Methodology for Genetic Studies of Twins and Families

Michael C. Neale
Department of Psychiatry
Virginia Institute for Psychiatric and Behavioral Genetics
Virginia Commonwealth University

Hermine H. M. Maes
Department of Human Genetics
Virginia Institute for Psychiatric and Behavioral Genetics
Virginia Commonwealth University

Kluwer Academic Publishers B.V.
Dordrecht, The Netherlands

CONTRIBUTING AUTHORS

Lindon J. Eaves	*Department of Human Genetics* *Medical College of Virginia*
John K. Hewitt	*Department of Human Genetics* *Medical College of Virginia*
Joanne M. Meyer	*Department of Human Genetics* *Medical College of Virginia*
Michael C. Neale	*Department of Human Genetics* *Medical College of Virginia*
Hermine H. M. Maes	*Institute of Physical Education* *Catholic University Leuven*
Dorret I. Boomsma	*Department of Experimental Psychology* *Free University, Amsterdam*
Conor V. Dolan	*Department of Psychology* *University of Amsterdam*
Peter C. M. Molenaar	*Department of Psychology* *University of Amsterdam*
Karl G. Jöreskog	*Department of Statistics* *University of Uppsala*
Nicholas G. Martin	*Queensland Institute of Medical* *Research*
Lon R. Cardon	*Institute for Behavioral Genetics* *University of Colorado*
David W. Fulker	*Institute for Behavioral Genetics* *University of Colorado*
Andrew C. Heath	*Department of Psychiatry* *Washington University School of Medicine*

Contents

List of Tables

List of Figures

Chapter 1

The Scope of Genetic Analyses

1.1 Introduction and General Aims

This book has its origin in a week-long intensive course on methods of twin data analysis taught between 1987 and 1997 at the Katholieke Universiteit of Leuven in Belgium, the University of Helsinki, Finland, and the Institute for Behavioral Genetics, Boulder, in Colorado. Our principal aim here is to help those interested in the genetic analysis of individual differences to realize that there are more challenging questions than simply "Is trait X genetic?" or "What is the heritability of X?" and that there are more flexible and informative methods than those that have been popular for more than half a century. We shall achieve this goal primarily by considering those analyses of data on twins that can be conducted with the Mx program. There are two main reasons for this restriction: 1) the basic structure and logic of the twin design is simple and yet can illustrate many of the conceptual and practical issues that need to be addressed in any genetic study of individual differences; 2) the Mx program is well-documented, freely available for personal computers and Unix workstations, and can be used to apply all of the basic ideas we shall discuss. We believe that the material to be presented will open many new horizons to investigators in a wide range of disciplines and provide them with the tools to begin to explore their own data more fruitfully.

The four main aims of this introductory chapter are:

1. to identify some of the scientific questions which have aroused the curiosity of investigators and led them to develop the approaches we describe

2. to trace part of the intellectual tradition that led us to the approach we are to present in this text

3. to outline the overall logical structure of the approach

4. to accomplish all of these with the minimum of statistics and mathematics.

Before starting on what we are going to do, however, it is important to point out what we are not going to cover. There will be almost nothing in this book about detecting the contribution of individual loci of large effect against the background of other genetic and environmental effects ("segregation analysis"). In contrast to the first edition, there will be a chapter on linkage analysis concerning the location on the genome of individual genes of major effect, if they exist. These issues have been treated extensively elsewhere (see e.g., Ott, 1985, Sham, 1998, Lange, 1997, Lynch & Walsh, 1998) — often to the exclusion of issues that may still turn out to be equally important, such as those outlined in this chapter. When the history of genetic epidemiology is written, we believe that the approaches described here will be credited with revealing the naivete of many of the simple assumptions about the action of genes and environment that are usually made in the search for single loci of large effect. Our work may thus be seen in the context of exploring those parameters of the coaction of genes and environment which are frequently not considered in conventional segregation and linkage analysis.

1.2 Heredity and Variation

Genetic epidemiology is impelled by three basic questions:

1. Why isn't everyone the same?

2. Why are children like their parents?

3. Why aren't children from the same parents all alike?

These questions address variation within individuals and covariation between relatives. As we shall show, covariation between relatives can provide useful information about variation within individuals.

1.2.1 Variation

In this section we shall examine the ubiquity of variability, and its distinction from mean levels in populations and sub-populations.

Variation is Everywhere

Almost everything that can be measured or counted in humans shows variation around the mean value for the population. Figure 1.1 shows the pattern of variation for self-reported weight (lb.) in a U.S. sample. The observation that individuals differ is almost universal and covers the entire spectrum of measurable traits,

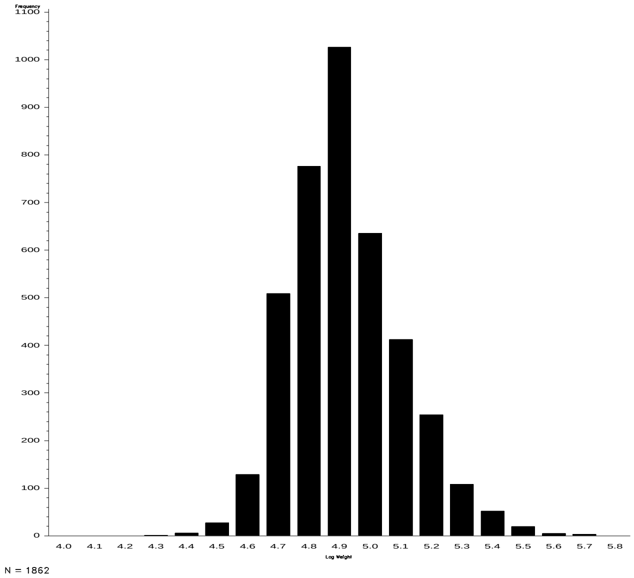


Figure 1.1: Variability in self reported weight in a sample of US twins.

whether they be physical such as stature or weight, physiological such as heart rate or blood pressure, or psychological such as personality, abilities, mental health, or attitudes. The methods we shall describe are concerned with explaining why these differences occur.

Beyond the *a priori* Approach

As far as possible, the analyses we use are designed to be agnostic about the causes of variation in a particular variable. Unfortunately, the same absence of *a priori* bias is not always found among our scientific peers! A referee once wrote in a report on a manuscript describing a twin study:

> *It is probably alright to use the twin study to estimate the genetic contribution to variables which you know are genetic like stature and weight, and it's probably alright for things like blood pressure. But it certainly can't be used for behavioral traits which we know are environmental like social attitudes!*

Such a crass remark nevertheless serves a useful purpose because it illustrates an important principle which we should strive to satisfy, namely to find methods that are *trait-independent*; that is, they do not depend for their validity on investigators *a priori* beliefs about what causes variation in a particular trait. Such considerations may give weight to choosing one study design rather than another, but they cannot be used to decide whether we should believe our results when we have them.

Biometrical Genetical and Epidemiological Approaches

Approaches that use genetic manipulation, natural or artificial, to uncover latent (i.e. unmeasured) genetic and environmental causes of variation are sometimes called *biometrical genetical* (see e.g. Mather and Jinks, 1982). The methods may be contrasted to the more conventional ones used in individual differences, chiefly in the areas of psychology, sociology and epidemiology. The conventional approaches try to explain variation in one set of measures (the *dependent variables*) by references to differences in another set of measures (*independent variables*). For example, the risk for cardiovascular and lung diseases might be assumed to be dependent variables, and cigarette smoking, alcohol use, and life stress independent variables. A fundamental problem with this "epidemiological approach" is that its conclusions about causality can be seriously misleading. Erroneous inferences would be made if both the dependent and independent variables were caused by the same latent genetic and environmental variables (see e.g., Chapters 6 and 10).

Not Much Can Be Said About Means

It is vital to remember that almost every result in this book, and every conclusion that others obtain using these methods, relate to the causes of human differences, and may have almost nothing to do with the processes that account for the development of the mean expression of a trait in a particular population. We are necessarily concerned with what makes people vary around the mean of the population, race or species from which they are sampled. Suppose, for example, we were to find that differences in social attitudes had a very large genetic component of variation among U.S. citizens. What would that imply about the role of culture in the determination of social attitudes? It could imply several things. First, it might mean that culture is so uniform that only genetic effects are left to account for differences. Second, it might mean that cultural changes are adopted so rapidly that environmental effects are not apparent. A trivial example may make this clear. It is possible that understanding the genetic causes of variation in stature among humans may identify the genes responsible for the difference in stature between humans and chimpanzees, but it is by no means certain. Neither would a demonstration that all human variation in stature was due to the environment lead us to assume that the differences between humans and chimpanzees were not genetic. This point is stressed because, whatever subsequent genetic research on population and species differences might establish, there is no necessary connection between what is true of the typical human and what causes variation around the central tendency. For this reason, it is important to avoid such short-hand expressions as "height is genetic" when really we mean "individual differences in height are mainly genetic."

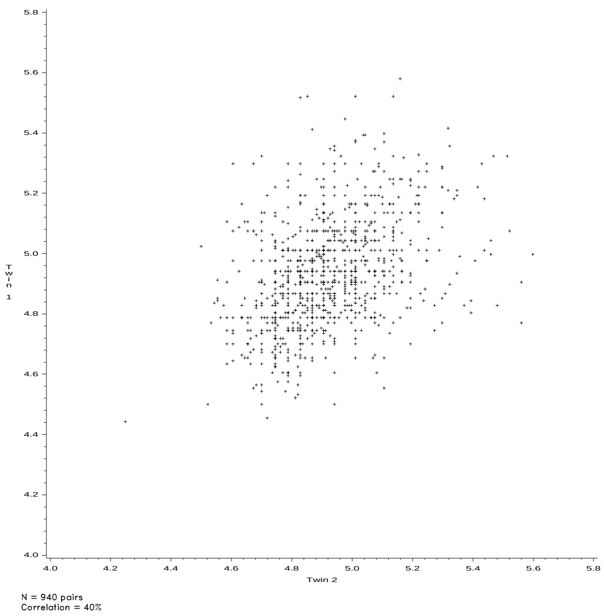
Variation and Modification

What has been said about means also extends to making claims about intervention. The causes of variation that emerge from twin and family studies relate to a particular population of genotypes at a specific time in its evolutionary and cultural history. Factors that change the gene frequencies, the expression of gene effects, or the frequencies of the different kinds of environment may all affect the outcome of our studies. Furthermore, if we show that genetic effects are important, the possibility that a rare but highly potent environmental agent is present cannot entirely be discounted. Similarly, a rare gene of major effect may hold the key to understanding cognitive development but, because of its rarity, accounts for relatively little of the total variation in cognitive ability. In either case, it would be foolhardy to claim too much for the power of genetic studies of human differences. This does not mean, however, that such studies are without value, as we shall show. Our task is to make clear what conclusions are justified on the basis of the data and what are not.

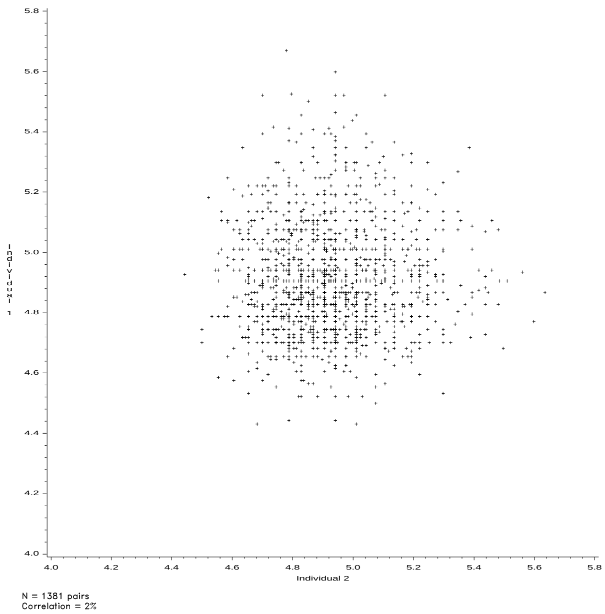
1.2.2 Graphing and Quantifying Familial Resemblance

Look at the two sets of data shown in Figure 1.2. The first part of the figure is a scatterplot of measurements of weight in a large sample of non-identical (fraternal, dizygotic, DZ) twins. Each cross in the diagram represents a single twin pair. The second part of the figure is a scatterplot of pairs of completely unrelated people from the same population. Notice how the two parts of the figure differ. In the unrelated pairs the pattern of crosses gives the general impression of being circular; in general, if we pick a particular value on the X axis (first person's weight), it makes little difference to how heavy the second person is on average. This is what it means to say that measures are *uncorrelated* — knowing the score of the first member of a pair makes it no easier to predict the score of the second and *vice-versa*. By comparison, the scatterplot for the fraternal twins (who are related biologically to the same degree as brothers and sisters) looks somewhat different. The pattern of crosses is slightly elliptical and tilted upwards. This means that as we move from lighter first twins towards heavier first twins (increasing values on the X axis), there is also a general tendency for the average scores of the second twins (on the Y axis) to increase. It appears that the weights of twins are somewhat correlated. Of course, if we take any particular X value, the Y values are still very variable so the correlation is not perfect. The *correlation coefficient* (see Chapter 2) allows us to quantify the degree of relationship between the two sets of measures. In the unrelated individuals, the correlation turns out to be 0.02 which is well within the range expected simply by chance alone if the measures were really independent. For the fraternal twins, on the other hand, the correlation is 0.44 which is far greater than we would expect in so large a sample if the pairs of measures were truly independent.

The data on weight illustrate the important general point that relatives are



(a)



(b)

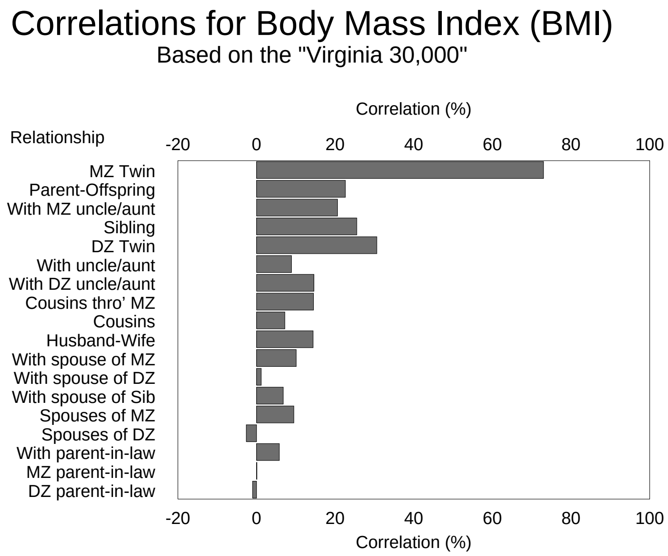
Figure 1.2: Two scatterplots of weight in: a) a large sample of DZ twin pairs, and b) pairs of individuals matched at random.

usually much more alike than unrelated individuals from the same population. That is, although there are large individual differences for almost every trait than can be measured, we find that the trait values of family members are often quite similar. Figure 1.3 gives the correlations between relatives in large samples of nuclear families for body mass index (BMI), and conservatism. One simple way of interpreting the correlation coefficient is to multiply it by 100 and treat it as a percentage. The correlation ($\times$ 100) is the "percentage of the total variation in a trait which is caused by factors shared by members of a pair." Thus, for example, our correlation of 0.44 for the weights of DZ twins implies that, of all the factors which create variation in weight, 44% are factors which members of a DZ twin pair have in common. We can offer a similar interpretation for the other kinds of relationship. A problem becomes immediately apparent. Since the DZ twins, for example, have spent most of their lives together, we cannot know whether the 44% is due entirely to the fact that they shared the same environment *in utero*, lived with the same parents after birth, or simply have half their genes in common — and we shall never know until we can find another kind of relationship in which the degree of genetic similarity, or shared environmental similarity, is different.

Figure 1.4 gives a scattergram for the weights of a large sample of identical (monozygotic, MZ) twins. Whereas DZ twins, like siblings, on average share only half their genes, MZ twins are genetically identical. The scatter of points is now much more clearly elliptical, and the 45° tilt of the major axis is especially obvious. The correlation in the weights in this sample of MZ twins is 0.77, almost twice that found for DZ's. The much greater resemblance of MZ twins, who are expected to have completely identical genes establishes a strong *prima facie* case for the contribution of genetic factors to differences in weight. One of the tasks to be addressed in this book is how to interpret such differential patterns of family resemblance in a more rigorous, quantitative, fashion.

1.2.3 Within Family Differences

At a purely anecdotal level, when parents hear about the possibility that genes create differences between people, they will sometimes respond "Well, that's pretty obvious. I've raised three sons the same way and they've all turned out differently." At issue here is not whether their conclusions are soundly based on their data, so much as to indicate that not all variation is due to factors that family members share in common. No matter how much parents contribute genetically to their children and, it seems, no matter how much effort they put into parenting, a large part of the outcome appears beyond their immediate control. That is, there are large differences even within a family. Some of these differences are doubtless due to the environment since even identical twins are not perfectly alike. Figure 1.5 is a bar chart of the (absolute) weight differences within pairs of twins. The darker, left-hand column of each pair gives the percentage of the DZ sample falling in the indicated range of differences, and the lighter, right-hand column shows the corre-



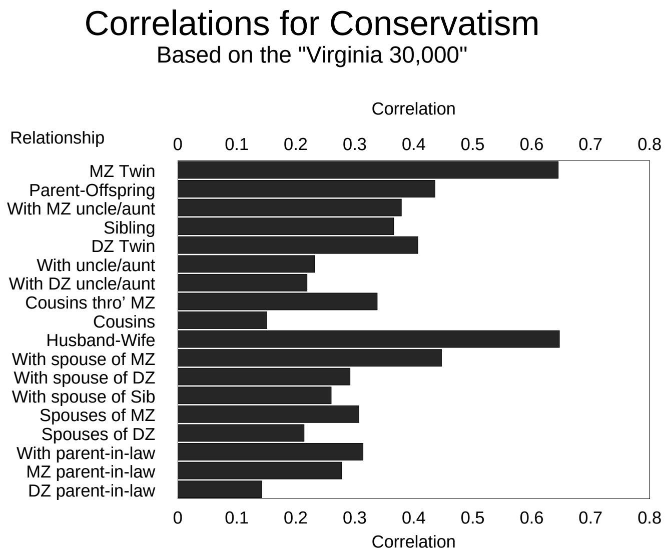


Figure 1.3: Correlations for body mass index (weight/height^2) and conservatism between relatives. Data were obtained from large samples of nuclear families ascertained through twins.

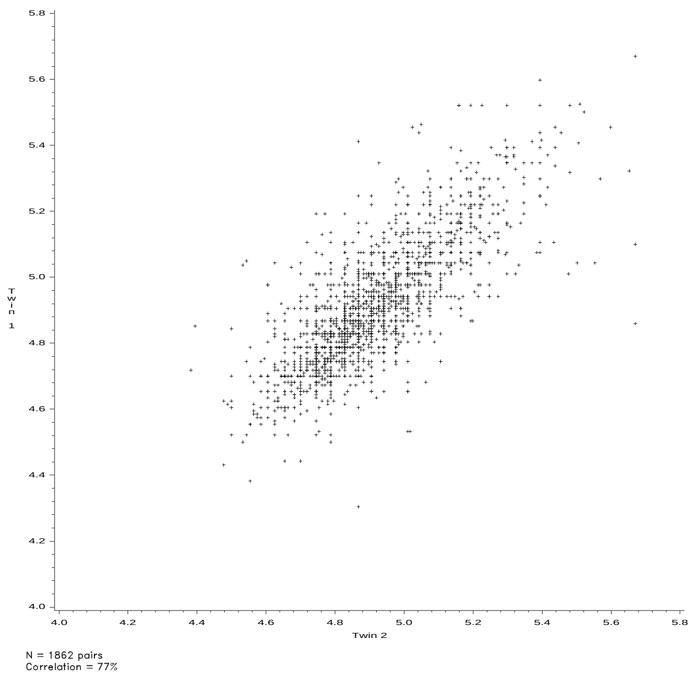


Figure 1.4: Scatterplot of weight in a large sample of MZ twins.

sponding percentages for MZ pairs. For MZ twins, these differences must be due to factors in the environment that differ even within pairs of genetically identical individuals raised in the same home. Obviously the differences within DZ pairs are much greater on average. The known mechanisms of Mendelian inheritance can account for this finding since, unlike MZ twins, DZ twins are not genetically identical although they are genetically related. DZ twins represent a separate sample from the genetic pool "cornered" by the parents. Thus, DZ twins will be correlated because their particular parents have their own particular selection of genes from the total gene pool in the population, but they will not be genetically identical because each DZ twin, like every other sibling in the same family, represents the result of separate, random, meioses[1] in the parental germlines.

[1] meiosis is the process of gametogenesis in which either sperm or ova are formed

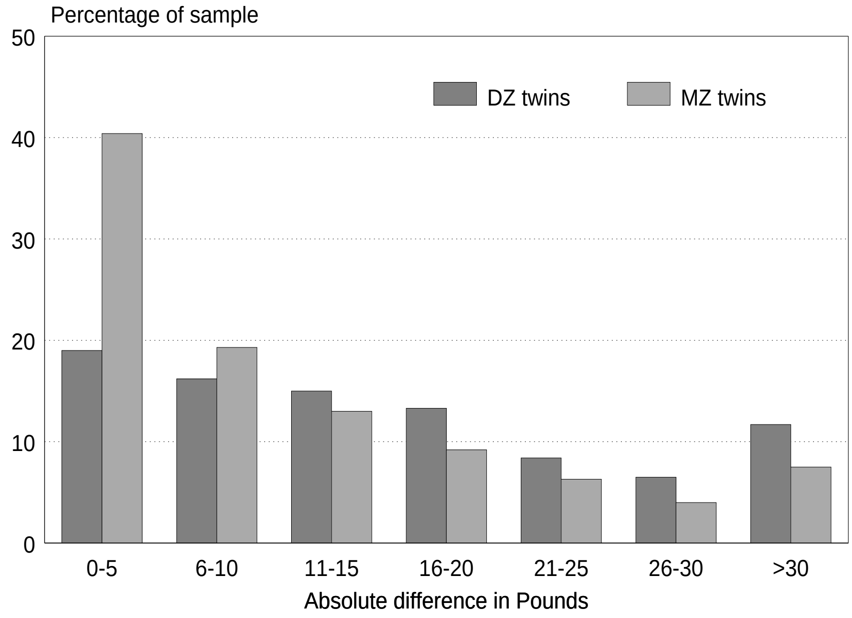


Figure 1.5: Bar chart of absolute differences in weight within MZ and DZ twin pairs.

1.3 Building and Fitting Models

As long as we study random samples of unrelated individuals, our understanding of what causes the differences we see will be limited. The total population variation is simply an aggregate of all the various components of variance. One practical approach to the analysis of variation is to obtain several measures of it, each known to reflect a different proportion of genetic and environmental components of the differences. Then, if we have a model for how the effects of genes and environment contribute differentially to each distinct measure of variation, we can solve to obtain estimates of the separate components. Figure 1.6 shows the principal stages in this process. There are two aspects: *theory* and *data.* The *model* is a formal, in our case mathematical, statement which mediates between the logic of the theory and the reality of the data. Once a model is formulated consistently, the predictions implied for different sets of data can be derived by a series of elementary math-

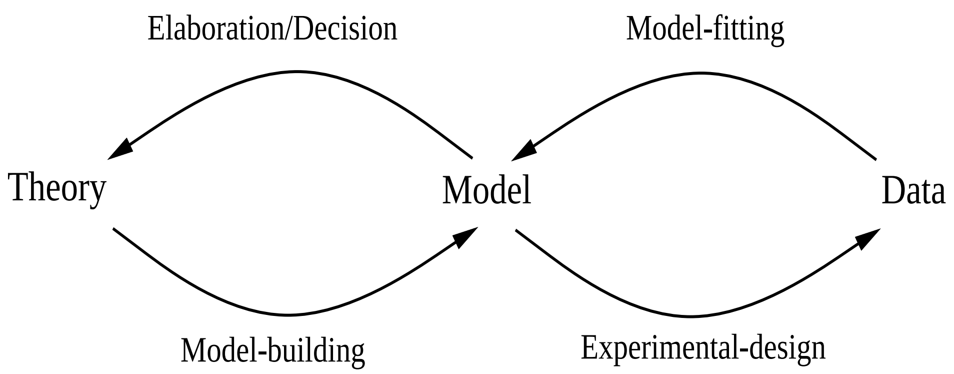


Figure 1.6: Diagram of the interrelationship between theory, model and empirical observation.

ematical operations. *Model building* is the task of translating the ideas of theory into mathematical form. A large part of this book is devoted to a discussion of model building. Inspection of the model, sometimes aided by computer simulation (see Chapters 7 and **??**), may suggest appropriate study designs which can be used to generate critical data to test some or all parts of a current model. The statistical processes of *model fitting* allow us to compare the predictions of a particular model with the actual observations about a given population. If the model fails, then we are forced to revise all or some of our theory. If, on the other hand, the model fits then we cannot know it is "right" in some ultimate sense. However, we might now wish to base new testable conjectures on the theory in order to enlarge the scope of observations it can encompass.

1.4 The Elements of a Model: Causes of Variation

No model is built in isolation. Rather it is built upon a foundation of what is either already known or what might be a matter for fertile conjecture. Part of the difficulty, but also the intrinsic appeal, of genetic epidemiology is the fact that it seeks either to distinguish between major sets of theoretical propositions, or to integrate them into an overall framework. From biology, and especially through knowledge of genetics, we have a detailed understanding of the intricacies of gene expression. From the behavioral and social sciences we have strong proposals about the importance of the environment, especially the social environment, for the development of human differences. One view of our task is that it gives a common conceptual and mathematical framework to both genetic and environmental theories so that we may decide which, if any, is more consistent with the facts in particular cases.

1.4.1 Genetic Effects

A complete understanding of genetic effects would need to answer a series of questions:

1. How important are genetic effects on human differences?
2. What kinds of action and interaction occur between gene products in the pathways between genotype and phenotype?
3. Are the genetic effects on a trait consistent across sexes?
4. Are there some genes that have particularly outstanding effects when compared to others?
5. Whereabouts on the human gene map are these genes located?

Questions 4 and 5 are clearly very important, but are not the immediate concern of this book. On the other hand, we shall have a lot to say about 1, 2, and 3. It is arguable that we shall not be able to understand 4 and 5 adequately, if we do not have a proper appreciation of these other issues.

The importance of genes is often expressed relative to all the causes of variation, both genetic and environmental. The proportion of variation associated with genetic effects is termed the *broad heritability.* However, the complete analysis of genetic factors does not end here because, as countless experiments in plant and animal genetics have shown (well in advance of modern molecular genetics; see e.g., Mather and Jinks, 1982), genes can act and interact in a variety of ways before their effects on the phenotype appear.

Geneticists typically distinguish between *additive* and *non-additive* genetic effects (these terms will be defined more explicitly in Chapter 3). These influences have been studied in detail in many non-human species using selective breeding experiments, which directly alter the frequencies of particular genotypes. In such experiments, the bulk of genetic variation is usually explained by additive genetic effects. However, careful studies have shown two general types of non-additivity that may be important, especially in traits that have been subject to strong directional selection in the wild. The two main types of genetic non-additivity are *dominance* and *epistasis.*

The term dominance derives initially from Mendel's classical experiments in which it was shown that the progeny of a cross between two pure breeding lines often resembled one parent more than the other. That is, an individual who carries different alleles at a locus (the *heterozygote*) is not exactly intermediate in expression between individuals who are pure breeding (*homozygous*) for the two alleles. While dominance describes the interaction between alleles at the same locus, *epistasis* describes the interaction between alleles at different loci.

Epistasis is said to occur whenever the effects of one gene on individual differences depend on which genotype is expressed at another locus. For example, suppose that at locus A/a individuals may have genotype AA, Aa or aa, and at locus B/b genotype BB, Bb or bb[2]. If the difference between individuals with genotype AA and those with genotype aa depends on whether they are BB or bb, then there would be *additive* × *additive* epistatic interactions. Experimental studies have shown a rich variety of possible epistatic interactions depending on the number and effects of the interacting loci. However, their detailed resolution in humans is virtually impossible unless we are fortunate enough to be examining a trait which is influenced by a small number of known genetic loci. Therefore we acknowledge their conceptual importance and model them if they are identified. Failure to take non-additive genetic effects into account may be one of the main reasons studies of twins give different heritability estimates from studies of adoptees and nuclear families (Eaves *et al.*, 1992; Plomin *et al.*, 1991).

As studies in genetic epidemiology become larger and better designed, it is becoming increasingly clear that there are marked sex differences in gene expression. An important factor in establishing this view has been the incorporation of unlike-sex twin pairs in twin studies (Eaves *et al.*, 1990). However, comparison of statistics derived from any relationship of individuals of unlike sex with those of like sex would yield a similar conclusion (see Chapter 9). We shall make an important distinction between two types of *sex-limited gene expression.* In the simpler case, the same genes affect both males and females, but their effects are consistent across sexes and differ only by some constant multiple over all the loci involved. We shall refer to this type of effect as *scalar sex-limitation.* In other cases, however, we shall discover that genetic effects in one sex are not just a constant multiple of their effects in the other. Indeed, even though a trait may be measured in exactly the same way in males and females, it may turn out that quite different genes control its expression in the two sexes. A classic example would be the case of chest-girth since at least some of the variation expressed in females may be due to loci that, while still present in males, are only expressed in females. In this case we shall speak of *non-scalar sex-limitation.* None of us likes the term very much, but until someone suggests something better we shall continue to use it!

1.4.2 Environmental Effects

Paradoxically, one of the greatest benefits of studies that can detect and control for genetic effects is the information they can provide about the sources of environmental influence. We make an important distinction between identifying which are the best places to look for specific environmental agents and deciding what those specific agents are. For example, it may be possible to show that variation in diastolic blood pressure is influenced by environmental effects shared by family members long before it is possible to demonstrate that the salient environmental factor is the amount of sodium in the diet. We make a similar distinction between estimating the overall contribution of genetic effects and identifying specific

[2] This notation is described more fully in Chapter 3.

loci that account for a significant fraction of the total genetic variation. Using some of the methods we shall describe later in this book it may indeed be possible to estimate the contribution of specific factors to the environmental component of variation (see Chapter 10). However, using the *biometrical genetical* approach which relies only on the complex patterns of family resemblance, it is possible to make some very important statements about the structure of the environment in advance of our ability to identify the specific features of the environment that are most important. Although the full subtlety for analyzing the environment cannot be achieved with data on twins alone, much less on twins reared together, it is nevertheless possible to make some important preliminary statements about the major sources of environmental influence which can provide a basis for future studies.

We may conceive of the total environmental variation in a trait as arising from a number of sources. The first major distinction we make is between environmental factors that operate within families and those which create differences between families. Sometimes the environment within families is called the *unique environment* or the *specific environment* or the *random environment*. Different authors may refer to it as V_E, V_{SE}, E_1, E_W or e^2, but the important thing is to understand the concept behind the symbols. The within-family environment refers to all those environmental influences which are so random in their origin, and idiosyncratic in their effects, as to contribute to differences between members of the same family. They are captured by Hamlet's words from the famous 'to be or not to be' soliloquy:

> *...the slings and arrows of outrageous fortune.*

The within-family environment will even contribute to differences between individuals of the same genotype reared in the same family. Thus, the single most direct measure of their impact is the variation within pairs of MZ twins reared together.

Obviously, if a large proportion of the total variation is due to environmental differences within families we might expect to look more closely at the different experiences of family members such as MZ twins in the hope of identifying particular environmental factors. However, we have to take account of a further important distinction, namely that between "long-term" and "short-term" environmental effects, even within families. If we only make a single measurement on every individual in a study of MZ twins, say, we cannot tell whether the observed phenotypic differences between members of an MZ twin pair are due to some lasting impact of an early environmental trauma, or due to much more transient differences that influence how the twins function on the particular occasion of measurement. Many of the latter kinds of influence are captured by the concept of "unreliability" variance in measurement theory. There is, of course, no hard and fast distinction between the two sources of variation because how far one investigator is prepared to treat short-term fluctuations as "unreliability" is largely a matter of his or her frame of reference. In depression, which is inherently episodic, short term fluctuations in behavior may point to quite specific environmental factors that trigger specific episodes (see, e.g., Kendler *et al.*, 1986). The main thing to realize is that what a single cross-sectional study assigns to the "within-family" environment may or may not be resolved into specific non-trivial environmental causes. How far to proceed with the analysis of within-family environment is a matter for the judgement and ingenuity of the particular investigator, aided by such data on repeated measures as he or she may gather.

The between-family environment would seem to be the place that many of the conceptually important environmental effects might operate. Any environmental factors that are shared by family members will create differences between families and make family members relatively more similar. The environment between families is sometimes called the *shared environment,* the *common environment* or just the *family environment.* Sometimes it is represented by the symbols E_2, EB, EC, CE, c^2 or V_{EC}. Again, all these symbols denote the same underlying processes.

In twin studies, the shared environment is expected to contribute to the correlation of both MZ and DZ twins as long as they are reared together. Just as we distinguish short-term and long-term effects of the within-family environment, so it is conceptually important to note that the effects of the shared environment may be more or less permanent and may persist even if family members are separated later in life, or they may be relatively transient in that they are expressed as long as individuals are living together, perhaps as children with their parents, but are dissipated as soon as the source of shared environmental influence is removed. Such effects can be detected by comparing the analyses of different age groups in a cross-sectional study, or by tracing changes in the contribution of the shared environment in a longitudinal genetic study (see Chapter **??**).

It is a popular misconception that studies of twins reared together can offer no insight about the effects of the shared environment. As we shall see in the following chapters, this is far from the case. Large samples of twins reared together can provide a strong *prima facie* case for the importance of between-family environmental effects that account for a significant proportion of the total variance. The weakness of twin studies, however, is that the various sources of the shared environment cannot be discriminated. It is nevertheless essential for our understanding of what the twin study can achieve, to recognize some of the reasons why this design can never be a "one-shot," self-contained investigation and why investigators should be open to the possibility of significant extensions of the twin study (see Chapter **??**).

The environmental similarity between twins may itself be due to several distinct sources whose resolution would require more extensive studies. First, we may identify the environmental impact of parents on their children. That is, part of the common environment effect in twins, can be traced to the fact that children learn from their parents. Formally, this implies that some aspect of variation in the maternal or paternal phenotypes (or both) creates part of the environmental variation between pairs of children. An excellent starting point for exploring some of these effects is the extension of the classical twin study to include data on the parents of twins (see Chapter **??**). In principle, we might find that parents do not contribute equally to the shared family environment. The effect of mothers on the environment of their offspring is usually called the "maternal effect" and the impact of fathers is called the "paternal effect." Although these effects can be resolved by parent-offspring data, they cannot be separated from each other as long as we only have twins in the sample.

Following the terms introduced by Cavalli-Sforza and Feldman (1981), the environmental effects of parent on child are often called *vertical cultural transmission* to reflect the fact that non-genetic information is passed vertically down the family tree from parents to children. However, the precise effects of the parental environment on the pattern of family resemblance depend on which aspect of the parental phenotype is affecting the offspring's environment. The shared environment of the children may depend on the same trait in the parents that is being measured in the offspring. For example, the environment that makes offspring more or less conservative depends directly on the conservatism of their parents. In this case we normally speak of "phenotype-to-environment ('P to E')" transmission. It is quite possible, however, that part of the shared environment of the offspring is created by aspects of parental behavior that are different from those measured in the children, although the two may be somewhat correlated. Thus, for example, parental income may exercise a direct effect on offspring educational level through its effect on duration and quality of schooling. Another example would be the effect of parental warmth or protectiveness on the development of anxiety or depression in their children. In this case we have a case of *correlated variable transmission.* Haley, Jinks and Last (1981) make a similar distinction between the "one character" and "two character" models for maternal effects. The additional feature of the parental phenotype may or may not be measured in either parents or children. When such additional traits are measured in addition to the trait of primary interest we will require *multivariate genetic models* to perform the data analysis properly. Some simple examples of these methods will be described in later chapters. Two extreme examples of correlated variable transmission are where the variable causing the shared environment is:

1. an index purely of the environmental determinants of the phenotype — "environment-to-environment ('E to E')" transmission

2. purely genetic — "genotype-to-environment ('G to E')" transmission.

