a reliable source of prospective prediction, we do recognize that genome-wide information provides researchers access to the G portion of our $P = G + E$ equation. Because research has repeatedly shown that G accounts for nearly half of the observed variation in many human traits (Polderman et al., 2015), we might begin our discussion by considering that genome-wide data *potentially* gives researchers a way to explain a considerable proportion of human individual differences. But the genome is unlikely to be the only predictor for traits of interest to social scientists. Indeed, recall that E explains the other half (in most cases) and that G and E are often intertwined in

complicated relationships (interactions and correlations, which will be covered in the next chapter). Thus, simply having a printout of your genome is unlikely to tell you much about the way your life will turn out.

Genomic risks, as we have mentioned several times, are *probabilistic* risks. This means that carrying an A where most other folks have a C might confer a slightly higher risk of developing some outcome. But it does not guarantee it.

Also relevant is Turkheimer's (1998) discussion of *weak biologism*, which can be characterized as the recognition that human behaviors will, for the most part, all rely on the same biological pathways. This allows for the recognition that most human outcomes will result from a combination of genetic and environmental influences. But it also complicates the study of genetic influences on those behaviors because it reveals that any given outcome is unlikely to have its own unique developmental mechanisms. For Turkheimer (1998), weak biologism was a way to signal that genetic factors do not work deterministically for complex outcomes like those of interest to social scientists. We echo this point and encourage social scientists to integrate genomic data into their work, but we encourage caution when doing so—the genome will not reveal deterministic pathways to human behavior, but by combining that data with information about our social worlds may help us better understand ourselves. This provides a nice segue to the next chapter, which will consider the ways in which our genes combine with our environments to impact phenotypic development.

Chapter 9

Genes & Environments I: Gene-environment Interplay

Testing for gene-environment interplay has become one of the fastest growing areas of modern behavioral genetics. Indeed, since the turn of the 21st century, there have been thousands of studies testing all manner of gene-environment interplay including ($G \times E$) and gene-environment correlation (rGE). With that many studies, it is easy to imagine that the landscape of analytic strategies has also expanded considerably. And indeed this has been the case. There are countless ways a researcher could analyze data "looking" for a $G \times E$ or an rGE (or both). Thus, our goal here is not to introduce you to *all* of the approaches one might take. Rather, our goal is to introduce you to these concepts and to consider the most common techniques that have been applied when testing for gene-environment interplay.

Before we move on, it is important to anchor this chapter in our central equations for the phenotypic score P and the variance in the phenotype V_P that have formed the backbone of the text since they were first introduced in chapter ??. Recall that we can represent the phenotypic score P with:

$$P_i = G_i + E_i$$

But we have noted throughout this text that this equation "hides" many forms of gene-environment (and even gene-gene or environment-environment) interplay. That is to say, up to this point we have assumed that $G \times E$ do not play a role in the etiology of P. This chapter will discuss ways that one can relax these assumptions by directly estimating certain parameters that are thought to capture gene-environment interplay. This is no trivial exercise. Indeed, recent evidence strongly supports the idea that $G \times E$ underlie much of the phenotypic variance in human complex traits. For example, Zuk and colleagues (2012) recently showed that much of the "missing" h^2 problem *could* be explained by the fact that GWAS overlooks $G \times E$s. Additionally, Del Giudice (2016) recently showed that the biometrical models described in chapters 5 and 6 may "hide" $G \times E$s that are consistent with theories of human plasticity (see differential susceptibility theory [Belsky, 1997; 2005] and biological sensitivity to context theory [Boyce and Ellis, 2005; Ellis and Boyce, 2008]).

Doing so requires that we alter the above equation to account for arbitrary $G \times E$ by forming a more generalized version of the phenotypic score equation (see Golan et al., 2014; Zuk et al., 2012):

$$P_i = \Psi(G_i, E_i)$$

where Ψ represents any arbitrary function between the genetic components G and the environmental components E for person i. This might include—but is not limited to—gene-gene interactions ($G \times G$), gene-environment interactions ($G \times E$), and gene-environment correlations (rGE). Because we have discussed $G \times G$ in previous chapters (e.g., see our discussion of epistasis from chapter 3), we will omit it from the present discussion. Thus, this chapter will focus on $G \times E$ and rGE. This chapter will consider some of the ways that a researcher who is interested in studying $G \times E$ may go about estimating the degree to which a gene moderates the influence of an environmental input. As you will see in the section that follows, estimating $G \times E$ relies on many of the same techniques and concerns that arose in the candidate gene chapter (chapter??). But modeling $G \times E$ requires several additional considerations be kept in mind. This means there are many permutations of issues that may arise in any given analysis of a $G \times E$. For this reason, we will forgo a simulation and demonstration of $G \times E$ for this chapter. Our primary reason for doing so is that we do not want to send the impression that there is only one approach to analyzing $G \times E$, which is undoubtedly what would happen if we were to demonstrate a simple $G \times E$.

The second part of this chapter will consider how genes might correlate with environmental inputs in systematic ways (i.e., rGE). As we will discuss in that section, studying rGE—like studying $G \times E$—can take many forms and, for that reason, we will omit a simulation/demonstration of rGE. For instance, a researcher could draw on a candidate gene dataset to model rGE by estimating the effect of the candidate gene on the environmental factor in question. Or, s/he could analyze twin data to estimate the heritability (h^2) of an environmental variable (Kendler and Baker, 2007). At first it may seem counterintuitive to say that one can estimate the h^2 of an environmental source of variance. But hopefully the theoretical processes will become more clear as we move through that discussion. Readers who feel they need a more thorough introduction to the theoretical underpinnings of rGE are encouraged to see our discussion in chapter 3. Also, we direct interested readers to the seminal work of Sandra Scarr and Kathleen McCartney (1983) for detailed discussion.

9.1 Gene-environment Interaction ($G \times E$)

The very first paragraph of James Tabery's (2014) book, *Beyond Versus: The Struggle to Understand the Interaction of Nature and Nurture*, tells the story of this chapter:

> We have moved beyond *versus*. Whether it is medical traits like clinical depression, behavioral traits like criminality, or cognitive traits like intelligence, it is now widely recognized that "nature versus nurture" does not apply. There are no genes for depression such that having the gene ensures the development

> of depression and lacking the gene ensures resilience to depression. Likewise, there are no environments for depression such that all differences in depression can be explained by pointing to those differences in environment. Rather, it is a truism that these complex human traits arise from both nature and nurture, and differences in those traits arise from both differences in nature and differences in nurture. Since it is both, the challenge of explaining the development of a trait like depression or accounting for differences in depression must involve understanding the interaction of nature and nurture.

In a similar vein, VanderWeele (2015: 249) notes: "It is not uncommon for the effect of one exposure [read: "gene"] on an outcome to depend in some way on the presence or absence of another exposure [read: "environment"]". In other words, when complex human outcomes are the focus—which is the case for this text—it is anticipated that the causes will interact during phenotypic development. Even our intuitive understanding of causality has a built in recognition that multicausality—which is a term Rothman and Greenland (2005) use to indicate an outcome of focus is part of a complicated system that has more than one cause—implies that causes do not act independently of one another. This to say that a basic understanding of causality recognizes that the influence of component A on P might depend on the presence of component B. And this, as it turns out, is precisely what behavioral geneticists are referring to when they talk of gene-environment interaction ($G \times E$).

So what does it mean to say two things "interact" in the etiology of a phenotype P? The short answer is "it could mean a few things". Colloquially, people often speak of two things "interacting" when they appear at the same time and in the same space. Thus, two individuals might "interact" at a social gathering. This is *not* what is meant when a behavioral geneticists says $G \times E$. On the contrary, a behavioral geneticist who is seeking to understand when/if/how genes and environments interact is interested in testing whether the presence/absence of one thing alters the impact of a second thing on phenotypic development. Put more directly and in the context of genes and environments: $G \times E$ tells us whether the presence/absence of an environment (gene) affects the impact of a gene (environment) on the phenotype. With this in mind, Tabery (2014: 135) defines an interaction as "*...the interdependence of actual difference makers in the causal mechanisms responsible for a phenomenon*" (emphasis in original).

A simple thought experiment might help clarify. Let us take an example popularized by the epidemiologists Kenneth Rothman and Sander Greenland (2005: s145):

> Suppose that someone experiences a traumatic injury to the head that leads to a permanent disturbance in equilibrium. Many years later, the faulty equilibrium plays a causal role in a fall that occurs while the person is walking on an icy path. The fall results in a broken hip.

In this case, one might say that the head injury that resulted in "faulty equilibrium" *interacted* with the icy path to produce the outcome of the fall, which resulted in a broken hip.

These two components are said to interact because the effect of one—say, the icy path—on the outcome was directly impacted by the presence of the other—say, the faulty equilibrium. In this example, both were necessary to result in the broken hip. The mere presence of one without the other would not have resulted in the same outcome.

The broken hip example is, of course, an oversimplified version of a much more complex reality. For starters, we know that broken hips can happen for any number of reasons, not just because of faulty equilibrium and icy paths. But that is not the point. Rather, the point of focus here is that human outcomes almost always are best described as resulting from a multicausal system. That simply means there is more than one factor that leads to most human outcomes. When this occurs, it is no longer appropriate to look for necessary or sufficient causes. Instead, causes are thought to be part of a bigger causal mechanism where the impact of cause A (e.g., faulty equilibrium) can be heightened when cause B (e.g., an icy path) is present.

Perhaps the most important point that is buried in this discussion is that two causes that interact in the production of an outcome may have *both* independent and interaction effects. In other words, it is possible that cause A contributes to some non-zero portion of the causal mechanism and so too does cause B. But it may be the case that neither alone is sufficient to explain the outcome. Also, it may be the case that a simple combination of both causes is not sufficient to explain the outcome. Continuing with the example above, a broken hip may not result if the icy path precedes the faulty equilibrium. But, when *both* causes A and B are present at the same time, the probability of the outcome is increased substantially.

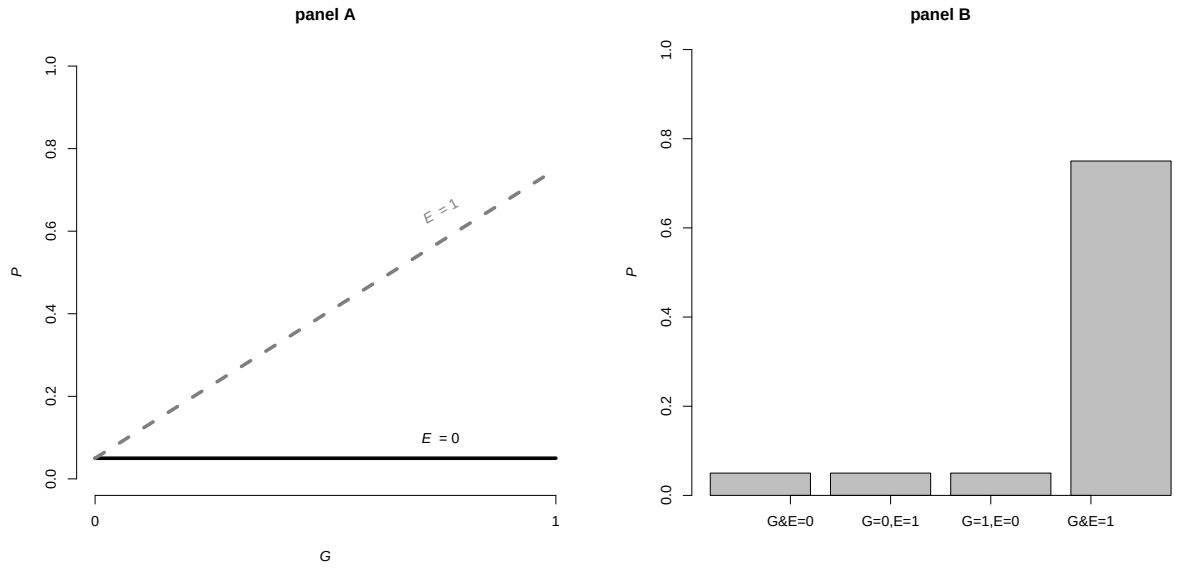
Rather than referring to causes A and B, let us assume we are conducting a candidate gene study (see chapter 7) and we wish to explore whether there is a $G \times E$ between our focal candidate gene and some environmental exposure E. The focus here on candidate gene research reveals that all the assumptions and limitations that prevail in that methodological realm will carry-over to the present context. That is to say, testing for $G \times E$ with a candidate gene (i.e., c$G \times E$) does not allow one to circumvent any known limitations of the larger candidate gene approach. We will return to a discussion of the assumptions and limitations of c$G \times E$ research at the end of this portion of the chapter.[1]

Graphical depiction of c$G \times E$ interactions may facilitate a better understanding. There are at least two ways to visualize an interaction between two causes in the etiology of an outcome. Let us return to thinking about the causes as being a gene G and an environment E in the etiology of a phenotype P. It will make this discussion more tractable if you consider just *one* G and *one* E. Though, this should not be taken to mean that one must study the development of a phenotype P one gene/environment at a time. Indeed, we will return to this latter point when we discuss the assumptions and limitations of $G \times E$ analysis.

[1]It should be noted that one is not restricted in any theoretical/conceptual sense to analyzing $G \times E$ in candidate gene studies. To be sure, it is entirely possible to test for $G \times E$ in the context of GWAS. Yet, to date, the vast majority of the $G \times E$ literature has been conducted with candidate genes (c$G \times E$). Thus, we focus our attention on that body of literature.

Two ways of visualizing a $G \times E$ in the etiology of P are displayed in Figure 9.1. Although these two plots may appear to be showing different things, in fact, they are displaying the exact same information and both reveal signs of a $G \times E$. Let us take a moment to walk through the two panels and reveal how they are just different ways of presenting the same information.

Figure 9.1: Two Ways of Visualizing a $G \times E$



First look at the panel on the left (i.e., panel A). In this figure, we see that G is a binary trait: it is either absent (=0) or it is present (=1). Similarly, E is treated as a binary exposure variable. In panel A, we are looking to see if the effect of G on P is moderated by the presence of E. We see evidence that this is the case because the impact of G on P differs according to the presence of E. In fact, there is no effect of G on P when $E = 0$ (i.e., when E is absent); the predicted phenotypic score P is 0.05 regardless of the presence/absence of G. This is reflected by the flat bold line. But there *is* an effect of G on P when $E = 1$ (i.e., when it is present). When $G = 0$ and $E = 1$, we see that the predicted score for $P = 0.05$. But when *both* G and E are present (i.e., both =1), we see that the predicted phenotypic score is $P = 0.75$.

The exact same information is displayed in the bar chart presented in panel B. As can be seen, the predicted outcome for $P = 0.05$ for the first three categories, which respectively index situations where both G and $E = 0$, when $G = 0$ and $E = 1$, and when $G = 1$ and $E = 0$. Only when *both* G and E =1 does our prediction change.

Because all three variables P, G, and E in this heuristic example are binary, we can take the predicted proportions of P from the graphical displays and put them into a table as shown below:

	$E = 0$	$E = 1$
$G = 0$	0.05	0.05
$G = 1$	0.05	0.75

Using the convention set forth by VanderWeele (2015, chapter 9), let p_{ge} reflect the predicted proportion of P when $G = g$ and $E = e$ such that:

	$E = 0$	$E = 1$
$G = 0$	p_{00}	p_{01}
$G = 1$	p_{10}	p_{11}

From here we can calculate a numerical estimate of the degree to which there is $G \times E$ by looking for *additive* interaction. Specifically, VanderWeele (2015) showed that an additive interaction could be measured by calculating[2]:

$$\begin{aligned} \text{additive interaction} &= (p_{11} - p_{00}) - [(p_{10} - p_{00}) + (p_{01} - p_{00})] \\ &= p_{11} - p_{10} - p_{01} + p_{00} \end{aligned}$$

where p_{11} is the predicted proportion of cases that present with the phenotype P when both G and E are present (i.e., $G = 1, E = 1$); p_{10} is the predicted proportion when $G = 1, E = 0$; and so on.

Note how the first equation calculates the difference between p_{11} and p_{00}, and then the difference between this value and the independent effects of G and E are calculated. Thus, one can interpret an additive interaction as the "...extent to which the effect of the two factors together exceeds the effect of each considered individually" (VanderWeele, 2015: 250).

Working through the heuristic data presented earlier, we can calculate the additive interaction effect as $0.75 - 0.05 - 0.05 + 0.05 = 0.70$. Because the resulting value is greater than zero, we can interpret the results as evidence of a *positive* additive interaction. When the additive interaction value is below zero, we can interpret the results as evidence of a *negative* additive interaction. Such a case would arise when the probability (i.e., the risk) of the outcome was reduced for cases with $G = 1$ and $E = 1$ compared to the other scenarios. For example: $0.25 - 0.60 - 0.60 + 0.70 = -0.25$.

9.1.1 Theoretical Models for $G \times E$

At this point, it is important to pause and consider the *substantive* interpretation of $G \times E$. In a broad sense, there are three models for considering how $G \times E$ may come about in the

[2]Many of the mathematical details presented in this chapter are drawn from VanderWeele's (2015) discussion of interaction analysis in chapter 9. Thus, readers who desire more detailed information may wish to consult his text.

etiology of human outcomes: 1) the diathesis-stress model (Shanahan and Hofer, 2005); 2) the social push model (Raine, 2002); and 3) the differential susceptibility model (Belsky, 1997; 2005). Let us consider each of them in turn.

Diathesis-stress

The diathesis-stress model hypothesizes that $G \times E$s work in a "for worse" fashion. In other words, the diathesis-stress model hypothesizes that someone with genotype A may be resilient to environmental stressors, but someone with genotype B will respond negatively (i.e., "for the worse"). In neutral or positive environments, the two genotypes are not expected to produce different outcomes. It is only under conditions where the environment is negative that we see a difference between the genotypes.

Shanahan and Hofer (2005: 66) put it like this: "That is, the phenotype is triggered when the contextual feature and specific genotype are combined. In the strong variant of the triggering interaction, neither the genotype nor the context influences the likelihood of the phenotype: Their combination is necessary. In the weak variant, the genotype and/or context have an additive effect on the likelihood or intensity of the phenotype, but their combination significantly, additionally influences the manifestation of the phenotype." And Boardman, Daw, and Freese (2013: s67) describe diathesis-stress this way: "The diathesis-stress model suggests that the genetic differences that are associated with negative outcomes in risky environments will have either an attenuated or entirely muted relationship in low-risk environments." As can be seen, the diathesis-stress model is characterized by an increased risk of a negative outcome when *both* environmental and genetic risk are present.