Although we can almost never claim to have a direct measure of the genotype for any quantitative trait, the latter conception recognizes that there may be a *genetic environment* (see e.g. Darlington, 1971), that is, genetic differences between some members of a population may be part of the environment of others. One consequence of the genetic environment is the seemingly paradoxical notion that *different genetic relationships also can be used to tease out certain important aspects of the environment.* For example, the children of identical twins can be used to provide a test of the environmental impact of the maternal genotype on the phenotypes of their children (see e.g., Nance and Corey, 1976). A concrete example of this phenomenon would be the demonstration that a mother's genes affect the birthweight of her children.

Although researchers in the behavioral sciences almost instinctively identify the parents as the most salient feature of the shared environment, we must recognize that there are other environmental factors shared by family members that do not depend directly on the parents. There are several factors that can create residual (non-parental) shared environmental effects. First, there may be factors that are shared between all types of offspring, twin or non-twin; these may be called *sibling shared environments.* Second, twins may share a more similar pre-and postnatal environment than siblings simply because they are conceived, born and develop at the same time. This additional correlation between the environments of twins is called the *special twin environment* and is expected to make both MZ and DZ twins more alike than siblings even in the absence of genetic effects. It is important to note that even twins separated at birth share the same pre-natal environment, so a comparison of twins reared together and apart is only able to provide a simple test of the *post-natal* shared environment[3].

A further type of environmental partition, the *special MZ twin environment* is sometimes postulated to explain the fact that MZ twins reared together are more correlated than DZ twins. This is the most usual environmental explanation offered as an alternative to genetic models for individual differences because the effects of the special MZ environment will tend to mimic those of genes in twins reared together. It is because of concern that genetic effects may be partly confounded with any special MZ twin environments that we stress the importance of thinking beyond the twin study to include other relationships. It becomes increasingly difficult to favor a purely non-genetic explanation of MZ twin similarity when the genetic model is able to predict the correlations for a rich variety of relationships from a few fairly simple principles. Since the special twin environment, however, would increase the correlation of MZ twins, its effects may often resemble those of non-additive genetic effects (dominance and epistasis) in models for family resemblance.

1.4.3 Genotype-Environment Effects

It has long been realized that the distinction we make for heuristic purposes between "genotype" and "environment" is an approximation which ignores several processes that might be important in human populations. Three factors defy the simple separation of genetic and environmental effects, but are likely to be of potential significance from what we know of the way genes operate in other species, and from the logical consequences of the grouping of humans into families of self-determining individuals who share both genes and environment in common.

[3]Twins born serially by embryo implantation are currently far too rare for the purposes of statistical distinction between pre- and post-natal effects!

The factors we need to consider are:

1. assortative mating
2. genotype-environment covariance (CovGE, or genotype-environment correlation, CorGE)
3. genotype $\times$ environment interaction (G$\times$E).

Each of these will be discussed briefly.

Assortative Mating.

Any non-random pairing of mates on the basis of factors other than biological relatedness is subsumed under the general category of *assortative mating.* Mating based on relatedness is termed *inbreeding,* and will not be examined in this book. We discuss assortative mating under the general heading of genotype-environmental effects for two main reasons. First, when assortment is based on some aspect of the phenotype, it may be influenced by both genetic and environmental factors. Second, assortative mating may affect the transmission, magnitude, and correlation of both genetic and environmental effects.

In human populations, the first indication of assortative mating is often a correlation between the phenotypes of mates. Usually, such correlations are positive. Positive assortment is most marked for traits in the domains of education, religion, attitudes, and socioeconomic status. Somewhat smaller correlations are found in the physical and cognitive domains. Mating is effectively random, or only very slightly assortative, in the personality domain. We are not aware of any replicated finding of a significant negative husband-wife correlation, with the exception of gender!

Assortative mating may not be the sole source of similarity between husband and wife — social interaction is another plausible cause. *A priori,* we might expect social interaction to play a particularly important role in spousal resemblance for habits such as cigarette smoking and alcohol consumption. Two approaches are available for resolving spousal interaction from strict assortative mating. The first depends on tracing the change in spousal resemblance over time, and the second requires analyzing the resemblance between the spouses of biologically related individuals (see Heath, 1987). Although the usual treatment of assortative mating assumes that spouses choose one another on the basis of the trait being studied (*primary phenotypic assortment*), we should understand that this is only one model of a process that might be more complicated in reality. For example, mate selection is unlikely to be based on an actual psychological test score. Instead it is probably based on some related variable, which may or may not be measured directly. If the variable on which selection is based is something that we have also measured, we call it *correlated variable assortment.* If the correlated trait is not measured directly we have *latent variable assortment.* In the simplest case, the latent variable may simply be the *true value* of trait of which the actual measure is just a more or less unreliable index. We then speak of *phenotypic assortment with error*.

Once we begin to consider latent variable assortment, we recognize that the latent variable may be more or less genetic. If the latent variable is due entirely to the social environment we have one form of *social homogamy* (e.g., Rao *et al.*, 1974). We can conceive of a number of intriguing mechanisms of latent variable assortment according to the presumed causes of the latent variable on which mate selection is based. For example, mating may be based on one or more aspects of the phenotypes of relatives, such as parents' incomes, culinary skills, or siblings' reproductive history. In all these cases of correlated or latent variable assortment, mate selection may be based on variables that are more reliable indices of the genotype than the measured phenotype. This possibility was considered by Fisher (1918) in what is still the classical treatment of assortative mating.

Clearly, the resolution of these various mechanisms of assortment is beyond the scope of the conventional twin study, although multivariate studies that include the spouses of twins, or the parents and parents-in-law of twins may be capable of resolving some of these complex issues (see, e.g., Heath *et al.*, 1985).

Even though the classical twin study cannot resolve the complexities of mate selection, we have to keep the issue in mind all the time because of the *effects* of assortment on the correlations between relatives, including twins. When mates select partners like themselves phenotypically, they are also (indirectly) choosing people who resemble themselves genetically and culturally. As a result, positive phenotypic assortative mating increases the genetic and environmental correlations between relatives. Translating this principle into the context of the twin study, we will find that assortative mating tends to increase the similarity of DZ twins relative to MZ twins. As we shall see, in twins reared together, the genetic effects of assortative mating will artificially inflate estimates of the family environmental component. This means, in turn, that estimates of the genetic component based primarily on the difference between MZ correlations and DZ correlations will tend to be biased downwards in the presence of assortative mating.

Genotype-Environment Correlation.

Paradoxically, the factors that make humans difficult to study genetically are precisely those that make humans so interesting. The experimental geneticist can control matings and randomize the uncontrolled environment. In many human societies, for better or for worse, consciously or unconsciously, people likely decide for themselves on the genotype of the partner to whom they are prepared to commit the future of their genes. Furthermore, humans are more or less free living organisms who spend a lot of time with their relatives. If the problem of mate selection gives rise to fascination with the complexities of assortative mating, it is the fact that individuals create their own environment and spend so much time with their relatives that generates the intriguing puzzle of genotype-environment correlation.

As the term suggests, *genotype-environment correlation (CorGE)* refers to the fact that the environments that individuals experience may not be a random sample of the whole range of environments but may be caused by, or correlated with, their genes. Christopher Jencks (1972) spoke of the "double advantage" phenomenon in the context of ability and education. Individuals who begin life with the advantage of genes which increase their ability relative to the average may also be born into homes that provide them with more enriched environments, having more money to spend on books and education and being more committed to learning and teaching. This is an example of positive CorGE. Cattell (1963) raised the possibility of negative CorGE by formulating a principle of "cultural coercion to the biosocial norm." According to this principle, which has much in common with the notion of *stabilizing selection* in population genetics, individuals whose genotype predisposes them to extreme behavior in either direction will tend to evoke a social response which will "coerce" them back towards the mean. For example, educational programs that are designed specifically for the average student may increase achievement in below average students while attenuating it in talented pupils.

Many taxonomies have been proposed for CorGE. We prefer one that classifies CorGE according to specific detectable consequences for the pattern of variation in a population (see Eaves *et al.*, 1977). The first type of CorGE, *genotype-environment autocorrelation* arises because the individual creates or evokes environments which are functions of his or her genotype. This is the "smorgasbord" model which views a given culture as having a wide variety of environments from which the individual makes a selection on the basis of genetically determined preferences. Thus, an intellectually gifted individual would invest more time in mentally stimulating activities. An example of possible CorGE from a different context is provided by an ethological study of 32 month-old male twins published a number of years ago (Lytton, 1977). The study demonstrated that parent-initiated interactions with their twin children are more similar when the twins are MZ rather than DZ. Of course, like every other increased correlation in the environment of MZ twins, it may not be clear whether it is truly a result of a treatment being elicited by genotype rather than simply a matter of identical individuals being treated more similarly. That is, the direction of causation is not clear.

Insofar as the genotypes of individuals create or elicit environments, cross-sectional twin studies will not be able to distinguish the ensuing CorGE from any other effects of the genes. That is, positive CorGE will increase estimates of all the genetic components of variance and negative CorGE will decrease them. However, we will have no direct way of knowing which genetic effects act *directly* on the phenotype and which result from the action of environmental variation caused initially by genetic differences. In this latter case, the environment may be considered as part of the "extended phenotype" (see Dawkins, 1982). If the process we describe were to accumulate during behavioral development, positive CorGE would lead to an increase in the relative contribution of genetic factors with age, but a constant genetic correlation across ages (see Chapter **??**). However, finding this pattern of developmental change would not necessarily imply that the actual mechanism of the change is specifically genotype-environment autocorrelation.

The second major type of CorGE is that which arises because the environment in which individuals develop is provided by their biological relatives. Thus, one individual's environment is provided by the phenotype of someone who is genetically related. Typically, we think of the correlated genetic and environmental effects of parents on their children. For example, a child who inherits the genes that predispose to depression may also experience the pathogenic environment of rejection because the tendency of parents to reject their children may be caused by the same genes that increase risk to depression. As far as the offspring are concerned, therefore, a high genetic predisposition to depression is correlated with exposure to an adverse environment because both genes and environment derive originally from the parents. We should note (i) that this type of CorGE can occur only if parent-offspring transmission comprises both genetic factors and vertical cultural inheritance, and (ii) that the CorGE is broken in randomly adopted individuals since the biological parents no longer provide the salient environment. Adoption data thus provide one important test for the presence of this type of genotype-environment correlation.

Although most empirical studies have focused on the parental environment as that which is correlated with genotype, parents are not the only relatives who may be influential in the developmental process. Children are very often raised in the presence of one or more siblings. Obviously, this is always the case for twin pairs. In a world in which people did not interact socially, we would expect the presence or absence of a sibling, and the unique characteristics of that sibling, to have no impact on the outcome of development. However, if there is any kind of social interaction, the idiosyncrasies of siblings become salient features of one another's environment. Insofar as the effect of one sibling or twin on another depends on aspects of the phenotype that are under genetic control, we expect there to be a special kind of *genetic environment* which can be classified under the general category of *sibling effects.* When the trait being measured is partly genetic, and also responsible for creating the sibling effects, we have the possibility for a specific kind of CorGE. This CorGE arises because the genotype of one sibling, or twin, is genetically correlated with the phenotype of the other sibling which is providing part of the environment. When above average trait values in one twin tend to increase trait expression in the other, we speak of *cooperation* effects (Eaves, 1976b) or *imitation* effects (Carey, 1986b). An example of imitation effects would be any tendency of deceptive behavior in one twin to reinforce deception in the other. The alternative social interaction, in which a high trait value in one sibling tends to act on the opposite direction in the other, produces *competition* or *contrast* effects We might expect such effects to be especially marked in environments in which there is competition for limited resources. It has sometimes been argued that contrast effects are an important source of individual differences in extraversion (see Eaves *et al.*, 1989) with the more extraverted twin tending to engender introversion in his or her cotwin and vice-versa.

Sibling effects typically have two kinds of detectable consequence. First, they produce differences in trait mean and variance as a function of sibship size and density. One of the first indications of sibling effects may be differences in variance between twins and singletons. Second, the genotype-environment correlation created by sibling effects depends on the biological relationship between the socially interacting individuals. So, for example, the CorGE is greater in pairs of MZ twins because each twin is reared with an cotwin of identical genotype. If there are co-operation (imitation) effects we expect the CorGE to make the total variance of MZ twins significantly greater than that of DZ's, which in turn would exceed that of singletons (Eaves, 1976b). Competition (contrast) effects will tend to make the MZ variance less than that of DZ's. Other effects ensue for the covariances between relatives, as discussed in Chapter 8. Sibling effects may conceivably be reciprocal, if siblings influence each other, or non-reciprocal, if an elder sibling, for example, is a source of social influence on a younger sibling.

Genotype × Environment Interaction

The interaction of genotype and environment ("G × E") must always be distinguished carefully from CorGE. Genotype-environment correlation reflects a non-random distribution of environments among different genotypes. "Good" genotypes get more or less than their fair share of "good" environments. By contrast, G × E interaction has nothing to do with the distribution of genetic and environmental effects. Instead, it relates to the actual way genes and environment affect the phenotype. G × E refers to the genetic control of sensitivity to differences in the environment. The old adage "sauce for the goose is sauce for the gander" describes a world in which G × E is absent, because it implies that the same environmental treatment has the same positive or negative effect regardless of the genotype of the individual upon whom it is imposed.

An obvious example of G × E interaction is that of inherited disease resistance. Genetically susceptible individuals will be free of disease as long as the environment does not contain the pathogen. Resistant individuals will be free of the disease even in a pathogenic environment. That is, changing the environment by introducing the pathogen will have quite a different impact on the phenotype of susceptible individuals than on resistant ones. More subtle examples may be the genetic control of sensitivity to the pathogenic effects of tobacco smoke or genetic differences in the effects of sodium intake on blood pressure.

The analysis of G × E in humans is extremely difficult in practice because of the difficulty of securing large enough samples to detect effects that may be small compared with the main effects of genes and environment. Studies of G × E in experimental organisms (see, e.g., Mather and Jinks, 1982) illustrate a number of issues which are also conceptually important in thinking about G×E in humans. We consider these briefly in turn.

The genes responsible for sensitivity to the environment are not always the same as those that control average trait values. For example, one set of genes may control overall liability to depression and a second set, quite distinct in their location and mode of action, may control whether individuals respond more or less to stressful environments. Another way of thinking about the issue is to consider measurements made in different environments as different traits which may or may not be controlled by the same genes. By analogy with our earlier discussion of sex-limitation, we distinguish between "scalar" and "non-scalar" G $\times$ E interaction. When the same genes are expressed consistently at all levels of a salient environmental variable so that only the amount of genetic variance changes between environments, we have "scalar genotype $\times$ environment interaction." If, instead of, or in addition to, changes in genetic variance, we also find that different genes are expressed in different environments we have "non-scalar G $\times$ E."

G $\times$ E interaction may involve environments that can be measured directly or whose effects can be inferred only from the correlations between relatives. Generally, our chances of detecting G $\times$ E are much greater when we can measure the relevant environments, such as diet, stress, or tobacco consumption. The simplest situation, which we shall discuss in Chapter 9, arises when each individual in a twin pair can be scored for the presence or absence of a particular environmental variable such as exposure to severe psychological stress. In this case, twin pairs can be divided into those who are concordant and those discordant for environmental exposure and the data can be tested for different kinds of G $\times$ E using relatively simple methods.

One "measurable" feature of the environment may be the phenotype of an individual's parent. A problem frequently encountered, however, is the fact that many measurable aspects of the environment, such as smoking and alcohol consumption, themselves have a genetic component so that the problems of mathematical modelling and statistical analysis become formidable. If we are unable to measure the environment directly, our ability to detect and analyze G $\times$ E will depend on the same kinds of data that we use to analyze the main effects of genes and environment, namely the patterns of family resemblance and other, more complex, features of the distribution of trait values in families. Generally, the detection of any interaction between genetic effects and unmeasured aspects of the between-family environment will require adoption data, particularly separated MZ twins. Interaction between genes and the within-family environment will usually be detectable only if the genes controlling sensitivity are correlated with those controlling average expression of the trait (see, e.g., Jinks and Fulker, 1970).

1.5 Relationships between Variables

Many of the critics of the methods we are to describe argue that, for twin studies at least, the so-called traditional methods such as taking the difference between the MZ and DZ correlations and doubling it as a heritability estimate give much the same answer as the more sophisticated methods taught here. In the final analysis, it must be up to history and the consumer to decide, but in our experience there are several reasons for choosing the methods presented here. First, as we have already shown, the puzzle of human variation extends far beyond testing whether genes play any role in variation. The subtleties of the environment and the varieties of gene action call for methods that can integrate many more types of data and test more complex hypotheses than were envisioned fifty or a hundred years ago. Only a model building/model fitting strategy allows us to trace the implications of a theory across all kinds of data and to test systematically for the consistency of theory and observation. But even if the skeptic is left in doubt by the methods proposed for the interpretation of variables considered individually, we believe that the conventional approaches of fifty years ago pale utterly once we want to analyze the genetic and environmental causes of correlation between variables.

The genetic analysis of multiple variables will occupy many of the succeeding chapters, so here it is sufficient to preview the main issues. There are three kinds of "multivariate" questions which are generic issues in genetic epidemiology, although we shall address them in the context of the twin study. Each is outlined briefly.

1.5.1 Contribution of Genes and Environment to the Correlation between Variables

The question of what causes variables to correlate is the usual entry point to multivariate genetic analysis. Students of genetics have long been familiar with the concept of pleiotropy, i.e., that one genetic factor can affect several different phenotypes. Obviously, we can imagine environmental advantages and insults that affect many traits in a similar way. Students of the psychology of individual differences, and especially of factor analysis, will be aware that Spearman introduced the concept of the "general intelligence factor" as a latent variable accounting for the pattern of correlations observed between multiple abilities. He also introduced an empirical test (the method of tetrad differences) of the consistency between his general factor theory and the empirical data on the correlations between abilities. Such factor models however, only operate at the descriptive phenotypic level. They aggregate into a single model genetic and environmental processes which might be quite separate and heterogeneous if only the genetic and environmental causes of inter-variable correlation could be analyzed separately. Cattell recognized this when he put forward the notion of "fluid" and "crystallized" intelligence. The former was dependent primarily on genetic processes and would tend to increase the correlation between measures that index genetic abilities. The latter was determined more by the content of the environment (an "environmental mold" trait) and would thus appear as loading more on traits that reflect the cultural environment. An analysis of multiple symptoms of anxiety and depression by Kendler *et al.* (1986) illustrates very nicely the point that the pattern of genetic and environmental effects on multiple measures may differ very markedly. They showed that twins' responses to a checklist of symptoms reflected a single underlying genetic dimension which influenced symptoms of both anxiety and depression. By contrast, the effects of the environment were organized along two dimensions ("group factors") — one affecting only symptoms of anxiety and the other symptoms of depression. More recently, this finding has been replicated with psychiatric diagnoses (**?**; **?**), which suggests that the liability to either disorder is due to a single common set of genes, while the specific expression of that liability as either anxiety or depression is a function of what kind of environmental event triggers the disorder in the vulnerable person. Such insights are impossible without methods that can analyze the correlations between multiple measures into their genetic and environmental components.

1.5.2 Analyzing Direction of Causation

Students of elementary statistics have long been made to recite "correlation does not imply causation" and rightly so, because a premature assignment of causality to a mere statistical association could waste scientific resources and do actual harm if treatment were to be based upon it. However, one of the goals of science is to analyze complex systems into elementary processes which are thought to be *causal* or more *fundamental* and, when actual experimental intervention is difficult, it may be necessary to look to the nexus of intercorrelations among measures for clues about causality.

The claim that correlation does not imply causality comes from a fundamental indeterminacy of any general model for the correlation between a single pair of variables. Put simply, if we observe a correlation between A and B, it can arise from one or all of three processes: A causing B (denoted $A \longrightarrow B$), B causing A, or latent variable C causing A and B. A general model for the correlation between A and B would need constants to account for the strength of the causal connections between A and B, B and A, C and A, C and B. Clearly, a single correlation cannot be used to determine four unknown parameters.

When we have more than two variables, however, matters may look a little different. It may now become possible to exclude some causal hypotheses as clearly inconsistent with the data. Whether or not this can be done will depend on the complexity of the causal nexus being analyzed. For example, a pattern of correlations of the form $r_{AC} = r_{AB} \times r_{BC}$ would support one or other of the causal sequences $A \longrightarrow B \longrightarrow C$ or $C \longrightarrow B \longrightarrow A$ in preference to orders that place A or C in the middle.

The fact that causality implies *temporal priority* has been used in some applications to advocate a longitudinal strategy for its analysis. One approach is the *cross-lagged panel study* in which the variables A and B are measured at two points in time, t_0 and t_1. If the correlation of A at t_0 with B at t_1 is greater than the correlation of B at t_0 with A at t_1, we might give some credence to the causal priority of A over B. Methods for the statistical assessment of such relative priorities are known as "cross-lagged panel analysis" (?) and may assessed within structural equation models (?).

The cross-lagged approach, though strongly suggestive of causality in some circumstances, is not entirely foolproof. With this fact in view, researchers are always on the look-out for other approaches that can be used to test hypotheses about causality in correlational data. It has recently become clear that the cross-sectional twin study, in which multiple measures are made only on one occasion, may, under some circumstances, allow us to test hypotheses about direction of causality without the necessity of longitudinal data. The potential contribution of twin studies to resolving alternative models of causation will be discussed in Chapter ??. At this stage, however, it is sufficient to give a simple insight about one set of circumstances which might lead us to prefer one causal hypothesis over another.

Consider the ambiguous relationship between exercise and body weight. In free-living populations, there is a significant correlation between exercise and body weight. How much of that association is due to the fact that people who exercise use up more calories and how much to the fact that fat people don't like jogging? In the simplest possible case, suppose that we found variation in exercise to be purely environmental (i.e., having no genetic component) and variation in weight to be partly genetic. Then there is no way that the direction of causation can go from body weight to exercise because, if this were the case, some of the genetic effects on body weight would create genetic variation in exercise. In practice, things are seldom that simple. Data are nearly always more ambiguous and hypotheses more complex. But this simple example illustrates that the genetic studies, notably the twin study, may sometimes yield valuable insight about the causal relationships between multiple variables.

1.5.3 Analyzing Developmental Change

Any cross-sectional study is a slice at one time point across the continuing ontogenetic dialogue between the organism and the environment. While such studies help us understand outcomes, they may not tell us much about the process of "becoming". For example, the longitudinal genetic study involving repeated measures of twins may be thought of as a multivariate genetic study in which the multiple occasions of measurement correspond to multiple traits in the conventional cross-sectional study. In the conventional multivariate study we ask such questions as "How much do genes create the correlation between different variables?", so in the longitudinal genetic study we ask "How far do genes (or environment) account for the developmental consistency of behavior?" and "To what extent are there specific genetic and environmental effects expressed at each point in time?". These are but two of a rich variety of questions which can be addressed with the methods we shall describe. One indication of the insight that can ensue from such an approach to longitudinal measures on twins comes from some of the data on cognitive growth obtained in the ground-breaking Louisville Twin Study. In a reanalysis by model fitting methods, Eaves *et al.* (1986) concluded that such data as had been published strongly suggested the involvement of a single common set of genes which were active from birth to adolescence and whose affects persisted and accumulated through time. By contrast, the shared environment kept changing during development. That is, parents who provided a better environment at one age did not necessarily do so at another, even though whatever they did had fairly persistent effects. The unique environment of the individual, however, was age-specific and very ephemeral in its effect. Such a model, based as it was on only that part of the data available in print, may not stand the test of more detailed scrutiny. Our aim here is not so much to defend a particular model for cognitive development as to indicate that a model fitting approach to longitudinal kinship data can lead to many important insights about the developmental process.

1.6 The Context of our Approach

Figure 1.7 summarizes the main streams of the intellectual tradition which converge to yield the ideas and methods we shall be discussing here. The streams divide and merge again at several places. The picture is not intended to be a comprehensive history of statistical or behavioral genetics, so a number of people whose work is extremely important to both disciplines are not mentioned. Rather, it tries to capture the main lines of thought and the "cast of characters" who have been especially influential in our own intellectual development. Not all of us would give the same weight to all the lines of descent.

1.6.1 Early History

To our knowledge, the first use of twin resemblance as a means of resolving alternative hypotheses about the causes of human individual differences appears in 426 A.D. by Augustine of Hippo in Book V of the City of God. Augustine argued that since twins, who were highly correlated in their times of birth, nevertheless had such discrepant life histories, there was little empirical support for planetary influence on human destiny. For Augustine's purpose, it was sufficient that at least some twin pairs showed markedly different life histories, despite being born at the same time. To go beyond testing the astrological hypothesis and use twins to answer the nature-nurture question required recognition of the fact that there are two types of twins, identical and fraternal, and some way of distinguishing between them.

1.6.2 19th Century Origins

Two geniuses of the last century provided the fundamental principles on which much of what we do today still depends. Francis Galton's boundless curiosity,

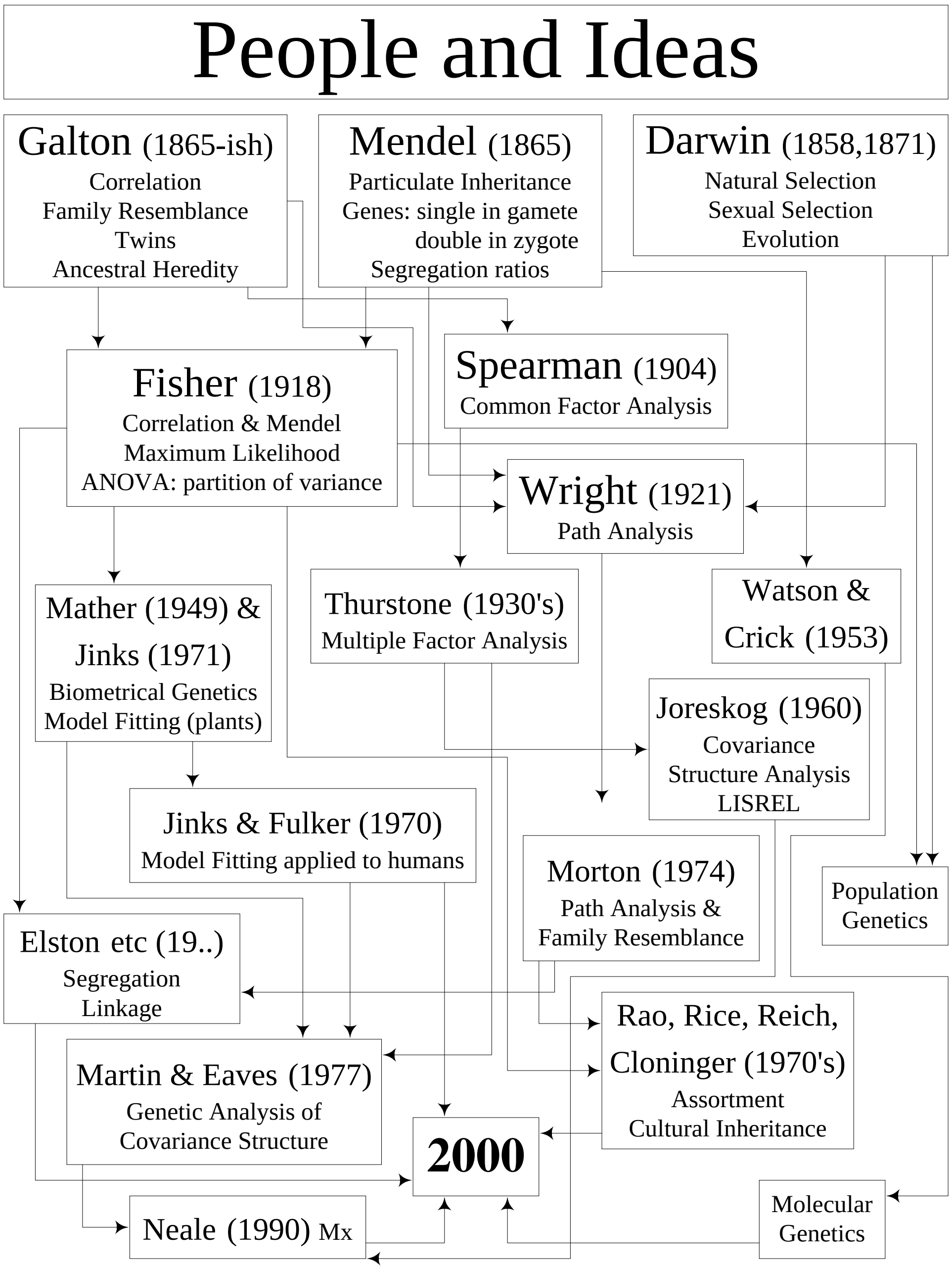


Figure 1.7: Diagram of the intellectual traditions leading to modern mathematical genetic methodology.

ingenuity and passion for measurement were combined in seminal insights and contributions which established the foundations of the scientific study of individual differences. Karl Pearson's three-volume scientific biography of Galton is an enthralling testimony to Galton's fascination and skill in bringing a rich variety of intriguing problems under scientific scrutiny. His *Inquiry into the Efficacy of Prayer* reveals Galton to be a true "child of the Enlightenment" to whom nothing was sacred. To him we owe the first systematic studies of individual differences and family resemblance, the recognition that the difference between MZ and DZ twins provided a valuable point of departure for resolving the effects of genes and culture, the first mathematical model (albeit inadequate) for the similarity between relatives, and the development of the correlation coefficient as a measure of association between variables that did not depend on the units of measurement.

The specificity that Galton's theory of inheritance lacked was supplied by the classical experiments of Gregor Mendel on plant hybridization. Mendel's demonstration that the inheritance of model traits in carefully bred material agreed with a simple theory of particulate inheritance still remains one of the stunning examples of how the alliance of quantitative thinking and painstaking experimentation can predict, in advance of any observations of chromosome behavior or molecular science, the necessary properties of the elementary processes underlying such complex phenomena as heredity and variation.

1.6.3 Genetic, Factor, and Path Analysis

The conflict between those, like Karl Pearson, who followed a Galtonian model of inheritance and those, like Bateson, who adopted a Mendelian model, is well known to students of genetics. Although Pearson appeared to have some clues about how Galton's data might be explained on Mendelian principles, it fell to Ronald Fisher, in 1918, to provide the first coherent and general account of how the "correlations between relatives" could be explained "on the supposition of Mendelian inheritance." Fisher assumed what is now called the *polygenic model,* that is, he assumed the variation observed for a trait such as stature was caused by a large number of individual genes, each of which was inherited in strict conformity to Mendel's laws. By describing the effects of the environment, assortative mating, and non-additive gene action mathematically, Fisher was able to show remarkable consistency between Pearson's own correlations between relatives for stature and a strictly Mendelian mechanism of inheritance. Some of the ideas first expounded by Fisher will be the basis of our treatment of *biometrical genetics* (Chapter 3).

In the same general era we witness the seeds of two other strands of thought which continue to be influential today. Charles Spearman, adopting Galton's idea that a correlation between variables might reflect a common underlying causal factor, began to explore the pattern of correlations between multiple measures of ability. So began the tradition of multivariate analysis which was, for much of psychology at least, embodied chiefly in the method of factor analysis which sought the latent variables responsible for the observed patterns of correlation between multiple variables. The notion of multiple factors, introduced through the work of Thurstone, and the concept of factor rotation to *simple structure,* provided much of the early conceptual and mathematical foundation for the treatment of multivariate systems to be discussed in this book.

Sewall Wright, whose long and distinguished career spans all of the six decades which have seen the explosion of genetics into the most influential of the life sciences, was the founding father of American population genetics. His seminal paper on path analysis, published in 1921 established a parallel stream of thought to that created by Fisher in 1918. The emphasis of Fisher's work lay in the formulation of a mathematical theory which could reconcile observations on the correlation between relatives with a model of particulate inheritance. Wright, on the other hand, was less concerned with providing a theory which could integrate two views of genetic inheritance than he was with developing a method for exploring ways in which different causal hypotheses could be expressed in a simple, yet testable, form. It is not too gross an oversimplification to suggest that the contributions of Fisher and Wright were each stronger where the other was weaker. Thus, Fisher's early paper established an explicit model for how the effects and interaction of large numbers of individual genes could be resolved in the presence of a number of different theories of mate selection. On the other hand, Fisher showed very little interest in the environment, choosing rather to conceive of environmental effects as a random variable uncorrelated between relatives. Fisher's environment is what we have called the "within family" environment, which seems appropriate for the kinds of anthropometric variables that Fisher and his predecessors chose to illustrate the rules of quantitative inheritance. However, it seems a little less defensible, on *a priori* grounds, as a model for the effects of environment on what Pearson (1904) called "the mental and moral characteristics of man" or those habits and lifestyles that might have a significant impact on risk for disease. By contrast, Wright's approach virtually ignored the subtleties of gene action, considering only additive genetic effects and treating them as a statistical aggregate which owed little to the laws of Mendel beyond the fact that offspring received half their genes from their mother and half from their father. On the other hand, Wright's strategy made it much easier to specify familial environmental effects, especially those derived from the social interaction of family members.

1.6.4 Integration of the Biometrical and Path-Analytic Approaches

These different strengths and weaknesses of the traditions derived from Fisher and Wright persisted into the 1970's. The biometrical genetical approach, derived from Fisher through the ground-breaking studies of Kenneth Mather and his student John Jinks established what became known as the "Birmingham School." The emphasis of this tradition was on the detailed analysis of gene action through carefully designed and properly randomized breeding studies in experimental organisms. Except where the environment could be manipulated genetically (e.g., in the study of the environmental effects of the maternal genotype), the biometrical genetical approach treated the environment as a random variable. Even though the environment might sometimes be correlated between families as a result of practical limitations on randomization, it was independent of genotype. Thus, the Birmingham School's initial treatment of the environment in human studies allowed for the partition of environmental components of variance into contributions within families (EW) and between families (EB) but was very weak in its treatment of genotype-environment correlation. Some attempt to remedy this deficiency was offered by Eaves (1976a; 1976b) in his treatment of vertical cultural transmission and sibling interaction, but the value of these models was restricted by the assumption of random mating.

The rediscovery of path analysis in a series of papers by Morton and his coworkers in the early 70's showed how many of the more realistic notions of how environmental effects were transmitted, such as those suggested by Cavalli-Sforza and Feldman (1981), could be captured much better in path models than they could by the biometrical approach. However, these early path models assumed that assortative mating to be based on homogamy for the social determinants of the phenotype. Although the actual mechanism of assortment is a matter for empirical investigation, this strong assumption, being entirely different from the mechanisms proposed by Fisher, precluded an adequate fusion of the Fisher and Wright traditions.

A crucial step was achieved in 1978 and 1979 in a series of publications describing a more general path model by Cloninger, Rice, and Reich which integrated the path model for genetic and environmental effects with a Fisherian model for the consequences of assortment based on phenotype. Since then, the approach of path analysis has been accepted (even by the descendants of the Birmingham school) as a first strategy for analyzing family resemblance, and a number of different nuances of genetic and environmental transmission and mate selection have now been translated into path models. This does not mean that the method is without limitations in capturing non-additive effects of genes and environment, but it is virtually impossible today to conceive of a strategy for the analysis of a complex human trait that does not include path analysis among the battery of techniques to be considered.

1.6.5 Development of Statistical Methods

Underlying all of the later developments of the biometrical-genetical, path-analytic and factor-analytic research programs has been a concern for the statistical problems of estimation and hypothesis-testing. It is one thing to develop models; to attach the most efficient and reliable numerical values to the effects specified in a model, and to decide whether a particular model gives an adequate account of the empirical data, are completely different. All three traditions that we have identified as being relevant to our work rely heavily on the statistical concept of likelihood, introduced by Ronald Fisher as a basis for developing methods for parameter estimation and hypothesis testing. The approach of "maximum likelihood" to estimation in human quantitative genetics was first introduced in a landmark paper by Jinks and Fulker (1970) in which they first applied the theoretical and statistical methods of biometrical genetics to human behavioral data. Essential elements of their understanding were that:

1. complex models for human variation could be simplified under the assumption of polygenic inheritance

2. the goodness-of-fit of a model should be tested before waxing lyrical about the substantive importance of parameter estimates

3. the most precise estimates of parameters should be obtained

4. possibilities exist for specifying and analyzing gene action and genotype $\times$ environment interaction

It was the confluence of these notions in a systematic series of models and methods of data analysis which is mainly responsible for breaking the intellectual gridlock into which human behavioral genetics had driven itself by the end of the 1960's.

Essentially the same statistical concern was found among those who had followed the path analytic and factor analytic approaches. Rao, Morton, and Yee (1974) used an approach close to maximum likelihood for estimation of parameters in path models for the correlations between relatives, and earlier work on the analysis of covariance structures by Karl Jöreskog had provided some of the first workable computer algorithms for applying the method of maximum likelihood to parameter estimation and hypothesis-testing in factor analysis. Guided by Jöreskog's influence, the specification and testing of specific hypotheses about factor rotation became possible. Subsequently, with the collaboration of Dag Sörbom, the analysis of covariance structures became elaborated into the flexible model for Linear Structural Relations (LISREL) and the associated computer algorithms which, over two decades, have passed through a series of increasingly general versions.

The attempts to bring genetic methods to bear on psychological variables naturally led to a concern for how the psychometrician's interest in multiple variables could be reconciled with the geneticist's methods for separating genetic and environmental effects. For example, several investigators (Vandenberg, 1965; Loehlin and Vandenberg, 1968; Bock and Vandenberg, 1968) in the late 1960's began to ask whether the genes or the environment was mainly responsible for the general ability factor underlying correlated measures of cognitive ability. The approaches that were suggested, however, were relatively crude generalizations of the classical methods of univariate twin data analysis which were being superseded by the biometrical and path analytic methods. There was clearly a need to integrate the model fitting approach of biometrical genetics with the factor model which was still the conceptual framework of much multivariate analysis in psychology. In discussion with the late Owen White, it became clear that Jöreskog's analysis of covariance structures provided the necessary statistical formulation. In 1977, Martin and Eaves reanalyzed twin data on Thurstone's Primary Mental Abilities using their own FORTRAN program for a multi-group extension of Jöreskog's model to twin data and, for the first time, used the model fitting strategy of biometrical genetics to test hypotheses, however simple, about the genetic and environmental causes of covariation between multiple variables. The subsequent wide dissemination of a multi-group version of LISREL (LISREL III) generated a rash of demonstrations that what Martin and Eaves had achieved somewhat laboriously with their own program could be done more easily with LISREL (Boomsma and Molenaar, 1986, Cantor, 1983; Fulker *et al.*, 1983; Martin *et al.*, 1982; McArdle *et al*, 1980). After teaching several workshops and applying LISREL to everyday research problems in the analysis of twin and family data, we discovered that it too had its limitations and was quite cumbersome to use in several applications. This led to the development of Mx, which began in 1990 and which has continued throughout this decade. Initially devised as a combination of a matrix algebra interpreter and a numerical optimization package, it has simplified the specification of both simple and complex genetic models tremendously.

In the 1980's there were many significant new departures in the specification of multivariate genetic models for family resemblance. The main emphasis was on extending the path models, such as those of Cloninger *et al.*, (1979a,b) to the multivariate case (Neale & Fulker, 1984; Vogler, 1985). Much of this work is described clearly and in detail by Fulker (1988) . Many of the models described could not be implemented with the methods readily available at the time of writing of the first edition this book. Furthermore, several of the more difficult models were not addressed in the first edition because of the lack of suitable data. Since that time many of the problems of specifying complex models have been solved using Mx, and this edition presents some of these developments. In addition, several research groups have now gathered data on samples large and diverse enough to exploit most of the theoretical developments now in hand.

The collection of large volumes of data in a rich variety of twin studies from around the world in the last ten years, coupled with the rocketing growth in the power of micro-computers, offer an unprecedented opportunity. What were once ground-breaking methods, available to those few who knew enough about statistics and computers to write their own programs, can now be placed in the hands of teachers and researchers alike.

Chapter 2

Data Preparation

2.1 Introduction

By definition, the primary focus of the study of human individual differences is on variation. As we have seen, the covariation between family members can be especially informative about the causes of variation, so we now turn to the statistical techniques used to measure both variation within and covariation between family members. We start by reviewing the calculation of variances and covariances by hand, and then illustrate how one may use pograms such as SAS, SPSS and PRELIS (SAS, 1988; SPSS, 1988; **?**) to compute these summary statistics in a convenient form for use with Mx. Our initial treatment assumes that we have well-behaved, normally-distributed variables for analysis (see Section 2.2). However, almost all studies involve some measures that are certainly not normal because they consist of a few ordered categories, which we call *ordinal* scales. In Section 2.3, we deal with the summary of these cruder forms of measurement, and discuss the concepts of degrees of freedom and goodness-of-fit that arise in this context.

During this decade advances in computer software and hardware have made the direct analysis of raw data quite practical. As we shall see, this method has some advantages over the analysis of summary statistics, especially when there are missing data. Section 2.4 describes the preparation of raw data for analysis with Mx.

2.2 Continuous Data Analysis

Biometrical analyses of twin data often make use of summary statistics that reflect differences, or variability, between and within members of twin pairs. Some early studies used mean squares and products, derived from an analysis of variance (Eaves et al., 1977; Martin and Eaves, 1977; Fulker et al., 1983; Boomsma and

Molenaar, 1987; Molenaar and Boomsma, 1987), but work over the past 15 years has embraced variance-covariance matrices as the summary statistics of choice. This approach, often called *covariance structure analysis*, provides greater flexibility in the treatment of some of the processes underlying individual differences, such as genotype × sex or genotype × environment interaction. In addition, variances and covariances are a more practical data summary for data that include the relatives of twins, such as parents or spouses (Heath et al., 1985). Because of the greater generality afforded by variances and covariances, we focus on these quantities rather than mean squares.

2.2.1 Calculating Summary Statistics by Hand

The variances and covariances used in twin analyses often are computed using a statistical package such as SPSS (SPSS, 1988) or SAS (SAS, 1988), or by PRELIS (**?**). Nevertheless, it is useful to examine how they are calculated in order to ensure a comprehensive understanding of one's observed data. In this section we describe the calculation of means, variances, covariances, and correlations.

Some simulated measurements from 16 MZ and 16 DZ twin pairs are presented in Table 2.1. The observed values in the columns labelled *Twin 1* and *Twin 2* have

Table 2.1: Simulated measurements from 16 MZ and 16 DZ Twin Pairs.

MZ		*DZ*	
Twin 1	*Twin 2*	*Twin 1*	*Twin 2*
3	2	0	1
3	3	2	3
2	1	1	2
1	2	4	3
0	0	3	1
2	2	2	2
2	2	2	2
3	2	1	3
3	3	3	4
2	3	1	0
1	1	1	1
1	1	2	1
4	4	3	3
2	3	3	2
2	1	2	2
1	2	2	2

been selected to illustrate some elementary principles of variation in twins[1].

In order to obtain the summary statistics of variances and covariances for genetic analysis, it is first necessary to compute the average value for a set of measurements, called the *mean*. The mean is typically denoted by a bar over the variable name for a group of observations, for example $\overline{X}$ or $\overline{Twin1}$ or $\overline{Twin2}$. The formula for calculation of the mean is:

$$\begin{aligned}\overline{X} &= \frac{X_1 + X_2 + \cdots + X_n}{n} \\ &= \frac{\sum_{i=1}^{n} X_i}{n}, \end{aligned} \tag{2.1}$$

in which X_i represents the i^{th} observation and n is the total number of observations. In the twin data of Table 2.1, the mean of the measurements on Twin 1 of the MZ pairs is

$$\begin{aligned}\overline{Twin1} &= \frac{3 + 3 + 2 + \cdots + 2 + 2 + 1}{16} \\ &= 32/16 \\ &= 2.0\end{aligned}$$

The mean for the second MZ twin ($\overline{Twin2}$) also is 2.0, as are the means for both DZ twins.

The *variance* of the observations represents a measure of dispersion around the mean; that is, how much, on average, observations differ from the mean. The variance formula for a sample of measurements, often represented as s^2 or V_{MZ} or V_{DZ}, is

$$\begin{aligned}s^2 &= \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} \\ &= \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} \end{aligned} \tag{2.2}$$

We note two things: first, the difference between each observation and the mean is squared. In principle, absolute differences from the mean could be used as a measure of variation, but absolute differences have a greater variance than squared differences (Fisher, 1920), and are therefore less efficient for use as a summary statistic. Likewise, higher powers (e.g. $\sum_{i=1}^{n}(X_i - \overline{X})^4$) also have greater variance. In fact, Fisher showed that the square of the difference is the most informative

[1] These data are for illustration *only*; they would normally be treated as ordinal, not continuous, and would be summarized differently, as described in Section 2.3. Note also that we do not need to have equal numbers of pairs in the two groups.

measure of variance, i.e., it is a *sufficient statistic*. Second, the sum of the squared deviations is divided by $n-1$ rather than n. The denominator is $n-1$ in order to compensate for an underestimate in the sample variance which would be obtained if s^2 were divided by n. (This arises from the fact that we have already used one parameter — the mean — to describe the data; see Mood & Graybill, 1963 for a discussion of bias in sample variance). Again using the twin data in Table 2.1 as an example, the variance of MZ Twin 1 is

$$\begin{aligned} V_{MZT1} &= \frac{(3-2)^2+(3-2)^2+\cdots+(2-2)^2+(1-2)^2}{15} \\ &= \frac{1+1+0+\cdots+0+0+1}{15} \\ &= 16/15 \end{aligned}$$

The variances of data from the second MZ twin, DZ Twin 1, and DZ Twin 2 also equal 16/15.

Covariances are computationally similar to variances, but represent mean deviations which are shared by two sets of observations. In the twin example, covariances are useful because they indicate the extent to which deviations from the mean by Twin 1 are similar to the second twin's deviations from the mean. Thus, the covariance between observations of Twin 1 and Twin 2 represents a scale-dependent measure of twin similarity. Covariances are often denoted by $s_{x,y}$ or Cov_{MZ} or Cov_{DZ}, and are calculated as

$$\begin{aligned} s_{x,y} &= \frac{(X_1-\overline{X})(Y_1-\overline{Y})+(X_2-\overline{X})(Y_2-\overline{Y})+\cdots+(X_n-\overline{X})(Y_n-\overline{Y})}{n-1} \\ &= \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n-1} \end{aligned} \tag{2.3}$$

Note that the variance formula shown in Eq. 2.2 is just a special case of the covariance when $Y_i = X_i$. In other words, the variance is simply the covariance between a variable and itself.

For the twin data in Table 2.1, the covariance between MZ twins is

$$\begin{aligned} \mathrm{Cov}_{MZ} &= \frac{(3-2)(2-2)+(3-2)(3-2)+\cdots+(1-2)(2-2)}{15} \\ &= \frac{0+1+0+0+\cdots+4+0+0+0}{15} \\ &= 12/15 \end{aligned}$$

The covariance between DZ pairs may be calculated similarly to give 8/15.

The *correlation coefficient* is closely related to the covariance between two sets of observations. Correlations may be interpreted in a similar manner as covariances, but are rescaled to give a lower bound of -1.0 and an upper bound of 1.0.

The correlation coefficient, r, may be calculated using the covariance between two measures and the square root of the variance (the *standard deviation*) of each measure:

$$r = \frac{\mathrm{Cov}_{x,y}}{\sqrt{V_x V_y}} \tag{2.4}$$

For the simulated MZ twin data, the correlation between twins is

$$\begin{aligned} r_{MZ} &= \frac{12/15}{\sqrt{(16/15)(16/15)}} \\ &= 12/16 = .75, \end{aligned}$$

and the DZ twin correlation is

$$\begin{aligned} r_{DZ} &= \frac{8/15}{\sqrt{(16/15)(16/15)}} \\ &= 8/16 = .50 \end{aligned}$$

Although variances and covariances typically define the observed information for biometrical analyses of twin data, correlations are useful for comparing resemblances between twins as a function of genetic relatedness. In the simulated twin data, the MZ twin correlation ($r = .75$) is greater than that of the DZ twins ($r = .50$). This greater similarity of MZ twins may be due to several sources of variation (discussed in subsequent chapters), but at the least is suggestive of a heritable basis for the trait, as increased MZ similarity could result from the fact that MZ twins are genetically identical, whereas DZ twins share only 1/2 of their genes on average.

2.2.2 Using SAS or SPSS to Summarize Data

The statistical packages SAS and SPSS are probably the most widely-used ways to store data collected in twin studies. In some cases relational databases such as Oracle, DB2, Paradox and Ingres may be used to store data collected from relatives because these offer powerful ways to maintain data in a consistent fashion according to normal form (**?**). Normal form is a way of storing data that avoids duplication of information; this is very important to avoid inconsistencies in the data. The general strategy may then be to use SAS or SPSS to extract the data from the database, to do preliminary data cleaning, to compute scales scores and transform them as necessary, and finally to dump the data in a format suitable for analysis with Mx. Here we discuss the advantages and disadvantages of this approach, and illustrate it with sample SAS and SPSS scripts.

By creating intermediate files for Mx to read, we are violating an elementary database principle to keep data in one place and one place only. This principle arises from the observation that almost as soon as there are two copies of data they become inconsistent and the updating chore requires more than double the effort as both sets must be updated and inconsistencies must be resolved. For that reason, it is best to consider the database as a master and to make updates to that dataset and that dataset only. Data analysis then involves creation of the intermediate data files using the same SAS or SPSS script. There are some very important advantages to this procedure. First, we know that the intermediate, file is not going to be updated by anyone else during our analysis — especially important in a multi-user environment. We want the comparison of models to be conducted on the same data, not on data that have changed from one analysis to the next! Second, the computation time taken to extract the data from the database may be non-trivial and it does not have to be repeated for every analysis.

SAS scripts to compute covariance matrices

This is not the place to describe in detail the workings of SAS; the thousands of pages in the manuals are quite adequate! All we aim to do here is to get the data in and get the covariance matrix and means out. SAS has a useful procedure, PROC CORR, which will print the required statistics, which can be cut and pasted into a file for Mx use. However, as is commonly the case with computer tasks, investing a little extra initial work on automation will save labor in the long run, and will be more error-proof.

It often happens that data are stored at the individual subject level rather than at the family level. Typically, each subject has a family number and an 'id' number to mark their position in the family (first or second twin). A necessary step to analyse the covariance between relatives is to 'glue' the data from family members together so that the family becomes the unit of measurement and covariances between family members may be computed. In SAS this is a relatively simple operation although care must be taken to supply labels for the variables that do not exceed the SAS maximum length of eight characters. The SAS script in Appendix **??** shows the case for twin data, and goes beyond the initial requirement by taking the sex of the twins into account. Five groups are created, being MZ male, DZ male, MZ female, DZ female and opposite DZ. The covariances are computed and output to .dat files which contain the number of observations (`Nobservations`), the number of input variables (`NInput`), labels, and the covariance matrices (CMatrix). These .dat files may be used directly in Mx in a diagram, or in a script using the `Include` statement.

Note that the assignment of the twins as 1 or 2 is usually arbitrary for the same sex groups, but in the opposite sex group the male (or female) twin is always first, and the female (or male) twin second. Strictly speaking, when there is no inherent order to the observations the variance-covariance matrix is not the best summary statistic to use. The intraclass correlation is the most appropriate summary for observations that do not have any order; it uses a joint estimate of the variance of twin 1 and twin 2, and partitions this into within pairs and between pairs components. However, the intraclass correlation is more difficult to generalize to the multivariate and multiple classes of relatives situations so we stay with covariance matrices here. Sometimes data on birth order or some other characteristic may be used to distinguish more formally between twin 1 and twin 2 within a pair, thereby giving some rationality to the ordering and use of covariance matrices. Should such an approach be taken, it is necessary to split the DZ opposite sex twin group into two groups according to whether the first twin is female or male.

Appendix ?? shows a SAS macro for creating an Mx .dat file, which fully describes the data: the variable labels, the sample size, the means and covariances. Comments, beginning with ! indicate the date the file was created. The resulting .dat file might look like this:

```
!
! Mx dat file created by SAS on 03FEB1998
!
Data NInputvars=4 NObservations=844
CMatrix Full
       1.0086       -0.0148       -0.0317       -0.0443
      -0.0148        1.0169       -0.0062        0.0068
      -0.0317       -0.0062        0.9342        0.0596
      -0.0443        0.0068        0.0596        0.9697
Means
       0.0139       -0.0729        0.0722        0.0159
Labels T1F1 T1F2 T2F1 T2F2
```

As will be seen in later chapters, this file is ready for immediate use for drawing path diagrams in the Mx GUI or in an Mx script with the `#include` command.

2.2.3 Using PRELIS to Summarize Continuous Data

PRELIS was developped by Karl Jöreskog and Dag Sörbom as a preprocessor for LISREL(?). Here we apply PRELIS to the simulated MZ twin data, and briefly discuss some of the further features of the software. In practice, data on MZ and DZ twins may be placed in separate files, often with one or more lines of data per twin pair[2]. It is easy to use PRELIS to generate summary statistics such as means and covariances for structural equation model fitting.

Suppose that the MZ twin data in Table 2.1 are stored in a file called `MZ.RAW` in the following way:

```
3  2
3  3
.  .
.  .
```

[2]It is possible to use data files that contain both types of twins and some code to discriminate between them, but it is less efficient.

```
.  .
2  1
1  2
```

We can use "free format" to read these data. Free format means that there is at least one space or end-of-line character between consecutive data items. These data could be entered using any simple text editor. If a wordprocessor such as Wordperfect or Microsoft Word were used, it would be necessary to save the file as a DOS or ASCII text file. Next, we would prepare an ASCII file containing the PRELIS commands to read these data and compute the means and covariances. We refer to files containing program commands as 'scripts'; the PRELIS script in this case might look like this:

```
Simple prelis example to compute MZ covariances
DA NI=2 NO=0
LA
Twin1 Twin2
COntinuous Twin1 Twin2
RAw FIle=MZ.RAW
OU SM=MZ.COV MA=CM
```

The first line is simply a title. PRELIS will treat all lines as part of the title until a line beginning with `DA` is encountered. The `DAta` line is used to specify basic features of the input (raw) data such as the number of input variables (`NI`) and the number of observations (`NO`). Here we have specified the number of observations as zero (`NO=0`), which asks PRELIS to count the number of cases for us. The next two lines of the script supply labels (`LA`) for the variables; these are optional but highly recommended when more than a few variables are to be read. Next, we define the variables Twin1 and Twin2 as *continuous.* By default, PRELIS 2 will treat any variable with less than 15 categories as ordinal. Although this is a reasonable statistical approach, it is not what we want for the purposes of this example. The next line in the script (beginning `RAw`) tells PRELIS where to find the data, and the `OUtput` line signifies the end of the script, and requests the covariance matrices (`MA=CM`) to be saved in the file `MZ.COV`. This output file is created by PRELIS — it is also ASCII format and looks like this:

```
(6D13.6)
  .106667D+01  .800000D+00  .106667D+01
```

The first line of the file contains a FORTRAN format for reading the data. The reader is referred to almost any text on FORTRAN, including User's Guides, for a detailed description of formats. The format used here is D format, for double precision. The 3 characters after the D give the power of 10 by which the printed number should be multiplied, so our .106667D+01 is really $.106667 \times 10^1 = 1.06667$. This number is part of the *lower triangle* of the covariance matrix. Since covariance matrices are always symmetric, only the lower triangle is needed. The file may in turn be read by Mx for the purposes of structural equation model fitting using syntax such as

```
CMatrix File=MZ.COV
```

within an Mx script — Mx by default expects only the lower triangle of covariance matrices to be supplied.

Suppose that, instead of just two variables, we had a data file with 20 variables per subject, with two lines for a twin pair. Also suppose that one of the variables identifies the zygosity of the pair, we wish to select only those pairs where zygosity is 1, and we only want the covariance of four of the variables. We could read these data into PRELIS using a FORTRAN format statement explicitly given in the PRELIS script. The script might look like this:

```
PRELIS script to select MZ's and compute covariances of 4 variables
DA NI=40 NO=0
LA
Zygosity Twin1P1 Twin1P2 Twin2P1 Twin2P2
RA FIle=MZ.RAW FO
(3X,F1.0,2x,F5.0,12X,F5.0/6X,F5.0,12X,F5.0)
SD Zygosity=1
OU SM=MZ.COV MA=CM
```

Note the `FOrtran` keyword at the end of the raw data line, indicating that the next line contains a Fortran format statement. The `SD` command selects cases where zygosity is 1, and deletes zygosity from the list of variables to be analyzed. Note that the FORTRAN format implicitly skips all the irrelevant variables, retaining only five (as specified by the F1.0 and F5.0 fields). Although we could have started with a more complete list of variables, read them in with an appropriate FORMAT, and used the PRELIS command `SD` to delete those we did not want, it is more efficient to save the program the trouble of reading these data by adjusting our `NI` and format statement. On the other hand, if the data file is not large or if a powerful computer is available, it may be better to use `SD` to save user time spent modifying the script.

2.3 Ordinal Data Analysis

Suppose that instead of making measurements on a continuous scale, we are able to discriminate only a few ordered categories with our measuring instrument. This situation is commonly encountered when assessing the presence or absence of disease, or responses to a single item on a questionnaire. Although it is possible to calculate a covariance matrix from these data, *the correlations usually will be biased.* The degree of bias depends on factors such as the number of categories and

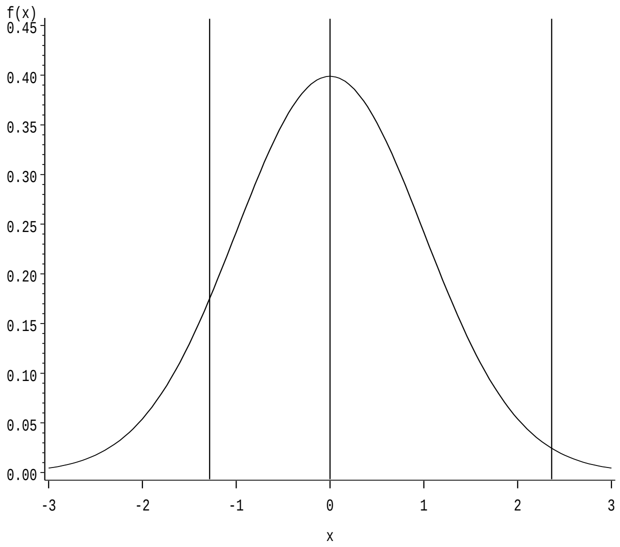


Figure 2.1: Univariate normal distribution with thresholds distinguishing ordered response categories.

the number of observations in each category, and usually results in an underestimate of the true liability correlation in the population. In this section we describe methods for summarizing ordinal data.

2.3.1 Univariate Normal Distribution of Liability

One approach to the analysis of ordinal data is to assume that the ordered categories reflect imprecise measurement of an underlying normal distribution of liability. A second assumption is that the liability distribution has one or more threshold values that discriminate between the categories (see Figure 2.1). This model has been used widely in genetic applications (Falconer, 1960; Neale *et al.,* 1986; Neale, 1988; Heath *et al* 1989a). As long as we consider one variable at a time, it is always possible to place the thresholds so that the proportion of the distribution lying between adjacent thresholds *exactly* matches the observed proportion of the sample that is found in each category. For example, suppose we had an item with four possible responses: 'none', 'a little', 'quite a lot', and 'a great deal'. In a sample of 200 subjects, 20 say 'none', 80 say 'a little', 98 say 'quite a lot' and 2 say 'a great deal'. If our assumed underlying normal distribution has mean 0 and variance 1, then placing thresholds at z-values of -1.282, 0.0 and 2.326 would partition the normal distribution as required. In mathematical terms, if there are p categories, $p-1$ thresholds are needed to divide the distribution. The expected proportion lying in category i is

$$\int_{t_{i-1}}^{t_i} \phi(x)\, dx$$

where $t_0 = -\infty$, $t_p = \infty$, and $\phi(x)$ is the unit variance normal probability density function (pdf), given by

$$\phi(x) = \frac{e^{-.5x^2}}{\sqrt{2\pi}}$$

This formulation is really a *parametric model* for the distribution of ordinal responses.

2.3.2 Bivariate Normal Distribution of Liability

When we have only one variable, there is no goodness-of-fit test for the liability model because it always gives a perfect fit. However, this is not necessarily so when we move to the multivariate case. Consider first, the example where we have two variables, each measured as a simple 'yes/no' binary response. Data collected from a sample of subjects could be summarized as a contingency table:

	Item 1	
Item 2	No	Yes
Yes	13	55
No	32	15

It is at this point that we encounter the crucial statistical concept of degrees of freedom (df). Fortunately, though important, calculating the number of df for a model is usually very easy; it is simply the difference between the number of observed statistics and the number of parameters in the model. In the present case we have a 2×2 contingency table in which there are four observed frequencies. However, if we take the total sample size as given and work with the proportion of the sample observed in each cell, we only need three proportions to describe the table completely, because the total of the cell proportions is 1 and the last cell proportion always can be obtained by subtraction. Thus in general for a table with r rows and c columns we can describe the data as rc frequencies or as $rc - 1$ proportions and the total sample size. The next question is, how many parameters does our model contain?

The natural extension of the univariate normal liability model described above is to assume that there is a continuous, bivariate normal distribution underlying the distribution of our observations. Given this model, we can compute the expected proportions for the four cells of the contingency table[3]. The model is illustrated graphically as contour and 3-D plots in Figure 2.2. The figures con-

[3]Mathematically these expected proportions can be written as double integrals. We do not explicitly define them here, but return to the subject in the context of ascertainment discussed in Chapter **??**

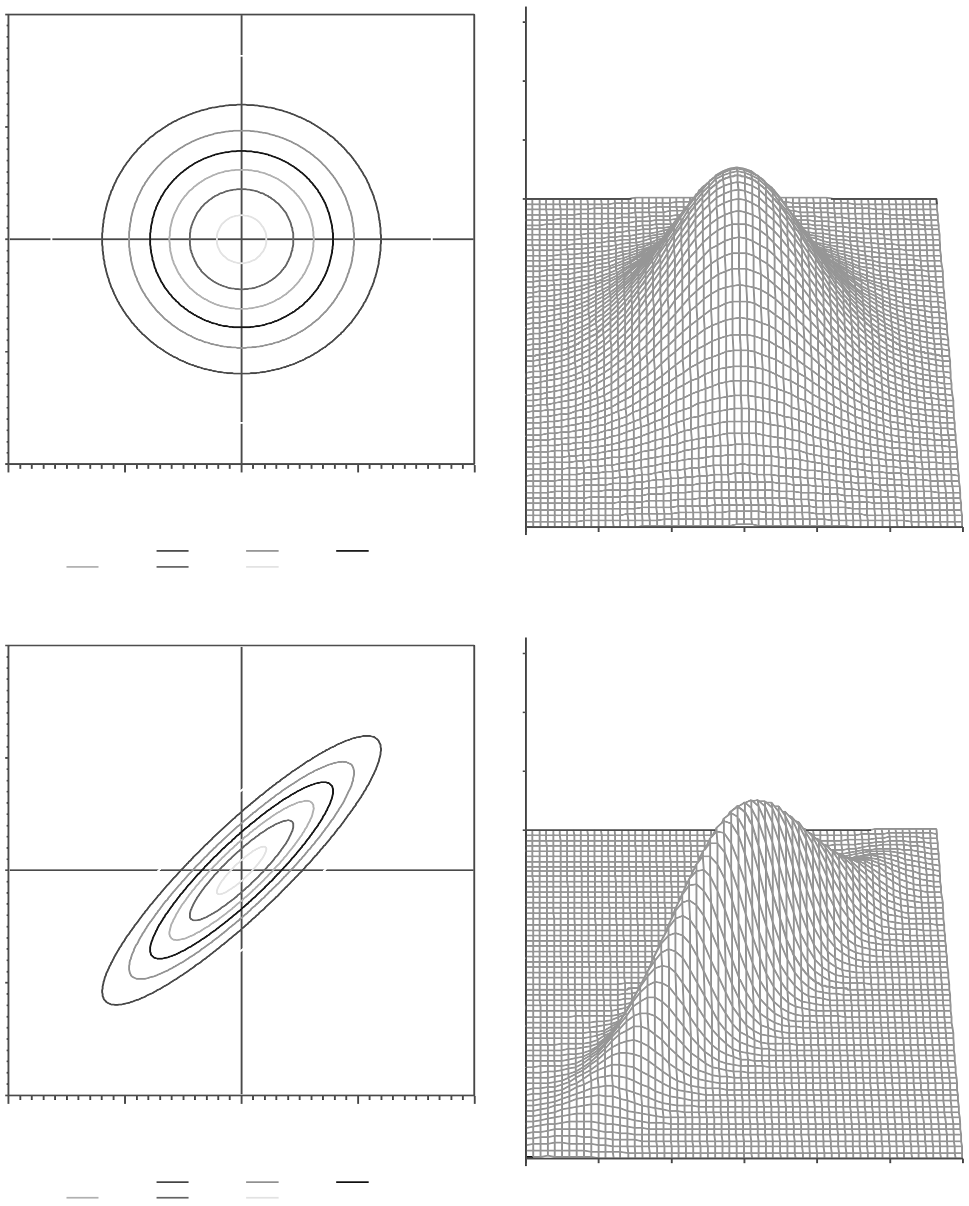


Figure 2.2: Contour and 3-D plots of the bivariate normal distribution with thresholds distinguishing two response categories. Contour plot in top left shows zero correlation in liability and plot in bottom left shows correlation of .9; the panels on the right shows the same data as 3-D plots.

trast the uncorrelated case ($r = 0$) with a high correlation in liability ($r = .9$) and are dramatically similar to the scatterplots of data from unrelated persons and from MZ twins, shown in Figures 1.2 and 1.4 on pages 6 and 9. By adjusting the correlation in liability and the two thresholds, the model can predict any combination of proportions in the four cells. Because we use 3 parameters to predict the 3 observed proportions, there are no degrees of freedom to test the goodness of fit of the model. This can be seen when we consider an arbitrary non-normal distribution created by mixing two normal distributions, one with $r = +.9$ and the second with $r = -.9$, as shown in Figure 2.3. With thresholds imposed as shown, equal proportions are expected in each cell, corresponding to a zero correlation and zero thresholds, not an unreasonable result *but with just two categories we have no knowledge at all that our distribution is such a bizarre non-normal example.* The case of a 2×2 contingency table is really a 'worst case scenario' for no degrees of freedom associated with a model, since absolutely any pattern of observed frequencies could be accounted for with the liability model. Effectively, all the model does is to transform the data; it cannot be *falsified.*

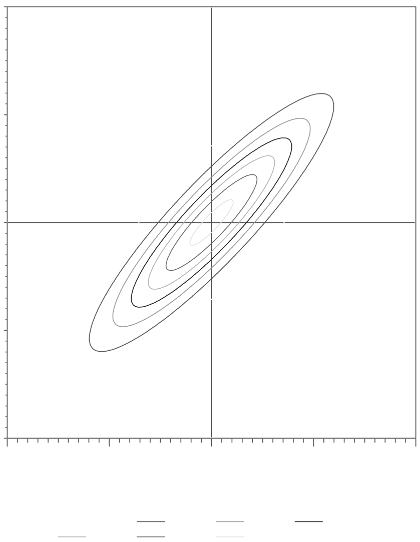
2.3.3 Testing the Normal Distribution Assumption

The problem of having no degrees of freedom to test the goodness of fit of the bivariate normal distribution to two binary variables is solved when we have at least three categories in one variable and at least two in the other. To illustrate this point, compare the contour plots shown in Figure 2.4 in which two thresholds have been specified for the two variables. With the bivariate normal distribution, there is a very strong pattern imposed on the relative magnitudes of the cells on the diagonal and elsewhere. There is a similar set of constraints with the mixture of normals, but quite different predictions are made about the off-diagonal cells; all four corner cells would have an appreciable frequency given a sufficient sample size, and probably in excess of that in each of the four cells in the middle of each side [e.g., (1,2)]. The bivariate normal distribution could never be adjusted to perfectly predict the cell proportions obtained from the mixture of distributions.

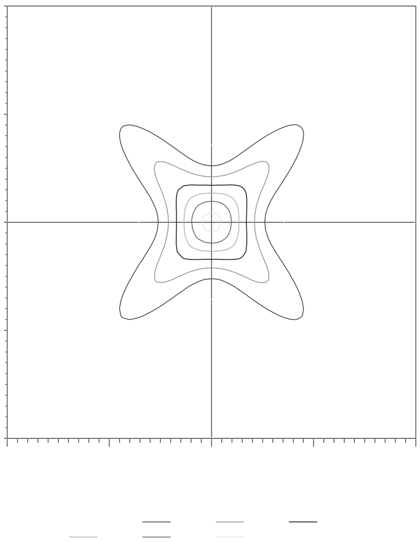
This intuitive idea of *opportunities for failure* translates directly into the concept of degrees of freedom. When we use a bivariate normal liability model to predict the proportions in a contingency table with r rows and c columns, we use $r-1$ thresholds for the rows, $c-1$ thresholds for the columns, and one parameter for the correlation in liability, giving $r+c-1$ in total. The table itself contains $rc-1$ proportions, neglecting the total sample size as above. Therefore we have degrees of freedom equal to:

$$\mathrm{df} = rc - 1 - (r + c - 1)\mathrm{df} = rc - r - c \tag{2.5}$$

The discrepancy between the frequencies predicted by the model and those actually

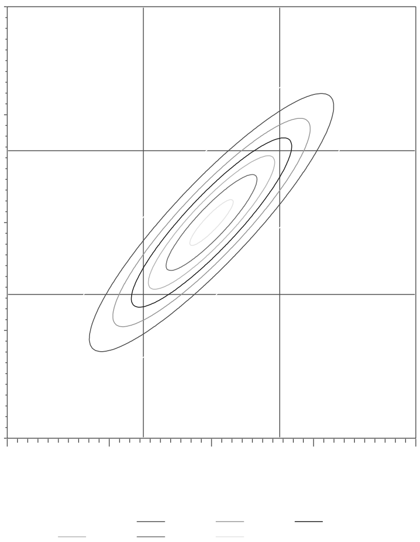


(a)

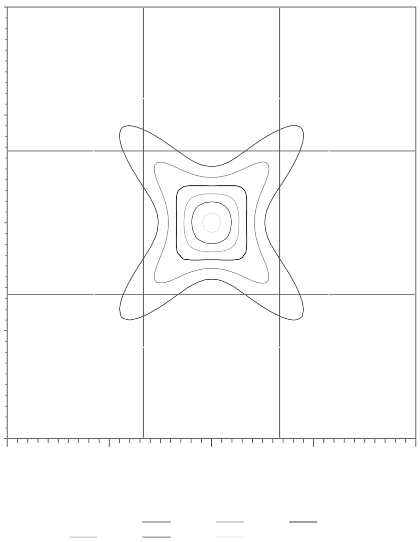


(b)

Figure 2.3: Contour plots of a bivariate normal distribution with correlation .9 (top); and of a mixture of bivariate normal distributions (bottom), one with .9 correlation and the other with -.9 correlation. One threshold in each dimension is shown.



(a)



(b)

Figure 2.4: Contour plots of a bivariate normal distribution with correlation .9 (top) and a mixture of bivariate normal distributions, one with .9 correlation and the other with -.9 correlation (bottom). Two thresholds in each dimension are shown.

observed in the data can be measured using the χ^2 statistic given by:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

Given a large enough sample, the model's failure to predict the observed data would be reflected in a significant χ^2 for the goodness of fit.

In principle, models could be fitted by maximum likelihood directly to contingency tables, employing the observed and expected cell proportions. This approach is general and flexible, especially for the multigroup case — the programs LISCOMP (Muthén, 1987) and Mx (Neale, 1991) use the method — but it is currently limited by computational considerations. When we move from two variables to larger examples involving many variables, integration of the multivariate normal distribution (which has to be done numerically) becomes extremely time-consuming, perhaps increasing by a factor of ten or so for each additional variable.

An alternative approach to this problem is to use PRELIS 2 to compute each correlation in a pairwise fashion, and to compute a weight matrix. The weight matrix is an estimate of the variances and covariances of the correlations. The variances of the correlations certainly have some intuitive appeal, being a measure of how precisely each correlation is estimated. However, the idea of a correlation correlating with another correlation may seem strange to a newcomer to the field. Yet this covariation between correlations is precisely what we need in order to represent how much *additional* information the second correlation supplies over and above that provided by the first correlation. Armed with these two types of summary statistics — the correlation matrix and the covariances of the correlations, we may fit models using a structural equation modeling package such as Mx or LISREL, and make statistical inferences from the goodness of fit of the model.