One of the best ways to gain traction with the different models is to consider what sort of results they would produce if the "real-world" operated according to the tenets of the explanation. We can draw on the data from the toy example introduced earlier for this discussion. Recall that we set up the scenario like this:

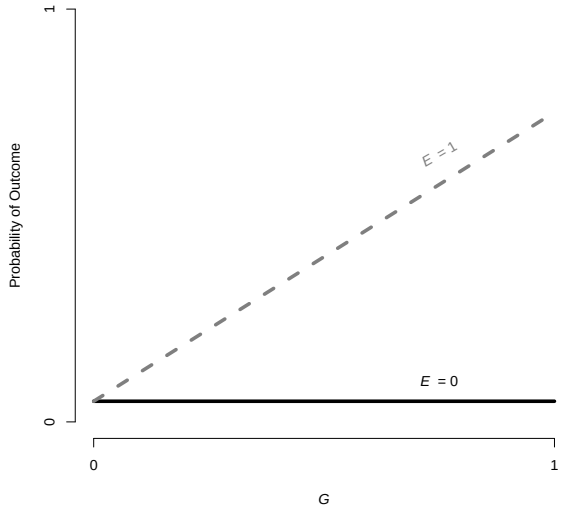
	$E = 0$	$E = 1$
$G = 0$	p_{00}	p_{01}
$G = 1$	p_{10}	p_{11}

Let us imagine that the environmental factor E represents the absence (coded 0) or presence (coded 1) of some risk factor. Perhaps it represents exposure to extreme poverty. Or it might represent the presence of an environmental toxin like lead or even an environmental source of bacteria. Whatever E is, let us assume that it's presence is—on balance—considered a risk factor for the development of an undesirable phenotype.

According to the diathesis-stress model, the genetic influence G (the "diathesis") should only emerge under conditions where the environmental risk factor E (the "stress") is also present. In other words, evidence consistent with the diathesis-stress model emerges when

p_{11} is larger than p_{00}, p_{01}, and p_{10}. A visual representation might look like the figure we saw earlier, which is reproduced in Figure 9.2. Notice how the effect of the environmental risk

Figure 9.2: The Diathesis-Stress Model



factor is only observable when the genetic "risk" is also present. This is the hallmark of the diathesis-stress model.

At this point, it is important to point out that $G \times E$ can be graphically presented in one of two ways. They can be presented 1) with the genotype on the x-axis or 2) with the environment on the x-axis. You might be wondering, what the difference is. Mathematically, there is no difference. One can (typically) pull all the same information out of a figure regardless of which way the data are represented. Typically, though, researchers will choose one of these two approaches based on which element—G or E—has been conceptualized as the causal risk factor and which has been conceptualized as the moderator. If the genetic influence G has been framed as the causal risk and the environment E is the moderator—the factor that impacts how G affects the outcome—then the researcher will most likely choose to place G on the x-axis as we did above. In the reverse scenario—where E is cast as the causal risk factor and G is the moderator, then the researcher may choose to place E on the x-axis.

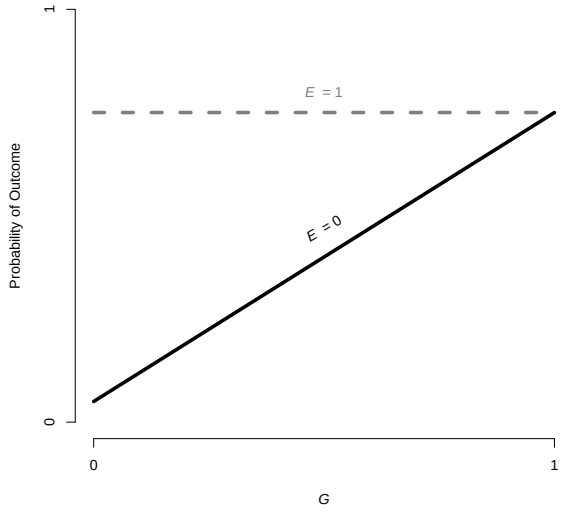
Social Push

The social push model differs from the diathesis-stress model by arguing that the impact of genotype may not emerge in "risky" environments, but instead will emerge when the environment is stable and/or benign. Adrian Raine (2002: 314) argued that, "A number of studies have found that psychophysiological factors show stronger relationships to antisocial behavior in those from benign social backgrounds that lack the classic psychosocial risk

factors for crime." In other words, it may be that genetic influences are "drowned out" when environmental stressors are present. This could occur if the environmental influence(s) were powerful enough to move all members of a group (or society) in one direction or another. Thus, when faced with environmental stressors that have large effects, we should not expect to see individual differences in the outcome of focus. Rather, we should expect most members of society to develop the phenotype under study. It is only when environmental conditions are benign—or, at least, not extreme—that we should expect individual differences to emerge.

Considering an extreme example might help to clarify. Imagine a society that approves—or even encourages—child abuse. In that society, we might not expect individual differences at the genetic level to explain individual differences in, say, depression. It is likely that many members of the group will develop depression at some point in their lifetime. It is only under conditions where the environment is more benign that we should expect to see genetic influences. A graphical depiction might look like Figure 9.3

Figure 9.3: The Social Push Model



Differential Susceptibility

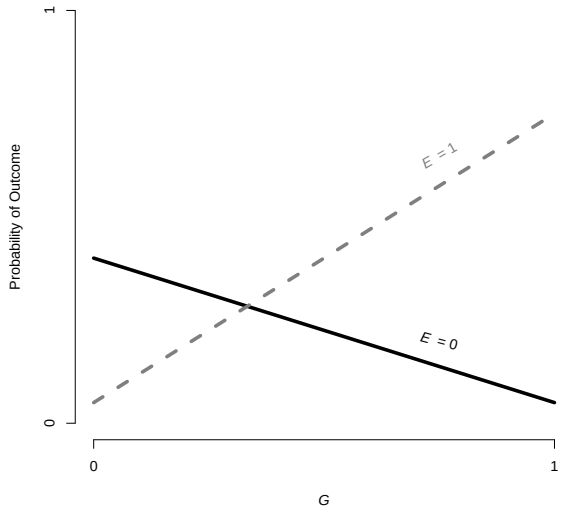
The alternative to the diathesis-stress and social push models is broadly known as the differential susceptibility model (see, generally, Del Giudice, 2016). Within this broad focus are two related but somewhat different theories known as differential susceptibility theory (Belsky, 1997; 2005) and biological sensitivity to context theory (Boyce and Ellis, 2005; Ellis and Boyce, 2008), as well as the integrated form that includes elements of both (Ellis, Boyce, Belsky, Bakermans-Kranenburg, and van Ijzendoorn, 2011). Briefly, the differential susceptibility approach to $G \times E$ allows for a change in the rank of the genotypes in terms

of their expected values of P as the environment changes. This approach allows for the possibility that individuals with a certain genotype may respond in a "for better and for worse" pattern to the environment. In this sense, the genotypes that underlie this susceptibility to the environment are thought to be "plasticity" genes such that they allow the individual to be adaptive to the environment. For others, their genotypes allow them to be resilient to the environmental influences, meaning they are less plastic.

Boardman and colleagues (2013: s67) noted that, "...the differential susceptibility hypothesis implies that alleles associated with negative outcomes in adverse environments may be associated with positive outcomes in the most salutary environments. ." A popular metaphor involving orchids and dandelions is often used to make sense of the differential susceptibility model. Briefly, the orchid is a very sensitive flower (indeed, one author's wife cannot keep them alive!) that responds well to nourishing environments and it withers in adverse conditions. The dandelion, however, is quite resilient and can seem to grow under even the most caustic conditions (they appear every year in that same author's backyard despite regular lawn treatments!). This metaphor—like any great comparative attempt—breaks down when one realizes there are certain strains of orchids that grow in high altitude, mountain climates (see Bakermans-Kranenburg and van Ijzendoorn, 2015). Nonetheless, the parallels are quite obvious and the broader points are easy to grasp.

Graphical representation of a $G \times E$ under the differential susceptibility model will typically reveals a cross-over effect, where the genetic influence works to make individuals less likely—compared to those who do not have the genetic "risk"—to develop the phenotype when the environment is positive (or benign) but more likely to develop the phenotype when the environment is negative. This can be seen in Figure 9.4

Figure 9.4: The Differential Susceptibility Model



A classic case of potential differential susceptibility can be found in the work of Caspi et al. (2002). These researchers found that a certain allele of the *MAOA* gene was associated with increased rates of antisocial behavior when environmental risk was present. But that same allele was associated with *lower* risk of antisocial behavior when environmental risk was absent.

These theories are well situated in broader evolutionary principles and, indeed, make sense when one considers the ways in which natural selection must work on human variation. The eminent scholar Michael Rutter (2006: 192) noted "In short, genetic variation in response to the environment is the raw material for natural selection." Rutter's point serves to remind us that any hypothesized relationship that involves genes and environments must make sense in a broader, evolutionary sense. At least for $G \times E$—and differential susceptibility—it appears that requirement has been met.

Important for the present focus is that several tests for differential susceptibility have been developed. Unfortunately, though, to the best of the our knowledge, none of these tests offer definitive evidence that can be used to declare the presence (or absence) of differential susceptibility compared to some other model such as diathesis-stress. For this reason, we do not focus much attention on these methods and, instead, direct the interested reader to the source material. As discussed by Bakermans-Kranenburg and van Ijzendoorn (2015: 390-391), there are methods that assess the shape of the interaction effect in an attempt to empirically distinguish between cases of differential susceptibility and cases of diathesis-stress. Those methods can be found in Kochanska et al. (2011) and Roisman et al. (2012). Others have developed model-fitting approaches that require the researcher to center his/her data on the interaction term and then different forms are tested against models that would be expected under diathesis-stress and differential susceptibility. This approach can be found in Widaman et al. (2012) (and see Belsky et al., 2013).

9.1.2 Empirically Detecting $G \times E$

Let us now turn our attention to more specific methodological/analytical issues. We will consider how one might analyze a dataset looking for a $G \times E$. Typically, behavioral geneticists interested in studying $G \times E$s often do so using regression models. Our focus here, therefore, will be exclusively on regression-based techniques for estimating the impact of a $G \times E$ on P.

Given that most $G \times E$s are analyzed using regression-based tactics, it is important to consider how the regression model used will impact the results that are gleaned. As epidemiologists have noted (ROTHMAN TEXTBOOK), researchers can differentiate between additive interactions and multiplicative interactions. And in many cases, the researcher can attain an additive interaction simply by estimating one type of model and a multiplicative interaction by estimating a different type of model. We will now consider both.

Additive $G \times E$

Detecting additive interactions is relatively straightforward *if one estimates a linear probability model* where P is the outcome and G, E, and $G \times E$ (i.e., the multiplied value of G and E) are predictors:

$$\mathbb{E}(P_i = 1|G = g_i, E = e_i) = \beta_0 + \beta_G(g_i) + \beta_E(e_i) + \beta_{G\times E}(g_i \times e_i)$$

Here, it can easily be shown that $\beta_0 = p_{00}$; $\beta_G = p_{10} - p_{00}$; $\beta_E = p_{01} - p_{00}$; and $\beta_{G\times E} = p_{11} - p_{10} - p_{01} + p_{00}$. That is, under the linear probability model—which is simply an OLS model where the outcome is a binary variable (see, Long, 1997)—$\beta_{G\times E}$ provides an estimate of additive interaction. The corresponding standard error can be used to construct a 95% confidence interval and/or to perform a hypothesis test. If one is able to verify that all of the normal confounding assumptions have been met (more on this in the Assumptions and Limitations portion of this chapter), then it is possible to speak to the causal effects of the $G \times E$ on the development of P.

As noted, though, these properties only hold when one estimates the $G \times E$ with a linear model. Put a different way, when an OLS model is estimated, the coefficient for the $G \times E$ (i.e., $\beta_{G\times E}$) provides an estimate of the additive interaction. But you may be thinking, "isn't it more appropriate to estimate a $G \times E$ in a logit model when P is binary?" For a host of different reasons, the answer is often "yes." The logit model is more appropriate because it relaxes some of the assumptions of the OLS model when the outcome of focus is a binary variable (see, generally, Long, 1997; Wooldridge, 2014). But, as we will reveal in the next section, the coefficient for the $G \times E$ in a logit model does *not* mean the same thing as it does from an OLS model. To be specific, estimating a $G \times E$ with logit leads to something known as a *multiplicative* interaction.

Multiplicative $G \times E$

If, rather than estimating the linear probability model, one were to estimate a logit model, the result would be:

$$\log\Big\{\frac{\mathbb{P}(P_i = 1|G = g_i, E = e_i)}{[1 - \mathbb{P}(P_i = 1|G = g_i, E = e_i)]}\Big\} = \pi_0 + \pi_G(g_i) + \pi_E(e_i) + \pi_{G\times E}(g_i \times e_i)$$

It is well-known that exponentiating both sides of the logit equation results in a substantively interpretable value for the right-hand side estimates. Specifically, the parameter estimates (i.e., the πs) become odds ratios (OR) when they are exponentiated. For example, e^{π_G} reveals the degree to which the odds that $P = 1$ change with a one-unit increase in G. More formally:

$$OR_G = \Big\{\frac{\mathbb{P}(P = 1|G = 1, E = 0)/[1 - \mathbb{P}(P = 1|G = 1, E = 0)]}{\mathbb{P}(P = 1|G = 0, E = 0)/[1 - \mathbb{P}(P = 1|G = 0, E = 0)]}\Big\}$$

which, using the notation from above is:

$$OR_G = \frac{p_{10}/(1 - p_{10})}{p_{00}/(1 - p_{00})}$$

Similarly, OR_E can be defined as:

$$OR_E = \left\{ \frac{\mathbb{P}(P=1|G=0,E=1)/[1-\mathbb{P}(P=1|G=0,E=1)]}{\mathbb{P}(P=1|G=0,E=0)/[1-\mathbb{P}(P=1|G=0,E=0)]} \right\}$$

$$= \frac{p_{01}/(1-p_{01})}{p_{00}/(1-p_{00})}$$

But, $OR_{G\times E}$ provides an estimate of something known as the *multiplicative interaction*, which is:

$$OR_{G\times E} = \frac{\frac{p_{11}/(1-p_{11})}{p_{00}/(1-p_{00})}}{\left[\frac{p_{10}/(1-p_{10})}{p_{00}/(1-p_{00})}\right]\left[\frac{p_{01}/(1-p_{01})}{p_{00}/(1-p_{00})}\right]}$$

$$= \frac{OR_{GE}}{(OR_G OR_E)}$$

This value—the multiplicative interaction—reveals the extent to which the presence of both G and E exceeds the product of G and E taken separately. Just as with the additive interaction, the multiplicative interaction can be positive or negative. But, because it is calculated on an odds ratio scale, values greater than 1 are positive (i.e., $OR_{G\times E} > 1$ are positive) and values below 1 are negative (i.e., $OR_{G\times E} < 1$ are negative).

Let us consider the data that were presented earlier, where $p_{00} = p_{01} = p_{10} = 0.05$ and $p_{11} = 0.75$. In this case:

$$\begin{aligned} OR_{G\times E} &= \frac{\frac{0.75/(1-0.75)}{0.05/(1-0.05)}}{\left[\frac{0.05/(1-0.05)}{0.05/(1-0.05)}\right]\left[\frac{0.05/(1-0.05)}{0.05/(1-0.05)}\right]} \\ &= \frac{\frac{3}{0.053}}{1} \\ &= 57 \end{aligned}$$

For these heuristic data, we find an extremely large multiplicative interaction. This is not surprising given the huge discrepancy that we built into this example.

We chose these values because they provide an example of a situation where both the additive and the multiplicative interactions point us toward the same substantive conclusion: the impact of G *and* E exceeds the impact of G absent E and vice versa. Put a slightly different way, in this heuristic example, we see that the probability the phenotype $P = 1$ is heightened considerably when G and E are both present compared to scenarios where only G (or E) is present.

Additive $G \times E$, Multiplicative $G \times E$, & $RERI_{OR}$

At this point, you are probably wondering whether you should present additive interactions or multiplicative interactions in your own research. If you are like us, you may have even been surprised to learn that there are different types of interaction. We are encouraged by the fact that VanderWeele (2015) gives such an accessible overview, but we are also concerned that many researchers remain unaware of the differences between the two.

Perhaps most troubling, is a point that we have not yet addressed. To be direct, we will now reveal that in some (if not *many*) "real world" circumstances, the substantive conclusions drawn from a $G \times E$ analysis will be contingent on the form of interaction that is chosen. Put a different way: it can be shown that in many common scenarios, an additive interaction will be positive (negative) but the multiplicative interaction will be negative (positive).[3]

Indeed, Greenland and colleagues (2008) have shown that if the two exposures G and E have an effect on P and there is no additive interaction, then the multiplicative interaction will be present. The same holds for the reverse scenario: if G and E have an effect on P, then the absence of a multiplicative interaction means that an additive interaction is present. The logic here is that if G and E have an impact on P, then there *must* be interaction on some scale (VanderWeele, 2015: 252).

This issue reveals the importance of defining the *scale* (i.e., additive or multiplicative) of the $G \times E$ of focus. Many scholars present $G \times E$ without acknowledging the scale to which their findings speak. This also raises an all important question: which scale should a researcher use? Unfortunately, there is no unilateral answer to this question, but we can offer certain recommendations. First, in keeping with VanderWeele (2015), we recommend that researchers present *both* additive and multiplicative findings when testing for $G \times E$. Second, it is our position that the *additive* $G \times E$ is the scale that is often (though, not always) desired. Thus, if one must choose between additive or multiplicative $G \times E$, we recommend the former.

Recall the additive scale is automatically assessed if one estimates a linear probability model (i.e., if OLS is estimated). But, this is not the case for the logit model. Rather, the latter automatically estimates the multiplicative scale. This, then, raises a second question. How is one to generate an estimate of the additive $G \times E$ if, say, a logit model is estimated due to the fact that P is a binary outcome? Researchers in this situation are encouraged to estimate a quantity known as the "relative excess risk due to interaction" ($RERI_{OR}$; Rothman, 1986; also referred to as the interaction contrast ratio by Greenland et al., 2008; and see, generally, VanderWeele, 2015: 254-257). The $RERI_{OR}$ can be calculated using odds ratios as an *approximation* of the additive interaction. $RERI_{OR}$ is appropriate when the outcome is rare (10% or fewer of the cases have a 1 on the outcome) because the ORs will approximate risk ratios (RR) from a log-linear model. If the outcome is not rare, then the

[3] A notable exception is with cases of "qualitative" [VanderWeele, 2015: 279] or what Tabery [2014] refers to as interaction where the groups change rank. In short, when there is a cross-over effect, *both* additive and multiplicative interaction will arrive at the same substantive conclusion.

*OR*s will depart from *RR* and, thus, $RERI_{OR}$ will no longer provide a close approximation of an additive $G \times E$.

Researchers who estimate a logit model—again, assuming the outcome is rare—can derive an estimate of the additive interaction from the *OR*s by calculating:

$$RERI_{OR} = \frac{p_{11}/(1-p_{11})}{p_{00}/(1-p_{00})} - \frac{p_{10}/(1-p_{10})}{p_{00}/(1-p_{00})} - \frac{p_{01}/(1-p_{01})}{p_{00}/(1-p_{00})} + 1$$

$$= OR_{GE} - OR_G - OR_E + 1$$

Note that OR_{GE} is *not* the odds ratio for $\pi_{G\times E}$. Rather, it is as defined above (i.e., the ratio of the odds for cases where both G and E are present compared to the odds for cases where both G and E are absent). Thus, with this in mind, estimating $RERI_{OR}$ from a logit model (see above) involves carrying out the following calculation:

$$RERI_{OR} = e^{\pi_G + \pi_E + \pi_{G\times E}} - e^{\pi_G} - e^{\pi_E} + 1$$

Attributing Effects to G or E

$G \times E$ raises the question of how we can define the effects of G (or E) on Y. We could simply focus on the "main effect" of G, but that would ignore the part of G that is wrapped up with E in the $G \times E$. As VanderWeele (2015: 284) notes, "In some sense, then, we can attribute the total effect of E on Y to the part that would be present still if G were 0 ... as well as to a part that has to do with the interaction between G and E." This statement reveals that, if one were to simply focus on the main effect term for G (or E), then a—perhaps substantial—portion of the effect that is actually attributable to that factor may go overlooked. This, of course, raises an important question: how can we decompose the various influences such that we can then say how much of P is due to G (or E)?

VanderWeele (2015: 281) shows that—when we estimate an additive interaction—we can decompose the total effect of G *and* E on P as follows:

$$p_{11} - p_{00} = (p_{10} - p_{00}) + (p_{01} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})$$

Notice that the left-hand side term reveals the total impact of G and E (including $G \times E$) on P by taking the baseline proportion of cases where $P = 1$ that would be expected when $G = 0$ and $E = 0$ (i.e., p_{00}) and subtracting it from the proportion of cases where $P = 1$ when $G = 1$ and $E = 1$ (i.e., p_{11}). Using our heuristic data from above, we would arrive at an estimate of the total impact of G and E as: $0.75 - 0.05 = 0.70$. So, if we know the total impact of G and E collectively, we could then use it to calculate the proportion of that total

effect that is attributable to G alone, E alone, and $G \times E$ like so:

$$\text{portion due to } G \text{ alone} = \frac{p_{10} - p_{00}}{p_{11} - p_{00}}$$

$$\text{portion due to } E \text{ alone} = \frac{p_{01} - p_{00}}{p_{11} - p_{00}}$$

$$\text{portion due to } G \times E = \frac{p_{11} - p_{10} - p_{01} + p_{00}}{p_{11} - p_{00}}$$

If a multiplicative interaction is estimated using the following logit model:

$$\log\Big\{\frac{\mathbb{P}(P_i = 1|G = g_i, E = e_i)}{[1 - \mathbb{P}(P_i = 1|G = g_i, E = e_i)]}\Big\} = \pi_0 + \pi_G(g_i) + \pi_E(e_i) + \pi_{G\times E}(g_i \times e_i)$$

then, the decompositions become:

$$\text{portion due to } G \text{ alone} \approx \frac{e^{\pi_G} - 1}{e^{\pi_G + \pi_E + \pi_{G\times E}} - 1}$$

$$\text{portion due to } E \text{ alone} \approx \frac{e^{\pi_E} - 1}{e^{\pi_G + \pi_E + \pi_{G\times E}} - 1}$$

$$\text{portion due to } G \times E \approx \frac{e^{\pi_G + \pi_E + \pi_{G\times E}} - e^{\pi_G} - e^{\pi_E} + 1}{e^{\pi_G + \pi_E + \pi_{G\times E}} - 1}$$

9.1.3 Sources of Bias

Like with any analytical approach, it will be critical to allow domain knowledge and theory to drive any and all tests for $G \times E$. Critical thinking will always trump "brute force" estimation techniques when it comes to identifying the complex ways that genetic and environmental influences synergistically combine to affect human behavior.

Additionally, when searching for $G \times E$ with candidate genes (i.e., c$G \times E$), it may not be safe to assume that the gene of focus is the only genetic influence on the phenotype. Recall the fourth law of behavior genetics (Chabris et al., 2015) points out that most human complex traits are affected by many genetic variants. This will be important to keep in mind when estimating a c$G \times E$ because if the candidate gene is in linkage disequilibrium with another genetic variant, then it may bias our results. Similarly, one must assume no epistasis when searching for c$G \times E$. If these assumptions are untenable, then the results produced by any empirical analysis may be biased.

With these points in mind, we refer the reader to the assumptions and limitations section of chapter ?? for more detail about the potential pitfalls of candidate gene research and,

by extension, the potential pitfalls of c$G \times E$. In addition to the assumptions and limitations discussed there, we have three additional topics that require careful attention here: 1) the issue of statistical power; 2) the importance of controlling for all interactions; and 3) unmeasured confounding of the environment.

Low Statistical Power

Without going into too much detail, suffice it to say that the levels of statistical power for detecting $G \times E$ is often considerably lower than it is for detecting the "main" effects of G and/or E individually. Recall from chapter ?? that statistical power is a function of the actual effect size that prevails in reality (the very thing we are trying to estimate), the sample size of the study trying to estimate the effect size, and the α level that has been chosen by the researcher.

Statistical power ranges between 0.00 and 1.00, and, for all intents and purposes, researchers (should) seek to maximize statistical power in any given study. When statistical power is low (e.g., let us use the arbitrary cut-off of 0.80 as the line between "acceptable" [at/above 0.80] and "low" [below 0.80] power), all sorts of problems can arise. As Ioannidis (2005) revealed, low levels of statistical power can reduce the probability that a researcher will detect an effect, it will reduce the probability that a statistically significant effect reflects a true effect, and it increases the probability that a researcher will suffer from the "winner's curse." The idea of the winner's curse comes from modern day auctions. In an auction, a particular item or asset is sold to the highest bidder. When a particular item is sold after a bidding war, it is quite likely that the new owner has over-paid. In other words, the "winner" of the auction is the one who was most willing to over pay. The same sort of problem can arise when a researcher performs a study (i.e., enters an auction house) with relatively low levels of statistical power. In order to be able to observe a statistically significant effect and claim the study was "successful", the researcher will need to observe an effect size that is large enough to push the test statistic into the region of statistical significance. This might only occur when the effect is over-stated, leading to what Gelman and Carlin (2014) coined "Type M" error.

Returning to our focus on $G \times E$, it turns out that—as a general rule—there will be less statistical power to detect a $G \times E$ than there will be to detect "main" effects of G and E independently. Now let us connect the dots. The fourth law of behavior genetics (Chabris et al. 2015) tells us that the influence of any particular genetic influence G is likely to be small. This means that statistical power to detect a $G \rightarrow Y$ (i.e., a "main") effect is likely to be small. Unless an effort is made to increase sample sizes to a level that will allow for adequate statistical power (which might necessitate sample sizes in the tens or hundreds of thousands for many genetic effects), then it is unlikely that any given study will have adequate statistical power to detect a likely $G \times E$ effect.

All of this means that, as a consumer of $G \times E$ research, you should be cautious and skeptical. As with any limitation of a statistical method, this concern does not mean that

all studies are wrong. Rather, it means that the probability that any given study will be wrong is something that can be computed if we have an idea about how much statistical power that study has. It appears that there is much room for improvement, at least as far as candidate $G \times E$ research goes (Duncan and Keller, 2011). As we noted in the chapter on GWAS, statistical power has been at the heart of the development of full-genome studies. So it may be that the concerns we have raised here will subside over time and as large-scale consortia gather more and more data to study human phenotypes.

Including All Interactions

Keller (2014) drew attention to an often overlooked complication with $G \times E$ research. His concern was that most $G \times E$ studies did not adequately control for confounding by simply including covariates as right-hand side predictors. To be more explicit, Keller (2014) revealed that the typical strategy for controlling away the influence of a vector of covariates C was to include them in the regression model along with G, E, and $G \times E$ like so:

$$\mathbb{E}(P_i = 1|G = g_i, E = e_i, C = c_i) = \beta_0 + \beta_G(g_i) + \beta_E(e_i) + \beta_{G\times E}(g_i \times e_i) + \beta_C(c_i)$$

The problem is that simply controlling for C does not fully account for the possibility that C confounds the $G \times E$. Perhaps the easiest way to understand Keller's (2014) argument is to consider that C could represent the actual cause of P and, for the sake of argument, let us assume C also interacts with G to predict P. Thus, the $G \times E$ is actually between G and C, not between G and E. If this were the case and the above regression model were estimated, the researcher might erroneously conclude that the $G \times E$ was a causal factor for P when in fact the causal interaction was actually $G \times C$. To avoid this problem, the researcher must properly control for C and all the potential interaction effects it may take:

$$\begin{aligned}\mathbb{E}(P_i = 1|G = g_i, E = e_i, C = c_i) = \beta_0 + \beta_G(g_i) + \beta_E(e_i) + \beta_{G\times E}(g_i \times e_i) \\ + \beta_C(c_i) + \beta_{G\times C}(g_i \times c_i) + \beta_{C\times E}(c_i \times e_i)\end{aligned}$$

Unmeasured Confounding of the Environment

Unmeasured confounding is always a concern for behavioral researchers because in most cases it is impossible to control for all possible sources of confounding. Luckily, if one is willing to assume the genetic variant is not in linkage disequilibrium, then unmeasured confounding of the genetic variant is not a problem.

So let us assume the researcher has reason to believe there is no confounding of the genetic effect. This only leaves the environmental influence that can be confounded. So what happens when a researcher finds a $G \times E$ but s/he cannot rule out all sources of confounding for the environmental variable E?

It turns out that under certain conditions, estimates of $G \times E$ will not be biased even if there is unmeasured confounding of E. VanderWeele (2015) explains that this will be the

case whenever a) the unmeasured confounder U is a confounder for E but not G; b) E and G are statistically independent after control is made for covariates C; and c) G does not interact with U. When these conditions are met, the $G \times E$ estimate will not be biased even if there is unmeasured confounding for the effect of E on Y.

This may, at first, seem counterintuitive. How could the effect of E on Y be confounded but the $G \times E$ not be? One way to think of it is that the $G \times E$ actually reflects an interaction between the G and *some* E, it just may be that the environment driving the interaction is U instead of E (i.e., the $G \times E$ could actually be a $G \times U$). In either case—assuming U is itself an environmental factor—the researcher will have found evidence of a $G \times E$.

9.1.4 Conclusion

It is critically important to note that the techniques discussed in this section can *potentially* identify and reveal where $G \times E$ take place. What *cannot* be revealed, though, is the degree to which a $G \times E$ identifies an actual functional biological pathway/mechanism. In short, as VanderWeele (2015: 318) notes, "...we have no way of going from any of these forms of interaction which we can assess with data directly to the underlying biology itself." Moffitt and Capsi (XXXX) have also cautioned researchers to only search for $G \times E$ when the putative biological mechanism is plausible and—at least to some extent—already understood. And, because we have leaned heavily on James Tabery's (2014: 157) account throughout this chapter, it is fitting to include his input on the matter.

> The survey of the empirical evidence for interaction above attests to the fact that the reality is much more complex. Many studies of interaction—be it heredity-environment interaction in nature, genotype-environment interaction in humans, or gene-environment interaction in humans—have taken place over the last century. Some of these studies turned up evidence for interaction. Other studies came back negative. And quite a few of the studies turned up evidence for interaction in one trait and, in the very same population, lack of evidence for interaction in another trait. My answer to the evidential question is thus: *it's a mixed bag, and we should not assume one way or the other whether interaction exists for any particular trait or any particular gene-environment relationship.* Ask yourself: why should interaction exist for lint yield but not fiber strength in cotton? Why should interaction exist for body weight but not fleece weight in sheep? Why should interaction exist for alcohol consumption but not smoking in esophageal cancer? I see no way to justify assuming ahead of time why interaction will turn up in any given case. There is plenty of empirical evidence for it in some cases, and there is plenty of empirical evidence against it in other cases. And so when it comes to any particular case of interaction, that case simply needs to be judged on its own empirical merits.

The important point embedded in this discussion is that human quantitative/behavioral

genetics is complex. It does *not* lend itself to simple explanations. Nothing is cut-and-dry. Instead, modern genetics research has revealed the importance of understanding the nuance of human complex traits. For the most part, *there is no gene for X*, where X is a human trait/behavior/outcome. It turns out that single-gene disorders are the exception, not the rule. This makes sense when one considers the simple mathematics of the issue. With only 20,000 genes, how could there be a gene for every minor piece of human life? The numbers do not add up. But note that this is not the same as saying that genes do not affect X. In other words, acknowledging the complexity of human life does not discount the possibility that genetic influences play a role. On the contrary, it simply reveals something scientists have known for a long time. Human outcomes are complex and they likely interact with the environment (i.e., non-genetic properties) in the development of phenotypic phenomena. Thus, for most human phenotypes, $G \times E$ are a non-ignorable part of the equation.

Recognizing this complexity also reveals a potential avenue toward more clarity. Specifically, realizing the importance of $G \times E$ could eventually lead to treatments and interventions that are tailored to one's genotype. This may sound like science fiction at first, but there is already budding discussion of these possibilities in many areas of social science. For example, Gajos et al. (2016) recently sparked a discussion of this possibility in criminology, where the focus was on increasing the efficacy of interventions with criminal offenders. There are, of course, many ethical and philosophical issues that must be worked out. We do not pretend to have the answers to those questions, but we are certain that discussion surrounding these points are sure to take center stage in the near future.

On this point, Bakermans-Kranenburg and van Ijzendoorn (2015) reported meta-analytic results that appear promising for those interested in tailoring interventions (in a broad sense, not necessarily in the criminological context) according to genotype. Specifically, they reviewed the findings from randomized experiments that explored $G \times E$ and reported evidence suggest ing that carriers of "susceptibility" genes experienced stronger intervention effects compared to those who did not carry those genes. In short, their meta-analytic results provide evidence for the proof of concept that $G \times E$ could someday be used to target intervention efforts much like specialized medicine is targeted to medical patients. $G \times E$ might even afford scientists an opportunity to reduce some of the most resilient and long-standing inequalities in social outcomes.

9.2 Gene-environment correlation (rGE)

We close this chapter with a brief discussion of gene-environment correlation (rGE). At first blush, it may seem odd that we only devote a section of a chapter to the expansive issue of rGE. This choice will, hopefully, make more sense when you consider that analyzing data in search of rGEs really does not necessitate anything unique in the sense of methodological or analytical strategy. The only element of the research process that needs to be considered unique to the study of rGE is the way the researcher conceptualizes the outcome under examination. To this point we have assumed—or explicitly noted—that the outcome is a

human complex trait such as a behavioral outcome. The study of rGEs requires a shift in this focus such that the outcome is treated as an environmental exposure. Thus, once can consider just about any methodological technique presented throughout this entire text as an appropriate strategy for studying rGE.

Perhaps the only unique piece of information necessary to study rGE is the theoretical distinctions between the three "types" of rGE. The three types of rGE were covered in a previous chapter (chapter 4, section 4.2), so we will refer the reader to that discussion for more detail. For now, it will suffice to simply remind the reader that the different forms of rGE include passive rGE, active rGE, and evocative rGE. Passive rGE recognizes that parents do not simply pass along their environment to their children, they also pass along their genes. Because parents pass along their genes and their environments to their children, researchers have hypothesized that children's genotypes will be correlated with their childhood environments even if there is no causal link between the two. This is, essentially, the idea behind passive rGE. Of course, though, applied researchers know that one will rarely find a situation where the childhood environment is completely spuriously related to the child's developmental outcomes. Instead, reality is such that a portion of any association between gene and environment might be partially attributable to passive rGE and partially attributable to other things, a causal relationship being but one of those possible "other things."

The next type of rGE is known as active rGE. The important piece of information to keep in mind here is that individuals do not randomly sort themselves into environments. Rather, humans are highly selective of their environments. Humans are quite adept at choosing environments that match their personality profiles so that they can maximize their potential output. For example, individuals with athletic propensities often select into environments that challenge their physical abilities. Individuals with higher levels of intelligence are more likely to select into higher education (though, of course, that is not the only reason humans go to college).

The third type of rGE is known as evocative rGE. This version of rGE recognizes the possibility that individuals' genetic propensities may evoke certain responses from their environment. Friendly people often "evoke" more friendly responses from strangers. Musically gifted individuals often "evoke" requests to perform and intellectually gifted individuals tend to "evoke" requests to continue their studies (e.g., through scholarship or grant offers). The point to each of these types of rGE is that individuals are not randomly distributed across environments. We shape and react to our environments based—at least partially—on our genetic propensities. We pursue the things we are good at either directly through our own choices or indirectly through the encouragement at the requests of others. Our childhood environments, too, are linked to our genetic profiles because of the selection and evocation of environmental responses that our parents experienced.

Recognizing these points leads to two broad ways that rGE might be of interest to researchers. First, one can conceptualize the environment as the outcome of study, such that

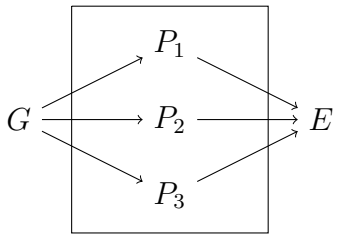
our analytic model becomes:

$$E_i = G_i + \epsilon_i$$

where ϵ_i represents an individual-level error term that we will treat as random and, from this point forward, will ignore. Notice that the above equation sets the environmental exposure as the outcome under study. This means researchers can study the sources of variation that might lead person A to an environmental exposure that differs from person B. The three types of rGE outlined above provide the theoretical backdrop necessary to see how this might come about.