It is also possible to use the bivariate normal liability distribution to infer the patterns of statistics that would be observed if ordinal and continuous variables were correlated. Essentially, there are specific predictions made about the expected mean and variance of the continuous variable in each of the categories of the ordinal variable. For example, the continuous variable means are predicted to increase monotonically across the categories if there is a correlation between the liabilities. An observed pattern of a high mean in category 1, low in category 2 and high again in category 3 would not be consistent with the model. The number of parameters used to describe this model for an ordinal variable with r categories is $r + 2$, since we use $r - 1$ for the thresholds, one each for the mean and variance of the continuous variable, and one for the covariance between the two variables. The observed statistics involved are the proportions in the cells (less one because the final proportion may be obtained by subtraction from 1) and the mean and variance of the continuous variable in each category. Therefore we have:

$$\begin{aligned} \mathrm{df}_{\mathrm{OC}} &= (r-1) + 2r - (r+2) \\ &= 2r - 3 \end{aligned} \qquad (2.6)$$

Table 2.2: Classification of correlations according to their observed distribution.

Measurement	Two Categories	Three or more Categories	Continuous
Two	Tetrachoric	Polychoric	Biserial
Three or more	Polychoric	Polychoric	Polyserial
Continuous	Biserial	Polyserial	Product Moment

So the number of degrees of freedom for such a test is $2r-3$ where r is the number of categories.

2.3.4 Terminology for Types of Correlation

One of the difficulties encountered by the newcomer to statistics is the use of a wide variety of terms for correlation coefficients. There are many measures of association between variables; here we confine ourselves to the parametric statistics computed by normal theory. These statistics correspond most naturally to our genetic theory, in which we assume that a large number of independent genetic and environmental factors give rise to variation — "multifactorial inheritance"[4].

Table 2.2 shows the name given to the correlation coefficient calculated under normal distribution theory, according to whether each variable has: two categories (dichotomous); several categories (polychotomous); or an infinite number of categories (continuous). If both variables are dichotomous, then the correlation is called a tetrachoric correlation *as long as it is calculated using the bivariate normal integration approach described in Section 2.3 above.* If we simply use the Pearson product moment formula (described in Section 2.2.1 above) then we have computed a phi-coefficient which will probably underestimate the population correlation in liability. Because the tetrachoric and polychoric are calculated with the same method, some authors refer to the tetrachoric as a polychoric, and the same is true of the use of polyserial instead of biserial. As we shall see, the theory behind all these statistics is essentially the same.

2.3.5 Using PRELIS with Ordinal Data

Here we give a PRELIS script to read only two from a long list of psychiatric diagnoses, coded as 1 or 0 in these data.

```
Diagnoses and age MZ twins: VARIABLES ARE:
```

[4]In fact quite a small number of genetic factors may give rise to a distribution which is for almost all practical purposes indistinguishable from a normal distribution (Kendler and Kidd, 1986).

```
 DEPLN4 DEPLN2 DEPLN1 DEPLB4 DEPLB2 DEPLB1 GADLN6 GADLN1
 GADLB6 GADLB1 GAD88B GAD88N PANN PANB PHON PHOB ETOHN
 ETOHB ANON ANOB BULN BULB DEPLN4T2 DEPLN2T2 DEPLN1T2
 DEPLB4T2 DEPLB2T2 DEPLB1T2 GADLN6T2 GADLN1T2 GADLB6T2
 GADLB1T2 GAD88BT2 GAD88NT2 PANNT2 PANBT2 PHONT2 PHOBT2
 ETOHNT2 ETOHBT2 ANONT2 ANOBT2 BULNT2 BULBT2/
 FORMAT IN FULL IS:
 (2X, F8.2,F1.0, 43(1X,F1.0)

Diagnoses and age MZ twins
DA NI=3 NO=0
LA; DOB DEPLN4 DEPLN4T2
RA FI=DIAGMZ.DAT FO
(2X, F8.2,F1.0, 43x,F1.0)
OR DEPLN4-DEPLN4T2
OU MA=PM SM=DEPLN4MZ.COR SA=DEPLN4MZ.ASY PA

Diagnoses and age DZ twins
DA NI=3 NO=0
LA; DOB DEPLN4 DEPLN4T2
RA FI=DIAGdZ.DAT FO
(2X, F8.2,F1.0, 43x,F1.0)
OR DEPLN4-DEPLN4T2
OU MA=PM SM=DEPLN4dZ.COR SA=DEPLN4dZ.ASY PA
```

Note that again we have used the FORTRAN format to control which variables are read. One key difference from the continuous case is the use of `MA=PM`, which requests calculation of a matrix of polychoric, polyserial and product moment correlations. The program uses product moment correlations when both variables are continuous, a polyserial (or biserial) when one is ordinal and the other continuous, and a polychoric (or tetrachoric) when both are ordinal. Running the script produces four output files `DEPLN4MZ.COR`, `DEPLN4MZ.ASY`, `DEPLN4DZ.COR` and `DEPLN4DZ.ASY` which may be read directly into Mx using `PMatrix` and `ACov` commands. Notice that we have 'stacked' two scripts in one file, one to read and compute statistics from the MZ data file (`FI=DIAGMZ.DAT`) and a second to do the same thing for the DZ data. Also notice that the `SM` command is used to output the correlation matrix and `SA` is to save the asymptotic weight matrix. In fact, PRELIS saves the weight matrix multiplied by the sample size which is what Mx expects to receive when the `ACov` command is used. The `PA` command requests that the asymptotic weight matrix itself be printed in the output. However, PRELIS saves this file in a binary format which must be converted to ASCII for use with Mx. The utility bin2asc, supplied with PRELIS, can be used for this purpose.

In the PRELIS output, there are a number of summary statistics for continuous variables (means and standard deviations, and histograms) and frequency distributions with bar graphs, for the ordinal variables. To provide the user with some guide to the origin of statistics describing the covariance between variables, PRELIS prints means and standard deviations of continuous variables separately for each category of each pair of ordinal variables, and contingency tables between each ordinal variables. Towards the end of the output there is a table printed with the following format:

```
                                        TEST OF MODEL
                     CORRELATION  CHI-SQU.  D.F. P-VALUE
                     -----------  --------  ---- -------

  DEPLN4 VS.      DOB  -.233 (PS)   5.067     1     .024
DEPLN4T2 VS.      DOB   .010 (PS)   6.703     1     .010
```

There are two quite different chi-squared tests printed on the output. The first, under `TEST OF MODEL` is a test of the goodness of fit of the bivariate normal distribution model to the data. In the case of two ordinal variables with r and c categories in each, there are $rc - r - c$ df as described in expression 2.5 above. Likewise there will be $2r - 3$ df for the continuous by ordinal statistics, as described in expression 2.6. If the p-value reported by PRELIS is low (e.g. $< .05$), then concern arises about whether the bivariate normal distribution model is appropriate for these data. For a polyserial correlation (correlations between ordinal and continuous variables), it may simply be that the continuous variable is not normally distributed, or that the association between the variables does not follow a bivariate normal distribution. For polychoric correlations, there is no univariate test of normality involved, so failure of the model would imply that the latent liability distributions do not follow a bivariate normal. Remember however that significance levels for these tests are not often the reported p-value, because we are performing multiple tests. If the tests were independent, then with n such tests the α significance level would not be the reported p-value but $1 - (1 - p)^n$. Therefore concern would arise only if p was very small and a large number of tests had been performed. In our case, the tests are not independent because, for example, the correlation of A and B is not independent of the correlation of A and C, so the attenuation of the α level of significance is not so extreme as the $1-(1-p)^n$ formula predicts. The amount of attenuation will be application specific, but would often be closer to $1 - (1 - p)^n$ than simply to p.

The second chi-squared statistic printed by PRELIS (not shown in the above sample of output) tests whether the correlation is significantly different from zero. A similar result should be obtained if the summary statistics are supplied to Mx, and a chi-squared difference test (see Chapter ??) is performed between a model which allows the correlation to be a free parameter, and one in which the correlation is set to zero.

The use of weight matrices as input to Mx is described elsewhere in this book. Here we have described the generation of a weight matrix for a correlation matrix, but it is also possible to use weight matrices for covariance matrices[5]. Both methods are part of the asymptotically distribution free (ADF) methods pioneered by Browne (1984). It is not yet clear whether maximum likelihood or ADF methods are generally better for coping with data that are not multinormally distributed; further simulation studies are required. The ADF methods require more numerical effort and become cumbersome to use with large numbers of variables. This is so because the size of the weight matrix rapidly increases with number of variables. The number of elements on and below the diagonal of a matrix is a *triangular number* given by $k(k+1)/2$. The number of elements in this weight matrix is a triangular number of a triangular number, or

$$\frac{k^4 + 2k^3 + 3k^2 + 2k}{8}$$

In the case of correlation matrices, the number of elements is somewhat less, but still increases as a quadratic function:

$$\frac{k^4 - 2k^3 + 3k^2 - 2k}{8}$$

As a compromise when the number of variables is large, Jöreskog and Sörbom suggest the use of diagonal weights, i.e. just the variances of the correlations and not their covariances. However, tests of significance are likely to be inaccurate with this method and estimates of anything other than the full or true model would be biased.

2.4 Preparing Raw Data

Almost by definition, raw data does not need to be prepared for analysis. However, computer programs rarely communicate with each other without some form of translation of data format, and getting data out of datasets maintained in popular statistical packages such as SAS or SPSS and into Mx is no exception. In this section we briefly describe SAS and SPSS scripts that output data into a file suitable for Mx to read.

Mx has two main ways to read individual scores. First, and most straightforward, is 'rectangular' format, with one case per line, with variables separated by one or more spaces. A case is a collection of possibly correlated observations, such as several variables assessed on an individual, or on both members of a twin pair, or on a whole family. Because family members correlate, it is necessary to consider the whole family as a 'case'. Separate cases are assumed to be uncorrelated, which is important for statistical purposes. Certain new methods available in programs

[5]The number of elements in a weight matrix for a covariance matrix is greater than that for a correlation matrix. For this reason, it is necessary to specify `Matrix=PMatrix` on the `Data` line of a Mx job that is to read a weight matrix.

such as Sudaan, SAS proc mixed, and Stata make it possible to account for some correlation between different cases, usually when data are grouped, e.g., subjects in the same school. These methods can prove useful for running standard statistical analyses at the individual level (multiple regression, survival analysis) by taking into account the covariation between family members. However, they do not help with the preparation of data for modeling genetic and environmental factors which is the primary objective here.

The default code that Mx recognises as indicating missing data is a dot '.' which is the same as SAS and SPSS. A sample SAS script to produce rectangular data is shown in Appendix **??**. Mx's missing command can be used to declare a different string as the missing value, and it is important to note that this is a string and not a numeric value, as 1.0 and 1.00 will be considered to be different.

The second main format for raw data that Mx accepts is variable length, or 'vl'.

2.5 Summary

We have described in detail the statistical operations involved in, and the use of SAS and PRELIS for, the measurement of variation and covariation. When we have continuous measures, the calculations are quite simple and can be done by hand, but for ordinal data the process is more complex. We obtain estimates of polychoric and polyserial correlations by using software that numerically integrates the bivariate normal distribution. In the process, we are effectively fitting a model of continuous multivariate normal liability with abrupt thresholds to the contingency table. This model cannot be rejected when there are only two categories for each measure, but may fail as the number of cells in the table increases.

While ordinal data are far more common than continuous measures in the behavioral sciences, we note that as the number of categories gets large (e.g., more than 15) the difference between the continuous and the ordinal treatments gets small. In general, the researcher should try to obtain continuous measures if possible, since considerable statistical power can be lost when only a few response categories are used, as we shall show in Chapter 7.

Chapter 3

Biometrical Genetics

3.1 Introduction and Description of Terminology

The principles of biometrical and quantitative genetics lie at the heart of virtually all of the statistical models examined in this book. Thus, an understanding of biometrical genetics is fundamental to our statistical approach to twin and family data. Biometrical models relate the "latent," or unobserved, variables of our structural models to the functional effects of genes. It is these effects, based on the principles of Mendelian genetics, that give our structural models a degree of validity quite unusual in the social sciences. The purpose of this chapter is to provide a brief introduction to biometrical models. Extensive treatments of the subject have been provided by Mather and Jinks (1982) and Falconer (1990). Here we employ the notation of Mather and Jinks.

Before we begin our discussion of biometrical genetics, we must describe some of the terms that are encountered frequently in biometrical and classical genetic discourse. For the present purposes, we use the term *gene* in reference to a "unit factor of inheritance" that influences an observable trait or traits, following the earlier usage by Fuller and Thompson (1978). Observable characteristics are referred to as *phenotypes*. The site of a gene on a chromosome is known as the *locus*. *Alleles* are alternative forms of a gene that occupy the same locus on a chromosome. They often are symbolized as A and a or B and b or A_1 and A_2. The simplest system for a segregating locus involves only two alleles (A and a), but there also may be a large number of alleles in a system. For example, the HLA locus on chromosome 6 is known to have 18 alleles at the A locus, 41 alleles at the B locus, 8 at C, about 20 at DR, 3 at DQ, and 6 at DP (Bodmer 1987). Nevertheless, if one or two alleles are much more frequent than the others, a two-allele system provides a useful approximation and leads to an accurate account for the phenotypic variation and covariation with which we are concerned. The *genotype* is the chromosomal complement of alleles for an individual. At a single locus (with two alleles) the genotype may be symbolized AA, Aa, or aa; if we consider multiple loci the genotype of an individual may be written as $AABB$, $AABb$, $AAbb$, $AaBB$, $AaBb$, $Aabb$, $aaBB$, $aaBb$, or $aabb$, in the case of two loci, for example. *Homozygosity* refers to a state of identical alleles at corresponding loci on homologous chromosomes; for example, AA or aa for one locus, or $AABB$, $aabb$, $AAbb$, or $aaBB$ for two loci. In contrast, *heterozygosity* refers to a state of unlike alleles at corresponding loci, Aa or $AaBb$, for example. When numeric or symbolic values are assigned to specific genotypes they are called *genotypic values.* The *additive value* of a gene is the sum of the average effects of the individual alleles. *Dominance deviations* refer to the extent to which genotypes differ from the additive genetic value. A system in which multiple loci are involved in the expression of a single trait is called *polygenic* ("many genes"). A *pleiotropic* system ("many growths") is one in which the same gene or genes influence more than one trait.

Biometrical models are based on the measurable effects of different genotypes that arise at a segregating locus, which are summed across all of the loci that contribute to a continuously varying trait. The number of loci generally is not known, but it is usual to assume that a relatively large number of genes of equivalent effect are at work. In this way, the categories of Mendelian genetics that lead to binomial distributions for traits in the population tend toward continuous distributions such as the normal curve. Thus, the statistical parameters that describe this model are those of continuous distributions, including the first moment, or the mean; second moments, or variances, covariances, and correlation coefficients; and higher moments such as measures of skewness where these are appropriate. This polygenic model was originally developed by Sir Ronald Fisher in his classic paper "*The correlation between relatives on the supposition of Mendelian inheritance*" (Fisher, 1918), in which he reconciled Galtonian biometrics with Mendelian genetics. One interesting feature of the polygenic biometrical model is that it predicts normal distributions for traits when very many loci are involved and their effects are combined with a multitude of environmental influences. Since the vast majority of biological and behavioral traits approximate the normal distribution, it is an inherently plausible model for the geneticist to adopt. We might note, however, that although the normality expected for a polygenic system is statistically convenient as well as empirically appropriate, none of the biometrical expectations with which we shall be concerned depend on how many or how few genes are involved. The expectations are equally valid if there are are only one or two genes, or indeed no genes at all.

In the simplest two–allele system (A and a) there are two parameters that define the measurable effects of the three possible genotypes, AA, Aa, and aa. These parameters are d, which is twice the measured difference between the homozygotes AA and aa, and h, which defines the measured effect of the heterozygote Aa, insofar as it does not fall exactly between the homozygotes. The point between the two homozygotes is m, the mean effect of homozygous genotypes. We refer to the parameters d and h as *genotypic effects.* The scaling of the three genotypes is

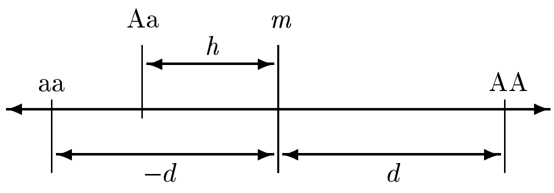


Figure 3.1: The d and h increments of the gene difference $A - a$. Aa may lie on either side of m and the sign of h will vary accordingly; in the case illustrated h would be negative. (Adapted from Mather and Jinks, 1977, p. 32).

shown in Figure 3.1.

To make the simple two–allele model concrete, let us imagine that we are talking about genes that influence adult stature. Let us assume that the normal range of height for males is from 4 feet 10 inches to 6 feet 8 inches; that is, about 22 inches[1]. And let us assume that each somatic chromosome has one gene of roughly equivalent effect. Then, roughly speaking, we are thinking in terms of loci for which the homozygotes contribute $\pm\frac{1}{2}$ inch (from the midpoint), depending on whether they are AA, the increasing homozygote, or aa, the decreasing homozygotes. In reality, although some loci may contribute greater effects than this, others will almost certainly contribute less; thus we are talking about the kind of model in which any particular polygene is having an effect that would be difficult to detect by the methods of classical genetics. Similarly, while the methods of linkage analysis may be appropriate for a number of quantitative loci, it seems unlikely that the majority of causes of genetic variation would be detectable by these means. The biometrical approach, being founded upon an assumption that inheritance may be polygenic, is designed to elucidate sources of genetic variation is these types of systems.

3.2 Breeding Experiments: Gametic Crosses

The methods of biometrical genetics are best understood through controlled breeding experiments with inbred strains, in which the results are simple and intuitively

[1]Note: 1 inch = 2.54cm; 1 foot = 12 inches.

Table 3.1: Punnett square for mating between two heterozygous parents.

		Male Gametes	
		$\frac{1}{2}A$	$\frac{1}{2}a$
Female Gametes	$\frac{1}{2}A$	$\frac{1}{4}AA$	$\frac{1}{4}Aa$
	$\frac{1}{2}a$	$\frac{1}{4}Aa$	$\frac{1}{4}aa$

obvious. Of course, in the present context we are dealing with continuous variation in humans, where inbred strains do not exist and controlled breeding experiments are impossible. However, the simple results from inbred strains of animals apply directly, albeit in more complex form, to those of free mating organisms such as humans. We feel an appreciation of the simple results from controlled breeding experiments provides insight and lends credibility to the application of the models to human beings.

Let us consider a cross between two inbred parental strains, P_1 and P_2, with genotypes AA and aa, respectively. Since individuals in the P_1 strain can produce gametes with only the A allele, and P_2 individuals can produce only a gametes, all of the offspring of such a mating will be heterozygotes, Aa, forming what Gregor Mendel referred to as the "first filial," or F_1 generation. A cross between two F_1 individuals generates what he referred to as the "second filial" generation, or F_2, and it may be shown that this generation comprises $\frac{1}{4}$ individuals of genotype AA, $\frac{1}{4}$ aa, and $\frac{1}{2}$ Aa. Mendel's first law, the *law of segregation*, states that parents with genotype Aa will produce the gametes A and a in equal proportions. The pioneer Mendelian geneticist Reginald Punnett developed a device known as the *Punnett square*, which he found useful in teaching Mendelian genetics to Cambridge undergraduates, that gives the proportions of genotypes that will arise when these gametes unite at random. (Random unions of gametes occur under the condition of random mating among individuals). The result of other matings such as $P_1 \times F_1$, the first backcross, B_1, and more complex combinations may be elucidated in a similar manner. A simple usage of the Punnett square is shown in Table 3.1 for the mating of two heterozygous parents in a two–allele system. The gamete frequencies in Table 3.1 (shown outside the box) are known as *gene or allelic frequencies*, and they give rise to the *genotypic frequencies* by a simple product of independent probabilities. It is this assumption of independence based on random mating that makes the biometrical model straightforward and tractable in more complex situations, such as random mating in populations where the gene frequencies are unequal. It also forms a simple basis for considering the more complex effects of non-random mating, or assortative mating, which are known to be important in human populations.

In the simple case of equal gene frequencies as we have in an F_2 population, it is easily shown that random mating over successive generations changes neither the gene nor genotype frequencies of the population. Male and female gametes of the type A and a from an F_2 population are produced in equal proportions so that random mating may be represented by the same Punnett square as given in Table 3.1, which simply reproduces a population with identical structure to the F_2 from which we started. This remarkable result is known as *Hardy–Weinberg equilibrium* and is the cornerstone of quantitative and population genetics. From this result, the effects of non-random mating and other forces that change populations, such as natural selection, migration, and mutation, may be deduced. Hardy-Weinberg equilibrium is achieved in one generation and applies whether or not the gene frequencies are equal and whether or not there are more than two alleles. It also holds among polygenic loci, linked or unlinked, although in these cases joint equilibrium depends on a number of generations of random mating.

For our purposes the genotypic frequencies from the Punnett square are important because they allow us to calculate the simple first and second moments of the phenotypic distribution that result from genetic effects; namely, the mean and variance of the phenotypic trait. The genotypes, frequencies, and genotypic effects of the biometrical model in Table 3.1 are shown below, and from these we can calculate the mean and variance.

Genotype (i)	AA	Aa	aa
Frequency (f)	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
Genotypic effect (x)	d	h	$-d$

The mean effect of the A locus is obtained by summing the products of the frequencies and genotypic effects in the following manner:

$$\begin{aligned}\mu_A &= \sum f_i x_i \\ &= \frac{1}{4}d + \frac{1}{2}h - \frac{1}{4}d \\ &= \frac{1}{2}h \end{aligned} \tag{3.1}$$

The variance of the genetic effects is given by the sum of the products of the genotypic frequencies and their squared deviations from the mean[2]:

$$\begin{aligned}\sigma_A^2 &= \sum f_i(x_i - \mu_A)^2 \\ &= \frac{1}{4}(d - \frac{1}{2}h)^2 + \frac{1}{2}(h - \frac{1}{2}h)^2 + \frac{1}{4}(-d - \frac{1}{2}h)^2 \\ &= \frac{1}{4}d^2 - \frac{1}{4}dh + \frac{1}{16}h^2 + \frac{1}{8}h^2 + \frac{1}{4}d^2 + \frac{1}{4}dh + \frac{1}{16}h^2\end{aligned}$$

[2]This is an application of the method described in Section 2.2.1. It looks a bit more intimidating here because of (a) the multiplication by the frequency, and (b) the use of letters not numbers. To gain confidence in this method, the reader may wish to choose values for d and h and work through an example.

$$= \frac{1}{2}d^2 + \frac{1}{4}h^2 \tag{3.2}$$

For this single locus with equal gene frequencies, $\frac{1}{2}d^2$ is known as the additive genetic variance, or V_A, and $\frac{1}{4}h^2$ is known as the dominance variance, V_D. When more than one locus is involved, perhaps many loci as we envisage in the polygenic model, Mendel's *law of independent assortment* permits the simple summation of the individual effects of separate loci in both the mean and the variance. Thus, for (k) multiple loci,

$$\mu = \frac{1}{2}\sum_{i=1}^{k} h_i \ , \tag{3.3}$$

and

$$\begin{aligned} \sigma^2 &= \frac{1}{2}\sum_{i=1}^{k} d_i^2 + \frac{1}{4}\sum_{i=1}^{k} h_i^2 \\ &= V_A + V_D \ . \end{aligned} \tag{3.4}$$

It is the parameters V_A and V_D that we estimate using the structural equations in this book.

In order to see how this biometrical model and the equations estimate V_A and V_D, we need to consider the joint effect of genes in related individuals. That is, we need to derive expectations for MZ and DZ covariances in terms of the genotypic frequencies and the effects of d and h.

3.3 Derivation of Expected Twin Covariances

3.3.1 Equal Gene Frequencies

Twin correlations may be derived in a number of different ways, but the most direct method is to list all possible twin-pair genotypes (taken as deviations from the population mean) and the frequency with which they arise in a random-mating population. Then, the expected covariance may be obtained by multiplying the genotypic effects for each pair, weighting them by the frequency of occurrence, and summing across all possible pairs. By this method the covariance among pairs is calculated directly. The overall mean for such pairs is, of course, simply the population mean, $\frac{1}{2}h$, in the case of equal gene frequencies, as shown in the previous section. There are shorter methods for obtaining the same result, but these are less direct and less intuitively obvious.

The covariance calculations are laid out in Table 3.2 for MZ, DZ, and Unrelated pairs of siblings, the latter being included in order to demonstrate the expected zero covariance for genetically unrelated individuals. The nine possible combinations of genotypes are shown in column 1, with their genotypic effects, x_{1i} and x_{2i}, in

Table 3.2: Genetic covariance components for MZ, DZ, and Unrelated siblings with equal gene frequencies at a single locus ($u = v = \frac{1}{2}$).

Genotype	Effect					Frequency		
Pair	x_{1i}	x_{2i}	$x_{1i} - \mu_1$	$x_{2i} - \mu_2$	$(x_{1i} - \mu_1)(x_{2i} - \mu_2)$	MZ	DZ	U
AA, AA	d	d	$d - \frac{1}{2}h$	$d - \frac{1}{2}h$	$d^2 - dh + \frac{1}{4}h^2$	$\frac{1}{4}$	$\frac{9}{64}$	$\frac{1}{16}$
AA, Aa	d	h	$d - \frac{1}{2}h$	$\frac{1}{2}h$	$\frac{1}{2}dh - \frac{1}{4}h^2$	-	$\frac{3}{32}$	$\frac{1}{8}$
AA, aa	d	$-d$	$d - \frac{1}{2}h$	$-d - \frac{1}{2}h$	$-d^2 + \frac{1}{4}h^2$	-	$\frac{1}{64}$	$\frac{1}{16}$
Aa, AA	h	d	$\frac{1}{2}h$	$d - \frac{1}{2}h$	$\frac{1}{2}dh - \frac{1}{4}h^2$	-	$\frac{3}{32}$	$\frac{1}{8}$
Aa, Aa	h	h	$\frac{1}{2}h$	$\frac{1}{2}h$	$\frac{1}{4}h^2$	$\frac{1}{2}$	$\frac{5}{16}$	$\frac{1}{4}$
Aa, aa	h	$-d$	$\frac{1}{2}h$	$-d - \frac{1}{2}h$	$-\frac{1}{2}dh - \frac{1}{4}h^2$	-	$\frac{3}{32}$	$\frac{1}{8}$
aa, AA	$-d$	d	$-d - \frac{1}{2}h$	$d - \frac{1}{2}h$	$-d^2 + \frac{1}{4}h^2$	-	$\frac{1}{64}$	$\frac{1}{16}$
aa, Aa	$-d$	h	$-d - \frac{1}{2}h$	$\frac{1}{2}h$	$-\frac{1}{2}dh - \frac{1}{4}h^2$	-	$\frac{3}{32}$	$\frac{1}{8}$
aa, aa	$-d$	$-d$	$-d - \frac{1}{2}h$	$-d - \frac{1}{2}h$	$d^2 + dh + \frac{1}{4}h^2$	$\frac{1}{4}$	$\frac{9}{64}$	$\frac{1}{16}$

$\mu_{x_1} = \mu_{x_2} = \frac{1}{2}h$ in all cases; genetic covariance $= \sum_i f_i(x_{1i} - \mu_1)(x_{2i} - \mu_2)$

columns 2 and 3. From these values the mean of all pairs, $\frac{1}{2}h$, is subtracted in columns 4 and 5. Column 6 shows the products of these mean deviations. The final three columns show the frequency with which each of the genotype pairs occurs for the three kinds of relationship. For MZ twins, the genotypes must be identical, so there are only three possibilities and these occur with the population frequency of each of the possible genotypes. For unrelated pairs, the population frequencies of the three genotypes are simply multiplied within each pair of siblings since genotypes are paired at random. The frequencies for DZ twins, which are the same as for ordinary siblings, are more difficult to obtain. All possible parental types and the proportion of paired genotypes they can produce must be enumerated, and these categories collected up across all possible parental types. These frequencies and the method by which they are obtained may be found in standard texts (e.g., Crow and Kimura, 1970, pp. 136-137; Falconer, 1960, pp. 152-157; Mather and Jinks, 1971, pp. 214-215).

The products in column 6, weighted by the frequencies for the three sibling types, yield the degree of genetic resemblance between siblings. In the case of MZ twins, the covariance equals

$$\begin{aligned}
\text{Cov(MZ)} &= d^2(\frac{1}{4}+\frac{1}{4}) + dh(-\frac{1}{4}+\frac{1}{4}) + \frac{1}{4}h^2(\frac{1}{4}+\frac{2}{4}+\frac{1}{4}) \\
&= \frac{1}{2}d^2 + \frac{1}{4}h^2 \ , \qquad (3.5)
\end{aligned}$$

which is simply expression 3.2, the total genetic variance in the population. If we sum over loci, as we did in expression 3.4, we obtain $V_A + V_D$, the additive and dominance variance, as we would intuitively expect since identical twins share all genetic variance. The calculation for DZ twins, with terms in d^2, dh, and h^2 initially separated for convenience, and collected together at the end, is

$$\begin{aligned}
\text{Cov(DZ)} &= d^2(\frac{9}{64}-\frac{1}{64}-\frac{1}{64}+\frac{9}{64}) \\
&+ dh(-\frac{9}{64}+\frac{3}{64}+\frac{3}{64}-\frac{3}{64}-\frac{3}{64}+\frac{9}{64}) \\
&+ \frac{1}{4}h^2(\frac{9}{64}-\frac{6}{64}+\frac{1}{64}-\frac{6}{64}+\frac{20}{64}-\frac{6}{64}+\frac{1}{64}-\frac{6}{64}+\frac{9}{64}) \\
&= \frac{1}{4}d^2 + \frac{1}{16}h^2 \qquad (3.6)
\end{aligned}$$

When summed over all loci, this expression gives $\frac{1}{2}V_A + \frac{1}{4}V_D$. The calculation for unrelated pairs of individuals yields a zero value as expected, since, on average, unrelated siblings have no genetic variation in common at all:

$$\begin{aligned}
\text{Cov(U)} &= d^2(\frac{1}{16}-\frac{1}{16}-\frac{1}{16}+\frac{1}{16}) \\
&= dh(-\frac{1}{16}+\frac{1}{16}+\frac{1}{16}-\frac{1}{16}-\frac{1}{16}+\frac{1}{16}) \\
&= \frac{1}{4}h^2(\frac{1}{16}-\frac{2}{16}+\frac{1}{16}-\frac{2}{16}+\frac{4}{16}-\frac{2}{16}+\frac{1}{16}-\frac{2}{16}+\frac{1}{16}) \\
&= 0 \qquad (3.7)
\end{aligned}$$

It is the fixed coefficients in front of V_A and V_D, 1.0 and 1.0 in the case of MZ twins and $\frac{1}{2}$ and $\frac{1}{4}$, respectively, for DZ twins that allow us to specify the Mx model and estimate V_A and V_D, as will be explained in subsequent chapters. These coefficients are the correlations between additive and dominance deviations for the specified twin types. This may be seen easily in the case where we assume that dominance is absent. Then, MZ and DZ genetic covariances are simply V_A and $\frac{1}{2}V_A$, respectively. The variance of twin 1 and twin 2 in each case, however, is the population variance, V_A. For example, the DZ genetic correlation is derived as

$$r_{\text{DZ}} = \frac{\text{Cov(DZ)}}{\sqrt{V_{T1}V_{T2}}} = \frac{\frac{1}{2}V_A}{\sqrt{V_A V_A}} = \frac{1}{2}$$

3.3.2 Unequal Gene Frequencies

The simple results for equal gene frequencies described in the previous section were appreciated by a number of biometricians shortly after the rediscovery of Mendel's work (Castle, 1903; Pearson, 1904; Yule, 1902). However, it was not until Fisher's remarkable 1918 paper that the full generality of the biometrical model was elucidated. Gene frequencies do not have to be equal, nor do they have to be the same for the various polygenic loci involved in the phenotype for the simple fractions, 1, $\frac{1}{2}$, $\frac{1}{4}$, and 0 to hold, providing we define V_A and V_D appropriately. The algebra is considerably more complicated with unequal gene frequencies and it is necessary to define carefully what we mean by V_A and V_D. However, the end result is extremely simple, which is perhaps somewhat surprising. We give the flavor of the approach in this section, and refer the interested reader to the classic texts in this field for further information (Crow and Kimura, 1970; Falconer, 1990; Kempthorne, 1960; Mather and Jinks, 1982). We note that the elaboration of this biometrical model and its power and elegance has been largely responsible for the tremendous strides in inexpensive plant and animal food production throughout the world, placing these activities on a firm scientific basis.

Consider the three genotypes, AA, Aa, and aa, with genotypic frequencies P, Q, R:

Genotypes	AA	Aa	aa
Frequency	P	Q	R

The proportion of alleles, or gene frequency, is given by

$$\begin{aligned} \text{gene frequency } (A) &= P + \frac{Q}{2} = u \\ (a) &= R + \frac{Q}{2} = v \; . \end{aligned} \tag{3.8}$$

These expressions derive from the simple fact that the AA genotype contributes only A alleles and the heterozygote, Aa, contributes $\frac{1}{2}$ A and $\frac{1}{2}$ a alleles. A Punnett square showing the allelic form of gametes uniting at random gives the genotypic frequencies in terms of the gene frequencies:

		Male Gametes	
		$u\ A$	$v\ a$
Female Gametes	$u\ A$	u^2AA	$uvAa$
	$v\ a$	$uvAa$	v^2aa

which yields an alternative representation of the genotypic frequencies

Genotypes	AA	Aa	aa
Frequency	u^2	$2uv$	v^2

That these genotypic frequencies are in Hardy-Weinberg equilibrium may be shown by using them to calculate gene frequencies in the new generation, showing them to be the same, and then reapplying the Punnett square. Using expression 3.8, substituting u^2, $2uv$, and v^2, for P, Q, and R, and noting that the sum of gene frequencies is 1 ($u+v=1.0$), we can see that the new gene frequencies are the same as the old, and that genotypic frequencies will not change in subsequent generations

$$\begin{aligned} u_1 &= u^2 + \frac{1}{2}2uv = u^2 + uv = u(u+v) = u \\ v_1 &= v^2 + \frac{1}{2}2uv = v^2 + uv = v(u+v) = v\ . \end{aligned} \tag{3.9}$$

The biometrical model is developed in terms of these equilibrium frequencies and genotypic effects as

Genotypes	AA	Aa	aa
Frequency	u^2	$2uv$	v^2
Genotypic effect	d	h	$-d$

(3.10)

The mean and variance of a population with this composition is obtained in analogous manner to that in 3.1. The mean is

$$\begin{aligned} \mu &= u^2d + 2uvh - v^2d \\ &= (u-v)d + 2uvh \end{aligned} \tag{3.11}$$

Because the mean is a reasonably complex expression, it is not convenient to sum weighted deviations to express the variance as in 3.2, instead, we rearrange the variance formula

$$\begin{aligned} \sigma^2 &= \sum f_i(x_i-\mu)^2 \\ &= \sum f_i(x_i^2 - 2x_i\mu + \mu^2) \\ &= \sum f_i x_i^2 - 2\mu \sum f_i x_i + \mu^2 \\ &= \sum f_i x_i^2 - 2\mu^2 + \mu^2 \\ &= \sum f_i x_i^2 - \mu^2 \end{aligned} \tag{3.12}$$

Applying this formula to the genotypic effects and their frequencies given in 3.10 above, we obtain

$$\begin{aligned} \sigma^2 &= u^2d^2 + 2uvh^2 + v^2d^2 - [(u-v)d + 2uvh]^2 \\ &= u^2d^2 + 2uvh^2 + v^2d^2 - [(u-v)^2d^2 + 4uvdh(u-v) + 4u^2v^2h^2] \\ &= u^2d^2 + 2uvh^2 + v^2d^2 - [(u^2-2uv-v^2)d^2 + 4uvdh(u-v) + 4u^2v^2h^2] \\ &= 2uv[d^2 + 2(v-u)dh + (1-2uv)h^2] \\ &= 2uv[d^2 + 2(v-u)dh + (v-u)h^2 + 2uvh^2] \\ &= 2uv[d + (v-u)h]^2 + 4u^2v^2h^2\ . \end{aligned} \tag{3.13}$$

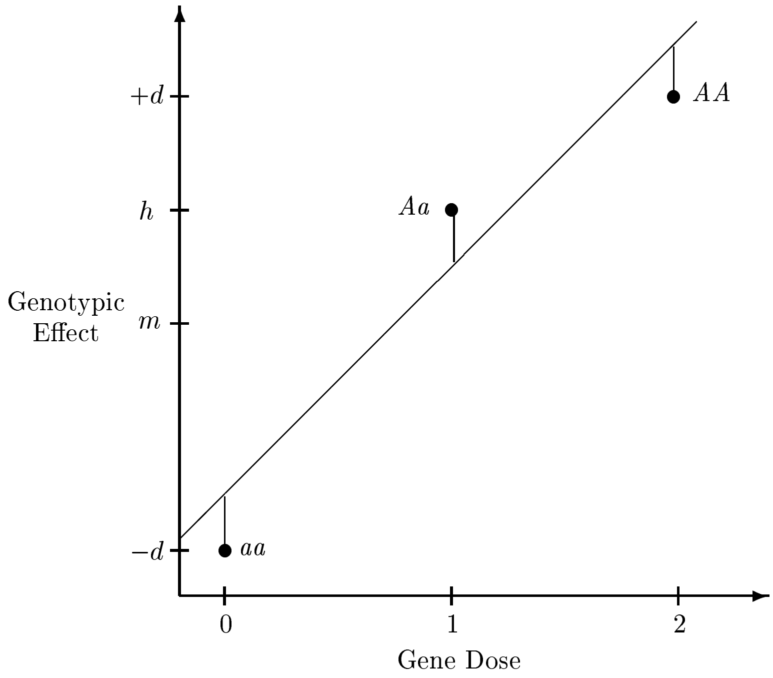


Figure 3.2: Regression of genotypic effects on gene dosage showing additive and dominance effects under random mating. The figure is drawn to scale for $u = v = \frac{1}{2}$, $d = 1$, and $h = \frac{1}{2}$.