A researcher interested in estimating the genetic influences on an environmental exposure has many tools at his/her disposal. To be sure, it is possible to use the biometrical models described in chapters 5 and 6, to use the candidate gene approach described in chapter 7, or to use the genome-wide approaches described in chapter 8. In other words, for the researcher studying rGE, the analytic approaches do not change. Only the outcome(s) under examination—and the theoretical explanations of how the genetic influences affects the outcome—differ from what we have described throughout this text.

On the latter point, it is important to recognize that the genetic influence on any environmental exposure is unlikely to be direct. Instead, it is almost always going to be the case that the genetic influence of G on E will work through some individual-level phenotype (P) such as a personality feature. It is appropriate, therefore, to think of rGE as a mediation model where $G \rightarrow P \rightarrow E$. If a researcher were to estimate the impact of G on E with, say, an ACE model (chapter 5), it would likely be implied that the parameter estimate represented the total effect (i.e., all the indirect effects) that worked through all the potential mediators (phenotypes, P) that could possible come in between G and E. One might represent these points graphically like so:



where all of the parameters that are inside the box will be collapsed into the single point estimate of the influence of G on E (e.g., as the heritability estimate that would be gleaned from a biometrical model).

The second reason rGE may be of interest is that they can complicate the study of any given environmental source of variation (E) as a potential cause of variation in a phenotype (P). This is the primary concern to which we want to draw your attention in the chapter 10. Specifically, imagine you were interested in estimating the causal effect of E on P and you carried out a regression-based analysis:

$$P_i = \beta_0 + \beta_E(E_i) + \epsilon_i$$

This equation makes the—somewhat subtle—assumption that the influence of E on P can be estimated freely without considering the effect of G. That assumption, of course, will only hold if *both* $G \times E$ *and* rGE are zero. Put differently, if there is any sort of gene-environment interplay involved in causing variation in P, then the above equation will produce a biased estimate of β_E. The only way to produce an unbiased estimate of β_E is to explicitly model or control away the gene-environment interplay that is involved in producing variation in P. We dealt with the ways to model $G \times E$ in the previous section of this chapter. So one could include explicit controls for any $G \times E$ that was thought to play a role in producing variation in P. Strategies for modeling rGE—as we have noted repeatedly in this section—have been presented throughout this text.

One concern, however, quickly arises. What happens if the extent of gene-environment interplay is too complicated to reliably model in a regression-based framework? For example, what happens if there are too many $G \times E$s to account for. Same thing for rGE, what if there are too many possible ways that rGE might confound the relationship between E and P to be reliably modeled? When this occurs, and it is our position that it does in fact occur regularly, the researcher is limited in his/her options. Limited, but not left completely stranded. We will cover some of the most powerful designs and analytic strategies available to deal with these very concerns in the next chapter.

Chapter 10

Genes & Environments II: Modeling the Effect(s) of the Environment

10.1 Conceptual Overview

Up to this point in the text, we have focused almost exclusively on the G part of our central equation: $P = \Psi(G, E)$. This chapter will depart from that focus and instead will home in on the E portion of the equation. Specifically, our central question for this chapter is "what is E?" One might be inclined to answer that E is already well understood. For example, we already know much about the development of many phenotypes of interest to readers of this text. And most of our information about that development comes from research that can be thought of as being exclusively focused on E. Take, for instance, sociological research. It is probably fair to characterize sociology as a discipline focused on the socio-environmental inputs to human behavior. Thus, sociologists study E and how it affects P.

But this simple characterization overlooks a key issue. Most behavioral research that does not come out of behavioral genetics attempts to estimate the relationship between E and P while ignoring the influence of G. This may not be problematic if G is zero *or* if E is independent of G. Put into the context of our central equation, scholars are safe to study how $E \rightarrow P$ as long as a) the $G \rightarrow P$ relationship is zero or b) the correlation between G and E (i.e., rGE) is zero and there is no interaction between G and E (i.e., there is no $G \times E$). If *both* a or b is true, then the researcher interested in studying E is safe to proceed without considering the role of G. But, if either a and b is not true, then the researcher must attempt to model the role of (e.g., control for) G in order to arrive at an unbiased estimate of the influence of E on P.

This point can be thought of as a simple discussion of the role of a confounder variable, where G plays the role of the confounder. When there is a possibility that G operates as a confounder variable, the researcher will have several strategies s/he can rely upon to arrive

at an unbiased estimate of the relationship between E and P. Those researcher strategies are the focus of the present chapter. Thus, placed into the context of our central equation, this chapter provides methods that one can use to understand the role of E in the development of P, controlling for G.

10.1.1 The Discordant Twin Design

One of the first—and easiest to grasp—modeling strategies to develop has been referred to broadly as the "between-twins" study design. The logic of the design is simple to grasp: if one has reason to suspect that G might confound the estimate of E on P, then the simplest way to rule out the confounding influence of G is to perform an analysis that can control away the influence of G. This is possible if one has access to a sample of monozygotic (MZ) twins. Recall from chapter 5 that MZ twins have—for the most part—identical genotypes. Thus, if one were to study MZ twins who were discordant for some environmental exposure E, then the researcher could estimate the impact of E on P by studying between twin differences among twin pairs.

As you might imagine, this design is quite popular among researchers who study disease traits. It lends itself very well to a case-control type of design (GREENLAND ??), where the researcher identifies "cases" (e.g., those who present with the trait) and then simply matches them to their MZ twin who does not present with the disease.

In many ways, this is one of the most conservative and cleanest approaches to estimate the counterfactual condition. Recall that the potential outcomes/counterfactual tradition is a leading perspective to estimating causal effects. The basic idea goes like this: if you wanted to identify the causal effect of A on Y, then the best thing to do would be to simultaneously assign a person to receive $A = 0$ *and* $A = 1$. Then, sometime later, compare the person's outcome on Y. Did the person present with the phenotype (i.e., $Y = 1$) when $A = 1$ but not when $A = 0$? If so, then one could say that A causes Y.

The obvious problem is that one can never assign someone to *simultaneously* be in two conditions (i.e., $A = 0$ and $A = 1$) at one time. A person can only occupy one state (e.g., $A = 0$), leaving the other state an unobserved and thus unknown (e.g., $A = 1$). We call the outcome that would have been observed had the unobserved state actually occurred the counterfactual. Since we cannot directly observe counterfactual states, we are resigned to try and estimate this as best we can. Hence the popularity of the case-control design with MZ twins. Since MZ twins share their genotype, it is often argued that if twin A presents with the phenotype, then the best estimate of the counterfactual condition is twin B who does not present with the phenotype.

Note that there are at least two ways to approach this type of study. One could identify twins who have the phenotype in question and, therefore, they could attempt to identify the cause(s). In other words, one could look for the causes of the effect. The other approach would be to find MZ twins who present with an environmental exposure thought to be the

cause of a phenotype and, then, they could look forward to see if the phenotype in question actually develops. This might be referred to as searching for the effects of a cause(s). There are, of course, benefits and disadvantages to both approaches. Those need not concern us here, but readers who are interested may find guidance in Greenland et al. (TEXTBOOK ??) and Morgan and Winship (2014??).

As with the preceding chapters of this text, we are primarily interested in the modeling strategy that would be necessary to actually generate estimates from this type of study. So let us imagine we have a dataset like the one shown in the table below.

Pair ID	Pair Type	phenotype_1	phenotype_2	environment_1	environment_2
1	MZ	0	1	0	1
2	MZ	1	1	1	1
3	MZ	1	0	1	0
⋮					
100	MZ	0	0	0	0

With these heuristic data it is easy to see that the environmental exposure would emerge as a strong (in fact, the correlation would be 1.00 based on the data we've provided) predictor of the phenotype. This can be seen because every time the environmental risk is present (i.e., environment = 1) the phenotype is also present (i.e., phenotype = 1).

But note that we glean this information by looking at the within twin pair discordance for both the environment and the phenotype. Every time there is a discordance between the twins on the environmental risk, there is also the same pattern of discordance for the phenotype. But how does one glean an effect estimate from data like these? The simplest way to do so would be to generate difference scores for each twin pair and then to estimate the impact of the environmental difference score on the phenotypic difference score like so:

$$\text{phenotype}_1 - \text{phenotype}_2 = \beta_0 + \beta_E(\text{environment}_1 - \text{environment}_2)$$

This strategy would yield a consistent estimate of the influence of the environment on the phenotype. Perhaps most importantly, this strategy would have controlled away the influence of any genetic influences because the concordance between MZ twins in the same pair is always 1.00. Thus, perfect concordance—which would qualify as a constant if we were to include it in the statistical model–would not be able to explain *dis*cordance.

Although this strategy is useful and remains popular, there is a more general approach to study twin differences that can instead be implemented and, in some ways, is superior to the simple twin difference method described here. That approach is known as the fixed effects regression model. We will first explain the general fixed effects model and then will describe its application in the context of twin data in the next section.

10.1.2 The Fixed Effects Model

The fixed effects model is a general strategy developed to help rule out stable influences on a variable across units. Those units might be time—as is often the case when fixed effects estimation is carried out in disciplines like economics—or they might be families as is the case for the present focus. Imagine we have data like we saw in the table above. Those data would often be referred to as "wide" format, meaning we have the data arranged so that each *pair* is the row and the specific values for each twin appear as separate variables (e.g., phenotype$_1$ and phenotype$_2$). We could easily "reshape" those data to be "long" formatted, such that each *twin* appeared as the row and now the pair identifier was included as a separate variable. Reshaping from wide to long format is easily handled by most modern statistical packages, so we will not spend time demonstrating this procedure. But for those who want a quick reference using `R`, we recommend the following discussion: `https://stats.idre.ucla.edu/r/faq/how-can-i-reshape-my-data-in-r/`.

Let's reshape the data we presented earlier. Doing so results in:

Pair ID	Pair Type	Twin ID	phenotype	environment
1	MZ	1	0	0
1	MZ	2	1	1
2	MZ	1	1	1
2	MZ	2	1	1
3	MZ	1	1	1
3	MZ	2	0	0
⋮				
100	MZ	1	0	0
100	MZ	2	0	0

Notice that all the same information that was presented earlier also appears in this table. The only difference is the structure (i.e., "shape") of the table. Most important is to notice that a new variable—the twin ID—now appears. This will be important because we will use it as a covariate to help capture any random differences that might appear in the data between twin 1 and twin 2. Next, notice that there is only one variable for the phenotype and one variable for the environment. These variables now capture the phenotype information and the environment information for both twins. The information for twin 1 now appears on the first line for each pair and the information for twin 2 appears on the second line.

Now let us turn our attention to the structure of the statistical model. The simplest form of the fixed effects model is the linear model. Let us build up to the fixed effects model by first imagining we estimate a "standard" regression model and did not account for the clustering of twins within pairs. Doing so would result in the following:

$$P_{ij} = \beta_0 + \beta_E(E_{ij}) + U_j + \epsilon_{ij} \tag{10.1}$$

where P has been inserted—for brevity—for phenotype, E has been inserted for environment, and β_0 represents the intercept, which here captures the mean value of P for cases where i and j are equal to 0 (which may have little conceptual meaning, so we can ignore it now. We will return to a substantive discussion of the intercept momentarily). Notice the subscripts i and j appear in this equation; they identify the individual twin and the twin pair, respectively. Also note that there are two sources of error included in this equation. The first is U_j, which parameter represents the "fixed" effect. For our purposes, the fixed effect U_j represents all unobserved factors that impact P *that are shared between twins from the same family.* The second source of error is depicted as ϵ_{ij}, which captures sources of error that are unique to each twin in the dataset.

For now, let us turn our attention to U_j, because that will be the primary source of variance that the fixed effects model rules out by design. One can think of U_j as capturing all of the influences on P that are *fixed* across twins in the same family. Thus, U_j might capture shared environmental factors like parental income or socioeconomic status, it will capture stable factors like parent race and age, as well as factors related to the neighborhood the twins grow up in, and so on. In addition to shared environmental influences, U_j also captures any genetic influences on P. The logic here is that—as long as we restrict our analysis to MZ twins—any genetic influence on P will be captured by the fixed effect U_j because MZ twins share 100% of their genotype. Obviously, then, U_j is an important source of variance in P. Ideally, of course, we would be able to include measures of all the sources of variance that are captured by U_j. If that were the case, and we had measures of all these sources of variance, then we could include them in the model and omit U_j. But of course, that will almost never be the case. Instead, it is far more likely that a researcher will have data on P, E, and perhaps only a few sources of variance captured by U_j.

Our primary concern with U_j is that it may capture influences on P that are correlated with E. If we were to estimate the association between E and P without controlling for all sources of variance in U_j, then we would run the risk of misspecification. In other words, we may end up with an estimator for β_E that is biased due to uncontrolled confounding by U_j. The most straightforward way to eliminate this potential for confounding is to find a way to control for U_j. And that is exactly what the fixed effects model will do.

Another concern with the above equation is that it completely ignores the fact that our data are clustered and that twins appear within families. This means our standard errors are likely to be incorrect (likely too small) and need to be adjusted to reflect the nested nature of the data. One way to adjust for the clustering would be to estimate the model separately for twin 1 and then again for twin 2:

$$P_{1j} = \mu_1 + \beta_E(E_{1j}) + U_j + \epsilon_{1j} \tag{10.2}$$

$$P_{2j} = \mu_2 + \beta_E(E_{2j}) + U_j + \epsilon_{2j} \tag{10.3}$$

Notice that the equations share a few elements: both estimate β_E, which means we must assume the estimator is equivalent in both equations. This is one of the fundamental assumptions of the fixed effects model: we must assume the parameter estimates are, in fact, *fixed* across the units. Also note that U_j appears in both equations. Given our desire to

control away U_j, you may have already realized what we are working up to. But, before moving to the next step, it is also important to point out that both equations have their own *unique* sources of error, which are captured by ϵ_{1j} and ϵ_{2j}.

In order to rule out the impact of U_j, we could start by averaging the two equations across all j, which would result in:

$$\bar{P}_{ij} = \bar{\mu}_i + \beta_E(\bar{E}_{ij}) + \bar{U}_j + \bar{\epsilon}_{ij} \tag{10.4}$$

Because $\bar{U}_j = U_j$, we can re-write equation 10.4:

$$\bar{P}_{ij} = \bar{\mu}_i + \beta_E(\bar{E}_{ij}) + U_j + \bar{\epsilon}_{ij}$$

Now, if we subtract 10.4 from 10.1, we end up with:

$$P_{ij} - \bar{P}_{ij} = \mu_i - \bar{\mu}_i + \beta_E(E_{ij} - \bar{E}_{ij}) + U_j - U_j + \epsilon_{ij} - \bar{\epsilon}_{ij} \tag{10.5}$$

which, because $U_j - U_j = 0$, can be simplified to:

$$\ddot{P}_{ij} = \ddot{\mu}_i + \beta_E(\ddot{E}_{ij}) + \ddot{\epsilon}_{ij} \tag{10.6}$$

where, following Wooldridge (2010), we let $\ddot{P}_{ij} \equiv P_{ij} - \bar{P}_{ij}$, and so on.

There are just a few points to note before we move to a demonstration of this method in the next section. First, note that the fixed effects model—as was discussed above—is a widely used technique. It is often employed in macro-level researchers to rule out time stable factors that might explain phenomena of interest to, say, political scientists and economists. Criminologists, too, have capitalized on variants of the fixed effects model to test whether certain policies—like the death penalty—have any appreciable impact on crime rates (Kovandzic et al., 2009).

Second, the fixed effects model we have outlined above can be considered a general form of the broader fixed effects approach. In other words, there are certainly variations that one could introduce to address a known issue or to address a violation of a certain assumption (more on the modeling assumptions below). In this vein, you may have noticed that we restricted our example data to twin *pairs*. You may be wondering what to do if your data have more than two twins in some families. This issue can, of course, expand and complicate things if there are multiple different twin pairs within a family. If, for example, your data include many families that have multiple pairs of twins then an even more generalized version of the fixed effects model may be necessary. Interested readers should consult CITE (???).

But for most cases, the general form of the fixed effects model we introduce here will perform well. That is because the version we provide allows for any arbitrary number of twins within a twin set. Take, for example, a situation where most families have a "traditional" twin pair, meaning there are only two children who are twins. But what happens if some families have triplets or quadruplets? The fixed effects model we introduce handles these situations well because the twin pair (or twin set) *average* is being subtracted from each

individual twin's score. This was shown in equation 10.5. For families that include triplets (or other types of multiple births), equation 10.4 will be defined by the overall average and the differencing carried out in equation 10.5 will be conducted for every twin observation.

This reveals an interesting point—a point where logic and the math coalesce. In the special case where the number of siblings in all families is equal to two (i.e., $I = 2$ for all j in J), the demeaned value produced in equation 10.6 reduces to a simple difference score between the two siblings, just as long as one includes a dummy variable identifying sibling 1. In other words, when all pairs in a dataset have two and only two twins, the fixed effects model simplifies to the "simpler" difference score approach that was considered earlier.

A third point to keep in mind is that the fixed effect model allows researchers to control for genetic confounding when attempting to estimate the association between an environmental factor and a phenotypic outcome (Barnes et al., 2014; CITES). But this approach might be considered, by some, to be overly conservative because it also controls away all shared environmental influences. This point reveals itself in the math of the fixed effects model because the twin differences are what is under investigation. Thus, only factors that can lead to *differences* between twins are picked up by this model. Because twins, by definition, share the shared environment (i.e., c^2 from chapter ??), factors that underlie this source of variance are removed from the estimating equation. This is all to say that the demeaning of the twin scores that was shown in equation 10.5 works to rule out genetic influences (assuming only MZ twins are used) *and* shared environmental factors. This raises important questions about what the shared environment *is*, what goes into it, and whether it is an important source of variation to be considered directly. If the shared environment is known to be the predominant source of variation for the phenotype of focus, then the fixed effects model may be an inappropriate choice. Readers are thus cautioned against applying the fixed effects broadly and blindly. Consult the available literature prior to estimation so that you have an idea about how much variation your model will be eliminating. In this endeavor, the massive meta-analysis conducted by Polderman et al. (2015) and the accompanying webpage (??) should be helpful.

Fourth, it is commonly argued in behavioral science research that using within-person fixed effects (i.e., using longitudinal data and treating the individual as j and time as i, to use the notation from earlier) effectively rules out genetic influences. Thus, scholars sometimes draw on the within-person fixed effects model as a way to control away genetic influences and more accurately estimate the impact of environmental influences on a phenotype. But there is one problem with this logic: the within-person fixed effects model can only rule out genetic influences—and environmental influences—that are stable over all time periods T. In other words, the within-person fixed effects model rules out stable *between*-person differences. But it does not rule out *within*-person differences. Considering the mathematical structure of the model will help to clarify. Recall the "demeaning" part of the fixed effects model that was shown in equations 10.4 and 10.5. We noted that this portion of the model removes stable variance by subtracting the mean value $\bar{P}_{ij}$ from each observation P_{ij}. Thus, the fixed effects model rules out by design any influences that impact the mean value of P for all i in each j (i.e., it removes $\bar{P}_{ij}$). If we simply adjust the substantive meanings of the subscripts so that i

is used to indicate time points and j captures individuals in the study, then the fixed effects model can be used to remove all factors that influence the mean value of P for each person j over all time points i. This is precisely what the within-person fixed effects model does. This is a powerful technique because it rules out stable factors that affect *between*-person variation in P by subtracting $\bar{P}_{ij}$ from each observation of P_{ij}. But note that this model does not remove factors that influence *within*-person variation in P for each person j over time i. If genetic—or environmental—influences have an impact on changes in P over time, then the within-person fixed effects model will not eliminate all genetic influences.

This leads to the natural question of how genetic influences might impact variation in P over time. Most scholars think of genetic influences as fixed, only impacting the persons development of P to the point of fixity. But there are at least two reasons to believe genetic influences can affect both a person's mean level of P and his/her observed variation in P over time. The first reason to think this is that epigenetic changes can regulate how and when a genetic influences emerges. Epigenetic effects can emerge over time, where a gene is turned "on" at one point in development and then turned "off" at another (CITES). Epigenetic changes can also occur from one place to another. Socio-environmental context appears to affect gene regulation in ways that are only now starting to be explained (CITES). Thus, the science of epigenetics is beginning to reveal that thinking of the genome as a fixed influence on P is incorrect. The second reason to believe genetic influences can affect variation– -or changes—in P is to recall our discussion from chapter ?? about gene-environment interaction (i.e., $G \times E$). This line of theory and empirical work reveals how genetic influences are context specific in the sense that a genetic influence might be contingent (or moderated) on the realized environmental variation. In some situations, the genetic influence may be powerful and easy to identify. In others, the environmental sources of variation may be so powerful that all genetic influences—even though they may be present—are impossible to identify. Research and logic suggest that $G \times E$s are probably the rule and not the exception when it comes to the development of human complex traits (CITES). Thus, we must keep this in mind when trying to rule out genetic influences with the fixed effects model. If $G \times E$s work to make genetic influences "visible" in some situations but not others, it may not be possible to fully control for their influence by estimating a within-person fixed effects model.

This raises a fifth point that should be addressed directly. The question concerns the fixed effects model's ability to control for $G \times E$s. If one estimates the model we developed above, relying on data from MZ twins, then the parameter estimates gleaned from the estimating equation can be considered to have taken into account any arbitrary $G \times E$ for all G and for all E where E is a shared environment. Put a different—are more direct—way, the fixed effects model removes variance attributable to $G \times$ *shared environment*. It does not rule out $G \times non - shared\ environment$. Those interactions must be specified explicitly and included in the estimation model. This is an important point to keep in mind because it reveals that the parameter estimates gleaned from the fixed effects model may be biased if $G \times non - shared\ environment$ is not properly accounted for.

A sixth consideration concerns what happens when pairs other than MZs (e.g., DZs and full siblings) are included in the sample. The short response is that the parameter estimates

produced by the fixed effects equation may be biased in unknown directions if pairs other than MZs are included. It is important to emphasize that they *may* be biased. It is not certain that they will, just that they could. The reason is simple to see. Recall the elegance of the fixed effects approach is that it rules out unobserved sources of variance (U_j) that are shared by cases i in j. Which is to say the fixed effects model controls away all sources of variance that are shared between siblings/twins. When DZ twins, for example, are included then not all sources of genetic influences will be removed because DZ twins only share 50% of their genetic material on average. That means that we will only remove 50% of the genetic sources of variance in P for DZ twins, but we will still remove 100% of shared environmental sources of variance. To the extent that genetic influences on P are shared with genetic influences on our key independent variable of focus for the study (i.e., E from above), then the parameter estimate β_E will reflect this bias. This may or may not be a serious concern and, thus, must be assessed on a study-to-study basis. Although there is no statistical test we can offer to assess whether one has biased his/her results by including pairs other than MZs in the sample, it is often informative to estimate the fixed effects model separately for each type of pair available in the data. If, for example, the data includes both MZs and DZs, then the fixed effects model could be estimated for MZs and then again for DZs. If the parameter estimates do not substantively or statistically differ (e.g., one could perform a coefficient comparison test), then it may be reasonable to assume the biasing impact of genetic confounding is ignorable.

With these points in mind, we will now move to a short demonstration of the fixed effects model. Like with the previous chapters of this text, we will provide all code in `R` so that the data used in this demonstration can be reproduced and the examples can be replicated or adjusted to reflect the reader's needs.

10.2 Demonstration

To begin this demonstration, we must simulate a "long" twin dataset that includes a phenotype P, an environmental source of variance E, a random error component U_j that is fixed across twins within the same family but that can vary across families, and an individual-specific random error component ϵ_{ij}. It may help if we substantively label the phenotype and the environmental variable that is—at least for this demonstration—going to be assumed to have a causal effect on the phenotype.

Drawing on an empirical example from the labor economics literature, let us imagine we want to estimate the causal effect of educational attainment (E) on income (P) later in life. This very association has been the focus of much research. Many scholars have analyzed the association between educational attainment and income using the fixed effects design we advocate for here. In fact, a recent study by Sandewall, Cesarini, and Johannesson (2014) drew on the fixed effects model to test the robustness of the educational attainment–income relationship after including a control for IQ difference across twins. Their results showed that educational attainment was associated with higher income, but controlling for IQ differences

reduced the size of the effect by about 15%.

Let us imagine we want to simulate a dataset of $J = 100$ MZ twin pairs ($I = 200$ individual twins) where we have a measure of income (P) and educational attainment (E). We will worry about including a covariate for IQ later. Simulating a dataset that is amenable to fixed effects regression will require us to approach things a little differently than we have in previous chapters. It will be easiest if we start by defining certain parameters we would like R to take into account when simulating the file. Then we will define the error terms U_j and ϵ_{ij}. From there, we will simulate values for educational attainment and income.

```
# clear the workspace
remove(list=ls())

# set the seed for reproducibility
set.seed(007)

# specify the number of pairs to be simulated
J<-100

# generate a Pair ID
pairID<-as.integer(1:J)

# specify the number of twins per pair
I<-2

# simulate the pair-level error term, assume it's standard normal
U_j<-rnorm(J,0,1)

# expand the data so that it appears i times (once for each twin)
data<-data.frame(rep(pairID,I),rep(U_j,I))

# cleanup the names
colnames(data)<-c("pairID","U_j")

# identify all twins as MZ
data$MZ<-as.integer(1)

# generate a Twin ID
data$twinID<-as.integer(ave(data$MZ,data$pairID,FUN=seq_along))

# simulate a twin-level error term, assume it's for income ~N(60000,15000)
data$e_ij<-rnorm(I*J,60000,15000)

# simulate E, assume it's educational attainment ~Poisson(12)
data$E<-rpois(I*J,12)

# simulate P (income), allow E to have an effect
data$P<-1000*data$E+data$U_j+data$e_ij

# sort by pairID
data<-data[order(data$pairID),]

# re-order the variables
data<-data[c(1,4,3,7,6,2,5)]

# view the first 10 rows of the data file, print to xtable so readable by LaTeX
rownames(data)<-c()
library(xtable)
xtable(head(data,10))

# estimate the fixed effects model using plm package
#install.packages("plm")
library(plm)
fixed<-plm(P~E+factor(twinID),data=data,index=c("pairID","twinID"),model="within")
```

```
55
56 # reformat the "fixed" object so it's readable by xtable
57 xtable.plm<-xtable:::xtable.lm
58 xtable(fixed)
```

As always, the first few lines of code are simple housekeeping (as in line 2) and setting the seed so the data we produce is reproducible on your own machine (line 5). The unique elements for this demonstration begin on line 7, which is where we specify the number of pairs to be simulated as $J = 100$. Next, we generate a Pair ID (line 11, which also includes a request to create the values as integers. This will simply ensure that we get whole numbers in any output produced below.) and then we specify that each pair should have $I = 2$ twins (line 14). Thus, we should end up with $n = J \times I = 100 \times 2 = 200$ MZ twins (we will focus exclusively on MZ twins for this demonstration, for those interested in how to handle these types of analyses with MZ and DZ twins simultaneously, see Turkheimer and Harden, 2014).

The simulation begins to take shape on line 17, which is where we specify the fixed effect, U_j, that will ultimately be removed from estimating equation as was described above. For simplicity, we will assume U_j is normally distributed with mean = 0 and standard deviation = 1 ($U_j \sim N(0, 1)$). Assuming a standard normal distribution for U_j simplifies everything that runs through the simulation below because it will not affect the mean value of P, nor will it fundamentally affect the parameter effects we will soon specify. But it is important that we highlight that the assumption of the distribution for U_j will not always hold. On the contrary, it is likely to rarely hold in real-world contexts. But, as we have shown above, this source of variation is removed from the fixed effects equation, so our simulation is robust to different specifications, which means that fixed effects estimates are robust to different assumptions about the shape of the error distribution (CONFIRM WITH WOOLDRIDGE). In fact, it is not strictly necessary to include U_j in our simulation, but we decided to include it so that we could make these points and so that the reader could see how it might play a role if a modeling strategy other than the fixed effects model were undertaken.

The next line of code (line 20) will create the dataset using R's `data.frame` command. Here, we create a data file by replicating the Pair ID variable I times and same for the U_j value. Since we previously specified $I = 2$, the snippet of code on line 20 will simply duplicate the Pair ID values and the U_j values so that we now have a dataset that has two twins in each pair. (Line 23 renames the variables because the code on line 20 renames them so that the user knows they are replicated values).

Line 26 creates an MZ identifier. This line of code is not strictly necessary, but it is good practice so that you never forget whether the cases are MZ or DZ. Also, it simplifies our next task, which is to create a Twin ID variable. There are many ways to go about this task, but one of the simplest is to simply count up the values of the MZ identifier within each Pair ID. That is, essentially, what happens on line 29 such that we end up with a 1 for the first twin in each pair and a 2 for the second twin.

Line 32 simulates the twin-level error term (ϵ_{ij}) from above. Notice that we specify the error term to draw $I \times J$ values from a normal distribution with mean = 60,000 and standard

deviation = 15,000. Simulating the values for ϵ_{ij} in this way will ensure that we end up with values for P that are on a scale that is what we might expect for income data.

Line 35 simulates the values that will be used for E, which we have conceptualized as educational attainment. We will operationalize educational attainment as the highest degree completed, such that the values for E should be integers that are centered around 12 (high school diploma) but that vary on both sides of 12 such that some respondents do not finish high school and others will have completed many years of secondary education. In order to get a variable that has these properties, we simulate E by drawing from a Poisson distribution that has μ (mean) = 12. Note that in the Poisson distribution, $\mu = \sigma^2$, so we only need to specify the number of draws ($I \times J$) and μ on line 35.

Finally, on line 38 we simulate values for P, which we have conceptualized as income. Notice that we form values of P based on an effect of E (specifically, we specify that each unit increase in E—one additional year of education—should increase P by 1,000), the pair-level error term (U_j), and the twin-level error term (ϵ_{ij}).

The next few lines of code (lines 41, 44, 47) simply re-arrange and sort the data so that when we request a short output of the top of the file on line 48, we see a nicely arranged and sorted dataset. Specifically, we get the following (As we have seen in previous chapters, the `xtable` command will produce LaTeX readable output, which is why we load the `xtable` library on line 48 and then call it on line 49):

	pairID	twinID	MZ	P	E	U_j	e_ij
1	1	1	1	80788.16	13	2.29	67785.88
2	1	2	1	99352.45	9	2.29	90350.16
3	2	1	1	79811.90	11	-1.20	68813.10
4	2	2	1	85936.19	13	-1.20	72937.39
5	3	1	1	70809.31	12	-0.69	58810.00
6	3	2	1	66625.66	7	-0.69	59626.36
7	4	1	1	52384.17	10	-0.41	42384.58
8	4	2	1	80009.11	11	-0.41	69009.52
9	5	1	1	74629.86	10	-0.97	64630.83
10	5	2	1	93246.24	15	-0.97	78247.21

Now all that is left is to estimate the fixed effects model. `R` has many options for estimating relationships through fixed effects regression. We will use the model that is part of the `plm` package. You may need to install that package if this is your first time using it on your machine. If this is the case, then you will want to run line 52, which we have commented out to indicate that it may not be necessary for all readers. Once the `plm` package has been installed, it is necessary to load its library as we do on line 53. Then, as shown on line 54, the fixed effects model can be estimated on the simulated dataset by typing: `fixed<-plm(P~E+factor(twinID),data=data,index=c(''pairID'',''twinID''), model=''within'')`.

Let us breakdown this line of code one piece at a time. The first part of the code (`fixed<-plm`) is straightforward, we ask `R` to create a new object called `fixed` and in that object we ask it to place the output of the `plm` command. The `plm` command structure is similar to most other regression commands in `R`. Specifically, we first specify the dependent variable (`P`) followed by a tilde ($\sim$) and any predictor variables. In this case, we only have two: `E` and the Twin ID variable which we enter as a dummy variable (we tell `R` to treat it as a dummy variable by specifying `as.factor(twinID)`) to control for any random differences that might occur between twins labeled twin 1 and those labeled twin 2. The rest of the code simply specifies where the data are located (`data=data`), which variables in the data file are to be used to inform the structure of the file (`index=c(''pairID'',''twinID'')`), and which of the various multi-level models to run. On that last point, we specify that the `''within''` model should be estimated. The term "within" is just another term for the fixed effects model (because some have found the term "fixed" to be confusing, the term "within" has been used as an alternative—we agree that "within" is probably easier to remember and more intuitive. However, "fixed effects" remains the most popular—as best we can tell—which is why we have used it exclusively up to this point.).

The last two lines of code (lines 57 and 58) ask `R` to place the output from the fixed effects model into a format that can be read by LaTeX, which produces the following:

	Estimate	Std. Error	t-value	Pr(>|t|)
E	1001.9778	389.4704	2.57	0.0116
factor(twinID)2	-2482.6951	2025.2460	-1.23	0.2232

As can be seen, the estimates from the fixed effects model using our simulated data converged on a positive value for the estimate of the impact of E on P. Specifically, the model suggested that a one unit increase in E—which, recall, represents educational attainment—corresponds with an average increase in P of roughly \$1,002. The standard error was roughly 390, which results in a t-statistic of 2.57, which is statistically significant at the 0.05 α level but it is not statistically significant at the 0.01 α level. In other words, if these data reflected actual information from MZ twins, we would have found evidence to suggest that educational attainment causally impacts (positively) income later in life. The return on education would be estimated to be roughly \$1,000 for every additional year of study.

10.3 Assumptions & Limitations

Let us now consider the assumptions and limitations that must be considered anytime a researcher chooses to estimate a fixed effects model. In general, there are two "classes" of assumptions the researcher must make. The first class will be specific to the underlying form of regression that was chosen. In other words, if the researcher estimates a linear model, then the typical econometric assumptions specific to that model must be met in order for the estimates to be considered unbiased (e.g., assumptions about homoskedasticity and linearity

of the influence of the predictors). Readers interested in the different assumptions required by various forms of regression such as the linear model, the logit model, or the Poisson model are encouraged to see Wooldridge (2010).

The second class of assumptions concerns the sources of unmeasured confounding that might remain even after one carries out a fixed effects analysis. Recall we have built much of this chapter around our central equation: $P = \Psi(G, E)$. Let us now expand that equation a bit. If we assume the G component is captured by additive genetic influences alone (see chapter 3 for a review) and that the E component is made up of two types of environmental influences—shared environments C and nonshared environments E—then we can set the function to be $P = A + C + E$, as we did in earlier chapters (e.g., chapter 5).

Recall that shared environments C are any environmental influence that make two siblings more similar to one another and nonshared environments E are influences that make two siblings different. If we estimate the fixed effects model using a sample of MZ twins, we can safely assume that all sources of confounding attributable to A *and* C have been controlled. Put a different way, when a researcher uses a dataset made up of MZ twins, the fixed effects model will completely rule out biases due to genetic effects and shared environmental effects. This leaves only the nonshared environment as a source of confounding.

Being able to rule out two (likely major) sources of variation in a phenotype simply by design is a key feature of the fixed effects model. But note that the fixed effects model does not rule out *all* confounding. It is still subject to the usual concerns over confounding if the confounder is a nonshared environment. And this means that researchers should seek to include measures of any known (or hypothesized) sources of nonshared environmental confounding in their studies because they can be included in the statistical model as covariates.

But there is one notable exception to this discussion. Specifically, the fixed effects model will only rule out all confounding due to genetic effects if MZ twins are the only pairs analyzed. If pairs other than MZ twins are included in the sample, then the potential for genetic confounding will not fully be ruled out. The reason is simple: MZ twins share 100% of their DNA so genetic differences are a between-pair source of variance and are therefore ruled out by design. But if other types of siblings are in the sample, this rule will not hold and instead genetic differences will be *both* a between-pair and a within-pair source of variance. Because the fixed effects model only rules out between-pair sources of variances, it cannot completely account for genetic sources of confounding when siblings other than MZs form the sample.

Now this does not mean that genetic confounding will always plague fixed effects studies that are conducted, say, on non-twin sibling data. Quite the contrary, the fixed effects model will still account for a portion of those confounding influences. A general rule of thumb to keep in mind is that the fixed effects model can control for $R\%$ of genetic confounding; where R corresponds to the level of genetic overlap observed, on average, in the sample. So genetic confounding may not be a major concern if the study is 90% MZ twins and 10% DZ twins. It may be more of a concern if the sample is 50% non-twin siblings and 50%

unrelated pairs. The point is that researchers must understand the fixed effects model rules out genetic confounding as a function of the make-up of the sample under study.

10.4 Other Ways to Control for Genetic Influences

This brief section will introduce two relatively new approaches to controlling for genetic confounding. This first technique, referred to as the pedigree risk approach, was developed by Schwartz and colleagues (2015). These authors capitalized on a unique data source—the National Epidemiologic Survey of Alcohol and Related Conditions (NESARC)—to construct a latent measure of genetic risk for antisocial behavior. Their approach, simply put, was to create an index that would capture one's level of risk of intergenerational transmission of antisocial behavior.