When the variance is arranged in this form, the first term $(2uv[d+(v-u)h]^2)$ defines the additive genetic variance, V_A, and the second term $(4u^2v^2h^2)$ the dominance variance, V_D. Why this particular arrangement is used to define V_A and V_D rather than some other may be seen if we introduce the notion of gene dose and the regression of genotypic effects on this variable, which essentially is how Fisher proceeded to develop the concepts of V_A and V_D.

If A is the increasing allele, then we can consider the three genotypes, AA, Aa, aa, as containing 2, 1, and 0 doses of the A allele, respectively. The regression of genotypic effects on these gene doses is shown in Figure 3.2. The values that enter into the calculation of the slope of this line are

Genotype	AA	Aa	aa
Genotypic effect (y)	d	h	$-d$
Frequency (f)	u^2	$2uv$	v^2
Dose (x)	2	1	0

From these values the slope of the regression line of y on x in Figure 3.2 is given by $\beta_{y,x} = \sigma_{x,y}/\sigma_x^2$. In order to calculate σ_x^2 we need μ_x, which is

$$\begin{aligned} \mu_x &= 2u^2 + 2uv \\ &= 2u(u+v) \\ &= 2u \ . \end{aligned} \tag{3.14}$$

Then, σ_x^2 is

$$\begin{aligned} \sigma_x^2 &= 2^2u^2 + 1^2 2uv - 2^2u^2 \\ &= 4u^2 + 2uv - 4u^2 \\ &= 2uv \end{aligned}$$

using the variance formula in 3.12. In order to calculate $\sigma_{x,y}$ we need to employ the covariance formula

$$\sigma_{x,y} = \sum f_i x_i y_i - \mu_x \mu_y \ , \tag{3.15}$$

where μ_y and μ_x are defined as in 3.11 and 3.14, respectively. Then,

$$\begin{aligned} \sigma_{xy} &= 2u^2d + 2uvh - 2u[(u-v)d + 2uvh] \\ &= 2u^2d + 2uvh - 2u^2d + 2uvd - 4u^2vh \\ &= 2uvd + h(2uv - 4u^2v) \\ &= 2uvd + 2uvh(1-2u) \\ &= 2uvd + 2uvh(1-u-u) \\ &= 2uvd + 2uvh(v-u) \\ &= 2uv[d + (v-u)h] \ . \end{aligned} \tag{3.16}$$

Therefore, the slope is

$$\begin{aligned} \beta_{y,x} &= \frac{\sigma_{xy}}{\sigma_x^2} \\ &= 2uv[d + (v-u)h]/2uv \\ &= d + (v-u)h \ . \end{aligned} \tag{3.17}$$

Following standard procedures in regression analysis, we can partition σ_y^2 into the variance due to the regression and the variance due to residual. The former is equivalent to the variance of the expected y; that is, the variance of the hypothetical points on the line in Figure 3.2, and the latter is the variance of the difference between observed y and the expected values.

The variance due to regression is

$$\begin{aligned} \beta\sigma_{xy} &= 2uv[d + (v-u)h][d + (v-u)h] \\ &= 2uv[d + (v-u)h]^2 \\ &= V_A \end{aligned} \tag{3.18}$$

and we may obtain the residual variance simply by subtracting the variance due to regression from the total variance of y. The variance of genotypic effects (σ_y^2) was given in 3.13, and when we subtract the expression obtained for the variance due to regression 3.18, we obtain the residual variances:

$$\begin{aligned} \sigma_y^2 - \beta\sigma_{x,y} &= 4u^2v^2h^2 \\ &= V_D \ . \end{aligned} \tag{3.19}$$

In this representation, genotypic effects are defined in terms of the regression line and are known as genotypic values. They are related to d and h, the genotypic effects we defined in Figure 3.1, but now reflect the population mean and gene frequencies of our random mating population. Defined in this way, the genotypic value (G) is $G = A + D$, the additive (A) and dominance (D) deviations of the individual.

G	$=$	A	$+$	D	frequency
G_{AA}	$=$	$2v[d + h(v - u)]$	$-$	$2v^2h$	u^2
G_{Aa}	$=$	$(v - u)[d + h(v - u)]$	$+$	$2uvh$	$2uv$
G_{aa}	$=$	$-2u[d + h(v - u)]$	$-$	$2u^2h$	v^2

In the case of $u = v = \frac{1}{2}$, this table becomes

G	$=$	A	$+$	D	frequency
G_{AA}	$=$	d	$-$	$\frac{1}{2}h$	$\frac{1}{4}$
G_{Aa}	$=$			$\frac{1}{2}h$	$\frac{1}{2}$
G_{aa}	$=$	$-d$	$-$	$\frac{1}{2}h$	$\frac{1}{4}$

from which it can be seen that the weighted sum of all G's is zero ($\sum f_i G_i = 0$). In this case the additive effect is the same as the genotypic effect as originally scaled, and the dominance effect is measured around a mean of $\frac{1}{2}h$. This representation of genotypic value accurately conveys the extreme nature of unusual genotypes. Let $d = h = 1$, an example of complete dominance. In that case, $G_{AA} = G_{Aa} = \frac{1}{2}$ and $G_{aa} = -1\frac{1}{2}$ on our scale. Thus, aa genotypes, which form only $\frac{1}{4}$ of the population, fall far below the mean of 0, while the remaining $\frac{3}{4}$ of the population genotypes fall only slightly above the mean of 0. Thus, the bulk of the population appears relatively normal, whereas aa genotypes appear abnormal or unusual. When dominance is absent ($h = 0$), Aa genotypes, which form $\frac{1}{2}$ of the population, have a mean of 0 and the less frequent genotypes AA and aa appear deviant. This situation is accentuated as the gene frequencies depart from $\frac{1}{2}$. For example, with $u = \frac{3}{4}$, $v = \frac{1}{4}$, and $h = d = 1$, then AA and Aa combined form $\frac{15}{16}$ of the population with a genotypic value of $\frac{1}{8}$, just slightly above the mean of 0, whereas the aa genotype has a value of $-1\frac{7}{8}$. In the limiting case of a very rare allele, AA and Aa tend to 0, the population mean, while only aa genotypes take an extreme value. These values intuitively correspond to our notion of a rare disorder of extreme effect, such as untreated phenylketonuria (PKU).

Table 3.3: Genetic covariance components for MZ, DZ, and Unrelated Siblings with unequal gene frequencies at a single locus.

Genotype	Effect		Frequency		
Pair	x_{1i}	x_{2i}	MZ	DZ	U
AA, AA	d	d	u^2	$u^4 + u^3v + \frac{1}{4}u^2v^2$	u^4
AA, Aa	d	h	–	$u^3v + \frac{1}{2}u^2v^2$	$2u^3v$
AA, aa	d	$-d$	—	$\frac{1}{4}u^2v^2$	u^2v^2
Aa, AA	h	d	—	$u^3v + \frac{1}{2}u^2v^2$	$2u^3v$
Aa, Aa	h	h	$2uv$	$u^3v + 3u^2v^2 + uv^3$	$4u^2v^2$
Aa, aa	h	$-d$	—	$\frac{1}{2}u^2v^2 + uv^3$	$2uv^3$
aa, AA	$-d$	d	—	$\frac{1}{4}u^2v^2$	u^2v^2
aa, Aa	$-d$	h	—	$\frac{1}{2}u^2v^2 + uv^3$	$2uv^3$
aa, aa	$-d$	$-d$	u^4	$\frac{1}{4}u^2v^2 + uv^3 + v^4$	v^4

The genotypic values A and D that we employ in the Mx model have precisely the expectations given above in 3.18 and 3.19, but are summed over all polygenic loci contributing to the trait. Thus, the biometrical model gives a precise definition to the latent variables employed in Mx for the analysis of twin data.

3.4 Summary

Table 3.3 replicates Table 3.2 employing genotypic frequencies appropriate to random mating and unequal gene frequencies. Using the table to calculate covariances among sibling pairs of the three types, MZ twins, DZ twins, and unrelated siblings, gives

$$\begin{array}{lclcl} \text{Cov(MZ)} & = & 2uv[d + (v-u)h]^2 + 4u^2v^2h^2 & = & V_A + V_D \\ \text{Cov(DZ)} & = & uv[d + (v-u)h]^2 + u^2v^2h^2 & = & \frac{1}{2}V_A + \frac{1}{4}V_D \\ \text{Cov(U)} & = & 0 & = & 0 \end{array}$$

By similar calculations, the expectations for half-siblings and for parents and their offspring may be shown to be $\frac{1}{4}V_A$ and $\frac{1}{2}V_A$, respectively. That is, these relationships do not reflect dominance effects. The MZ and DZ resemblances are the primary focus of this text, but all five relationships we have just discussed may be analyzed in the extended Mx approaches we discuss in Chapter **??**.

With more extensive genetical data, we can assess the effects of *epistasis*, or non-allelic interaction, since the biometrical model may be extended easily to include such genetic effects. Another important problem we have not considered is that of assortative mating, which one might have thought would introduce insuperable problems for the model. However, once we are working with genotypic values such as A and D, the effects of assortment can be readily accommodated in the model by means of reverse path analysis (Wright, 1968) and the Pearson–Aitken treatment of selected variables (Aitken, 1934). Fulker (1988) describes this approach in the context of Fisher's (1918) model of assortment.

In this chapter, we have given a brief introduction to the biometrical model that underlies the model fitting approach employed in this book, and we have indicated how additional genetic complexities may be accommodated in the model. However, in addition to genetic influences, we must consider the effects of the environment on any phenotype. These may be easily accommodated by defining environmental influences that are common to sib pairs and those that are unique to the individual. If these environmental effects are unrelated to the genotype, then the variances due to these influences simply add to the genetic variances we have just described. If they are not independent of genotype, as in the case of sibling interactions and cultural transmission, both of which are likely to occur in some behavioral phenotypes, then the Mx model may be suitably modified to account for these complexities, as we describe in Chapters 8 and **??**.

Chapter 4

Matrix Algebra

4.1 Introduction

Many people regard journal articles and books that contain matrix algebra as prohibitively complicated and ignore them or shelve them indefinitely. This is a sad state of affairs because learning matrix algebra is not difficult and can reap enormous benefits. Science in general, and genetics in particular, is becoming increasingly quantitative. Matrix algebra provides a very economical language to describe our data and our models; it is essential for understanding Mx and other data analysis packages. In common with most languages, the way to make it "stick" is to *use* it. Those unfamiliar with, or out of practice at, using matrices will benefit from doing the worked examples in the text. Readers with a strong mathematics background may skim this chapter, or skip it entirely, using it for reference only. We do not give an exhaustive treatment of matrix algebra and operations but limit ourselves to the bare essentials needed for structural equation modeling. There are many excellent texts for those wishing to extend their knowledge; we recommend Searle (1982) and Graybill (1969).

In this chapter, we will introduce matrix notation in Section 4.2 and matrix operations in Section 4.3. The general use of matrix algebra is illustrated in Section 4.4 on equations and Section 4.5 on other applications.

4.2 Matrix Notation

Although matrices and certain matrix operations were used as long ago as 2000 BC in ancient China, it is only relatively recently that a comprehensive matrix algebra has been developed. During the 1850's, Cayley worked on general algebraic systems (Boyer, 1985 p. 627) and developed the basis of matrix algebra as it is used today. The concept of a matrix is a very simple one, being just a table of numbers or symbols laid out in *rows* and *columns*,

$$\text{e.g.,} \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \text{ or } \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

In most texts, the table is enclosed in brackets, either: curved, (); square, [ ]; or curly, {}.

It is conventional to specify the configuration of the matrix in terms of Rows × Columns and these are its *dimensions* or *order*. Thus the first matrix above is of order 3 by 2 and the second is a 3 × 3 matrix.

A common occurrence of matrices in behavioral sciences is the *data matrix* where the *rows are subjects* and the *columns are measures,* e.g.,

	Weight	*Height*
S_1	50	20
S_2	100	40
S_3	150	60
S_4	200	80

It is convenient to let a single letter symbolize a matrix. This is written in UPPERCASE **boldface.** Thus we might say that our data matrix is **A**, which in handwriting we would underline with either a straight or a wavy line. Sometimes a matrix is written ${}_4\mathbf{A}_2$ to specify its dimensions. The economy of using matrices is immediately apparent: we can represent a whole table by a single symbol, whether it contains just one row and one column, or a billion rows and a billion columns! There are several special terms for matrices with one row or one column or both. When a matrix consists of a single number, it is called a *scalar*; when it consists of single column (row) of numbers it is called a column (row) *vector.* Scalars are usually represented as lower case, non-bold letters. Vectors are normally represented as a **bold** lowercase letter. Thus, the weight measurements of our four subjects are

$$\begin{bmatrix} 50 \\ 100 \\ 150 \\ 200 \end{bmatrix} = \mathbf{a}$$

We can refer to the specific elements of matrix **A** as a_{ij} where i indicates the row number and j indicates the column number.

Certain special forms of matrices exist. We have already defined scalars and row and column vectors. A matrix full of zeroes is called a *null* matrix and a matrix full of ones is called a *unit matrix.* Matrices in which the number of rows is equal to the number of columns are called *square* matrices. Among square matrices, *diagonal matrices* have at least one non-zero diagonal element, with every off-diagonal element zero. By diagonal, we mean the 'leading diagonal' from the top left element of the matrix to the bottom right element. A special form of the diagonal matrix is the *identity* matrix, $\mathbf{I}$, which has every diagonal element one and every non-diagonal element zero. The identity matrix functions much like the number one in ordinary algebra.

4.3 Matrix Algebra Operations

Matrix algebra defines a set of operations that may be performed on matrices. These operations include addition, subtraction, multiplication, inversion (multiplication by the inverse is similar to division) and transposition. We may separate the operations into two mutually exclusive categories: *unary* and *binary*. Unary operations are performed on a single matrix, and binary operations combine two matrices to obtain a single matrix result. Binary operations will be described first.

4.3.1 Binary Operations

Addition and subtraction

Matrices may be *added* if and only if they have the *same dimension*. They are then said to be *conformable for addition*. Each element in the first matrix is added to the corresponding element in the second matrix to form the same element in the solution.

$$\text{e.g.} \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} + \begin{pmatrix} 8 & 11 \\ 9 & 12 \\ 10 & 13 \end{pmatrix} = \begin{pmatrix} 9 & 15 \\ 11 & 17 \\ 13 & 19 \end{pmatrix}$$

or symbolically,

$$\mathbf{A} + \mathbf{B} = \mathbf{C}.$$

One *cannot add*

$$\begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} + \begin{pmatrix} 8 & 10 \\ 9 & 11 \end{pmatrix}$$

because they have a different number of rows. Subtraction works in the same way as addition, e.g.

$$\begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} - \begin{pmatrix} 2 & 5 \\ 2 & 5 \\ 2 & 5 \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{pmatrix}$$

which is written

$$\mathbf{A} - \mathbf{B} = \mathbf{C}.$$

Matrix multiplication

Matrices are *conformable for multiplication* if and only if the number of columns in the first matrix equals the number of rows in the second matrix. This means that *adjacent columns and rows must be of the same order.* For example, the matrix product ${}_3\mathbf{A}_2 \times {}_2\mathbf{B}_1$ may be calculated; the result is a 3×1 matrix. In general, if we multiply two matrices ${}_i\mathbf{A}_j \times {}_j\mathbf{B}_k$, the result will be of order $i \times k$.

Matrix multiplication involves calculating a *sum of cross products* among *rows of the first matrix* and *columns of the second matrix* in all possible combinations.

$$\text{e.g.} \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 1 \times 1 + 4 \times 2 & 1 \times 3 + 4 \times 4 \\ 2 \times 1 + 5 \times 2 & 2 \times 3 + 5 \times 4 \\ 3 \times 1 + 6 \times 2 & 3 \times 3 + 6 \times 4 \end{pmatrix} = \begin{pmatrix} 9 & 19 \\ 12 & 26 \\ 15 & 33 \end{pmatrix}$$

This is written

$$\mathbf{AB} = \mathbf{C}$$

The only exception to the above rule is multiplication by a *single number* called a scalar. Thus, for example,

$$2 \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} = \begin{pmatrix} 2 & 8 \\ 4 & 10 \\ 6 & 12 \end{pmatrix}$$

by convention this is often written as

$$2\mathbf{A} = \mathbf{C}.$$

Although convenient and often found in the literature, we do not recommend this style of matrix formulation, but prefer use of the kronecker product. The kronecker product of two matrices, symbolized $\mathbf{A} \otimes \mathbf{B}$ is formed by multiplying each element of $\mathbf{A}$ by the matrix $\mathbf{B}$. If $\mathbf{A}$ is a scalar, every element of the matrix $\mathbf{B}$ is multiplied by the scalar.

The simplest example of matrix multiplication is to multiply a vector by itself. If we premultiply a column vector $(n \times 1)$ by its transpose[1], the result is a scalar called the *inner product*. For example, if

$$\mathbf{a}' = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$$

then the inner product is

$$\mathbf{a}'\mathbf{a} = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = 1^2 + 2^2 + 3^2 = 14$$

which is the sum of the squares of the elements of the vector $\mathbf{a}$. This has a simple graphical representation when $\mathbf{a}$ is of dimension 2×1 (see Figure 4.1).

[1]Transposition is defined in Section 4.3.2 below. Essentially the rows become columns and *vice versa.*

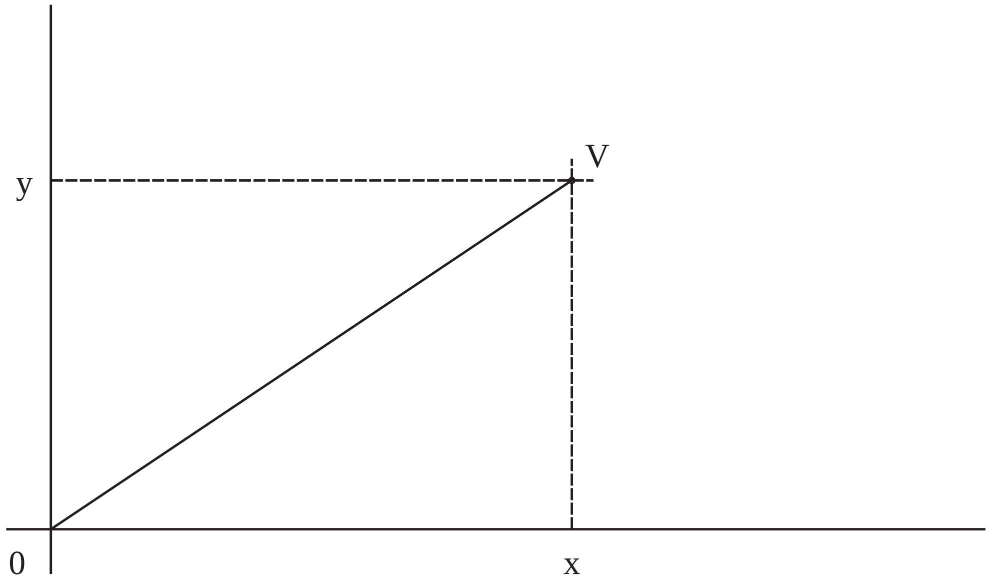


Figure 4.1: Graphical representation of the inner product $\mathbf{a}'\mathbf{a}$ of a (2×1) vector $\mathbf{a}$, with $\mathbf{a}' = (xy)$. By Pythagoras' theorem, the distance of the point V from the origin O is $\sqrt{x^2 + y^2}$, which is the square root of the inner product of the vector.

4.3.2 Unary Operations

Transposition

A matrix is transposed when the rows are written as columns and the columns are written as rows. This operation is denoted by writing $\mathbf{A}'$ or $\mathbf{A}^T$. For our example data matrix on page 74,

$$\mathbf{A}' = \begin{pmatrix} 50 & 100 & 150 & 200 \\ 20 & 40 & 60 & 80 \end{pmatrix}$$

a row vector is usually written

$$\mathbf{a}' = \begin{pmatrix} 50 & 100 & 150 & 200 \end{pmatrix}$$

Clearly, $(\mathbf{A}')' = \mathbf{A}$.

Determinant of a matrix

For a square matrix $\mathbf{A}$ we may calculate a scalar called the *determinant* which we write as $|\mathbf{A}|$. In the case of a 2×2 matrix, this quantity is calculated as

$$|\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21}.$$

We shall be giving numerical examples of calculating the determinant when we address matrix inversion. The determinant has an interesting geometric representation. For example, consider two standardized variables that correlate r. This situation may be represented graphically by drawing two vectors, each of length 1.0, having the same origin and an angle a, whose cosine is r, between them (see Figure 4.2).

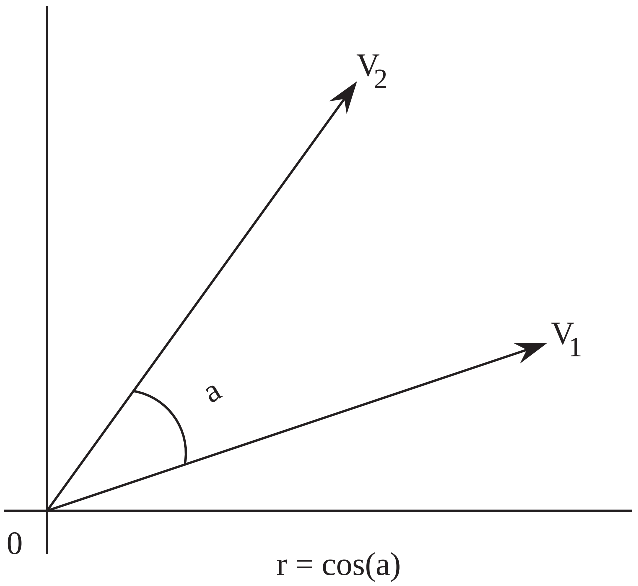


Figure 4.2: Geometric representation of the determinant of a matrix. The angle between the vectors is the cosine of correlation between two variables, so the determinant is given by twice the area of the triangle OV_1V_2.

It can be shown (the proof involves symmetric square root decomposition of matrices) that the area of the triangle OV_1V_2 is $.5\sqrt{|\mathbf{A}|}$. Thus as the correlation r increases, the angle between the lines decreases, the area decreases, and *the determinant decreases.* For two variables that correlate perfectly, the determinant of the correlation (or covariance) matrix is zero. Conversely, the determinant is at a maximum when $r = 0$; the angle between the vectors is 90°, and we say that the variables are *orthogonal.* For larger numbers of variables, the determinant is a function of the hypervolume in n-space; if any single pair of variables correlates perfectly then the determinant is zero. In addition, if one of the variables is a linear combination of the others, the determinant will be zero. For a set of variables with given variances, the determinant is maximized when all the variables are orthogonal, i.e., all the off-diagonal elements are zero.

Many software packages [e.g., Mx; SAS, 1985] and numerical libraries (e.g., IMSL, 1987; NAG, 1990) have algorithms for finding the determinant and inverse of a matrix. But it is useful to know how matrices can be inverted by hand, so we present a method for use with paper and pencil. To calculate the determinant of larger matrices, we employ the concept of a *cofactor*. If we delete row i and column j from an $n \times n$ matrix, then the determinant of the remaining matrix is called the *minor* of element a_{ij}. The cofactor, written A_{ij} is simply:

$$A_{ij} = (-1)^{i+j} \text{minor } a_{ij}$$

The determinant of the matrix $\mathbf{A}$ may be calculated as

$$|\mathbf{A}| = \sum_{i=1}^{n} a_{ij} A_{ij}$$

where n is the order of $\mathbf{A}$.

The determinant of a matrix is related to the concept of *definiteness* of a matrix. In general, for a null column vector $\mathbf{x}$, the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ is always zero. For some matrices, this quadratic is zero *only* if $\mathbf{x}$ is the null vector. If $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ for all non-null vectors $\mathbf{x}$ then we say that the matrix is *positive definite*. Conversely, if $\mathbf{x}'\mathbf{A}\mathbf{x} < 0$ for all non-null $\mathbf{x}$, we say that the matrix is *negative definite*. However, if we can find some non-null $\mathbf{x}$ such that $\mathbf{x}'\mathbf{A}\mathbf{x} = 0$ then the matrix is said to be *singular*, and its determinant is zero. As long as no two variables are perfectly correlated, and there are more subjects than measures, a covariance matrix calculated from data on random variables will be *positive definite*. Mx will complain (and rightly so!) if it is given a covariance matrix that is not positive definite. The determinant of the covariance matrix can be helpful when there are problems with model-fitting that seem to originate with the data. However, it is possible to have a matrix with a positive determinant yet which is negative definite (consider $-\mathbf{I}$ with an even number of rows), so the determinant is not an adequate diagnostic. Instead we note that all the eigenvalues of a positive definite matrix are greater than zero. Eigenvalues and eigenvectors may be obtained from software packages, including Mx, and the numerical libraries listed above[2].

Trace of a matrix

The trace of a matrix is simply the sum of its diagonal elements. Thus the trace of the matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = 1 + 5 + 9 = 15$$

[2] Those readers wishing to know more about the uses of eigenvalues and eigenvectors may consult Searle (1982) or any general text on matrix algebra.

Inverse of a matrix

In ordinary algebra the division operation $a \div b$ is equivalent to multiplication of the reciprocal $a \times \frac{1}{b}$. Thus one binary operation, division, has been replaced by two operations, one binary (multiplication) and one unary (forming $\frac{1}{b}$). In matrix algebra we make an equivalent substitution of operations, and we call the unary operation *inversion.* We write the inverse of the matrix $\mathbf{A}$ as $\mathbf{A}^{-1}$, and calculate it so that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

and

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}\,,$$

where $\mathbf{I}$ is the identity matrix. In general the inverse of a matrix is not simply formed by finding the reciprocal of each element (this holds only for scalars and diagonal matrices[3]), but is a more complicated operation involving the determinant.

There are many computer programs available for inverting matrices. Some routines are general, but there are often faster routines available if the program is given some information about the matrix, for example, whether it is symmetric, positive definite, triangular, or diagonal. Here we describe one general method that is useful for matrix inversion; we recommend undertaking this hand calculation at least once for at least a 3×3 matrix in order to fully understand the concept of a matrix inverse.

Procedure: In order to invert a matrix, the following four steps can be used:

1. Find the determinant
2. Set up the matrix of cofactors
3. Transpose the matrix of cofactors
4. Divide by the determinant

For example, the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 1 & 5 \end{pmatrix}$$

can be inverted by:

1.

$$|\mathbf{A}| = (1 \times 5) - (2 \times 1) = 3$$

2.

$$A_{ij} = \begin{bmatrix} (-1)^2 \times 5 & (-1)^3 \times 1 \\ (-1)^3 \times 2 & (-1)^4 \times 1 \end{bmatrix} = \begin{pmatrix} 5 & -1 \\ -2 & 1 \end{pmatrix}$$

[3]N.B. For a diagonal matrix one takes the reciprocal of only the diagonal elements!

3.

$$A'_{ij} = \begin{pmatrix} 5 & -2 \\ -1 & 1 \end{pmatrix}$$

4.

$$\mathbf{A^{-1}} = \frac{1}{3}\begin{pmatrix} 5 & -2 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} \frac{5}{3} & -\frac{2}{3} \\ -\frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

To verify this, we can multiply $\mathbf{AA}^{-1}$ to obtain the identity matrix:

$$\frac{1}{3}\begin{pmatrix} 5 & -2 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 1 & 5 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

The result that $\mathbf{AA}^{-1} = \mathbf{I}$ may be used to solve the pair of simultaneous equations:

$$\begin{aligned} x_1 + 2x_2 &= 8 \\ x_1 + 5x_2 &= 17 \end{aligned}$$

which may be written

$$\begin{pmatrix} 1 & 2 \\ 1 & 5 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 8 \\ 17 \end{pmatrix}$$

i.e.,

$$\mathbf{Ax} = \mathbf{y}$$

premultiplying both sides by the inverse of $\mathbf{A}$, we have

$$\begin{aligned} \mathbf{A}^{-1}\mathbf{Ax} &= \mathbf{A}^{-1}\mathbf{y} \\ \mathbf{x} &= \mathbf{A}^{-1}\mathbf{y} \\ &= \frac{1}{3}\begin{pmatrix} 5 & -2 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 8 \\ 17 \end{pmatrix} \\ &= \frac{1}{3}\begin{pmatrix} 6 \\ 9 \end{pmatrix} \\ &= \begin{pmatrix} 2 \\ 3 \end{pmatrix} \end{aligned}$$

which may be verified by substitution.

For a larger matrix it is more tedious to compute the inverse. Let us consider the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}$$

1. The determinant is

$$|\mathbf{A}| \quad = \quad +1\begin{vmatrix} 0 & 1 \\ -1 & 0 \end{vmatrix} - 1\begin{vmatrix} 1 & 1 \\ 1 & 0 \end{vmatrix} + 0\begin{vmatrix} 1 & 0 \\ 1 & -1 \end{vmatrix} = +1 + 1 + 0 = 2$$

2. The matrix of cofactors is:

$$A_{ij} = \begin{bmatrix} + \begin{vmatrix} 0 & 1 \\ -1 & 0 \end{vmatrix} & - \begin{vmatrix} 1 & 1 \\ 1 & 0 \end{vmatrix} & + \begin{vmatrix} 1 & 0 \\ 1 & -1 \end{vmatrix} \\ - \begin{vmatrix} 1 & 0 \\ -1 & 0 \end{vmatrix} & + \begin{vmatrix} 1 & 0 \\ 1 & 0 \end{vmatrix} & - \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} \\ + \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} & - \begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix} & + \begin{vmatrix} 1 & 1 \\ 1 & 0 \end{vmatrix} \end{bmatrix} = \begin{pmatrix} 1 & 1 & -1 \\ 0 & 0 & 2 \\ 1 & -1 & -1 \end{pmatrix}$$

3. The transpose is

$$A'_{ij} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & -1 \\ -1 & 2 & -1 \end{pmatrix}$$

4. Dividing by the determinant, we have

$$\mathbf{A}^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & -1 \\ -1 & 2 & -1 \end{pmatrix} = \begin{pmatrix} .5 & 0 & .5 \\ .5 & 0 & -.5 \\ -.5 & 1 & -.5 \end{pmatrix}$$

which may be verified by multiplication with $\mathbf{A}$ to obtain the identity matrix.

4.4 Equations in Matrix Algebra

Matrix algebra provides a very convenient short hand for writing sets of equations. For example, the pair of *simultaneous equations*

$$\begin{aligned} y_1 &= 2x_1 + 3x_2 \\ y_2 &= x_1 + x_2 \end{aligned}$$

may be written

$$\mathbf{y} \quad = \quad \mathbf{A}\mathbf{x}$$

i.e.,

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Also if we have the following pair of equations:

$$\begin{aligned} \mathbf{y} &= \mathbf{Ax} \\ \mathbf{x} &= \mathbf{Bz}, \end{aligned}$$

then

$$\begin{aligned} \mathbf{y} &= \mathbf{A}(\mathbf{Bz}) \\ &= \mathbf{ABz} \\ &= \mathbf{Cz} \end{aligned}$$

where $\mathbf{C} = \mathbf{AB}$. This is very convenient notation compared with direct substitution. The Mx structural equations are written in this general form, i.e.,

Real variables (y) = Matrix $\times$ hypothetical variables.

To show the simplicity of the matrix notation, consider the following equations:

$$\begin{aligned} y_1 &= 2x_1 + 3x_2 \\ y_2 &= x_1 + x_2 \\ x_1 &= z_1 + z_2 \\ x_2 &= z_1 - z_2 \end{aligned}$$

Then we have

$$\begin{aligned} y_1 &= 2(z_1 + z_2) + 3(z_1 - z_2) \\ &= 5z_1 - z_2 \\ y_2 &= (z_1 + z_2) + (z_1 - z_2) \\ &= 2z_1 + 0 \end{aligned}$$

Similarly, in matrix notation, we have $\mathbf{y} = \mathbf{ABz}$, where

$$\mathbf{A} = \begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

and

$$\mathbf{AB} = \begin{pmatrix} 5 & -1 \\ 2 & 0 \end{pmatrix},$$

or

$$\begin{aligned} y_1 &= 5z_1 - z_2 \\ y_2 &= 2z_2 \end{aligned}$$

4.5 Applications of Matrix Algebra

Matrix algebra is used extensively throughout multivariate statistics (see e.g., Graybill, 1969; Mardia *et al.*, 1979; Maxwell, 1977; Searle, 1982). Here we do not propose to discuss statistical methods, but simply to show two examples of the utility of matrices in expressing general formulae applicable to any number of variables or subjects.

4.5.1 Calculation of Covariance Matrix from Data Matrix

Suppose we have a data matrix **A** with rows corresponding to subjects and columns corresponding to variables. We can calculate a mean for each variable and replace the data matrix with a matrix of *deviations from the mean*. That is, each element a_{ij} is replaced by $a_{ij} - \mu_j$ where μ_j is the mean of the j^{th} variable. Let us call the new matrix **Z**. The covariance matrix is then simply calculated as

$$\frac{1}{N-1}\mathbf{Z}'\mathbf{Z}$$

where N is the number of subjects.

For example, suppose we have the following data:

X	Y	$X - \overline{X}$	$Y - \overline{Y}$
1	2	-2	-4
2	8	-1	2
3	6	0	0
4	4	1	-2
5	10	2	4

So the matrix of deviations from the mean is

$$\mathbf{Z} = \begin{pmatrix} -2 & -4 \\ -1 & 2 \\ 0 & 0 \\ 1 & -2 \\ 2 & 4 \end{pmatrix}$$

and therefore the covariance matrix of the observations is

$$\frac{1}{N-1}\mathbf{Z}'\mathbf{Z} = \frac{1}{4}\begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ -4 & 2 & 0 & -2 & 4 \end{pmatrix}\begin{pmatrix} -2 & -4 \\ -1 & 2 \\ 0 & 0 \\ 1 & -2 \\ 2 & 4 \end{pmatrix}$$

$$
\begin{aligned}
&= \frac{1}{4}\begin{pmatrix} 10 & 12 \\ 12 & 40 \end{pmatrix} \\
&= \begin{pmatrix} 2.5 & 3.0 \\ 3.0 & 10.0 \end{pmatrix} = \begin{pmatrix} S_x^2 & S_{xy} \\ S_{xy} & S_y^2 \end{pmatrix}
\end{aligned}
$$

The diagonal elements of this matrix are the variances of the variables, and the off-diagonal elements are the covariances between the variables. The *standard deviation* is the square root of the variance (see Chapter 2).