More specifically, NESARC participants were asked to first report the number of full brothers and sisters they had who had lived to be at least 10 years old. Then, they were asked to report the total number of aunts and uncles—maternal and paternal—they had in their family who had lived to be at least 10 years old. Next, respondents were asked, "... whether their blood/natural mother ever had behavior problems. These questions were repeated for a total of six relatives: mother, father, maternal grandfather, maternal grandmother, paternal grandmother, and paternal grandfather" (p. 775). Responses to these items were then factor analyzed and the factor loadings were fixed to reflect the average level of genetic overlap between the participant and the relative of focus. For example, when siblings were the focus, the factor loading was fixed to 0.50. When grandparents were the focus, the factor loading was fixed to 0.25. In this way, Schwartz and colleagues were able to create a latent genetic risk score—a pedigree risk—without having to rely on direct observation of twin or sibling data. What is more, this pedigree risk measure can be used in a regression-based framework just like any other covariate. Thus, genetic confounding can—perhaps only partially—be controlled with such a pedigree risk measure.

But before taking that step, it was first important to establish that the measure performed as expected. The most straightforward way to establish this was to include the pedigree risk measure as a covariate in a regression model and then to examine the proportion of variance (indexed by R^2) in the outcome that was explained by the pedigree risk variable. The authors analyzed several outcome measures that tapped into the respondent's level of antisocial behavior. Thus, the authors anticipated that R^2 would be close to 0.50. Their regression models tended to converge on R^2 values that hovered in the range of 0.50 (although most were below that mark, some were quite close with several models having $R^2 = 0.49$ or 0.48).

This suggests the pedigree risk measure might perform at a level that is on par with other ways of controlling for genetic influences like the fixed effects model discussed in this chapter. Yet, it is important to reiterate the authors' caution that was sounded: "Whenever possible, more traditional quantitative genetic methodologies should be favored, but the presented

strategy remains a viable alternative for more limited samples" (p. 772). With this caution in mind, we believe the pedigree risk measure is a strategy worth pursuing and developing further because it offers a relatively straightforward and inexpensive way to control for one of the most consistent influences on human outcomes—genetic factors.

The second strategy that can be used to account for genetic confounding is the polygenic risk score approach that was introduced in chapter ??. Recall that the polygenic risk score is a summary measure—think of it like a scale measure—of one's genetic risk across all loci that are tagged in a GWAS. The polygenic score therefore represents a direct measure of the G from our $P = \Psi(G, E)$ equation. But the polygenic score approach is relatively new, which means that only for certain phenotypes is the polygenic score a powerful enough predictor to be considered a measure that would capture the majority of sources that drive genetic confounding. Put a different way, a polygenic score for a complex outcome like human aggression remains a relatively weak predictor (see the GWAS by Pappa et al., 2016), but that does not mean it cannot be used by social scientists. It does mean, however, that including a polygenic score for aggression in a multivariate regression model is unlikely to account for all sources of genetic confounding.

But this is not the case for all polygenic scores that one might create. Some polygenic scores have been shown to account for a sizable portion of the estimated heritability for the phenotype of focus. For example, GWAS—and, thus, any polygenic score that was created from that GWAS—for anorexia nervosa boasts a SNP heritability (h^2) of 0.559, which means that a polygenic score for anorexia nervosa would likely account for $\approx 55\%$ of the variance in that outcome. The SNP h^2 for behavioral traits tend to be much lower. For instance, the SNP h^2 for, albeit impressive, is only 0.1885. Compare this against the h^2 estimate that typically ranges between 0.50 and 0.80 in twin studies and one can see how the polygenic score for something like intelligence may not fully capture all the sources of confounding that may be traced to the genome (Note that all SNP h^2 estimates discussed here were gleaned from the LD Hub webpage: `http://ldsc.broadinstitute.org/lookup/`).

Although polygenic scores are expected to increase in precision over time—indeed, the SNP h^2 for educational attainment has risen quite rapidly from the first GWAS in 2013 (Rietveld et al., 2013) to the most recent in 2018 (CITE)—the present state for most complex traits of interest to social scientists would suggest that a polygenic score may not fully account for all sources of genetic confounding. We would encourage researchers interested in controlling for genetic variance for the purposes of isolating other sources of variance to consider one of the other strategies covered in this chapter. But this is not to say that polygenic scores are not useful. On the contrary, we anticipate that polygenic scores will, in the not too distant future, represent a viable way to control for genetic confounding for many human phenotypes (see, generally, Visscher et al., 2017).

10.5 Conclusion

The bulk of this book has focused on the G portion of the $P = \Psi(G, E)$ equation that forms the foundation of behavioral genetic thought. But this chapter represents a (sometimes slight) departure from that focus. Where each of the chapters that have come before this one focused, in some way, on how a researcher can estimate G's impact on P or the degree to which variation in P is defined by variation in G, the present chapter focused on E. Our focus here has been to show how researchers can leverage the information available in G to better isolate the influence of E on P. In order to appreciate the points raised in this chapter, it may help if we remind you that Ψ in $P = \Psi(G, E)$ is intended to account for all sources of overlap—correlation and interaction—that may exist between G and E. This immediately reveals why it is important to account for G (e.g., control for it) when one wants to understand the link between E and P. If, for instance, there is correlation between G and E—what we introduced as gene-environment correlation in chapter ??—then estimating the correlation between E and P without accounting for the correlation between G and G may lead to bias in the estimate(s). This chapter, therefore, discussed one of the most popular and powerful strategies for ruling out such bias; the fixed effects model. We also briefly touched on two other strategies that we anticipate will gain in popularity over time. Indeed, we anticipate the polygenic score approach will eventually become the favored approach.

Part III

Practical Concerns

Chapter 11

Practical Issues, Ethical Concerns, & A Philosophical Discussion

A continued use and growth of quantitative genetics by social scientists has the potential to make significant contributions to the understanding of human phenotypes. Most of these potential contributions are widely known and have been covered extensively in this book, such as knowing the comparative effects of genetic and environmental contributors, unraveling the interconnections between genes and the environment, and estimating statistical models that are able to produce unbiased parameter estimates of environments without genetic confounding. At the same time, however, there are some emerging practical issues and ethical concerns that are often overlooked by social scientists using quantitative genetics, but need to be addressed more systematically. While not exhaustive, in this chapter we discuss some of these issues and concerns and offer some suggestions on how they may be dealt with effectively.

11.1 Practical Issues

There are a number of practical issues that have emerged and that will continue to emerge for social scientists using quantitative genetics. One of the most obvious and most pressing issues has to do with the social science ideology that permeates most disciplines. This ideology underscores the importance of sociological explanations and downplays, ridicules, or ignores outright explanations that incorporate individual differences, particularly those that focus on genetic variation. Perhaps that is why a quick perusal of most social science journals reveals that they focus almost exclusively on environmental contributors to phenotypes and rarely incorporate genetics. A recent analysis of all articles appearing in the most prestigious criminology journals revealed, for instance, that only about 1.6% of all articles in the top journals even mentioned genetics in the studies (Schwartz, 2014). There is little doubt that this same publication pattern would be detected in other fields of study, including sociology

and social work.

There are a growing number of social scientists who appear to be open to the role of genetic influences, but all too often they put their own sociological spin on the findings and their own writings. In short, the findings must comport with their own sociological beliefs about human behavior; genes matter to them, but only insofar as they remain consistent with the hegemony of environmental dominance. To illustrate what we mean by this, consider the recent research on the interpretation of gene-environment interactions ($G \times E$). There are two key explanations for $G \times E$s that we discussed in chapter 9: diathesis stress and differential susceptibility.

The most commonly employed explanation for $G \times E$s has been the diathesis-stress model (Gottesman, 1991). According to this model, genetic predispositions only emerge when they are paired with disadvantaged environments. This intersection of a genetic liability along with a disadvantaged environment ultimately leads to the creation of a particular phenotype. Importantly, the environmental pathogen acts as a "trigger" that causes the genetic potential to materialize. Without the environment, the genetic predisposition would remain nothing more than resting potential and would not lead the phenotype to develop. Diathesis stress has almost always been used to explain negative outcomes by focusing on genetic risk and disadvantaged or adverse environmental conditions. This model has been used widely to explain an array of phenotypes, including criminal behavior, depressive symptomologies, and numerous other types of psychopathologies.

Until relatively recently, the diathesis-stress model was the only model that was used to explain $G \times E$s. That was until Belsky advanced a new explanation for $G \times E$s that was, in many ways, a challenge to the diathesis-stress model (Belsky & Pluess, 2009). Known as the differential susceptibility model, this explanation posited that there was too much emphasis placed on the negative side of the equation. In other words, by focusing only on negative outcomes, risky genes, and disadvantaged environmental triggers, the diathesis-stress model was ignoring how $G \times E$s would interface with positive outcomes. Differential susceptibility cast a wider net and explored the full gamut of phenotypic outcomes (i.e., ranging from positive to negative) and, as a result, also examined how genes and environments could combine together (i.e., $G \times E$s) to explain the full spectrum of phenotypic outcomes. To do so, Belsky pointed out that instead of viewing genes as predisposing for risk, they should be viewed as priming for susceptibility to environmental conditions. Whereas diathesis-stress would identify alleles that would predispose for negative outcomes, the logic of differential susceptibility would hold that these same alleles would be better characterized as plasticity alleles. Plasticity alleles, according to the differential susceptibility model, simply index how susceptible each person is to the environment; the greater the number of plasticity alleles, the more likely the environment is to affect each person.

What is particularly unique about the differential susceptibility model—and what separates it from the diathesis-stress model—is that plasticity alleles are thought to interact with both negative and positive environments to predict both negative and positive outcomes. The reason for this dual focus is because plasticity alleles simply indicate how plastic

an individual is; that is, how easily the environment is able to impinge upon a person. A greater number of plasticity alleles indicates that all environments will have a greater impact because the individual is more plastic. So, a person with a comparatively large number of plasticity alleles will be more easily molded by environments compared with a person with a relatively smaller number of plasticity alleles. What is key, however, is that a larger number of plasticity alleles will mean that when exposed to positive environments, a positive outcome is more likely and, at the same time, when that same person is exposed to negative environments, a negative outcome is more likely. Belsky has dubbed this somewhat paradoxical effect, "for better and for worse" (Belsky & Pluess, 2009; Belsky, Bakermans-Kranenburg, & IJzendoorn, 2007).

The advancement of the differential-susceptibility model has sparked a considerable amount of research, particularly among psychologists, testing its utility against the utility of the diathesis-stress model. The results of these psychological studies have produced somewhat mixed results. Some studies have shown support for diathesis-stress (Nederhof et al., 2012), others have shown support for differential-susceptibility (Belsky & Beaver, 2011), and still others have shown partial support for both models (Kochanska et al., 2011). Based on the results of these studies, therefore, it appears that much more research needs to be undertaken that directly examines both of these models using methodologies and statistical techniques capable of ferretting out which one is the more accurate model (Stoltz et al., 2017).

In much of the social sciences there does not appear to be a concerted effort to test these two explanations against each other. Instead, the diathesis-stress model has been dumped in lieu of the differential-susceptibility model to explain $G \times E$s. In some social science circles, any evidence of $G \times E$s is now being equated with differential susceptibility without any formal test or concerted attempt to examine the role of diathesis stress. Of course, this is not the case for all $G \times E$ research and there are even teams of researchers who have proposed new ways of testing the diathesis-stress model against the differential susceptibility model (Roisman et al., 2012). The advancement of statistical techniques designed to test these two models are the types of approaches that need to be employed to further our understanding of $G \times E$s and the processes that underlie them.

Somewhere along the line, empirical evidence of $G \times E$s has become equated with epigenetics among some circles of social scientists. To understand what is meant by epigenetics, it is important to first reiterate that DNA is located in the nucleus of every cell except red blood cells and the information encoded into DNA is the exact same in all cells. What this means is that the DNA in a heart cell is the same DNA that is found in a lung cell. What distinguishes a heart cell from a lung cell (and all other cells for that matter) is that only the genes that are necessary for the functioning of the heart are "turned on" and all other genes are 'turned off." The genome is not responsible for determining which genes are "turned on" and which genes are "turned off"; rather, that is the responsibility of what is known as the epigenome.

The epigenome includes the chemical markers that are located on the strands of DNA, some of which have the capacity to influence gene activity by affecting the ability of DNA

to be duplicated on RNA. Some of these epigenetic chemical markers increase gene activity (e.g., histone acetylation) and some of these chemical markers decrease gene activity (e.g., methylation). In contrast to the genome, the epigenome is fluid and can change over the life course in response to, among other stimuli, environmental conditions. Alterations to the epigenome, in turn, affect which genes are "turned on" and which genes are "turned off." There is even some evidence indicating that epigenetic modifications that accrue over a person's life can then be transmitted to future generations (Holliday, 2006).

Epigenetics has become the most recent fad in the social sciences; a way to tip a cap to genetics and, at the same time, maintain that the environment matters more. As Moffitt and Beckley (2015, p. 124) note, "Many social scientists embrace the new epigenetics research because it has been billed as evidence that environment trumps genes." And, after all, that is what many social scientists care deeply about. This all-out embrace of epigenetics, however, has not been mirrored by the evidence. In fact, the evidence to date supporting the role of epigenetic processes in human behavioral and personality phenotypes is largely lacking from the literature. Certainly, there have been some high-profile studies showing epigenetic processes at play, but these studies have been somewhat isolated and have been largely confined to non-human animal studies or to studies that are based on very small numbers of clinical subjects. Against this backdrop, epigeneticists have pleaded for social scientists to use caution before championing the role of epigenetic processes in the development of human phenotypes (Heijmans & Mill, 2012; Mill & Heijmans, 2013).

Nonetheless, social science scholarship is replete with references to epigenetics, promises of what epigenetics can hold for the future of the social sciences, and discussions of how epigenetics underscores the fact that the environment is more potent than genetics. Epigenetics, moreover, can be invoked to explain virtually any genetic, environmental, or biosocial influence, including those that are currently unexplainable with the available technology. Why?—because at this point epigenetics is largely unfalsifiable because, until recently, epigenetics data were largely unavailable. Indeed, currently there are very few datasets with human phenotype data that also include epigenetic information. Also, because epigenetic markers are dynamic—meaning they can and do change over time—it is difficult to identify the causal ordering if one were to find an epigenetic "tag" was correlated with some behavioral outcome.

It is interesting to consider that while social scientists have been quick to embrace epigenetics—despite the fact that it is infinitely more complex than genetics and not nearly as well understood as genetics—many of these same social scientists have been just as quick to discount the role of genetic influences on human phenotypes. In fact, some social scientists have largely argued that genes do not have any direct, independent, and additive effects on human phenotypes, but rather only can exert their effects via epigenetic processes (Wright et al., 2015). This, of course, is an empirical question, but one that has been the subject of a lot of research and the findings consistently do not support this belief; rather, genes can—and indeed do—have additive, independent effects on virtually every human phenotype ever measured.

Our point is not to lay waste to epigenetics; we realize that it likely has utility in our understanding of human behavioral phenotypes. At this point, however, it is likely premature to place too much stock in a field of study that has been largely unproven when it comes to its application to the development of human traits and behaviors. Against this backdrop, we echo the call of epigeneticists and urge caution in highlighting the role of epigenetic processes in social science research until research findings that are replicated by independent research teams are published. As Marzi and colleagues stated, "We need to come to terms with the possibility that epigenetic epidemiology is not yet well matched to experimental, nonhuman models in uncovering the biological embedding of stress [the environment]."

Two additional practical concerns are worthy of discussion. First, there is now incontrovertible evidence indicating that shared and nonshared environments have vastly different influences on phenotypes (Polderman et al., 2015). Although some exceptions exist, most research indicates that almost all of the environmental variance is accounted for by nonshared environmental effects. Shared environmental influences typically flutter around 0.00-0.10, with the effects almost always being 0.00 for phenotypes measured in adulthood. Virtually all social science research, however, does not focus on estimating nonshared environmental effects, but rather proceeds without making any distinction between shared and nonshared environments. This is a serious oversight as to estimate nonshared environmental effects virtually requires a sample that contains at least two siblings in the same household. After all, how else can environmental differences and phenotypic differences be measured and analyzed without a sibling sample from which to difference them? Additionally, without distinguishing between these two environments, the true environmental effect is likely attenuated or obfuscated. Moving forward, then, researchers should, at the very least, be in a position to be able to measure environmental variables as both shared environmental and nonshared environmental measures. Failure to do so will likely produce biased results and will ultimately retard advancements in understanding environmental factors that are associated with the development of phenotypic variance.

Second, there has been a great deal of concern regarding replication (or non-replication) in studies and this has been particularly true for candidate gene research and for studies testing for c$G \times E$s. There are a number of different possibilities for such non-replication, but regardless of the real reason for it, safeguards need to be put into place to limit the number of false positives that are taken as gospel. Some journals, for instance, require two independent samples in order for a novel genetic association finding to be published. There are other ways that the issue of non-replication could be addressed, including publicly releasing the data used in the project, registering all study plans in advance of the study, and providing information about the types and number of analyses examined. Whatever approach is employed, transparency is key as it helps to eliminate concerns about questionable analytic approaches (more on these below).

Collectively these practical concerns stand in the way of allowing for an impartial and evidence-based approach to evaluating social science that includes quantitative genetic methodologies. While these are not the only practical issues, they do represent some of the more widespread ones that are encountered in the field. The good news, of course, is that most of

these obstacles can be overcome without the collection of any new data and without the need to learn any new statistical techniques. Rather, what is key to eliminating these practical issues is focusing on evidence, losing ideologies, and evaluating research using an objective, scientific approach that is free from any type of value judgments.

11.2 Ethical Concerns

No matter what type of methodology, statistic, or topic is being studied, there is always the possibility that ethical concerns could surface. All too often, however, this is not fully realized and social scientists point to genetic and biosocial research as the main areas where unethical outcomes are applicable. Perhaps nowhere is this truer than when it comes to the ethical concerns regarding how genetic information might be used as the scaffolding to justify the implementation of certain policies. Much of this concern likely stems from the history of eugenics and its ties to genetic research. What is interesting to point out, however, is that there is evidence that the eugenics movement was just as much a product of sociological research (e.g., Karl Marx's writings) as it was genetic research (Ridley, 1998). In fact, some of the most influential sociologists of the time were largely supportive of eugenics, seeing it as a way of eliminating some of the key social problems affecting society. Many social scientists find it surprising to learn that W. E. B. Du Bois (1932)—one of the most prominent social scholars—was in favor of eugenics and argued that when it comes to human reproduction, "quality and not mere quantity really counts" (p. 167).