The correlation is

$$\frac{S_{xy}}{\sqrt{S_x^2 S_y^2}} = \frac{S_{xy}}{S_x S_y}$$

In general, a correlation matrix may be calculated from a covariance matrix by pre- and post-multiplying the covariance matrix by a diagonal matrix $\mathbf{D}$ in which each diagonal element d_{ii} is $\frac{1}{S_i}$, i.e., the reciprocal of the standard deviation for that variable. Thus, in our two variable example, we have:

$$\begin{pmatrix} \frac{1}{S_x} & 0 \\ 0 & \frac{1}{S_y} \end{pmatrix} \begin{pmatrix} S_x^2 & S_{xy} \\ S_{xy} & S_y^2 \end{pmatrix} \begin{pmatrix} \frac{1}{S_x} & 0 \\ 0 & \frac{1}{S_y} \end{pmatrix} = \begin{pmatrix} 1.0 & R_{xy} \\ R_{xy} & 1.0 \end{pmatrix}$$

4.5.2 Transformations of Data Matrices

Matrix algebra provides a natural notation for *transformations*. If we premultiply the matrix ${}_i\mathbf{B}_j$ by another, say ${}_k\mathbf{T}_i$, then the rows of $\mathbf{T}$ describe linear combinations of the rows of $\mathbf{B}$. The resulting matrix will therefore consist of k rows corresponding to the linear transformations of the rows of $\mathbf{B}$ described by the rows of $\mathbf{T}$. A very simple example of this is premultiplication by the identity matrix, $\mathbf{I}$, which, as noted earlier, merely has 1's on the leading diagonal and zeroes everywhere else. Thus, the transformation described by the first row may be written as 'multiply the first row by 1 and add zero times the other rows.' In the second row, we have 'multiply the second row by 1 and add zero times the other rows,' and so the identity matrix transforms the matrix $\mathbf{B}$ into the same matrix. For a less trivial example, let our data matrix be $\mathbf{X}$, then

$$\mathbf{X}' = \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ -4 & 2 & 0 & -2 & 4 \end{pmatrix}$$

and let

$$\mathbf{T} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

then

$$
\begin{aligned}
\mathbf{Y}' &= \mathbf{TX}' \\
&= \begin{pmatrix} -6 & 1 & 0 & -1 & 6 \\ 2 & -3 & 0 & 3 & -2 \end{pmatrix}.
\end{aligned}
$$

In this case, the transformation matrix specifies two transformations of the data: the first row defines the sum of the two variates, and the second row defines the difference (row 1 − row 2). In the above, we have applied the transformation to the raw data, but for these linear transformations it is easy to apply the transformation to the covariance matrix. The covariance matrix of the transformed variates is

$$\begin{aligned} \frac{1}{N-1}\mathbf{Y}'\mathbf{Y} &= \frac{1}{N-1}(\mathbf{TX}')(\mathbf{TX}')' \\ &= \frac{1}{N-1}\mathbf{TX}'\mathbf{XT}' \\ &= \mathbf{T}(\mathbf{V_x})\mathbf{T}' \end{aligned}$$

which is a useful result, meaning that linear transformations may be applied directly to the covariance matrix, instead of going to the trouble of transforming all the raw data and recalculating the covariance matrix.

4.5.3 Further Operations and Applications

There exists a great variety of matrix operations and functions with much broader scope than the limited selection given in this chapter. For example, there are two other forms of matrix multiplication in common use, direct or kronecker products, and dot products. Similar extensions to addition and subtraction exist, and numerous matrix functions beyond determinant and trace can be defined. One place to study further operations is Searle (1982); applications and some definitions can be found in Neale (1997). We hope that the outline provided here will make understanding structural equation modeling of twin data much easier, and provide a starting point for those who wish to study the subject in more detail.

4.6 Exercises

If you find these exercises insufficient practice, more may be found in almost any text on matrix algebra. Further practice may be obtained by computing the expected covariance matrix of almost any model in this book, selecting a set of trial values for the parameters. The exercise can be extended by computing fit functions (Chapter **??**) for the model and parameter values selected. For the purposes of general introduction, however, the few given in this section should suffice.

4.6.1 Binary operations

Let

$$\mathbf{A} = \begin{pmatrix} 3 & 6 \\ 2 & 1 \end{pmatrix}, \ \mathbf{B} = \begin{pmatrix} 1 & 0 & 3 & 2 \\ 0 & -1 & -1 & 1 \end{pmatrix}$$

1. Form **AB**.
2. Form **BA**. (Careful, this might be a trick question!)

Let

$$\mathbf{C} = \begin{pmatrix} 3 & 6 \\ 2 & 1 \end{pmatrix}, \ \mathbf{D} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

1. Form **CD**.
2. Form **DC**.
3. In ordinary algebra, multiplication is *commutative*, i.e. $xy = yx$. In general, is matrix multiplication commutative?

Let

$$\mathbf{E}' = \begin{pmatrix} 1 & 0 & 3 \\ 1 & 2 & 1 \end{pmatrix}$$

1. Form $\mathbf{E}(\mathbf{C}+\mathbf{D})$.
2. Form $\mathbf{EC}+\mathbf{ED}$.
3. In ordinary algebra, multiplication is *distributive over addition*, i.e. $x(y+z) = xy+xz$. In general, is matrix multiplication distributive over matrix addition? Is matrix multiplication distributive over matrix subtraction?

4.6.2 Unary operations

1. Show for two (preferably non-trivial) matrices conformable for multiplication that
$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$$
2. If **C** is
$$\begin{pmatrix} 2 & 6 \\ .5 & 4 \end{pmatrix},$$
find the determinant of **C**.
3. What is the inverse of matrix **C**?
4. If **D** is
$$\begin{pmatrix} .2 & .3 \\ .4 & .6 \end{pmatrix},$$
find the determinant of **D**.
5. What is the inverse of **D**?
6. If $tr(\mathbf{A})$ means the trace of **A**, what is $tr(\mathbf{C})$ + $tr(\mathbf{D})$?

Chapter 5

Path Analysis and Structural Equations

5.1 Introduction

Path analysis was invented by the geneticist Sewall Wright (1921a, 1934, 1960, 1968), and has been widely applied to problems in genetics and the behavioral sciences. It is a technique which allows us to represent, in diagrammatic form, linear 'structural' models and hence derive predictions for the variances and covariances (the *covariance structure*) of our variables under that model. The books by Kenny (1979), Li (1975), or Wright (1968) supply good introductory treatments of path analysis, and general descriptions of structural equation modeling can be found in Bollen (1989) and Loehlin (1987). In this chapter we provide only the basic background necessary to understand models used in the genetic analyses presented in this text.

A path diagram is a useful heuristic tool to graphically display causal and correlational relations or the paths between variables. Used correctly, it is one of several mathematically complete descriptions of a linear model, which include less visually immediate forms such as (i) structural equations and (ii) expected covariances derived in terms of the parameters of the model. Since all three forms are mathematically complete, it is possible to translate from one to another for such purposes as applying it to data, increasing understanding of the model, verifying its identification, or presenting results.

The advantage of the path method is that it goes beyond measuring the degree of association by the correlation coefficient or determining the best prediction by the regression coefficient. Instead, the user makes explicit hypotheses about relationships between the variables which are quantified by path coefficients. Better still, the model's predictions may be statistically compared with the observed data, as we shall go on to discuss in Chapters **??** and **??**. Path models are in fact extremely general, subsuming a large number of multivariate methods, including (but not limited to) multiple regression, principle component or factor analysis, canonical correlation, discriminant analysis and multivariate analysis of variance and covariance. Therefore those that take exception to 'path analysis' in its broadest sense, should be aware that they dismiss a vast array of multivariate statistical methods.

We begin by considering the conventions used to draw and read a path diagram, and explain the difference between correlational paths and causal paths (Section 5.2). In Sections 5.3 and 5.4 we briefly describe assumptions of the method and tracing rules for path diagrams. Then, to illustrate their use, we present simple linear regression models familiar to most readers (Section 5.5). We define these both as path diagrams and as structural equations — some individuals handle path diagrams more easily, others respond better to equations! We also apply the method to two basic representations of a simple genetic model for covariation in twins (Section 5.6), with special reference to the identity between the matrix specification of a model and its graphical representation. Finally we discuss identification of models and parameters in Section 5.7.

5.2 Conventions Used in Path Analysis

A path diagram usually consists of boxes and circles, which are connected by arrows. Consider the diagram in Figure 5.1 for example.

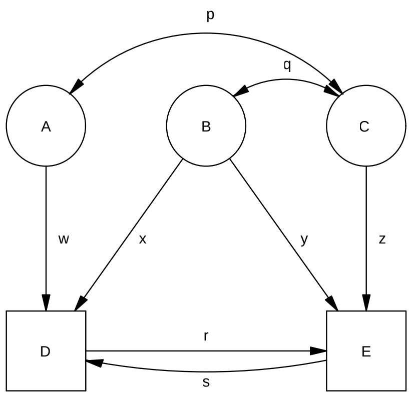


Figure 5.1: Path diagram for three latent (A, B and C) and two observed (D and E) variables, illustrating correlations (p and q) and path coefficients (r, s, w, x, y and z).

Squares or rectangles are used to enclose *observed* (manifest or measured) variables, and circles or ellipses surround *latent* (unmeasured) variables.

Single-headed arrows ('paths') are used to define causal relationships in the model, with the variable at the tail of the arrow causing the variable at the head. Omission of a path from one variable to another implies that there is no direct causal influence of the former variable on the latter. In the path diagram in (Figure 5.1) D is determined by A and B, while E is determined by B and C. When two variables cause each other, we say that there is a feedback-loop, or 'reciprocal causation' between them. Such a feedback-loop is shown between variables D and E in our example.

Double-headed arrows are used to represent a covariance between two variables, which might arise through a common cause or their reciprocal causation or both. In many treatments of path analysis, double-headed arrows may be placed *only* between variables that do not have causal arrows pointing at them. This convention allows us to discriminate between *dependent/endogenous* variables and *independent/ultimate/exogenous* variables.

Dependent variables are those variables we are trying to predict (in a regression model) or whose intercorrelations we are trying to explain (in a factor model). Dependent variables may be determined or caused by either independent variables or other dependent variables or both. In Figure 5.1, D and E are the dependent variables. *Independent* variables are the variables that explain the intercorrelations between the dependent variables or, in the case of the simplest regression models, predict the dependent variables. The causes of independent variables are not represented in the model. A, B and C are the independent variables in Figure 5.1.

Omission of a double-headed arrow reflects the hypothesis that two independent variables are uncorrelated. In Figure 5.1 the independent variables B and C correlate, C also correlates with A, but A does not correlate with B. This illustrates (i) that two variables which correlate with a third do not necessarily correlate with each other, and (ii) that when two factors cause the same dependent variable, it does not imply that they correlate. In some treatments of path analysis, a double-headed arrow from an independent variable to itself is used to represent its variance, but this is often omitted if the variable is standardized to unit variance. However, for completeness and mathematical correctness, we do recommend to always include the standardized variance arrows.

By convention, lower-case letters (or numeric values, if these can be specified) are used to represent the values of paths or double-headed arrows, in contrast to the use of upper-case for variables. We call the values corresponding to causal paths *path coefficients,* and those of the double-headed arrows simply *correlation coefficients* (see Figure 5.1 for examples). In some applications, subscripts identify the origin and destination of a path. The first subscript refers to the variable being caused, and the second subscript tells which variable is doing the causing. In most genetic applications we assume that the variables are scaled as deviations from the means, in which case the constant intercept terms in equations will be zero and can be omitted from the structural equations.

Each dependent variable usually has a *residual*, unless it is fixed to zero *ex-hypothesi*. The residual variable does not correlate with any other determinants of its dependent variable, and will usually (but not always) be uncorrelated with other independent variables.

In summary therefore, the conventions used in path analysis:

- Observed variables are enclosed in squares or rectangles. Latent variables are enclosed in circles or ellipses. Error variables are included in the path diagram, and may be enclosed by circles or ellipses or (occasionally) not enclosed at all.
- Upper-case letters are used to denote observed or latent variables, and lower-case letters or numeric values represent the values of paths or two-way arrows, respectively called path coefficients and correlation coefficients.
- A one-way arrow between two variables indicates a postulated direct influence of one variable on another. A two-way arrow between two variables indicates that these variables may be correlated without any assumed direct relationship.
- There is a fundamental distinction between independent variables and dependent variables. Independent variables are not caused by any other variables in the system.
- Coefficients may have two subscripts, the first indicating the variable to which arrow points, the second showing its origin.

5.3 Assumptions of Path Analysis

Sewall Wright (Wright, 1968, p. 299) described path diagrams in the following manner:

> "[In path analysis] *every included variable, measured or hypothetical, is represented by arrows as either completely determined by certain others* (the dependent variables), *which may in turn be represented as similarly determined, or as an ultimate variable* (our independent variables). *Each ultimate factor in the diagram must be connected by lines with arrowheads at both ends with each of the other ultimate factors, to indicate possible correlations through still more remote, unrepresented factors, except in cases in which it can safely be assumed that there is no correlation .... the strict validity of the method depends on the properties of formally complete linear systems of unitary variables.*"

Some assumptions of the method, implicit or explicit in Wright's description, are:

- Linearity: All relationships between variables are linear. The assumption of a linear model seems valid as a wide variety of non-linear functions are well approximated by linear ones particularly within a limited range. (Sometimes non-linearity can be removed by appropriate transformation of the data prior to statistical analysis; but some models are inherently non-linear).

- Causal closure: All direct influences of one variable on another must be included in the path diagram. Hence the non-existence of an arrow between two variables means that it is assumed that these two variables are not directly related. The formal completeness of the diagram requires the introduction of residual variables if they are not represented as one of the ultimate variables, unless there is reason to assume complete additivity and determination by the specified factors.

- Unitary Variables: Variables may not be composed of components that behave in different ways with different variables in the system, but they should vary as a whole. For example, if we have three variables, A, B, and C, but A is really a composite of A1 and A2, and A1 is positively correlated with B and C, but A2 is positively correlated with B but negatively correlated with C, we have a potential for disaster!

5.4 Tracing Rules of Path Analysis

One of the greatest advantages of path diagrams is their foundation upon standard rules for reading paths, called "tracing rules," which yield the expected variances and covariances among the variables in the diagram.

In this section we first describe the tracing rules for standardized variables, following Wright's (1934, 1968) development of the method, and then outline the rules for unstandardized variables. Although nearly all path diagrams may be traced using rules for unstandardized variables,[1] we present path derivations for standardized and unstandardized variables separately because the former are much easier to trace than the latter, and because rules for unstandardized variables are fairly simple generalizations of the principles used in tracing paths between standardized variables. An excellent resource for learning tracing rules is the program RAMPATH (McArdle and Boker, 1990), which has a 'draw_bridges' command that illustrates the rules for any model.

5.4.1 Tracing Rules for Standardized Variables

The basic principle of tracing rules is described by Sewall Wright (1934) with the following words:

[1] Multivariate path diagrams, including delta path (van Eerdewegh, 1982), copath (Cloninger, 1980), and conditional path diagrams (Carey, 1986a) employ slightly different rules, but are outside the scope of this book. See Vogler (1985) for a general description.

> *"Any correlation between variables in a network of sequential relations can be analyzed into contributions from all the paths (direct or through common factors) by which the two variables are connected, such that the value of each contribution is the product of the coefficients pertaining to the elementary paths. If residual correlations are present (represented by bidirectional arrows) one (but never more than one) of the coefficients thus multiplied together to give the contribution of the connecting path, may be a correlation coefficient. The others are all path coefficients."*

In general, the expected correlation between two variables in a path diagram of standardized variables may be derived by tracing all connecting routes (or "chains") between the variables, while adhering to the following conditions. One may:

1. Trace backward along an arrow and then forward, or simply forwards from one variable to the other but *never forward and then back*
2. Pass through each variable only once in each chain of paths
3. Trace through at most one two-way arrow in each chain of paths

A corollary of the first rule is that one may *never pass through adjacent arrowheads.*

The contribution of each chain traced between two variables to their expected correlation is the *product* of its standardized coefficients. The expected correlation between two variables is the sum of the contributions of all legitimate routes between those two variables. Note that these rules assume that there are no feedback loops; i.e., that the model is 'recursive'.

5.4.2 Tracing Rules for Unstandardized Variables

If we are working with unstandardized variables, the tracing rules of the previous section are insufficient to derive expected correlations. However, in the absence of paths from dependent variables to other dependent variables, expected *covariances*, rather than correlations, may be derived with only slight modifications to the tracing rules (see Heise, 1975):

1. At any change of direction in a tracing route which is not a two-way arrow connecting different variables in the chain, the expected variance of the variable at the point of change is included in the product of path coefficients; thus, any path from an dependent variable to an independent variable will include the double-headed arrow from the independent variable to itself, unless it also includes a double-headed arrow connecting that variable to another independent variable (since this would violate the rule against passing through adjacent arrowheads)
2. In deriving variances, the path from an dependent variable to an independent variable and back to itself is only counted once

Perhaps a simpler approach to unstandardized path analysis is to make certain that all residual variances are included explicitly in the diagram with double-headed arrows pointing to the variable itself. Then the chains between two variables are formed simply if we

1. Trace backwards, change direction at a two-headed arrow, then trace forwards.

As before, the expected covariance is computed by multiplying all the coefficients in a chain and summing over all possible chains. We consider chains to be different if either a) they don't have the same coefficients, or b) the coefficients are in a different order. For a clear and thorough mathematical treatment, see the RAMPATH manual (McArdle and Boker, 1990).

5.5 Path Models for Linear Regression

In this Section we attempt to clarify the conventions, the assumptions and the tracing rules of path analysis by applying them to regression models. The path diagram in Figure 5.2a represents a linear regression model, such as might be used, for example, to predict systolic blood pressure [SBP], Y_1 from sodium intake X_1. The model asserts that high sodium intake is a *cause*, or *direct effect*, of high blood pressure (i.e., sodium intake $\rightarrow$ blood pressure), but that blood pressure also is influenced by other, unmeasured ('residual'), factors. The regression equation represented in Figure 5.2a is

$$Y_1 = a_1 + b_{11}X_1 + E_1, \tag{5.1}$$

where a is a constant intercept term, b_{11} the regression or 'structural' coefficient, and E_1 the residual error term or disturbance term, which is uncorrelated with X_1. This is indicated by the absence of a double-headed arrow between X_1 and E_1 or an indirect common cause between them [$\mathrm{Cov}(X_1,E_1) = 0$]. The double-headed arrow from X_1 to itself represents the variance of this variable: $\mathrm{Var}(X) = s_{11}$; the variance of E_1 is $\mathrm{Var}(E) = z_{11}$. In this example SBP is the dependent variable and sodium intake is the independent variable.

We can extend the model by adding more independent variables or more dependent variables or both. The path diagram in Figure 5.2b represents a multiple regression model, such as might be used if we were trying to predict SBP (Y_1) from sodium intake (X_1), exercise (X_2), and body mass index [BMI] (X_3), allowing once again for the influence of other residual factors (E_1) on blood pressure. The double-headed arrows between the three independent variables indicate that correlations are allowed between sodium intake and exercise (s_{21}), sodium intake and BMI (s_{31}), and BMI and exercise (s_{32}). For example, a negative covariance between exercise and sodium intake might arise if the health-conscious exercised more and ingested less sodium; positive covariance between sodium intake and BMI

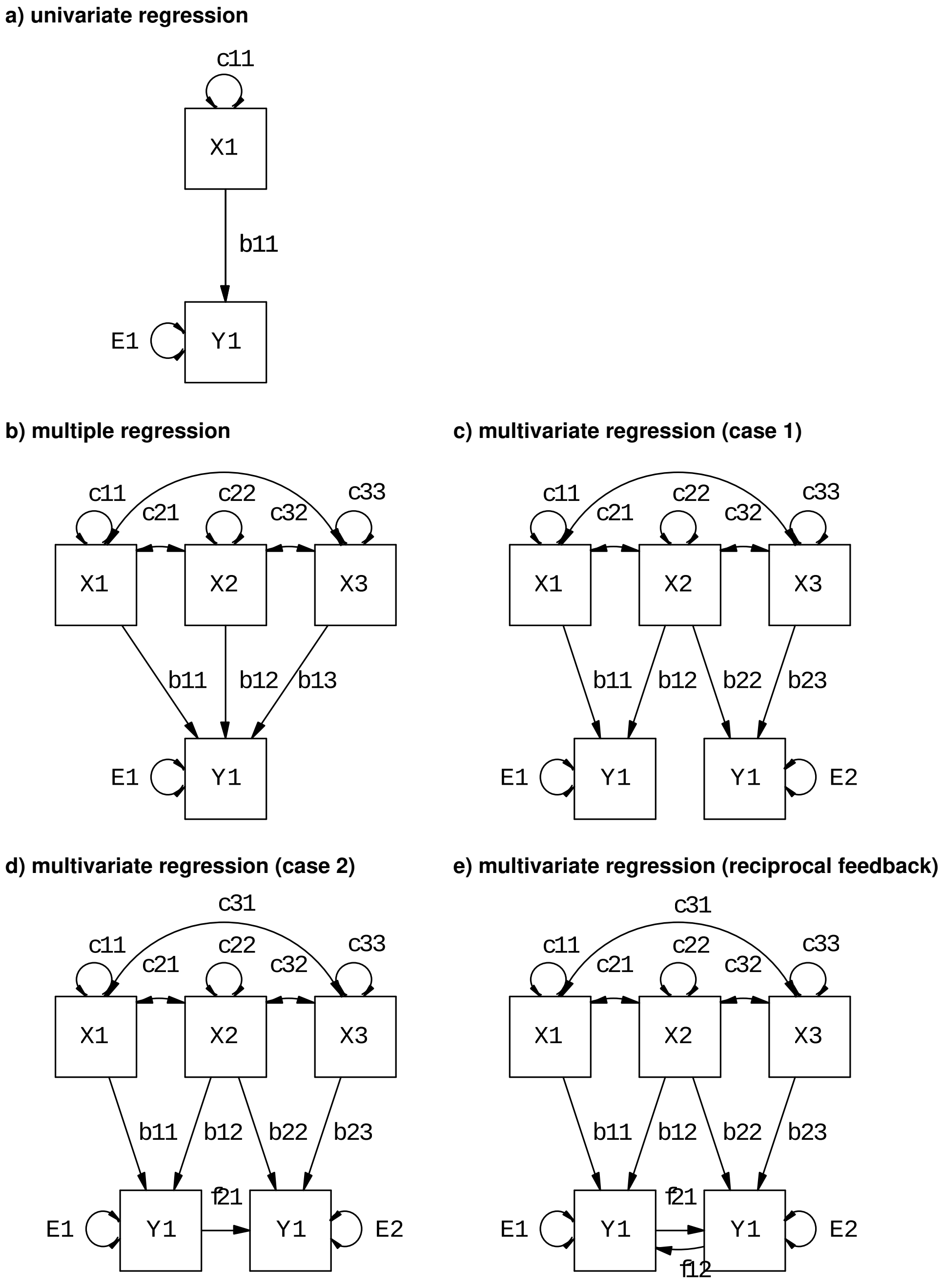


Figure 5.2: Regression path models with manifest variables.

could occur if obese individuals ate more (and therefore ingested more sodium); and a negative covariance between BMI and exercise could exist if overweight people were less inclined to exercise. In this case the regression equation is

$$Y_1 = a_1 + b_{11}X_1 + b_{12}X_2 + b_{13}X_3 + E_1. \tag{5.2}$$

Note that the estimated values for a_1, b_{11} and E_1 will not usually be the same as in equation 5.1 due to the inclusion of additional independent variables in the multiple regression equation 5.2. Similarly, the only difference between Figures 5.2a and 5.2b is that we have multiple independent or predictor variables in Figure 5.2b.

Figure 5.2c represents a multivariate regression model, where we now have two dependent variables (blood pressure, Y_1, and a measure of coronary artery disease [CAD], Y_2), as well as the same set of independent variables (case 1). The model postulates that there are direct influences of sodium intake and exercise on blood pressure, and of exercise and BMI on CAD, but no direct influence of sodium intake on CAD, nor of BMI on blood pressure. Because the X_2 variable, exercise, causes both blood pressure, Y_1, and coronary artery disease, Y_2, it is termed a *common cause* of these dependent variables. The regression equations are

$$Y_1 = a_1 + b_{11}X_1 + b_{12}X_2 + E_1$$

and

$$Y_2 = a_2 + b_{22}X_2 + b_{23}X_3 + E_2. \tag{5.3}$$

Here a_1 and E_1 are the intercept term and error term, respectively, and b_{11} and b_{12} the regression coefficients for predicting blood pressure, and a_2, E_2, b_{22}, and b_{23} the corresponding coefficients for predicting coronary artery disease. We can rewrite equation 5.3 using matrices (see Chapter 4 on matrix algebra),

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & 0 \\ 0 & b_{22} & b_{23} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}$$

or, using matrix notation,

$$\mathbf{y} = \mathbf{a} + \mathbf{Bx} + \mathbf{Ie},$$

where $\mathbf{y}$, $\mathbf{a}$, $\mathbf{x}$, and $\mathbf{e}$ are column vectors and $\mathbf{B}$ is a matrix of regression coefficients and $\mathbf{I}$ is an identity matrix. Note that each variable in the path diagram which has an arrow pointing to it appears exactly one time on the left side of the matrix expression.

Figure 5.2d differs from Figure 5.2c only by the addition of a causal path (f_{12}) from blood pressure to coronary artery disease, implying the hypothesis that high blood pressure increases CAD (case 2). The presence of this path also provides a link between Y_2 and X_1 ($Y_2 \leftarrow Y_1 \leftarrow X_1$); this type of process with multiple intervening variables is typically called an *indirect effect* (of X_1 on Y_2). Thus we see that dependent variables can be influenced by other dependent variables, as well as by independent variables. Figure 5.2e adds an additional causal path from CAD to blood pressure (f_{21}), thus creating a 'feedback-loop' (hereafter designated as $\Longleftrightarrow$) between CAD and blood pressure. If both f parameters are positive, the interpretation of the model would be that high SBP increases CAD and increased CAD in turn increases SBP. Such reciprocal causation of variables requires special treatment and is discussed further in Chapters 8 and ??. Figure 5.2e implies the structural equations

$$Y_1 = a_1 + f_{12}Y_2 + b_{11}X_1 + b_{12}X_2 + E_1$$

and

$$Y_2 = a_2 + f_{21}Y_1 + b_{22}X_2 + b_{23}X_3 + E_2 \tag{5.4}$$

In matrix form, we may write these equations as

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} 0 & f_{12} \\ f_{21} & 0 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & 0 \\ 0 & b_{22} & b_{23} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}$$

i.e.,

$$\mathbf{y} = \mathbf{a} + \mathbf{Fy} + \mathbf{Bx} + \mathbf{Ie}$$

Now that some examples of regression models have been described both in the form of path diagrams and structural equations, we can apply the tracing rules of path analysis to derive the expected variances and covariances under the models. The regression models presented in this chapter are all examples of unstandardized variables. We illustrate the derivation of the expected variance or covariance between some variables by applying the tracing rules for unstandardized variables in Figures 5.2a, 5.2b and 5.2c. As an exercise, the reader may wish to trace some of the other paths.

In the case of Figure 5.2a, to derive the expected covariance between X_1 and Y_1, we need trace only the path:

$$\text{(i)} \qquad X_1 \overset{s_{11}}{\longleftrightarrow} X_1 \xrightarrow{b_{11}} Y_1$$

yielding an expected covariance of $(s_{11}b_{11})$. Two paths contribute to the expected variance of Y_1,

$$\begin{aligned} &\text{(i)} \qquad Y_1 \xleftarrow{b_{11}} X_1 \overset{s_{11}}{\longleftrightarrow} X_1 \xrightarrow{b_{11}} Y_1, \\ &\text{(ii)} \qquad Y_1 \xleftarrow{1} E_1 \overset{z_{11}}{\longleftrightarrow} E_1 \xrightarrow{1} Y_1; \end{aligned}$$

yielding an expected variance of Y_1 of $(b_{11}^2 s_{11} + z_{11})$.

In the case of Figure 5.2b, to derive the expected covariance of X_1 and Y_1, we can trace paths:

$$
\begin{array}{ll}
\text{(i)} & Y_1 \xleftarrow{b_{11}} X_1 \xleftrightarrow{s_{11}} X_1, \\
\text{(ii)} & Y_1 \xleftarrow{b_{12}} X_2 \xleftrightarrow{s_{21}} X_1, \\
\text{(iii)} & Y_1 \xleftarrow{b_{13}} X_3 \xleftrightarrow{s_{31}} X_1,
\end{array}
$$

to obtain an expected covariance of $(b_{11}s_{11} + b_{12}s_{21} + b_{13}s_{31})$. To derive the expected variance of Y_1, we can trace paths:

$$
\begin{array}{ll}
\text{(i)} & Y_1 \xleftarrow{b_{11}} X_1 \xleftrightarrow{s_{11}} X_1 \xrightarrow{b_{11}} Y_1, \\
\text{(ii)} & Y_1 \xleftarrow{b_{12}} X_2 \xleftrightarrow{s_{22}} X_2 \xrightarrow{b_{12}} Y_1, \\
\text{(iii)} & Y_1 \xleftarrow{b_{13}} X_3 \xleftrightarrow{s_{33}} X_3 \xrightarrow{b_{13}} Y_1, \\
\text{(iv)} & Y_1 \xleftarrow{1} E_1 \xleftrightarrow{z_{11}} E_1 \xrightarrow{1} Y_1, \\
\text{(v)} & Y_1 \xleftarrow{b_{11}} X_1 \xleftrightarrow{s_{21}} X_2 \xrightarrow{b_{12}} Y_1, \\
\text{(vi)} & Y_1 \xleftarrow{b_{12}} X_2 \xleftrightarrow{s_{21}} X_1 \xrightarrow{b_{11}} Y_1, \\
\text{(vii)} & Y_1 \xleftarrow{b_{11}} X_1 \xleftrightarrow{s_{31}} X_3 \xrightarrow{b_{13}} Y_1, \\
\text{(viii)} & Y_1 \xleftarrow{b_{13}} X_3 \xleftrightarrow{s_{31}} X_1 \xrightarrow{b_{11}} Y_1, \\
\text{(ix)} & Y_1 \xleftarrow{b_{12}} X_2 \xleftrightarrow{s_{32}} X_3 \xrightarrow{b_{13}} Y_1, \\
\text{(x)} & Y_1 \xleftarrow{b_{13}} X_3 \xleftrightarrow{s_{32}} X_2 \xrightarrow{b_{12}} Y_1,
\end{array}
$$

yielding a total expected variance of $(b_{11}^2 s_{11} + b_{12}^2 s_{22} + b_{13}^2 s_{33} + 2b_{11}b_{12}s_{21} + 2b_{11}b_{13}s_{31} + 2b_{12}b_{13}s_{32} + z_{11})$.

In the case of Figure 5.2c, we may derive the expected covariance of Y_1 and Y_2 as the sum of

$$
\begin{array}{ll}
\text{(i)} & Y_1 \xleftarrow{b_{11}} X_1 \xleftrightarrow{s_{21}} X_2 \xrightarrow{b_{22}} Y_2, \\
\text{(ii)} & Y_1 \xleftarrow{b_{11}} X_1 \xleftrightarrow{s_{31}} X_3 \xrightarrow{b_{23}} Y_2, \\
\text{(iii)} & Y_1 \xleftarrow{b_{12}} X_2 \xleftrightarrow{s_{22}} X_2 \xrightarrow{b_{22}} Y_2, \\
\text{(iv)} & Y_1 \xleftarrow{b_{12}} X_2 \xleftrightarrow{s_{32}} X_3 \xrightarrow{b_{23}} Y_2,
\end{array}
$$

giving $[b_{11}(s_{21}b_{22} + s_{31}b_{23}) + b_{12}(s_{22}b_{22} + s_{32}b_{23})]$ for the expected covariance. This expectation, and the preceding ones, can be derived equally (and arguably more easily) by simple matrix algebra. For example, the expected covariance matrix $(\mathbf{\Sigma})$ for Y_1 and Y_2 under the model of Figure 5.2c is given as

$$\mathbf{\Sigma} = \mathbf{BSB}' + \mathbf{Z},$$

$$= \begin{pmatrix} b_{11} & b_{12} & 0 \\ 0 & b_{22} & b_{33} \end{pmatrix} \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{pmatrix} \begin{pmatrix} b_{11} & 0 \\ b_{12} & b_{22} \\ 0 & b_{23} \end{pmatrix} + \begin{pmatrix} z_{11} & 0 \\ 0 & z_{22} \end{pmatrix}$$

in which the elements of **B** are the paths from the **X** variables (columns) to the **Y** variables (rows); the elements of **S** are the covariances between the independent variables; and the elements of **Z** are the residual error variances.

5.6 Path Models for the Classical Twin Study

To introduce genetic models and to further illustrate the tracing rules both for standardized variables and unstandardized variables, we examine some simple genetic models of resemblance. The classical twin study, in which MZ twins and DZ twins are reared together in the same home is one of the most powerful designs for detecting genetic and shared environmental effects. Once we have collected such data, they may be summarized as observed covariance matrices (Chapter 2), but in order to test hypotheses we need to derive expected covariance matrices from the model. We first digress briefly to review the biometrical principles outlined in Chapter 3, in order to express the ideas in a path–analytic context.

In contrast to the regression models considered in previous sections, many genetic analyses of family data postulate independent variables (genotypes and environments) as *latent* rather than manifest variables. In other words, the genotypes and environments are not measured directly but their influence is inferred through their effects on the covariances of relatives. However, we can represent these models as path diagrams in just the same way as the regression models. The brief introduction to path-analytic genetic models we give here will be treated in greater detail in Chapter 6, and thereafter.

From quantitative genetic theory (see Chapter 3), we can write equations relating the phenotypes P_i and P_j of relatives i and j (e.g., systolic blood pressures of first and second members of a twin pair), to their underlying genotypes and environments. We may decompose the total genetic effect on a phenotype into that due to the additive effects of alleles at multiple loci, that due to the dominance effects at multiple loci, and that due to the epistatic interactions between loci (Mather and Jinks, 1982). Similarly, we may decompose the total environmental effect into that due to environmental influences shared by twins or sibling pairs reared in the same family ('shared', 'common', or 'between-family' environmental effects), and that due to environmental effects which make family members differ from one another ('random', 'specific', or 'within-family' environmental effects). Thus, the observed phenotypes, P_i and P_j, are assumed to be linear functions of the underlying additive genetic variance (A_i and A_j), dominance variance (D_i and D_j), shared environmental variance (C_i and C_j) and random environmental variance (E_i and

E_j). In quantitative genetic studies of human populations, epistatic genetic effects are usually confounded with dominance genetic effects, and so will not be considered further here. Assuming all variables are scaled as deviations from zero, we have

$$P_1 = e_1 E_1 + c_1 C_1 + a_1 A_1 + d_1 D_1$$

and

$$P_2 = e_2 E_2 + c_2 C_2 + a_2 A_2 + d_2 D_2$$

Particularly for pairs of twins, we do not expect the magnitude of genetic or environmental effects to vary as a function of relationship[2] so we set $e_1 = e_2 = e$, $c_1 = c_2 = c$, $a_1 = a_2 = a$, and $d_1 = d_2 = d$. In matrix form, we write

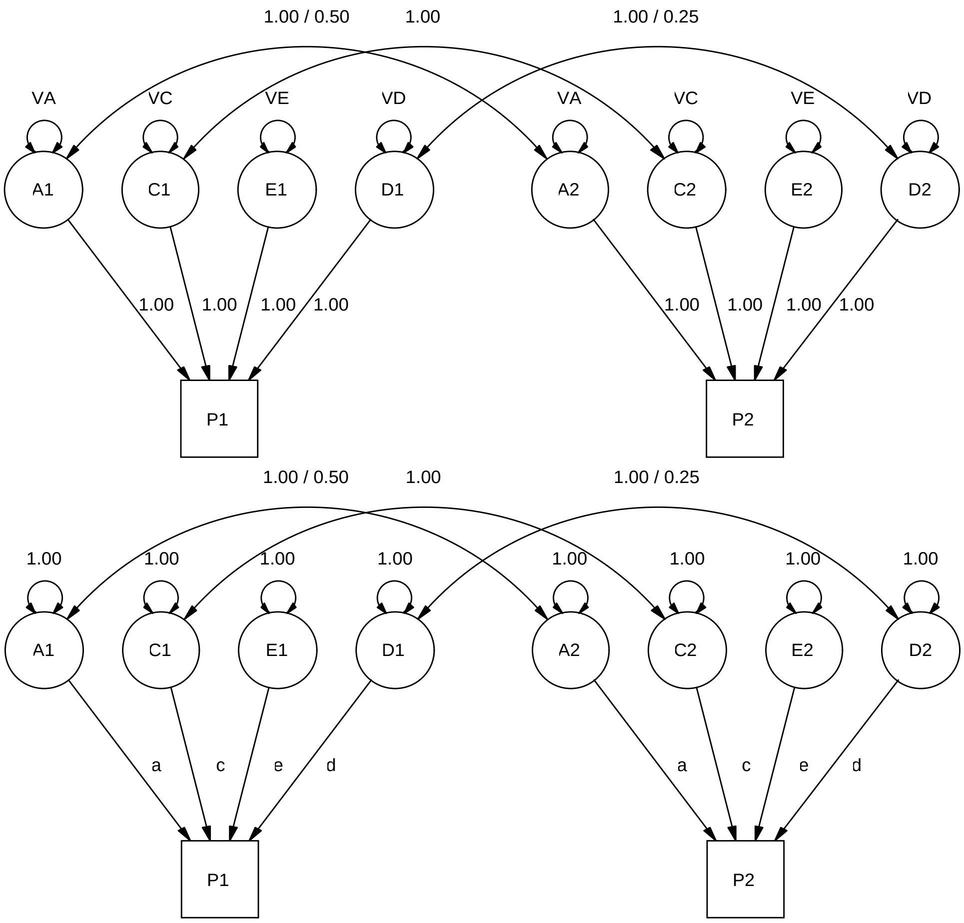
$$\begin{pmatrix} P_1 \\ P_2 \end{pmatrix} = \begin{pmatrix} e & c & a & d & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & e & c & a & d \end{pmatrix} \begin{pmatrix} E_1 \\ C_1 \\ A_1 \\ D_1 \\ E_2 \\ C_2 \\ A_2 \\ D_2 \end{pmatrix}.$$

Unless two or more waves of measurement are used, or several observed variables index the phenotype under study, residual effects are included in the random environmental component, and are not separately specified in the model.

Figures 5.3a and 5.3b represent two alternative parameterizations of the basic genetic model, illustrated for the case of pairs of monozygotic twins (MZ) or dizygotic twins (DZ), who may be reared together (MZT, DZT) or reared apart (MZA, DZA). In Figure 5.3a, the traditional *path coefficients model,* the variances of the latent variables A_1, C_1, E_1, D_1 and A_2, C_2, E_2, D_2 are standardized ($V_E = V_C = V_A = V_D = 1$, and the path coefficients e, c, a, or d — quantifying the paths from the latent variables to the observed variable, measured on both twins, P_1 and P_2 — are free parameters to be estimated. Figure 5.3b is called a *variance components model* because it fixes $e = c = a = d = 1$, and estimates separate random environmental, shared environmental, additive genetic and dominance genetic variances instead.

The traditional path model illustrates tracing rules for standardized variables, and is straightforward to generalize to multivariate problems; the variance components model illustrates an unstandardized path model. Provided all parameter estimates are non-negative, tracing the paths in either parameterization will give the same solution, with $V_A = a^2$, $V_D = d^2$, $V_C = c^2$ and $V_E = e^2$.

[2] i.e. we do not expect different heritabilities for twin 1 and twin 2; however for other relationships such as parents and children, the assumption may not be valid, as could be established empirically if we had genetically informative data in both generations.



a)

b)

Figure 5.3: Alternative representations of the basic genetic model: a) traditional path coefficients model, and b) variance components model.

5.6.1 Path Coefficients Model: Example of Standardized Tracing Rules

When applying the tracing rules, it helps to draw out each tracing route to ensure that they are neither forgotten nor traced twice. In the traditional path model of Figure 5.3a, to derive the expected twin covariance for the case of monozygotic twin pairs reared together, we can trace the following routes:

$$\begin{array}{ll} \text{(i)} & P1 \xleftarrow{c} C1 \xleftrightarrow{1} C2 \xrightarrow{c} P2 \\ \text{(ii)} & P1 \xleftarrow{a} A1 \xleftrightarrow{1} A2 \xrightarrow{a} P2 \\ \text{(iii)} & P1 \xleftarrow{d} D1 \xleftrightarrow{1} D2 \xrightarrow{d} P2 \end{array}$$

so that the expected covariance between MZ twin pairs reared together will be

$$r_{MZ} = c^2 + a^2 + d^2. \tag{5.5}$$

In the case of dizygotic twin pairs reared together, we can trace the following routes:

$$\begin{array}{ll} \text{(i)} & P1 \xleftarrow{c} C1 \xleftrightarrow{1} C2 \xrightarrow{c} P2 \\ \text{(ii)} & P1 \xleftarrow{a} A1 \xleftrightarrow{0.5} A2 \xrightarrow{a} P2 \\ \text{(iii)} & P1 \xleftarrow{d} D1 \xleftrightarrow{0.25} D2 \xrightarrow{d} P2 \end{array}$$

yielding an expected covariance between DZ twin pairs of

$$r_{DZ} = c^2 + 0.5a^2 + 0.25d^2. \tag{5.6}$$

The expected variance of a variable — again assuming we are working with standardized variables — is derived by tracing all possible routes from the variable back to itself, without violating any of the tracing rules given in Section 5.4.1 above. Thus, following paths from P1 to itself we have

$$\begin{array}{ll} \text{(i)} & P1 \xleftarrow{e} E1 \xrightarrow{e} P1 \\ \text{(ii)} & P1 \xleftarrow{c} C1 \xrightarrow{c} P1 \\ \text{(iii)} & P1 \xleftarrow{a} A1 \xrightarrow{a} P1 \\ \text{(iv)} & P1 \xleftarrow{d} D1 \xrightarrow{d} P1 \end{array}$$

yielding the predicted variance for $P1$ or $P2$ in Figure 5.3a of

$$V_P = e^2 + c^2 + a^2 + d^2. \tag{5.7}$$

An important assumption implicit in Figure 5.3 is that an individual's additive genetic deviation is uncorrelated with his or her shared environmental deviation (i.e., there are no arrows connecting the latent C and A variables of an individual). In Chapter **??** we shall discuss how this assumption can be relaxed. Also implicit in the coefficient of 0.5 for the covariance of the additive genetic values of DZ twins or siblings is the assumption of random mating, which we shall also relax in Chapter **??**.

5.6.2 Variance Components Model: Example of Unstandardized Tracing Rules

Following the unstandardized tracing rules, the expected covariances of twin pairs in the variance components model of Figure 5.3b, are also easily derived. For the case of monozygotic twin pairs reared together (MZT), we can trace the following routes:

$$
\begin{array}{ll}
\text{(i)} & P1 \xleftarrow{1} C1 \xleftrightarrow{V_C} C2 \xrightarrow{1} P2 \\
\text{(ii)} & P1 \xleftarrow{1} A1 \xleftrightarrow{V_A} A2 \xrightarrow{1} P2 \\
\text{(iii)} & P1 \xleftarrow{1} D1 \xleftrightarrow{V_D} D2 \xrightarrow{1} P2
\end{array}
$$

so that the expected covariance between MZ twin pairs reared together will be

$$\text{Cov}(MZT) = V_C + V_A + V_D.$$

Only the latter two chains contribute to the expected covariance of MZ twin pairs reared apart, as they do not share their environment. The expected covariance of MZ twin pairs reared apart (MZA)is thus

$$\text{Cov}(MZA) = V_A + V_D.$$

In the case of dizygotic twin pairs reared together (DZT), we can trace the following routes:

$$
\begin{array}{ll}
\text{(i)} & P1 \xleftarrow{1} C1 \xleftrightarrow{V_C} C2 \xrightarrow{1} P2 \\
\text{(ii)} & P1 \xleftarrow{1} A1 \xleftrightarrow{0.5V_A} A2 \xrightarrow{1} P2 \\
\text{(iii)} & P1 \xleftarrow{1} D1 \xleftrightarrow{0.25V_D} D2 \xrightarrow{1} P2
\end{array}
$$

yielding an expected covariance between DZ twin reared together of

$$\text{Cov}(DZT) = V_C + 0.5V_A + 0.25V_D.$$

Similarly, the expected covariance of DZ twin pairs reared apart (DZA) is

$$\text{Cov}(DZA) = 0.5V_A + 0.25V_D.$$

In deriving expected variances of unstandardized variables, any chain from a dependent variable to an independent variable will include the double-headed arrow from the independent variable to itself (unless it also includes a double-headed arrow connecting that variable to another independent variable) and each path from an dependent variable to an independent variable and back to itself is only counted once. In this example the expected phenotypic variance, for all groups of relatives, is easily derived by tracing all the paths from $P1$ to itself:

$$\begin{array}{ll} \text{(i)} & P1 \xleftarrow{1} E1 \xleftrightarrow{V_E} E1 \xrightarrow{1} P1 \\ \text{(ii)} & P1 \xleftarrow{1} C1 \xleftrightarrow{V_C} C1 \xrightarrow{1} P1 \\ \text{(iii)} & P1 \xleftarrow{1} A1 \xleftrightarrow{V_A} A1 \xrightarrow{1} P1 \\ \text{(iv)} & P1 \xleftarrow{1} D1 \xleftrightarrow{V_D} D1 \xrightarrow{1} P1 \end{array}$$

yielding the predicted variance for P1 or P2 in Figure 5.3b of

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

The equivalence between Figures 5.3a and 5.3b comes from the biometrical principles outlined in Chapter 3: a^2, c^2, e^2, and d^2 are defined as $\frac{V_A}{V_P}$, $\frac{V_C}{V_P}$, $\frac{V_E}{V_P}$, and $\frac{V_D}{V_P}$, respectively. Since correlations are calculated as covariances divided by the product of the square roots of the variances (see Chapter 2), the twin correlations in Figure 5.3a may be derived using the covariances and variances in Figure 5.3b. Thus, in Figure 5.3b, the correlation for MZ pairs reared together is

$$\begin{aligned} r_{\text{MZT}} &= \frac{V_C + V_A + V_D}{\sqrt{(V_C + V_A + V_D + V_E)}\sqrt{(V_C + V_A + V_D + V_E)}} \\ &= \frac{V_C + V_A + V_D}{V_P} \\ &= \frac{V_C}{V_P} + \frac{V_A}{V_P} + \frac{V_D}{V_P} \\ &= c^2 + a^2 + d^2 \end{aligned}$$

Similarly, the correlations for MZ twins reared apart, and for DZ twins together and apart are

$$\begin{aligned} r_{\text{MZA}} &= a^2 + d^2 \\ r_{\text{DZT}} &= c^2 + 0.5a^2 + 0.25d^2 \\ r_{\text{DZA}} &= 0.5a^2 + 0.25d^2, \end{aligned}$$

as in the case of Figure 5.3a.

5.7 Identification of Models and Parameters

One key issue with structural equation modeling is whether a model, or a parameter within a model is *identified*. We say that the free parameters of a model are either

(i) overidentified; (ii) just identified; or (iii) underidentified. If all of the parameters fall into the first two classes, we say that the model as a whole is identified, but if one or more parameters are in class (iii), we say that the model is not identified. In this section, we briefly address the identification of parameters in structural equation models, and illustrate how data from additional types of relative may or may not identify the parameters of a model.

When we applied the rules of standardized path analysis to the simple path coefficient model for twins (Figure 5.3a), we obtained expressions for MZ and DZ covariances and the phenotypic variance:

$$\begin{aligned}
\mathrm{Cov}(MZ) &= c^2 + a^2 + d^2 & (5.8)\\
\mathrm{Cov}(DZ) &= c^2 + .5a^2 + .25d^2 & (5.9)\\
V_P &= c^2 + a^2 + d^2 + e^2 & (5.10)
\end{aligned}$$

These three equations have four unknown parameters c, a, d and e, and illustrate the first point about identification. *A model is underidentified if the number of free parameters is greater than the number of distinct statistics that it predicts.* Here there are four unknown parameters but only three distinct statistics, so the model is underidentified.

One way of checking the identification of simple models is to represent the expected variances and covariances as a system of equations in matrix algebra:

$$\mathbf{Ax} = \mathbf{b}$$

where $\mathbf{x}$ is the vector of parameters, $\mathbf{b}$ is the vector of observed statistics, and $\mathbf{A}$ is the matrix of weights such that element A_{ij} gives the coefficient of parameter j in equation i. Then, if the inverse of $\mathbf{A}$ exists, the model is identified. Thus in our example we have:

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & .5 & .25 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} c^2 \\ a^2 \\ d^2 \\ e^2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}. \qquad (5.11)$$

where b_1 is Cov(MZ), b_2 is Cov(DZ), and b_3 is V_P. Now, what we would really like to find here is the left inverse, $\mathbf{L}$, of $\mathbf{A}$ such that $\mathbf{LA} = \mathbf{I}$. However, it is easy to show that left inverses may exist only when $\mathbf{A}$ has at least as many rows as it does columns (for proof see, e.g., Searle, 1982, p. 147). Therefore, if we are limited to data from a classical twin study, i.e. MZ and DZ twins reared together, it is necessary to assume that one of the parameters a, c or d is zero to identify the model. Let us suppose that we have reason to believe that c can be ignored, so that the equations may be rewritten as:

$$\begin{pmatrix} 1 & 1 & 0 \\ .5 & .25 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} a^2 \\ d^2 \\ e^2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

and in this case, the inverse of **A** exists[3]. Another, generally superior, approach to resolving the parameters of the model is to collect new data. For example, if we collected data from separated MZ or DZ twins, then we could add a fourth row to **A** in equation 5.11 to get (for MZ twins apart)

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & .5 & .25 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} c^2 \\ a^2 \\ d^2 \\ e^2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} \tag{5.12}$$

where b_4 is Cov(MZA), and again the inverse of **A** exists. Now it is not necessarily the case that adding another type of relative (or type of rearing environment) will turn an underidentified model into one that is identified! Far from it, in fact, as we show with reference to siblings reared together, and half-siblings and cousins reared apart. Under our simple genetic model, the expected covariances of the siblings and half-siblings are

$$\begin{aligned} \text{Cov}(Sibs) &= c^2 + .5a^2 + .25d^2 && (5.13)\\ \text{Cov}(Half - sibs) &= .25a^2 && (5.14)\\ \text{Cov}(Cousins) &= .125a^2 && (5.15)\\ V_P &= c^2 + a^2 + d^2 + e^2 && (5.16) \end{aligned}$$

as could be shown by extending the methods outlined in Chapter 3. In matrix form the equations are:

$$\begin{pmatrix} 1 & .5 & .25 & 0 \\ 0 & .25 & 0 & 0 \\ 0 & .125 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} c^2 \\ a^2 \\ d^2 \\ e^2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}. \tag{5.17}$$

where b_1 is Cov(Sibs), b_2 is Cov(Half-sibs), b_3 is Cov(Cousins), and b_4 is V_P. Now in this case, although we have as many types of relationship with different expected covariance as there are unknown parameters in the model, we still cannot identify all the parameters, because the matrix **A** is singular. The presence of data collected from cousins does not add any information to the system, because their expected covariance is exactly half that of the half-siblings. In general, if any row (column) of a matrix can be expressed as a linear combination of the other rows (columns) of a matrix, then the matrix is singular and cannot be inverted. Note, however, that just because we cannot identify the model as a whole, it does not mean that none of the parameters can be estimated. In this example, we can obtain a valid estimate of additive genetic variance a^2 simply from, say, eight times the difference of the half-sib and cousin covariances. With this knowledge and the

[3]The reader may like to verify this by calculating the determinant according to the method laid out in Section 4.3.2 or with the aid of a computer.

observed full sibling covariance, we could estimate the *combined* effect of dominance and the shared environment, but it is impossible to separate these two sources.

Throughout the above examples, we have taken advantage of their inherent simplicity. The first useful feature is that the parameters of the model only occur in linear combinations, so that, e.g., terms of the form c^2a are not present. While true of a number of simple genetic models that we shall use in this book, it is not the case for them all (see Table **??** for example). Nevertheless, some insight may be gained by examining the model in this way, since if we are able to identify both c and c^2a then both parameters may be estimated. Yet for complex systems this can prove a difficult task, so we suggest an alternative, numerical approach. The idea is to simulate expected covariances for certain values of the parameters, and then see whether a program such as Mx can recover these values from a number of different starting points. If we find another set of parameter values that generates the same expected variances and covariances, the model is not identified. We shall not go into this procedure in detail here, but simply note that it is very similar to that described for power calculations in Chapter 7.

5.8 Summary

In this chapter we have reviewed briefly the use of path analysis to represent certain linear and genetic models. We have discussed the conventions of path analysis, and shown how it may be used to derive the covariance matrices predicted under a particular model. We emphasize that the systems described here have been chosen as simple examples to illustrate elementary principles of path analysis. Although these examples are somewhat simplistic in the sense that they do not elucidate many of the characteristics of which structural equation models are capable, familiarity with them should provide sufficient skills for comprehension of other, more advanced, genetic models described in this text and for development of one's own path models.

However, one aspect of structural models which has not been discussed in this chapter is that of *multiple indicators*. While not strictly a feature of path analysis, multiple indicator models, — those with more than one measure for each dependent or independent variable — warrant some attention because they are used often in genetic analyses of twin data, and in analyses of behavioral data in general. Our initial regression examples from Figure 5.2 assumed that we had only a single measure for each variable (systolic blood pressure, sodium intake, etc), and could ignore measurement error in these observed variables. Inclusion of multiple indicators allows for explicit representation of assumptions about measurement error in a model. In our regression example of Figures 5.2d and e, for example, we might have several measures of our independent (x) variables, a number of measures of sodium intake (e.g., diet diary and urinary sodium), multiple measures of exercise (e.g., exercise diary and frequency checklist), and numerous measures of obesity (e.g., self-report body mass index, measures of skinfold thickness). Likewise, we might have many estimates of our dependent η variables, such as repeated measures of blood pressure, and several tests for coronary artery disease. Figure 5.4 expands Figure 5.2a by illustrating the cases of (a) one variable per construct, (b) two variables per construct, and (c) three or more observed variables per construct.

Covariance and variance expectations for multiple indicator models such as those shown in Figure 5.4 follow without exception from the path tracing rules outlined earlier in this chapter. However, the increase in number of variables in these models often results in substantial increases in model complexity. One of the important attractions of Mx is its flexibility in specifying models using matrix algebra. Various commands are available that allow changing the number of variables with relative ease. It is to the Mx model specification that we now turn.

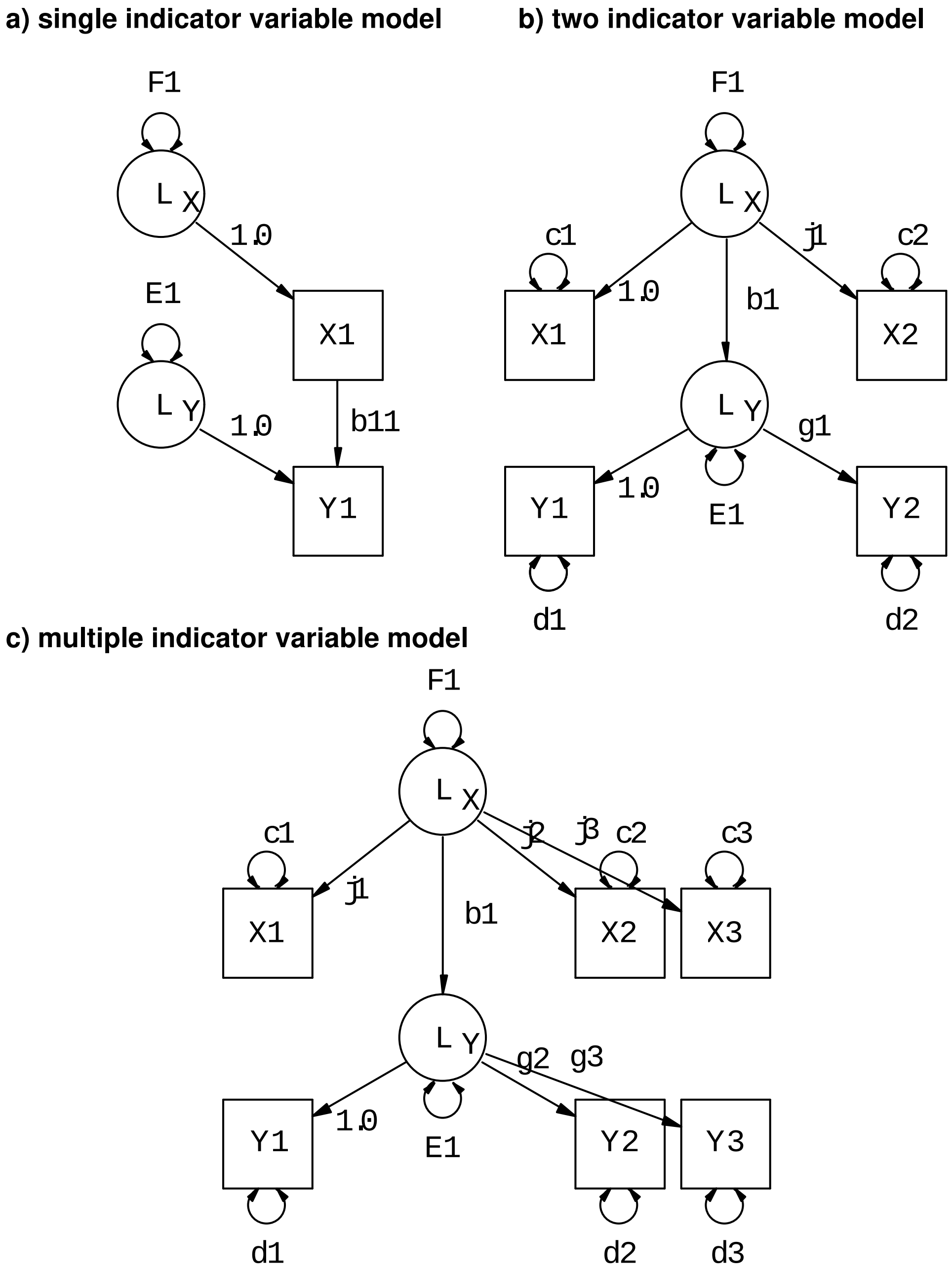


Figure 5.4: Regression path models with multiple indicators.

Chapter 6

Univariate Analysis

6.1 Introduction

In this chapter we take the univariate model as described in Chapter 5, and apply it to twin data. The main goals of this chapter are i) to enable the readers to apply the models to their own data, and ii) deepen their understanding of both the scope and the limitations of the method. In Section 6.2.1 a model of additive genetic (A), dominance genetic (D), common environment (C), and random environment (E) effects is presented although D and C are confounded when our data have been obtained from pairs of twins reared together. The first example concerns a continuous variable: body mass index (BMI), a widely used measure of obesity, and Section 6.2.2 describes how these data were obtained and summarized. In Section 6.2.3 we fit this model to authentic data, using Mx in a path coefficients approach. Section 6.2.5 illustrates the univariate model fitted with variance components. An alternative treatment which may be skipped without loss of continuity. The results of initial model-fitting to BMI data appear in Section 6.2.6 and two extensions to the model, the use of means (Section 6.2.7) and of unmatched twins (Section 6.2.8), are described before drawing general conclusions about the BMI analyses in Section 6.2.9. In Section 6.3 the basic model is applied to ordinal data. The second example (Section 6.3.1) describes the collection and analysis of major depressive disorder in a sample of adult female twins. This application serves to contrast the data summary and analysis required for an ordinal variable against those appropriate for a continuous variable. In most twin studies there is considerable heterogeneity of age between pairs. As shown in Section 6.4, such heterogeneity can give rise to inflated estimates of the effects of the shared environment. We, therefore, provide a method of incorporating age into the structural equation model to separate its effects from other shared environmental influences.

6.2 Fitting Genetic Models to Continuous Data

6.2.1 Basic Genetic Model

Derivations of the expected variances and covariances of relatives under a simple univariate genetic model have been reviewed briefly in the chapters on biometrical genetics and path analysis (Chapters 3 and 5). In brief, from biometrical genetic theory we can write structural equations relating the phenotypes, P, of relatives i and j (e.g., BMI values of first and second members of twin pairs) to their underlying genotypes and environments which are latent variables whose influence we must infer. We may decompose the total genetic effect on a phenotype into contributions of:

- Additive effects of alleles at multiple loci (A),
- Dominance effects at multiple loci (D),
- Higher-order epistatic interactions between pairs of loci (additive $\times$ additive, additive $\times$ dominance, dominance $\times$ dominance: AA, AD, DD), and so on.

In practice even additive $\times$ dominance and dominance $\times$ dominance epistasis are confounded with dominance in studies of humans, and the power of resolving genetic dominance and additive $\times$ additive epistasis is very low. We shall therefore limit our consideration to additive and dominance genetic effects.

Similarly, we may decompose the total environmental effect into that due to environmental influences shared by twins or sibling pairs reared in the same family (*shared, common*, or *between-family* environmental (C) effects), and that due to environmental effects that make family members differ from one another (*within-family, specific*, or *random environmental* (E) effects). Thus, the observed phenotypes, P_i and P_j, will be linear functions of the underlying additive genetic deviations (A_i and A_j), dominance genetic deviations (D_i and D_j), shared environmental deviations (C_i and C_j), and random environmental deviations (E_i and E_j). Assuming all variables are scaled as deviations from zero, we have

$$\begin{aligned} P_1 &= e_1E_1 + c_1C_1 + a_1A_1 + d_1D_1 \\ P_2 &= e_2E_2 + c_2C_2 + a_2A_2 + d_2D_2 \end{aligned} \tag{6.1}$$

In most models we do not expect the magnitude of genetic effects, or the environmental effects, to differ between first and second twins, so we set $e_1 = e_2 = e, c_1 = c_2 = c, a_1 = a_2 = a, d_1 = d_2 = d$. Likewise, we do not expect the values of e, c, a, and d to vary as a function of relationship. In other words, the effects of genotype and environment on the phenotype are the same regardless of whether one is an

MZ twin, a DZ twin, or not a twin at all. In matrix form, we may write

$$\begin{pmatrix} P_1 \\ P_2 \end{pmatrix} = \begin{pmatrix} a & c & e & d & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a & c & e & d \end{pmatrix} \begin{pmatrix} A_1 \\ C_1 \\ E_1 \\ D_1 \\ A_2 \\ C_2 \\ E_2 \\ D_2 \end{pmatrix}$$

As shown in Chapter 5, this model generates a predicted covariance matrix (Σ) which is equal to

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

Unless two or more waves of measurement are used, or several variables index the phenotype under study, residual effects (such as measurement error) will form part of the random environmental component, and are not explicitly included in the model.

To obtain estimates for the genetic and environmental effects in this model, we must also specify the variances and covariances among the latent genetic and environmental factors. Two alternative parameterizations are possible: 1) the variance components approach (Chapter 3), or 2) the path coefficients model (Chapter 5). The variance components approach becomes cumbersome for designs involving more complex pedigree structures than pairs of relatives, but it does have some numerical advantages (see Chapter **??**, p. **??**).

In the *variance components approach* we estimate variances of the latent nonshared and shared environmental and additive and dominance genetic variables, V_E, V_C, V_A, or V_D, and fix $a = c = e = d = 1$. Thus, the phenotypic variance is simply the sum of the four variance components. In the *path coefficients approach* we standardize the variances of the latent variables to unity ($V_E = V_C = V_A = V_D = 1$) and estimate a combination of a, c, e, and d as free parameters. Thus, the phenotypic variance is a weighted sum of standardized variables. In this volume we will often refer to models that have particular combinations of free parameters in the general path coefficients model. Specifically, we refer to an ACE model as one having only additive genetic, common environment, and random environment effects; an ADE model as one having additive genetic, dominance, and random environment effects; an AE model as one having additive genetic and random environment effects, and so on.

Figures 5.3a and 5.3b in Chapter 5 represent path diagrams for the two alternative parameterizations of the full basic genetic model, illustrated for the case of pairs of monozygotic twins (MZ) or dizygotic twins (DZ), who may be reared together (MZT, DZT) or reared apart (MZA, DZA). For simplicity, we make certain strong assumptions in this chapter, which are implied by the way we have drawn the path diagrams in Figure 5.3:

1. No genotype-environment correlation, i.e., latent genetic variables A are uncorrelated with latent environmental variables C and E;

2. No genotype $\times$ environment interaction, so that the observed phenotypes are a linear function of the underlying genetic and environmental variables;

3. Random mating, i.e., no tendency for like to marry like, an assumption which is implied by fixing the covariance of the additive genetic deviations of DZ twins or full sibs to $0.5V_A$;

4. Random placement of adoptees, so that the rearing environments of separated twin pairs are uncorrelated.

We discuss ways in which these assumptions may be relaxed in subsequent chapters, particularly Chapter 9 and Chapter **??**.

6.2.2 Body Mass Index in Twins

Table 6.1 summarizes twin correlations and other summary statistics (see Chapter 2) for untransformed BMI, defined as weight (in kilograms) divided by the square of height (in meters). BMI is an index of obesity which has been widely used in epidemiologic research (Bray, 1976; Jeffrey and Knauss, 1981), and has recently been the subject of a number of genetic studies (Grilo and Pogue-Guile, 1991; Cardon and Fulker, 1992; Stunkard et al., 1986). Values between 20–25 are considered to fall in the normal range for this population, with BMI < 20 taken to indicate underweight, BMI > 25 overweight, and BMI > 28 obesity (Australian Bureau of Statistics, 1977) though standards vary across nations. The data analyzed here come from a mailed questionnaire survey of volunteer twin pairs from the Australian NH&MRC twin register conducted in 1981 (Martin and Jardine, 1986; Jardine, 1985). Questionnaires were mailed to 5967 pairs age 18 years and over, with completed questionnaires returned by both members of 3808 (64%) pairs, and by one twin only from approximately 550 pairs, yielding an individual response rate of 68%.

The total sample has been subdivided into a young cohort, aged 18-30 years, and an older cohort aged 31 and above. This allows us to examine the consistency of evidence for environmental or genetic determination of BMI from early adulthood to maturity. For each cohort, twin pairs have been subdivided into five groups: monozygotic female pairs (MZF), monozygotic male pairs (MZM), dizygotic female pairs (DZF), dizygotic male pairs (DZM) and opposite-sex pairs (DZFM). We have avoided pooling MZ or like-sex DZ twin data across sex before computing summary statistics. Pooling across sexes is inappropriate unless it is known that there is no gender difference in mean, variance, or twin pair covariance, and no genotype $\times$ sex

Table 6.1: Twin correlations and summary statistics for untransformed BMI in twins concordant for participation in the Australian survey. BMI is calculated as kg/m^2. Notation used is N: sample size in pairs; r: correlation; $\bar{x}$: mean; σ^2: variance; skew: skewness; kurt: kurtosis. Groups consist of monozygotic (MZ) or dizygotic (DZ) twin pairs who are male (M) female (F) or opposite-sex (OS).

			First Twin†				Second Twin			
	N	r	$\bar{x}$	σ^2	skew	kurt	$\bar{x}$	σ^2	skew	kurt
MZF										
Young	534	.78	21.25	7.73	1.82	6.84	21.30	8.81	2.14	9.44
Older	637	.69	23.11	11.87	1.22	2.53	22.97	11.25	1.08	2.11
DZF										
Young	328	.30	21.58	8.56	1.75	6.04	21.64	9.84	2.38	12.23
Older	380	.32	22.77	10.93	1.40	4.03	22.95	12.63	1.26	2.43
MZM										
Young	251	.77	22.09	5.95	0.28	0.10	22.13	5.77	0.40	0.30
Older	281	.70	24.22	6.42	0.11	-0.05	24.30	7.85	0.43	0.63
DZM										
Young	184	.32	22.71	8.16	1.00	1.71	22.61	9.63	1.55	6.24
Older	137	.37	24.18	8.28	0.41	0.70	24.08	7.42	0.72	0.43
DZFM										
Young	464	.23	21.33	6.89	1.06	1.84	22.47	6.81	0.76	1.72
Older	373	.24	23.07	12.63	1.23	2.24	24.65	8.52	0.88	1.49

† Female twins are 'first twin' in opposite-sex pairs.

interaction; it should almost always be avoided. Among same-sex pairs, twins were assigned as first or second members of a pair at random. In the case of opposite-sex twin pairs, data were ordered so that the female is always the first member of the pair.

In both sexes and both cohorts, MZ twin correlations are substantially higher than like-sex DZ correlations, suggesting that there may be a substantial genetic contribution to variation in BMI. In the young cohort, the like-sex DZ correlations are somewhat lower than one-half of the corresponding MZ correlations, but this finding does not hold up in the older cohort. In terms of additive genetic (V_A) and dominance genetic (V_D) variance components, the expected correlations between MZ and DZ pairs are respectively $r_{MZ} = V_A + V_D$ and $r_{DZ} = 0.5V_A + 0.25V_D$, respectively (see Chapters 3 and **??**). Thus the fact that the like-sex DZ twin correlations are less than one-half the size of the MZ correlations in the young cohort suggests a contribution of genetic dominance, as well as additive genetic variance, to individual differences in BMI. Model-fitting analyses (e.g., (**?**) are needed to determine whether the data:

1. Are consistent with simple additive genetic effects
2. Provide evidence for significant dominance genetic effects
3. Enable us to reject a purely environmental model
4. Indicate significant genotype × age-cohort interaction.

Skewness and kurtosis measures in Table 6.1 indicate substantial non-normality of the marginal distributions for raw BMI. We have also computed the polynomial regression of absolute intra-pair difference in BMI values on pair sum[1] separately for each like-sex twin group. These are summarized in Table 6.2. If the joint distribution of twin pairs for BMI is bivariate normal, these regressions should be non-significant. Here, however, we observe a highly significant regression: on average, pairs with high BMI values also exhibit larger intra-pair differences in BMI. This is likely to be an artefact of scale, since using a log-transformation substantially reduces the magnitude of the polynomial regression (as well as reducing marginal measures of skewness and kurtosis).

In general, raw data or variance-covariance matrices, not correlations, should be used for model-fitting analyses with continuously distributed variables such as BMI. The simple genetic models we fit here predict no difference in variance between like-sex MZ and DZ twin pairs, but *the presence of such variance differences may indicate that the assumptions of the genetic model are violated.* This is an important point which we must consider in some detail. To many researchers the opportunity to expose an assumption as false may seem like something to be avoided if possible, because it may mean i) more work or b) difficulty publishing the results. But there

[1] i.e. the unsigned difference between twin 1 and twin 2 of each pair, $|\mathrm{BMI}_{\mathrm{twin}\ 1} - \mathrm{BMI}_{\mathrm{twin}\ 2}|$ with $\mathrm{BMI}_{\mathrm{twin}\ 1} + \mathrm{BMI}_{\mathrm{twin}\ 2}$

Table 6.2: Polynomial regression of absolute intra-pair difference in BMI ($|\mathrm{BMI}_{T1} - \mathrm{BMI}_{T2}|$) on pair sum ($\mathrm{BMI}_{T1} + \mathrm{BMI}_{T2}$), sum^2, and sum^3. The multiple regression on these three quantities is shown for raw and log-transformed BMI scores.

Sample	Raw BMI R^2	Log BMI R^2
Young MZF	0.11***	0.04***
Older MZF	0.16***	0.06***
Young MZM	0.10***	0.04*
Older MZM	0.09***	0.03*
Young DZF	0.34***	0.15***
Older DZF	0.27***	0.12***
Young DZM	0.15***	0.06*
Older DZM	0.03	0.01

***$p < .001$; *$p < .05$.

are better reasons not to use a technique that hides assumption failure? For sure, if we fitted models to correlation matrices, variance differences would never be observed, but to do so would be like, in physics, breaking the thermometer if a temperature difference did not agree with the theory. Rather, we should look at failures of assumptions as opportunities in disguise. First, a novel effect may have been discovered! Second, if the effect biases the parameters of intereset, it may be possible to contol for the effect statistically, and therefore obtain unbiased estimates. Third, we may have the opportunity to develop a new and useful method of analysis.

To return to the task in hand, we present summary twin pair covariance matrices in Table 6.3. These statistics have been computed for 7 ln (BMI), and means have been computed as (7ln (BMI) $-$ 21), to yield summary statistics with magnitudes of approximately unity. Rescaling the data in this way will often improve the efficiency of the optimization routines used in model-fitting analyses (Gill et al., 1981)[2]

6.2.3 Building a Path Coefficients Model Mx Script

With the introduction from the previous sections and chapters, we are now in a position to set up a simple genetic model using Mx. The script in Appendix **??** fits a simple univariate genetic model, estimating path coefficients, to covariance matrices for two like-sex twin groups: MZ twin pairs reared together, and DZ twin

[2]small observed variances ($< .5$) can be problematic as the predicted covariance matrix may become non-positive definite.