Regardless of whether genetic, environmental, or biosocial research was used to justify eugenics, the concern with how genetic research will be used to inform policy remains a key concern of social scientists. Take, for example, genetic research on criminal behavior. Critical readers of this line of research often argue that the findings of a genetic effect on criminal behavior would be used to justify oppressive policies that focus only on punitive approaches to crime control. Moreover, there is a concern that such research would lead to the greater use of capital punishment, to laws that restrict reproduction (e.g., the elimination of conjugal visits in prison), and to other forms of punishment that advocate extremely long prison sentences (e.g., life without the possibility of parole). Certainly genetic research could be used to justify these types of policies, but so could environmental research. After all, the get-tough movement towards crime, as well as the disproportionate rate of minority (particularly African-American) offenders who were incarcerated in the 1980s and 1990s was certainly not supported by genetic research as there was virtually no criminological research published during that time focusing on genetic influences. Instead, almost all of the criminological research published in the 1970s and early 2000s focused exclusively on environmental risk factors. Yet, there is little discussion of how environmental explanations have been used historically—and how they may be used in the future—to create policies that intentionally or unintentionally create inequalities and that in retrospect are seen as highly oppressive.

It is also important to note that genetic findings can lead to progressive policies and to the implementation of policies that might be more effective than policies that are based solely

on environmental research. Consider, for instance, the abolishment of the death penalty for juvenile murderers. The research that was used to overturn the use of the death penalty for these offenders was biological, neurobiological, and genetic; it had nothing to do with purely environmental explanations or sociological research. So, when it comes to perhaps the most progressive criminal justice policy decision in the past few decades, it was biosocial research findings, not sociological ones, that were the most important. We point this out to highlight the fact that while on the surface it might appear as though environmental explanations are safe whereas genetic explanations are dangerous, there are contemporary examples that paint a different picture.

There is even evidence that the public becomes more accepting of diversity when that diversity is grounded in genetic variance. For instance, the legal scholar, Richard Posner (1992), has provided some lines of argumentation indicating that the public is much more accepting of homosexuals now than in the past based, in part, on research revealing that sexual orientation may be tied to biological and even genetic influences. The public appears to view these findings in a way that casts a more positive light on homosexuals than was previously viewed when research findings appeared to show that sexuality was cultivated solely by socialization factors. Whether a genetic basis was found for other marginalized and historically oppressed groups would lead to them being more accepted remains to be determined. As for now, however, the findings in relation to homosexuality hint at the very real possibility that genetic research might lead to progressive beliefs and a greater openness to previously stigmatized groups.

In addition to the fears related to policy implications, there are other ethical concerns that need to be addressed. For example, scholars have a duty to ensure that their research is being conducted in an ethical fashion. Some of the ethical obligations to scholars are well-known and widely discussed, such as keeping data confidential and ensuring the anonymity of all study participants. These types of issues pertain to all research and are not unique to particular areas of study. When it comes to quantitative genetic research, however, there are some potential ethical issues that pertain more to this body of research than to others. We discuss three of these potential issues here.

First, for social scientists who engage in primary data collection, there are some unique issues that arise when collecting DNA samples. To begin with, the researchers have to ensure that all DNA samples are handled in a way to prevent them from being compromised. Confidentiality and anonymity take on all the more importance when collecting DNA samples. Why?—because if the samples become compromised, then the DNA has the potential to link subjects to crimes they may have committed, it could be used to establish paternity/maternity, and it could also reveal information regarding the genetic propensity for developing significant health problems, such as certain cancers or terminal diseases. This also leads to another ethical dilemma—that is, what should be done by researchers if they discover that a subject has a genetic liability for a particular serious phenotype, such as a cancer? Two points are worthy of mention. First, for most social scientists, the genes that would be collected are genes that have nothing to do with diseases and disorders, so this likely would not become an issue for the vast majority of all social science data collection.

Second, single genes tend to have relatively small effects and thus even if a subject possessed a genetic predisposition for a particular phenotype (e.g., violence) the risk that this single gene would confer would likely be relatively small. Even so, there must be great care taken in order to ensure that the samples are protected and that the results of the DNA tests remain confidential.

One potential way to help deal with these potential issues is to recruit scholars from other fields of study who have previous experience working with the collection of biological material. This type of collaborative effort would offer two advantages over working alone (or with another colleague in the same department). First, an interdisciplinary team would allow for each member of the research team to contribute to the data collection project by relying on their areas of expertise. This certainly is an advantage with any type of interdisciplinary data collection effort, but it takes on all the more importance with DNA collection because there are more specialized tasks and roles involved (e.g., collecting the biological material, genotyping the sample, destroying/retaining the DNA, analyzing the data, etc.). Second, by drawing on the experience of other scholars from other fields of study would likely allow for a more successful navigation of some of the potential ethical pitfalls that might be encountered. This approach would setup social scientists nicely for the ability to collect genetically sensitive data in the future with or without such an interdisciplinary team.

Second, P hacking, sometimes referred to as data dredging or fishing for findings, is a very real issue with molecular genetic research, particularly genetic association studies. P hacking occurs when researchers capitalize on chance findings by analyzing hundreds (or thousands) of potential associations and then cherry-picking the one significant finding and writing their study around that particular association (Head et al., 2015). For example, with molecular genetic association studies, a researcher could calculate a bivariate correlation matrix between a single genetic polymorphism and one hundred phenotypes. Perhaps one or two of these associations emerge as statistically significant. The researcher could then develop a manuscript around those two findings (or two separate papers, one devoted to each finding). Without disclosing to reviewers, editors, and potential readers of the study that they calculated ninety-some other correlations, nobody is privy to the fact that these findings are really nothing more than what would have been expected by chance along (about 5 statistically significant results for 100 correlations when using a $P < 0.05$ level of significance). As can be seen, researchers who engage in p-hacking are capitalizing on the chance significant findings that are inherently built into statistical tests.

These types of practices can be quite harmful. After all, to those unaffiliated with the research team producing the results, it appears as though the findings emerged from careful, thoughtful analyses. The error rate that is reported, moreover, is supposed to provide consumers of the study with a general idea of how likely the association is to be incorrect. Given, however, that the finding emerged from P hacking, not only is the error rate misleading, but the findings are also unlikely to be replicated. Why?—because the statistically significant association that was reported is unlikely to be a "real" significant finding. Indeed, part of the problem with the lack of replication for novel genetic findings is likely due, to

some extent, to P hacking. The same holds true for gene-environment interactions, where there are almost unlimited combinations of genes, environments, and phenotypes to examine in order to uncover statistically significant gene-environment interactions predicting human phenotypes.

Although there is no surefire way of preventing P hacking from occurring, there are certainly some ways to limit its influence and to deter researchers from engaging in such practices. Some journals, for instance, have dealt with this problem head-on, such as by requiring an independent replication sample for novel genetic effects. With this approach, a chance finding in one sample will not be detected in the independent sample. While there are certainly some drawbacks to requiring two samples, it certainly helps with issues related to p-hacking and non-replication of findings.

Perhaps the most effective, straightforward, and defensible approach to dealing with P hacking and some related issues comes from a recommendation made by the Center for Open Science. The mission of this organization "...is to increase openness, integrity, and reproducibility of research" (https://cos.io/about/mission/). This Center provides detailed insight into how to achieve their mission, and they provide numerous tools and depositories to help researchers be open and transparent in their research. But perhaps the single most important recommendation—at least when it comes to preventing P hacking with genetic research—is by pre-registering all study plans and procedures prior to the commencement of the study. In that way, the research team is allowing anyone interested in seeing what the plans of the study were even before data analysis began. Of course, there are likely ways around this as well, but it certainly represents a step in the right direction, and one that would likely encourage researchers to move away from P hacking and towards a more reliable, defensible, and objective way of conducting all studies, including quantitative genetic studies.

11.3 A Philosophical Discussion: Spanning the Explanatory Divide

Finally, let us turn our attention to a matter that is probably best considered a philosophical issue. As we have shown throughout this text, quantitative and behavioral geneticists deserve credit for a simple, yet revelatory observation that a phenotype is made up of genetic and environmental components, such that:

$$P = G + E$$

This deceptively simple equation conceals many important philosophical and mathematical points. First, note that the G is listed before the E. Although it takes only a cursory understanding of simple arithmetic to realize that the left-hand side of the equation is identical regardless of whether G or E is listed first, there is a great philosophical divide that can be traced to this point. Specifically, behavioral geneticists and quantitative geneticists are trained to see the world such that G affects P and any influence of E can be thought of

as noise. E, in this framework, becomes a nuisance parameter. Something that must be accounted for but is not necessarily of primary interest.

Social scientists are often trained in the reverse—they focus on the environment E as the primary influence of phenotypic scores and genetic influences G are the noise that must be controlled/accounted for but that is not of any direct interest (see, generally, Rafter, Posick, & Rocque, 2016).

This is, of course, a gross oversimplification, but we believe it captures broadly an important "explanatory divide" (Tabery, 2014) that currently exists between "traditional" social science disciplines and those that have more heavily integrated with behavioral/quantitative genetics. We are concerned that there is a lack of interdisciplinary discussion, which has led to the explanatory divide where one group views G as a key component in need of focus and the other sees it as a nuisance parameter. This becomes quite problematic when we consider that most human complex traits cannot be neatly divided into a G component and an E component.

Rather, as we discussed in chapter 9, gene-environment interplay ($G \times E$ and rGE) seems to be the rule and not the exception. Thus, it becomes increasingly troubling to imagine a scenario where we can, with any fidelity, say a human outcome is due to G *or* E; we now know it is both. As Burt (2016, p. 114) recently noted, "Indeed, one important, and often overlooked, consideration in studies of environmental influences is that the 'environment' may not be genetically independent of the outcome variable?" This quote serves to remind us that rGE and $G \times E$ make it so that we cannot accurately estimate the impact of E on P unless we have made an effort to account for the influence of G.

Nonetheless, it is easy to understand how the explanatory divide has developed. As Tabery (2014: 5-6) notes, scholars have often found themselves on either side of an explanatory divide when debating "nature versus nurture":

> But they go about that study in two very different ways: they identify different things that need explaining; they ask different causal questions about the thing that needs explaining; they point to different things that do the explaining; and they utilize different methodologies to provide those explanations. An example will help convey this, so consider depression again. Members of both the variation-partitioning approach and the mechanism-elucidation approach are interested in studying the nature and nurture of depression, but that commonality belies important differences. On the variation-partitioning approach, the thing to be explained is variation in depression in some population; employers of this approach ask how-much questions about that variation—how much of the variation in depression is the result of differences in nature and how much is due to differences in nurture? The variation-partitioners answer those how-much questions by identifying and measuring the causes of variation responsible for the variation in depression, and they utilize statistical methodologies to generate that measurement by partitioning up the total variation in depression and

> allotting it to the assorted causes of variation. On the mechanism-elucidation approach, in contrast, the thing to be explained is the developmental process that gives rise to depression; employers of this approach ask how questions about that process—how do differences in nature and differences in nurture interact during the developmental process to give rise to differences in depression? The mechanism-elucidators answer those questions by elucidating the causal mechanisms responsible for depression, utilizing experimental methodologies that intervene on the mechanisms to generate that elucidation.

As you can see, Tabery (2014) classifies quantitative/behavior genetics research into two camps: a) the variation-partitioning approach and b) the mechanism-elucidation approach. This is more than a mere classification scheme. Indeed, Tabery shows that many of the most important controversies and debates in the genetics literature can be traced to the point that one side of a debate approached the question at hand from the variation-partitioning approach and the other came to it from the mechanism-elucidation approach. The two approaches have several philosophical differences (see Tabery, 2014: 37 for a helpful table). Perhaps most important is that the former (variation-partitioning) attempts to describe phenotypic development as a statistical problem where G and E are properties that are estimated and the goal is to closely match one's estimates to observed data. The latter (mechanism-elucidation) attempts to describe phenotypic development by understanding the developmental—and, in some cases hypothetical—process that involves G and E.

Research into the genetic causes of complex human outcomes has always—and probably will always—had its critics (Charney papers, Moore and Shenk, 2017) and its defenders (Barnes et al., 2014; Sesardic, 1993). As Thomas J. Bouchard (1987: 58) once noted, "The massive, and vituperous, attacks on hereditarian findings clearly signal how seriously the environmental program is challenged by this evidence."

One of the points we have tried to advocate for in this text is that such reasoning—that heritability/genetic explanations are in competition with environmental/social explanations—is mistaken. If we are to span the explanatory divide, it will be necessary to dissolve the artificial competition between nature and nurture. Instead, scholars must begin to think of human behavior as a complex musical arrangement that requires both the vocals (e.g., nature) and the instrumental (e.g., nurture) pieces to make the song.

This may strike some readers as an appeal to complexity in order to wiggle out of explaining the tough to explain. But we do not see it that way. True, we are ending this discussion with a call for researchers to embrace the complexity and to work toward trying to understand complicated causal mechanisms that may never be fully understood. But trying to explain something that may never be fully understood is not the same as throwing one's hands up and claiming it is impossible. Just like anything else in life, the details of this scientific endeavor are complicated and require careful attention. We must walk before we run. And, thus, it was critical that those who came before us studied the heritability and environmentality of human behavior. It is easy for us now—with the benefit of knowing how things would turn out—to look back and snicker at how simple-minded certain scholarly explanations seem to

be. But we must keep in mind, that we know how things would turn out. We know and fully embrace the fact that human complex traits are a complicated melody that involves G and E. And that need not bother us. But to our academic forebearers, that reality was only one of several possible explanations. All of which probably seemed likely to be true. Thus, we must not scoff at earlier attempts to reason through the nature and nurture of some human trait. The very principles of science require that we try to find the simplest explanation of a phenomenon before moving to more complicated ones. It just turns out that human behavior does not afford simple explanation. And that is a reality with which we must be satisfied.

Chapter 12

The Future of Quantitative and Behavioral Genetics in the Social Sciences

Although some form of quantitative genetics has been around for more than one hundred years, in many ways only relatively recently has it begun to penetrate the social sciences on a consistent basis. Given the newness of quantitative genetics in the social sciences, it is both an exciting and frustrating time to be a social scientist using quantitative genetics. It is an exciting time because social scientists using these quantitative approaches are at the cutting-edge of the discipline and they have the potential to uncover new findings at an exponential rate. It is a frustrating time because social scientists who use quantitative genetic approaches are frequently met with heavy criticism, not all of which is productive to the broader discussion. In our view, and the view of others, the future of the social sciences depends, in large part, on the widespread use and application of quantitative genetic approaches. In order for quantitative genetics to make its biggest impact, and in order for the social sciences to remain current with other fields of study, we touch on a number of issues that we see as the most pressing and that need to be addressed in the upcoming years.

12.1 Institutionalization of Quantitative Genetics

If social scientists wish to embrace quantitative genetics and make it an integral part of their disciplines, then the most pressing change that needs to occur is for quantitative genetics to be institutionalized across social science departments. This, by no means, would be an easy task. Organizational culture and management are difficult to transform and academic departments are notorious for maintaining the status quo and resisting change. Even so, looking across disciplines and within particular fields of study, we see that change does occur and typically unfolds in a slow and methodical fashion. Small incremental changes

frequently emerge in response to newer findings, ground-breaking theories, or cutting-edge data analytic techniques. Changes can also occur on a more widespread basis, such as when new administrators are put in charge or when a new faculty member is hired who import with them a fresh set of skills, ideas, and theories. The point is that change can occur and does occur from time-to-time in academic departments. Whether widespread change occurs in social science departments when it comes to the integration of quantitative genetics, only time will tell. With that said, we are optimistic and see the institutionalization of quantitative genetics to not only be possible, but be quite likely. For this change to be fully realized, we touch upon three unique, but interrelated, ways in which institutionalization is most likely to occur.

The first and perhaps most important change that needs to occur is for quantitative genetics to be integrated into graduate student education. This would not require a complete overhaul of the graduate curriculum or even a new set of requirements for graduation. The changes could only involve an additional class or two being offered annually and these courses could be offered on an elective basis at first. Over time, a course or two in quantitative genetics could be required or, in some departments, students may be able to declare quantitative genetics as an area of specialty. The point is that quantitative genetic courses should be available to graduate students at most, if not all, social science departments offering a doctorate degree. Otherwise these departments will be graduating students who are not fully prepared to read, understand, and critique research that uses some type of quantitative genetic approach.

Keep in mind that a significant amount of time and courses in graduate school are funneled into training graduate students in advanced methodological and statistical approaches. Most students emerge from graduate school knowing, at a minimum, ordinary least squares (OLS) regression, binary logistic regression, and factor analysis. It is not uncommon, moreover, for students to graduate with an intimate knowledge of more advanced statistical approaches, such growth-curve modeling, latent class analyses, multilevel modeling, and structural equation modeling, to name a few. All of this really underscores the fact that quantitative approaches tend to dominate the social sciences and so it would not be too taxing nor would it be out-of-the-norm to add a course or two on quantitative genetics. Doing so would not only likely fit within the purview of the curriculum, but it would also likely be not all that difficult to accomplish. After all, as this book has highlighted, quantitative genetics is based on the same statistical foundation that is learned in most standard social science statistical courses, meaning that it should be relatively straightforward for students to learn given their statistical/quantitative background.

Adding quantitative genetics to the graduate curriculum likely will produce benefits that extend beyond students simply having knowledge on how to use these statistical techniques. Students who are interested in quantitative genetics and are trained to estimate these statistics would be able to address a broader range of research questions in their graduate-student research as well as with their dissertation. They would be able to apply a quantitative genetic framework to answer research questions that had never been posed previously or to address issues using statistical techniques that had never been employed previously. This is not

just a "pie-in-the-sky" approach as other disciplines—namely, psychology—have been quite successful in this regard. It is commonplace for psychology graduate students to complete a dissertation that uses a quantitative genetic approach. Quantitative genetics allows students to produce dissertations and research that are at the cutting-edge of science without any threat to the discipline itself; quantitative genetics is just another statistical tool in their repertoire. Being well-versed in quantitative genetics may also make them more marketable when hitting the job market. Not only will have they have a larger swath of jobs to select from, but if quantitative genetics does become institutionalized, then departments might actively seek to hire faculty who could teach courses in this area.

The second way in which quantitative genetics can be institutionalized in the social sciences is for faculty members to make a concerted and conscious effort to build quantitative genetics into their departments. This represents a quintessential component to the institutionalization effort, as without it the student-training issues outlined above will fail before they even get off the ground. Exactly how this would be accomplished would rest, in large part, on the faculty of each department. They would have to recognize that quantitative genetics is important and agree to channel resources to building this particular area. This would likely not be an easy sell as it would necessarily mean that resources would be diverted from building or sustaining other areas, areas that are more conventional and traditional. Even so, the amount of resources that would need to be available to build quantitative genetics would likely come primarily in the form of faculty hiring and, to a lesser extent, to the educating of faculty about quantitative genetic approaches.

Another way to build quantitative genetics in social science departments would be to hire senior faculty from other disciplines who have expertise in the area of quantitative genetics. In many ways, this would be no different than attempting to hire established scholars who have a particular expertise, such as a methodologist or a statistician. The difference, of course, would be that the focus would be on a particular type of quantitative approach rather than a quantitative expertise more broadly speaking. Again, for this to occur successfully the culture of the entire department would have to shift, at least somewhat. No longer could there be a focus only on traditional standard social science approaches to analyzing data and no longer could there be a bifurcation between social science research and a more genetically informed (or biosocial) research agenda. Without tearing down these boundaries, how else would it be possible to lure established faculty from one department where their work is valued to another where they are forced to publish their work in a narrow set of journals?

To further embed quantitative genetics in the social sciences, there should be an explicit focus on educating current faculty about quantitative genetics. This would be particularly difficult because faculty might not have any interest in this area and would likely not want to spend the time learning about it (in much the same way that any faculty is unlikely to want to learn techniques and information that is, at best, only tangentially related to their own work). But as more and more quantitative genetic research makes its way into the social sciences, there is a greater need for all faculty to understand these approaches, even if they have no interest in actually calculating these models in their own research. Without this knowledge, they will be unable to understand the logic of the models, point out biases in the

way the models were executed, or fully understand the knowledge that is being generated from these approaches.

How could it be possible to educate faculty members about quantitative genetics? Certainly there is not only one way to educate social scientists about this, but perhaps one of the more effective ways would be to bring in speakers or develop workshops that faculty could attend once a semester. Additionally, there are a couple of workshops/conferences (e.g., Summer Institute in Social-Science Genomics; The Integrating Genetics and the Social Sciences Conference) that are geared towards training social scientists in quantitative genetics. Departments could encourage faculty to attend these conferences and fully fund the costs associated with attending them. Doing so would go a long way towards making social science faculty a lot more proficient in quantitative genetics, as they would become familiar with the strengths, the limitations, and the effective application of these analytical approaches.

While this may be one of the more difficult obstacles to overcome (largely because it rests within each individual faculty member who may have no motivation for learning about quantitative genetics), if successful it would also likely result in the largest gain for the social sciences. Recall that these faculty are the ones who are experts in the social sciences, likely have published widely, and, at the same time, probably have given very little thought to how quantitative genetics would fit in with their research agenda and with the perspectives that they use the most widely. With a new-found understanding of quantitative genetics, they would be in position to make some significant advancements to their own research niche within the social sciences. For instance, they would be able to modify and amend existing theories in a way that accord with findings from quantitative genetic analyses (more on this later in the chapter) and they could also use quantitative genetics to employ newer ways of testing age-old questions. Last, and perhaps most significantly, thinking from a quantitative genetic approach would likely open up newer avenues for collaboration that no longer necessarily break along disciplinary boundaries. These are just a few of the potential many ways that an understanding of quantitative genetics might affect the scholarly output of established social science faculty members.

The third main way to institutionalize quantitative genetics in the social sciences would be through the collection of data that would allow for quantitative genetic analyses to be conducted. As with any topic, theory, or explanation, there must be data available in order for social scientists to test it. The same holds true when it comes to quantitative genetics. To be sure, there are certainly datasets available that are of interest to social scientists studying a broad range of phenotypes. Table ??, for example, lists some of the most widely used and available datasets that can be analyzed by scientists interested in quantitative genetic analyses.

While these datasets are quite useful and contain a great deal of information that is of interest to social scientists, they are also limited in that they are typically not tailor made for the testing of theories and ideas that are unique to each discipline and even to each researcher. This problem is no different than the shortcomings that are inherent with using

any other secondary dataset, wherein social scientists are forced to repurpose the data and use measures that are not ideal and samples that may not match perfectly with who and what they are hoping to study. This is a large part of the reason why social scientists will skip over secondary data analysis in lieu of collecting their own data. In doing so, there is complete control over the variables and measures that are included, the sample of respondents that is included, and the time period during which the data are collected. At the same time, there is a trade-off in that primary data collection is time consuming and quite costly. Moreover, the data tend to focus on a narrower range of variables, the sample size is usually comparatively smaller, and the data are typically not nationally representative.

Even despite these drawbacks, there is a considerable amount of primary data collection among social scientists. The exact goals of each data collection vary widely, but there is one commonality that cuts across virtually all of them: there is little effort devoted to collecting data that is able to be analyzed via quantitative genetic analyses. Rarely do social scientists collect information from more than one sibling per household, there is usually not a concerted effort to oversample twin pairs, and it is rare for these samples to collect biological data that could be used to genotype the respondents. In short, there is a general lack of effort to take into account genetics or to model the genetic architecture to phenotypes.

This is somewhat surprising given the tremendous amount of research that has emphasized the role of genetic influences on most phenotypes of interest to social scientists. What is more is that the National Institute of Health has recently made a push for the collection of genetic data in social science datasets. There are RFPs that will not even consider for funding proposals that focus only on social science samples and thus the funding of any large-scale sample now virtually requires the integration of some type of genetic information. A similar process has played out with psychology, wherein there have been calls by large-scale funding agencies to understand the nexus between biology/neurobiology and personality. In short, outside of the social science departments, there has been a tremendous push for the collection of data that marries genetics and mainstream social science research. To date, however, this push has been met with resistance as most data collection by social scientists does not include genetic information even though doing so is no longer cost prohibitive. Keep in mind, too, that most college campuses have a biology and/or genetics department that can help with certain aspects of data collection.

Collecting data that is genetically sensitive and that could be analyzed in a quantitative genetic framework by social scientists in a way that maps with their research interests is of utmost importance. Doing so would add significantly to the knowledge base, as it would allow experts in particular social science niches to test out ideas that are guided by a genetic and/or biosocial framework. Heretofore, such ideas either would not have been generated or, if they had been, there would not have been a way to test them. With the availability of data, these ideas could be tested and the amount of new findings that would emerge would likely excel at a rapid pace. In addition, it would also allow graduate students to have access to a much broader array of genetic data that they could use for their dissertations and individual research projects. After all, if the other pieces of the "institutionalization puzzle" fall into place, then many more students will be interested in, and have training

with, quantitative genetic analyses. The data would simply allow them to pursue their research interests. Last, even for social scientists who are not interested in focusing upon genetics per se, these data would still be tantamount as it would allow for them to control for genetic confounding. The end result would be the reporting of parameter estimates that would be much more reliable, more accurate, and significantly more defensible. Against this backdrop, the looming (and perhaps rhetorical) question would be: are there are any disadvantages to the primary collection of genetic data?

12.2 Theoretical Implications

Quantitative genetics in and of itself is nothing more than a technique or set of techniques that can be used to produce statistical output and to test ideas or answer research questions. Indeed, quantitative genetics has produced reams of statistical output which have been published in thousands of studies. All too often, these findings are interpreted independently of theory and the link between social science theories and quantitative genetic findings is often insulated from each other completely. This is a serious limitation as quantitative genetics can be used to test theories, amend theories, and create new theories. Given that theory is really at the heart of the social sciences, below we list four potential ways that quantitative genetics can guide and inform social science theories.

First, quantitative genetics allows for the direct testing of $G \times E$s. The importance of $G \times E$s for social science theories cannot be overstated. Understanding the role of $G \times E$s for the outcome of interest can help to identity which environments matter for which people and why some environments matter more for some people, but not for others. Typically, this type of variation is either ignored or simply chalked up to error. $G \times E$s, however, will allow for directly modeling this type of variation. As an example, consider the role of neighborhoods on human development. Although neighborhood-level conditions are often tied to a wide range of developmental outcomes (e.g., delinquency, educational status, etc.), there tends to be tremendous variation in response to neighborhood conditions, with only a relatively small number of all children and youth exposed to disadvantaged neighborhoods displaying maladaptive phenotypes. Existing theories of neighborhoods rarely, if ever, attempt to explain why there is so much variation which essentially deflates the effects of neighborhoods. If these theories focused more on $G \times E$s, then it might be possible to determine why there is variation in response to neighborhoods and, more importantly, to identify who is most likely to be affected by neighborhood conditions. Without doing so, the understanding of how, and in what ways, neighborhoods matter for human development remains incomplete. Keep in mind that $G \times E$s can be applied to all environments that are thought to cause human phenotypes. Successfully integrating $G \times E$s into existing theories would increase the explanatory power of the theories and would also help to provide a much more exhaustive understanding of the development of human phenotypes.

Second, social science theories focus on how differential exposure to environments is linked to differential outcomes for individuals. Despite this concerted focus on differential environ-

mental exposure, there tends to be relatively little concern about what causes differential exposure to environments. For instance, why are some children reared by abusive parents, why are some youth embedded within antisocial peer groups, and why are some families faced with poverty? For the most part, differential exposure to environments is typically taken as a given and/or assumed to be a largely random process. In some cases, selection into environments is modeled, but this is usually only an afterthought and/or accomplished by including an incomplete set of covariates. Rather than assuming exposure to environments is random, or controlling for selection factors sloppily, quantitative genetics is able to directly model the potential reasons and explanations for differential exposure to environments via rGE. Recall that *r*GE essentially decomposes the variance in environments into a genetic, nonshared environmental, and shared environmental components. The findings thus reveal the extent to which genetic and/or environmental influences are able to account for differential exposure to environments.

This is a particularly important (and heretofore overlooked) aspect of social science theories. The results have two particularly noteworthy implications. First off, the results of *r*GE models can provide a great deal of information about avenues for successful intervention to reduce or eliminate exposure to certain environments. For instance, knowing that environmental exposure is largely grafted by genetic influences would lead to very different ways of intervening versus learning that environmental exposure is largely structured by shared environmental factors. Without fully understanding the causes of environmental exposure, there is relatively little that can be done to affect exposure to such environments. And, given that much of social science research is focused on negative environments that are linked to negative outcomes, it stands to reason that trying to eliminate exposure to such environments would be critically important in reducing the incidence of negative outcomes. Secondly, understanding the factors that cause exposure to environments can help to provide a more specified statistical model when testing cause-and-effect hypotheses about environments and phenotypes. Findings from *r*GE analyses can provide information about the self-selection into environments and the causes of environmental variation. If the causes of environmental variation overlap with the causes of phenotypic variation, then those shared causes of variation would need to be accounted for in the statistical models in order to begin to make claims about causality. Failure to do so would leave open the possibility that the findings are misspecified and completely spurious. Taken together, *r*GE research is quite critical when it comes to developing defensible statistical models that are capable of testing and establishing causal connections between environments and phenotypes.

Third, quantitative genetic research is one approach that can be used to help actually falsify theories. The social sciences have a bad habit of failing to falsify theories. Theories that lack much empirical support are rarely (if ever) discarded. Rather, they are analyzed with different samples, tested with different measurement approaches, or scrutinized with a different statistical technique. Variation in findings are used as calls for why additional tests of the theory are needed. The end result is that theories rarely die because the imprecise nature of social science means that it is possible that these theories could be correct, at least in some of their propositions. What is perhaps even more interesting to consider is that the theories that attempt to explain the same phenomenon—and that are entirely incompatible

with the other theory—are also not falsified. This is particularly telling because two theories that attempt to explain the same phenotype, but that are incompatible, coexist without either being ruled out. As a result, students and scholars are forced to learn, memorize, and regurgitate theories, even when at least one of these theories is wrong.

Part of the reason for failing to falsify theories is because there are not hard-and-fast propositions that can be easily tested. Usually, the hypotheses that flow from social science theories are fuzzy, making it easy for advocates of the theory to argue that the tests of the theory do not really match the essence of the theory. However, quantitative genetics provides a unique opportunity to falsify theories because most social science theories either ignore the role of genetics or argue outright that genes do not matter for phenotypes. These are really factual issues that have a correct or incorrect answer. If a theory posits that genes do not matter and the results of an ACE model show that genes are involved, then the theory is incorrect as stated. The methods described in this book also show how genetic confounding can be taken into account to test for environmental causation. If these methods are used and the posited environment does not have an effect on the outcome of interest, then theory is incorrect. Similarly, theories that do not specifically make the distinction between shared and nonshared environments are at-risk for being incorrect if these environments have differential effects on behavior. The key point to bear in mind is that quantitative genetics is an objective set of tools that can be used to test theories in a way that would allow for definitive statements regarding whether the theory is correct or incorrect as currently stated. If used in this capacity, then the falsification of theories—one of the bedrocks of science—may once again become a current practice rather than an outdated ideal.

Fourth, and lastly, quantitative genetics can be used to help create entirely new theories of social behavior. There have been thousands of studies examining the genetic and environmental architecture to virtually every behavior studied by social scientists as well as the environments that usually identified as being salient to their development. To date, there has been little systematic effort designed to use these findings to build new theories of social behavior. There are, however, enough consistent findings from these studies to be able to do just that—that is, to create theories based around the findings of quantitative genetics studies. Doing so does not mean that existing theories have to be completely eliminated; rather, they can be used as the scaffolding from which the findings of quantitative genetics can be built upon. The possibilities of theoretical development that revolve around the findings from quantitative genetics are virtually limitless and likely would result in a greater understanding of the etiology of most social behaviors.

12.3 Consilience

In 1998, E. O. Wilson published the highly influential and national bestselling book, *Consilience: The Unity of Knowledge*, in which he makes the call that knowledge from different fields of study in the social sciences and humanities should be pooled together. Consilience, according to Wilson, can be thought of as the unity of knowledge, wherein findings from

divergent fields of study converge on the same pattern of results. These consistent findings, then, can be summarized in principles and ultimately integrated into unified explanations that have application across fields of study. Rather than having disciplinary boundaries, with research from one field of study failing to penetrate and inform other fields of study, consilience focuses on the integration of findings across disciplines. Seen in this way, knowledge produced by one field of study can be linked together with knowledge produced by other fields of study. In doing so, explanations of phenomenon are informed by multiple lines of inquiry in a unified framework, allowing us to potentially span Tabery's (2014) "explanatory divide".

Wilson's argument for consilience rests, in large part, on the findings that have been generated in relation to the hard sciences, particularly biology, neurobiology, and genetics. He argues that the research coming out of these disciplines has shown that biological processes are intimately involved in most of the domains studied by social scientists. Particular emphasis is placed on the brain as it is the hub of virtually everything and links together these somewhat disparate bodies of knowledge. For instance, genes and systems of genes are responsible for building the brain. These genes—and the brain which they are responsible for creating—have evolved in such a way as to promote behaviors, including social behaviors, as well as virtually every other human phenotype. In a nutshell, then, genes are responsible for creating the neurobiological substrates that are responsible for creating human dispositions, dispositions that studied in large part by social scientists. When viewed in this way, it is clear that virtually all aspects of human development—ranging from genotype, biochemical processes, neurobiology, environmental input, and ultimately behavioral phenotypes—are linked together in such a way that in order to understand completely the outcome (i.e., phenotype) all other parts of the equation must be understood fully, too. As a result, perhaps the only real way to accomplish this is through consilience—that is, the "jumbling together" of findings from all of these relevant disciplines.

Since the publication of Wilson's book nearly two decades ago, the case can be made that there certainly has been some progress made in terms of consilience. Where biological, genetic, and biosocial explanations were almost nonexistent in the 1990s and early 2000s in most social science departments, today there is at least a small presence of them across fields of study. Moreover, some universities have advocated making "cluster hires" that incorporate professors from different disciplines, but who focus on the same (or a similar) phenotype. Multidisciplinary collaboration is promoted more today than perhaps ever in the past and these collaborative efforts have resulted in increasing our knowledge base. Even with all of these changes, a perusal of academic departments on any university campus or a glance through the articles published in disciplinary-specific journals quickly reveals that sharp boundaries continue to exist among them. In some disciplines these demarcations are stronger than they are in others.

From our viewpoint, the integration of quantitative and behavioral genetics approaches will ultimately promote consilience in the social sciences. Quantitative and behavioral genetics can be used in virtually every discipline to study virtually every outcome of interest. The findings, then, can easily be linked together or, at the very least, provide the scaffolding

from which to begin to build theories and explanations of human phenotypes. The use of quantitative genetics, moreover, will provide sort of a common language between various disciplines.

Upon close inspection it becomes readily apparent that quantitative and behavioral genetics can be thought of as the anchor to the idea of consilience. Quantitative and behavioral genetics encompasses the influences of environments, neurobiology, personality, and genetics, to name just a few. It makes no assumptions about the causes of phenotypes, but rather is data driven. It uses a common language that speaks to scholars across fields of study. It focuses on moderation, mediation, and longitudinal development over time. Quantitative genetic approaches have been used in virtually every field of study ranging from economics and sociology to criminology and psychiatry and so the seeds have already been planted in most field of inquiry. Indeed, there is perhaps no other analytical approach that has been used so widely and frequently as quantitative and behavioral genetics.

What is also particularly interesting—and what really gets at the heart of consilience—is that the findings across studies that use quantitative genetic approaches are remarkably consistent—precisely what is supposed to occur with consilience. The findings are so consistent, in fact, that they have been summarized in what are called the "three laws of behavioral genetics" (Turkheimer, 2000, p. 160). The laws are as follows:

1. All human behavioral traits are heritable.
2. The effect of being raised in the same family is smaller than the effects of genes.
3. A substantial portion of the variation in complex human behavioral traits is not accounted for by the effect of genes or families.

These laws are quite impressive as they are supported by thousands of studies that cut across almost every phenotype ever studied (Polderman et al., 2015). Indeed, one would be hard-pressed to locate any other set of findings in the social sciences that are so robust that they could be developed into laws. These laws conform almost perfectly with Wilson's view of consilience that highlights the fact that human nature can be accounted for by a series of laws or principles. In our view, quantitative genetics could be—indeed should be—the anchor to these collaborative approaches as it really is the key by which E.O. Wilson's vision of consilience could be obtained.