Table 6.3: Covariances of Twin Pairs for Body Mass Index: 1981 Australian Survey. BMI $= 7 \times \ln(\mathrm{kg}/(\mathrm{m}^2))$.

	Young Cohort (<30)			Older Cohort (≥ 30)		
	Covariance Matrix		Means[a]	Covariance Matrix		Means[a]
	Twin 1	Twin 2	$\bar{x}'$	Twin 1	Twin 2	$\bar{x}'$
MZ female pairs	(N=534 pairs)			(N=637 pairs)		
Twin 1	0.7247	0.5891	0.3408	0.9759	0.6656	0.9087
Twin 2	0.5891	0.7915	0.3510	0.6656	0.9544	0.8685
DZ female pairs	(N=328 pairs)			(N=380 pairs)		
Twin 1	0.7786	0.2461	0.4444	0.9150	0.3124	0.8102
Twin 2	0.2461	0.8365	0.4587	0.3124	1.0420	0.8576
MZ male pairs	(N=251 pairs)			(N=281 pairs)		
Twin 1	0.5971	0.4475	0.6248	0.5445	0.4128	1.2707
Twin 2	0.4475	0.5692	0.6378	0.4128	0.6431	1.2884
DZ male pairs	(N=184 pairs)			(N=137 pairs)		
Twin 1	0.7191	0.2447	0.8079	0.6885	0.2378	1.2502
Twin 2	0.2447	0.8179	0.7690	0.2378	0.5967	1.2281
Opposite-sex pairs	(N=464 pairs)			(N=373 pairs)		
Female twin	0.6830	0.1533	0.3716	1.0363	0.1955	0.8922
Male twin	0.1533	0.6631	0.7402	0.1955	0.6463	1.3860

[a] $\bar{x}' = \bar{x} - 21$.

pairs reared together. The script is written to ignore information on means. The full path diagram is given in Figure 6.1 We have drawn this figure to correspond to

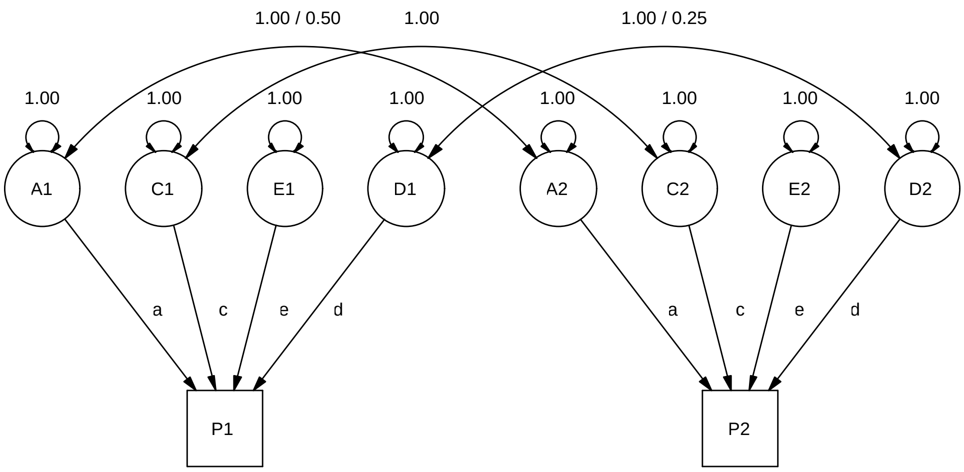


Figure 6.1: Univariate genetic model for data from monozygotic (MZ) or dizygotic (DZ) twins reared together. Genetic and environmental latent variables cause the phenotypes P_1 and P_2

the variables in the model. The latent genetic and environmental variables A, C, E and D cause the observed variables P_1 and P_2. The script is written to fit a model with free parameters e, a, and d, and fixing c to zero — implying that there are no effects of shared environment on BMI. The script is extensively documented using the comment facility in Mx: any line beginning with an exclamation mark is interpreted as a comment. We shall consider this first example Mx script in detail. Please note that reading this section is *not* a substitute for reading in detail the Mx manual (**?**), but merely a quick introduction to the essentials of a Mx script for genetic applications.

Each new statement in a script begins on a new line. For each group, we will have the following structure:

1. Title
2. Group type
3. Read and select any observed data, supply labels
4. Declare matrices to express the model
5. Specify parameters, (starting) values, equality constraints

6. Define matrix formulae for the model
7. Request fit functions, output and optimization options
8. End

We shall now examine the structure in greater detail, focusing on our BMI model. We plan to test hypotheses about the contributions of genetic and environmental factors to individual differences in BMI using data collected from MZ and DZ twins reared together. The Mx script therefore will have at least two groups. To simplify the structure of the script, we have added a calculation group at the beginning. We start the Mx script by indicating how many groups our job consists of with the `#NGroups 3` statement.

- *Title*

  ```
  Calculate genetic and environmental variance components
  ```

  A new title must be given at the start of each group.

- *Set the Group Type*

  ```
  Calculation NGroups=3
  ```

  where `NGroups` is the number of groups. This parameter is specified for the first group only. Calculation groups allow the specification of matrix operations in an algebra section which can greatly simplify the structure of the script. Here, we use the calculation group to specify the free and fixed parameters in the model, e, a, d, and c, and calculate their squared quantities, to be used in the expectations of the variances and the MZ and DZ covariances of the model (see Section ??).

- *Matrices Declaration*

  ```
  Begin Matrices;
   X Lower 1 1 Free
   Y Lower 1 1 Fixed
   Z Lower 1 1 Free
   W Lower 1 1 Free
   H Full 1 1
   Q Full 1 1
  End Matrices;
  ```

  The matrices declaration section begins with a `Begin Matrices;` line and ends with a `End Matrices;` line. Up to 26 matrices can be declared, each starting on a new line. Matrix names are restricted to one letter, from A to Z. The name is followed by the matrix type (see Mx manual for details on available matrix types), the number of rows and the number of colums. All matrix elements are fixed by default. If the keyword `Free` appears, each modifiable element has a free parameter specified to be estimated. In this example, four 1×1 matrices have been declared. Matrices `X`, `Z` and `W` represent free parameters a, e and d, respectively. The parameter c in matrix `Y` is fixed to zero (the word `Fixed` appears only for clarification). Two additional matrices, `H` and `Q`, are declared for fixed scalars to be used in the model specification.

- *Labels, Numbers and Parameters*

```
Label Row X add_gen
Label Row Y comm_env
Label Row Z spec_env
Label Row W dom_gen
Matrix H .5
Matrix Q .25
Start .6 All
```

Labels can be given for the row or column (or both) of any matrix. Values can be assigned to matrix elements using the `Matrix` command. If the matrix element is modifiable, the assigned value will be the starting value. The `Start` command here is used to assign the same starting value to `All` the free parameters in the model.

In genetic problems, we must assign starting values to parameters. In the present case, the only parameters to be estimated are a, d and e. In choosing starting values for twin data, a useful rule of thumb is to assume that the total variance is divided equally between the parameters that are to be estimated. In this case the predicted total variance is $3 \times .6^2 = 1.08$ which is close to the observed total variance in these data. For other data, other starting values may be required. Good starting values can save a significant amount of computer time, whereas bad starting values may cause any optimizer to fail to find a global minimum, or to hit a maximum number of iterations before converging.

- *Algebra Section*

```
Begin Algebra;
 A= X*X';
 C= Y*Y';
 E= Z*Z';
 D= W*W';
End Algebra;
```

The algebra section begins with a `Begin Algebra;` statement and ends with a `End Algebra;` statement. Each algebra operation starts on a new line and ends with a semi-colon (it may run over several lines so a ; is essential to mark the end of a formula). The matrix on the left side of the = sign is newly defined as the result of the matrix operation on the right side of the = sign. Matrices on the right have to be declared in the matrices declaration section or defined in a previous algebra statement. In this example the quantities a^2, c^2, e^2 and d^2 are calculated in matrices `A`, `C`, `E` and `D`, respectively.

- *End*

  Every group ends with an `End` statement.

The structure for the data groups for MZ and DZ twins is very similar. We will only discuss the first data group in detail. The first line gives the title for this group.

- *Data Section*

```
Data NInput_vars=2 NObservations=534
 Labels bmi_t1 bmi_t2
 CMatrix Symmetric File=ozbmimzf.cov
```

  where:

  1. `NInputvars` is number of input variables, i.e., $2n$, if there are n variables assessed for each member of a twin pair
  2. `NObservations` is number of observations or sample size, i.e., number of pairs used to compute the data matrix in this group.

  Mx allows the user the option of reading a list of names for the observed variables (`Labels`). This is *very* useful for clarification of the Mx output. Mx will read a covariance matrix (`CMatrix`), a correlation matrix (`KMatrix`), or a matrix of polychoric and polyserial correlations (`PMatrix`). The matrix may be read as a lower triangle in free format (the default, the keyword `Symmetric` is optional), or as a full matrix if the keyword `Full` is specified. It will also read means (`Means`) when these are needed. Summary statistics can be read from within the Mx script, for example,

```
CMatrix Symmetric
0.7247
0.5891 0.7915
```

  Alternatively, the data matrices can be read from separate files, e.g.,

```
CMatrix Symmetric File=ozbmimzf.cov
```

The lines referring to the actual data, `Data`, `Labels` and `CMatrix` can be saved in a `dat` file (e.g. `ozbmimzf.dat` which can then be included in the Mx script with the following statement:

```
#include ozbmimzf.dat
```

- *Matrices declaration*

```
Matrices= Group 1
```

The `= Group 1` command includes all the declared and defined matrices from the group 1 into the current group.

- *Model specification*

```
Covariances A+C+D+E | A+C+D_
            A+C+D   | A+C+D+E;
```

The expected covariance matrix is specified using a matrix formulation with the expected variances for twin 1 and twin 2 on the diagonal and the expected covariance between twins, in this case for MZ's, as the off-diagonal element. The expectation for the variance, $a^2+c^2+d^2+e^2$, is translated into `A+C+D+E`; that for the MZ covariance, $a^2+c^2+d^2$, into `A+C+D`. The resulting four 1×1 matrices are concatenated using the horizontal bar `|` and the vertical bar `_` operators to form the 2×2 expected covariance matrix, corresponding to the 2×2 observed covariance matrix. Note that the covariance statement needs to end with a semicolon.

- *Options*

```
Option RSiduals
```

Various options for statistical output and optimization can be specified. Usually, the choice of estimation procedure will be either maximum likelihood if covariance matrices are being analyzed, or weighted least squares if matrices of polychoric, polyserial, or product-moment correlations are being analyzed. The `RSiduals` option is very useful as it results in the printing of the observed, expected and residual matrices in the output.

The specification for the DZ group is very similar to that of the MZ group. Note the different number of observations, the new filename containing the DZ observed covariance matrix and the expected covariance matrix to match the expectation of the DZ covariance, $.5a^2+c^2+.25d^2$. A special form of matrix multiplication, the

Kronecker product, represented by the symbol `@`, is used to premultiply the matrix `A` by the scalar .5 and the matrix `D` by the scalar .25. The specification extends easily to the multivariate case (see Section ?).

After successfully running the Mx input script, by default, Mx prints the

1. User's input script
2. Parameter Specifications
3. Parameter Estimates
4. Measures of overall goodness-of-fit.

Other useful output can be requested by additional options, including:

- `NDecimals=`x – set number of decimals in printed output ($0 < x < 8$, default: x =4) — useful for simulation work.
- `Iterations=`xx – set maximum number of iterations (default: 1000).

The Mx manual should be consulted for a full description of the options.

6.2.4 Interpreting the Mx Output

We can run the example of Appendix **??** on a personal computer with Mx installed by typing:

```
Mx univar.mx univar.mxo
```

where `univar.mx` is the name of the script file, and `univar.mxo` is the name of the output file. We recommend `mx` and `mxo` as file extensions to make Mx input and output distinct from input and output of other programs. This example fits a model allowing for random environmental effects, additive genetic effects, and dominance genetic effects, to the young female like-sex MZ and DZ covariance matrices for log-transformed BMI. The Mx output includes:

1. Listing of the Mx script.
2. Parameter Specifications for each group, indicating the parameters to be estimated. Matrices are ordered alphabetically.

```
 MATRIX W
This is a LOWER matrix of order    1 by    1
      1
  1   3
```

```
 MATRIX X
This is a LOWER matrix of order    1 by    1
      1
  1   1

 MATRIX Y
This is a LOWER matrix of order    1 by    1
 It has no free parameters specified

 MATRIX Z
This is a LOWER matrix of order    1 by    1
      1
  1   2
```

If no labels are specified in the input script, Mx will use consecutive numbers for the rows and columns of each matrix. The matrix element 1 identifies the first free parameter to be estimated (a), referring to the first matrix element that was declared free (`Free`) in the matrices declaration section. Similarly, 2 identifies parameter e, and 3 identifies parameter d. It is important to check these to confirm that parameters have been correctly specified and that the total number of estimated parameters corresponds to the number of free parameters in the model to be fitted.

3. Mx Parameter Estimates for each group, obtained at the solution. In the case of Appendix **??**, for example, we obtain

```
 MATRIX W
This is a LOWER matrix of order    1 by    1
          1
  1   .5441

 MATRIX X
This is a LOWER matrix of order    1 by    1
          1
  1   .5621

 MATRIX Y
This is a LOWER matrix of order    1 by    1
 It has no free parameters specified

 MATRIX Z
This is a LOWER matrix of order    1 by    1
          1
  1   .4119
```

In other words, our maximum-likelihood parameter estimates are $a = 0.56$, $d = 0.54$, and $e = 0.41$ for these data.

4. If we include the option `RSiduals` in a group, the observed, and expected ('fitted') covariance matrix and residuals for that group are printed; Comparison of models should normally be based on likelihood-ratio chi-squared tests, since significance tests based on standard errors may be misleading for this example (Neale *et al.*, 1989b).

5. The goodness-of-fit chi-squared is reported. In this example, $\chi^2_3 = 3.71, p = 0.29$, indicating that the model gives a good fit to the data. A small p value (e.g. $< .05$) would indicate a lack of agreement between the data and the predictions of the model.

6. Finally, standardized parameter estimates can be calculated for each group. In this univariate case, we may standardize a^2 by computing $a^2/(a^2+c^2+e^2+d^2)$ to give the proportion of the total variance in BMI which is accounted for by additive genetic effects (40.4%). Similarly, we can calculate the proportion of variance accounted for by random environmental effects (21.7%), and by dominance genetic effects (37.9%). These analyses suggest that in young women age 30 and under, additive and non-additive genetic factors account for approximately 78% of the variance in BMI.

Discussion of these results continues in Section 6.2.6.

6.2.5 Building a Variance Components Model Mx Script

We include the variance components parameterization of the basic structural equation model for completeness. It will not be developed and applied in as great detail as the path coefficients parameterization because (i) it is difficult to generalize to more complex pedigree structures or multivariate problems, and (ii) doing so would contribute much by weight but little by insight to this volume. Readers seeking an easy introduction to twin models in Mx may skip this section and focus their attention on Section 6.2.3, the path coefficients parameterization.

For MZ and DZ twin pairs reared in the same family, the variance components parameterization is presented in (Figure 5.3b). Under the simplifying assumptions of the present chapter, the 2×2 expected covariance matrix of twin pairs (Σ) will be, in terms of variance components,

$$\begin{bmatrix} V_E + V_C + V_A + V_D & \omega_i V_C + \alpha_i V_A + \delta_i V_D \\ \omega_i V_C + \alpha_i V_A + \delta_i V_D & V_E + V_C + V_A + V_D \end{bmatrix}$$

where ω_i is 1 for twins, full sibs or adoptees reared in the same household, but 0 for separated twins or other biological relatives reared apart; a_i is 1 for MZ twin pairs, 0.5 for DZ pairs, full sibs, or parents and offspring, and 0 for genetically unrelated individuals; and δ_i is 1 for MZ pairs, 0.25 for DZ pairs or full sibs, and 0 for most other relationships. In terms of path coefficients, we need only substitute $V_E = e^2, V_C = c^2, V_A = h^2$, and $V_D = d^2$.

In data on twin pairs reared together the effects of shared environment and genetic dominance are confounded. If both additive genetic effects and shared environmental effects contribute to variation in a trait, the covariance of DZ twin pairs will be less than the MZ covariance, but greater than one-half the MZ covariance. If both additive genetic effects and dominance genetic effects contribute to variation in a trait, the covariance of DZ pairs will be less than one-half the MZ covariance. In terms of variance components, therefore, a substantial dominance genetic effect will lead to a negative estimate of the shared environmental variance component, if a model allowing for additive genetic and shared environmental variance components is fitted; while conversely a substantial shared environmental effect will lead to a negative estimate of the dominance genetic variance component, if a model allowing for additive and dominance genetic variance components is fitted (Martin *et al.*, 1978). In terms of path coefficients, however, since we are estimating parameters c or d, c^2 or d^2 can never take negative values, and so we will obtain an estimate of $c = 0$ in the presence of dominance, or $d = 0$ in the presence of shared environmental effects. Additional data on separated twin pairs (Jinks and Fulker, 1970) or on the parents or other relatives of twins (Fulker, 1982; Heath, 1983) are needed to resolve the effects of shared environment and genetic dominance when both are present.

Appendix **??** illustrates an example script for fitting a variance components model to twin pair covariance matrices for two like-sex twin pair groups. We estimate additive genetic, dominance genetic and random environmental variance components in the matrices `A`, `D` and `E`. The covariance statement is the same as for the path model example. The only change is in the calculation group, which does not square the estimates to construct A, C, E and D.

For the young male like-sex pairs, the estimates are $V_E = 0.14$, $V_A = 0.25$, and $V_D = 0.29$. We can calculate standardized variance components by hand, as $V_E^* = V_E/V_P$, $V_A^* = V_A/V_P$, and $V_D^* = V_D/V_P$, where $V_P = V_E + V_A + V_D = 0.6804$ (which can be read directly from the variance in the expected covariance matrix). In this example, random environmental effects account for 20.3% of the variance, additive genetic effects for 36.4% of the variance, and dominance genetic effects for 43.3% of the variance of BMI in young adult males. By χ^2 test of goodness-of-fit, our model gives only a marginally acceptable fit to the data ($\chi^2_3 = 7.28, p = 0.06$).

6.2.6 Interpreting Univariate Results

In model-fitting to univariate twin data, whether we use a variance components or a path coefficients model, we are essentially testing the following hypotheses:

1. No family resemblance ("E" model: $e > 0$: $a = c = d = 0$)

Table 6.4: Results of fitting models to twin pairs covariance matrices for Body Mass Index: Two-group analyses, complete pairs only.

	Females				Males			
	Young		Older		Young		Older	
Model (d.f.)	χ^2	p	χ^2	p	χ^2	p	χ^2	p
CE (4)	160.72	<.001	87.36	<.001	97.20	<.001	37.14	<.001
AE (4)	8.06	.09	2.38	.67	10.88	.03	5.03	.28
ACE (3)	8.06	<.05	2.38	.50	10.88	.01	5.03	.17
ADE (3)	3.71	.29	1.97	.58	7.28	.06	5.03	.17

2. Family resemblance solely due to additive genetic effects ("AE" model: $a > 0, e > 0, c = d = 0$)

3. Family resemblance solely due to shared environmental effects ("CE" model: $e > 0, c > 0, a = d = 0$)

4. Family resemblance due to additive genetic plus dominance genetic effects ("ADE" model: $a > 0, d > 0, e > 0, c = 0$)

5. Family resemblance due to additive genetic plus shared environmental effects ("ACE" model: $a > 0, c > 0, e > 0, d = 0$).

Note that we never fit a model that excludes random environmental effects, because it predicts perfect MZ twin pair correlations, which in turn generate a singular expected covariance matrix[3]. From inspection of the twin pair correlations for BMI, we noted that they were most consistent with a model allowing for additive genetic, dominance genetic, and random environmental effects. Model-fitting gives three important advantages at this stage:

1. An overall test of the goodness of fit of the model

2. A test of the relative goodness of fit of different models, as assessed by likelihood-ratio χ^2. For example, we can test whether the fit is significantly worse if we omit genetic dominance for BMI

3. Maximum-likelihood parameter estimates under the best-fitting model.

Table 6.4 tabulates goodness-of-fit chi-squares obtained in four separate analyses of the data from younger or older, female or male like-sex twin pairs. Let us

[3]A singular matrix cannot be inverted (see Chapter 4) and, therefore, the maximum likelihood fit function (see Chapter **??**) cannot be computed.

consider the results for young females first. The non-genetic model (CE) yields a chi-squared of 160.72 for 4 degrees of freedom[4], which is highly significant and implies a very poor fit to the data indeed. In stark contrast, the alternative model of additive genes and random environment (AE) is not rejected by the data, but fits moderately well ($p = .09$). Adding common environmental effects (the ACE model) does not improve the fit whatsoever, but the loss of a degree of freedom makes the χ^2 significant at the .05 level. Finally, the ADE model which substitutes genetic dominance for common environmental effects, fits the best according to the probability level. We can test whether the dominance variation is significant by using the likelihood ratio test. The difference between the χ^2 of a general model (χ^2_G) and the that of a submodel (χ^2_S) is itself a χ^2 and has $\mathrm{df}_S - \mathrm{df}_G$ degrees of freedom (where subscripts S and G respectively refer to the submodel and general model, in other words, the difference in df between the general model and the submodel). In this case, comparing the AE and the ADE model gives a likelihood ratio χ^2 of $8.06 - 3.71 = 4.35$ with $4 - 3 = 1$df. This is significant at the .05 level, so we say that there is significant deterioration in the fit of the model when the parameter d is fixed to zero, or simply that the parameter d is significant.

Now we are in a position to compare the results of model-fitting in females and males, and in young and older twins. In each case, a non-genetic (CE) model yields a significant chi-squared, implying a very poor fit to the data: the deviations of the observed covariance matrices from the expected covariance matrices under the maximum-likelihood parameter estimates are highly significant. In all groups, a full model allowing for additive plus dominance genetic effects and random environmental effects (ADE) gives an acceptable fit to the data, although in the case of young males the fit is somewhat marginal. In the two older cohorts, however, a model which allows for only additive genetic plus random environmental effects (AE) does *not* give a significantly worse fit than the full (ADE) model, by likelihood-ratio χ^2 test. In older females, for example, the likelihood-ratio chi-square is $2.38 - 1.97 = 0.41$, with degrees of freedom equal to $4 - 3 = 1$, i.e., $\chi^2_1 = 0.41$ with probability $p = 0.52$; while in older males we have $\chi^2_1 = 0.00, p = 1.00$. For the older cohorts, therefore, we find no significant evidence for genetic dominance. In young adults, however, significant dominance is observed in females (as noted above) and the dominance genetic effect is almost significant in males ($\chi^2_1 = 3.6, p = 0.06$).

Table 6.5 summarizes variance component estimates under the best-fitting models. Random environment accounts for a relatively modest proportion of the total variation in BMI, but appears to be having a larger effect in older than in younger individuals (30-31% versus 20-22%). Although the estimate of the narrow heritability (i.e., proportion of the total variance accounted for by additive genetic factors)

[4]The degrees of freedom associated with this test are calculated as the difference between the number of observed statistics (n_s) and the number of estimated parameters (n_p) in the model. Our data consist of two variances and a covariance for each of the MZ and DZ groups, giving $n_s = 6$ in total. The CE model has two parameters c and e, so $n_s - n_p = 6 - 2 = 4$df.

Table 6.5: Standardized parameter estimates under best-fitting model. Two-group analyses, complete pairs only.

	Estimate			
	a^2	c^2	e^2	d^2
Young females	0.40	0	0.22	0.38
Older females	0.69	0	0.31	0
Young males	0.36	0	0.20	0.44
Older males	0.70	0	0.30	0

is higher in the older cohort (69-70% vs 36-40%), the broad heritability (additive plus non-additive genetic variance) is higher in the young twins (78-80%).

6.2.7 Testing the Equality of Means

Applications of structural equation modeling to twin and other family data typically tend to ignore means. That is, observed measures are treated as deviations from the phenotypic mean (and are thus termed *deviation phenotypes*)[5], and likewise genetic and environmental latent variables are expressed as deviations from their means, which usually are fixed at 0. Most simple genetic models predict the same mean for different groups of relatives, so, for example, MZ twins, DZ twins, males from opposite-sex twin pairs, and males from like-sex twin pairs should have (within sampling error) equal means. Where significant mean differences are found, they may indicate sampling problems with respect to the variable under study or other violations of the assumptions of the basic genetic model. Testing for mean differences also may be important in follow-up studies, where we are concerned about the bias introduced by sample attrition, but can compare mean scores at baseline for those relatives who remain in a study with those who drop out. Fortunately, Mx facilitates tests for mean differences between groups.

For Mx to fit a model to means and covariances, both observed means and a model for them must be supplied. Appendix **??** contains a Mx script for fitting a univariate genetic model which also estimates the means of first and second twins from MZ and DZ pairs. The first change we make is to feed Mx the observed means in our sample, which we do with the `Means` command:

```
Means 0.9087 0.8685
```

Second, we declare a matrix for the means, e.g. `M Full 1 2` in the matrices declaration section. Third, we can equate parameters for the first and second twins by

[5]Except where explicitly noted, all models presented in this text treat observed variables as deviation phenotypes.

Table 6.6: Results of fitting models to twin pair covariance matrices and twin means for Body Mass Index: Two-group analyses, complete pairs only.

		Female				Male			
		Young		Older		Young		Older	
	df	χ^2	p	χ^2	p	χ^2	p	χ^2	p
Model I	6	7.84	.25	5.74	.57	12.81	.05	5.69	.58
Model II	5	3.93	.56	4.75	.58	7.72	.17	5.36	.50
Model III	3	3.71	.29	2.38	.67	7.28	.06	5.03	.17
Genetic Model		ADE		AE*		ADE		AE*	

* AE models have one more degree of freedom than shown in the df column

using a `Specify` statement such as

```
Specify M 101 101
```

where `101` is a parameter number that has not been used elsewhere in the script. By using the same number for the two means, they are constrained to be equal. Fourth, we include a model for the means:

```
Means M;
```

In the DZ group we also supply the observed means, and adjust the model for the means. We can then either (i) equate the mean for MZ twins to that for DZ twins by using the same matrix `M`, 'copied' from the MZ group or equated to that of the MZ group as follows:

```
M Full 1 2 = M2
```

where `M2` refers to matrix `M` in group 2; to fit a *no heterogeneity* model (Model I); or (ii) equate DZ twin 1 and DZ twin 2 means but allow them to differ from the MZ means by declaring a new matrix (possibly called M too; matrices are specific to the group in which they are defined, unless they are equated to a matrix or copied from a previous group) to fit a zygosity dependent means model ($\overline{MZ} \neq \overline{DZ}$, Model II); or (iii) estimate four means, i.e., first and second twins in each of the MZ and DZ groups; to fit the *heterogeneity* model (Model III). This third option gives a perfect fit to the data with regard to mean structure, so that the only contribution to the fit function comes from the covariance structure. Hence the four means model gives the same goodness-of-fit χ^2 as in the analyses ignoring means.

Table 6.6 reports the results of fitting models incorporating means to the like-sex twin pair data on BMI. In each analysis, we have considered only the best-fitting genetic model identified in the analyses ignoring means. Again we subtract the χ^2 of a more general model from the χ^2 of a more restricted model to get a likelihood ratio test of the difference in fit between the two. For the two older cohorts we find no evidence for mean differences either between zygosity groups or between first and second twins. That is, the model that assumes no heterogeneity of means (model 1) does not give a significantly worse fit than either (i) estimating separate MZ and DZ means (model 2), or (ii) estimating 4 means. For older females, likelihood-ratio chi-squares are $\chi^2_1 = 0.99, p = 0.32$ and $\chi^2_3 = 3.36, p = 0.34$; and for older males, $\chi^2_1 = 0.36, p = 0.55$ and $\chi^2_3 = 0.43, p = 0.33$. Maximum-likelihood estimates of mean log BMI in the older cohort are, respectively, 21.87 and 22.26 for females and males; estimates of genetic and environmental parameters are unchanged from those obtained in the analyses ignoring means. In the younger cohorts, however, we do find significant mean differences between zygosity groups, both in females ($\chi^2_1 = 3.91, p < 0.05$) and in males ($\chi^2_1 = 5.09, p < 0.02$). In both sexes, mean log BMI values are lower in MZ pairs (21.35 for females, 21.63 for males) than for DZ pairs (21.45 for females, 21.79 for males). As these data are not age-corrected, it is possible that BMI values are still changing in this age-group, and that the zygosity difference reflects a slight mean difference in age. We shall return to this question in Section 6.2.9.

6.2.8 Incorporating Data from Singleton Twins

In most twin studies, there are many twin pairs in which only one twin agrees to cooperate. We call these pairs *discordant-participant* as opposed to *concordant-participant* pairs, in which data are collected from both members of the pair. Sadly, data from discordant-participant pairs are often just thrown away. This is unfortunate not only because of the wasted effort on the part of the twins, researchers, and data entry personnel, but also because they provide valuable information about the representativeness of the sample for the variable under study. If sampling is satisfactory, then we would expect to find the same mean and variance in concordant-participant pairs as in discordant-participant pairs. Thus, the presence of mean differences or variance differences between these groups is an indication that biased sampling may have occurred with respect to the variable under investigation. To take a concrete example, suppose that overweight twins are less likely to respond to a mailed questionnaire survey. Given the strong twin pair resemblance for BMI demonstrated in previous sections, we might expect to find that individuals from discordant-participant pairs are on average heavier than individuals from concordant-participant pairs. Such sampling biases will have differential effects on the covariances of MZ and DZ twin pairs, and thus may lead to biased estimates of genetic and environmental parameters (Lykken *et al.*, 1987; Neale *et al.*, 1989b).

Table 6.7 reports means and variances for transformed BMI from individuals from discordant-participant pairs in the 1981 Australian survey. Zygosity assignment for MZ twins must be regarded as somewhat tentative, since most algorithms for zygosity diagnosis based on questionnaire data require reports from both members of a twin pair to confirm monozygosity (e.g., Eaves *et al.*, 1989b). In most

Table 6.7: Means and variances for BMI of twins whose cotwin did not cooperate in the 1981 Australian survey.

	Young Cohort (<=30)			Older Cohort (>30)		
Group	N	$\bar{x}'$	σ^2	N	$\bar{x}'$	σ^2
MZ Female Twins	33	0.1795	1.0640	44	0.6852	1.1461
DZ Female Twins	55	0.5836	0.8983	62	1.0168	1.7357
MZ Male Twins	24	1.3266	1.2477	36	1.3585	1.1036
DZ Male Twins	47	1.2705	1.5309	48	1.0379	1.6716
Opp-Sex Pair Females	65	0.6551	1.4390	81	0.9756	1.2690
Opp-Sex Pair Males	28	0.8724	0.9754	27	1.7149	1.0019

groups, comparing Table 6.7 to Table 6.3, we observe both higher means and higher variances in the discordant-participant pairs. It is clearly important to test whether these differences are statistically significant.

To fit a model simultaneously to the means, variances, and covariances of concordant-participant pairs and the means and variances of discordant-participant pairs, requires that we analyze data where there are different numbers of observed variables per group, which is easily done in Mx.

Appendix **??** presents a Mx script for testing for differences in mean or variance. We constrain the means of the responding twin in groups four (MZ discordant-participant) and five (DZ discordant-participant) to equal those of twins from the concordant-participant pairs. Our test for significant differences in means between the concordant-participant and discordant-participant groups is the improvement in goodness-of-fit obtained when we allow these latter, discordant-participant pairs, to take their own mean value.

Table 6.8 summarizes the results of model-fitting. Model I is the *no heterogeneity* model of means and variances between concordant-participant versus discordant-participant twins. Model II allows for heterogeneity of variances, Model III for heterogeneity of means. Finally, Model IV tests both differences in means and variances. For these analyses, we considered only the best-fitting genetic model based on the results of the analyses ignoring means, and allowed for zygosity differences in means only if these were found to be significant in the analyses of the previous Section (i.e., in the younger twin pairs; young female pairs are the *only* group in which we find no difference between concordant-participant pairs and discordant-participant pairs). In the two older cohorts a model allowing for heterogeneity of means (Model 3) gives a substantially better fit than one that assumes no heterogeneity of means or variances (Model 1: older females: $\chi^2_2 = 12.86, p < 0.001$; older males: $\chi^2_2 = 30.87, p < 0.001$). Specifying heterogeneity of variances in addition to heterogeneity of means does not produce a further improvement in fit (older females: $\chi^2_2 = 2.02, p = 0.36$; older males: $\chi^2_2 = 1.99, p = 0.37$). Such a result is not atypical because the power to detect differences in mean is much greater than that to detect a difference in variance.

Table 6.8: Results of fitting models to twin pair covariance matrices and twin means for Body Mass Index: Two like-sex twin groups, plus data from twins from incomplete pairs. Models test for heterogeneity of means or variances between twins from pairs concordant vs discordant for cooperation in 1981 survey.

		Female				Male			
		Young		Older		Young		Older	
	df	χ^2	p	χ^2	p	χ^2	p	χ^2	p
Model I	11	8.16	.70	20.62	.08	54.97	.001*	48.55	.001*
Model II	9	6.03	.74	17.84	.09	29.22	.001*	44.58	.001*
Model III	9	5.70	.77	7.76	.74	22.76	.01	7.68	.74
Model IV	7	3.93	.79	5.74	.77	7.72	.36	5.69	.77
Genetic Model		ADE		AE#		ADE		AE#	
Means Model		$\overline{MZ} \neq \overline{DZ}$		$\overline{MZ} = \overline{DZ}$		$\overline{MZ} \neq \overline{DZ}$		$\overline{MZ} = \overline{DZ}$	

* $p < .001$

\# AE models have two more degrees of freedom than shown in the df column

When considering these results, we must bear in mind several possibilities. Numbers of twins from the discordant-participant groups are small, and estimates of mean and variance in these groups will be particularly vulnerable to outlier-effects; that is, to inflation by one or two individuals of very high BMI. Further outlier analyses (e.g., Bollen, 1989) would be needed to determine whether this is an explanation of the variance difference. In the young males, it is also possible that age differences between concordant-participant pairs and discordant-participant pairs could generate the observed mean differences.

6.2.9 Conclusions: Genetic Analyses of BMI Data

The analyses of Australian BMI data which we have presented indicate a significant and substantial contribution of genetic factors to variation in BMI, consistent with other twin studies referred to at the beginning of Section 6.2.2. In the young cohort like-sex pairs, we find significant evidence for genetic dominance (or other genetic non-additivity), in addition to additive genetic effects, but in the older cohort non-additive genetic effects are non-significant. Further analyses are needed to determine whether genetic and environmental parameters are significantly different across cohorts, or indeed between males and females (see Chapter 9).

We have discovered unexpected mean differences between zygosity groups (in the young cohort), and between twins whose cotwin refused to participate in the 1981 survey, and twins from concordant-participant pairs. It is possible that these differences reflect only outlier effects caused by a handful of observations. In this case, if we recode BMI as an ordinal variable, we might expect to find no significant differences in the proportions of twins falling into each category[6]. Alternatively, it is possible that there is an overall shift in the distribution of BMI, in which case we must be concerned about the undersampling of obese individuals. If the latter finding were confirmed, further work would be needed to explore the degree to which genetic and environmental parameters might be biased (cf. Lykken et al., 1987; Neale *et al.*, 1989a; Neale and Eaves, 1992).

6.3 Fitting Genetic Models to Binary Data

It is very important to realize that binary or ordinal data do not preclude model-fitting. A large number of applications, from item analysis (e.g., Neale, *et al*, 1986; Kendler *et al.,* 1987) to psychiatric or physical illness (e.g., Kendler *et al*, 1992b,c) do not have measures on a quantitative scale but are limited to discontinuous forms of assessment. In Chapter 2 we discussed how ordinal data from twins could be summarized as contingency tables from which polychoric correlations and their asymptotic variances could be computed. Fitting models to this type of summary statistic or directly to the contingency table data themselves involves a number of additional considerations, which we illustrate here with data on major depressive disorder. Although details of the sample and measures used have been provided in several published articles (Kendler *et al.*, 1991a,b; 1992a), we briefly reiterate the methods to emphasize some of the practical issues involved with an interview study of twins.

6.3.1 Major Depressive Disorder in Twins

Data for this example come from a study of genetic and environmental risk factors for common psychiatric disorders in Caucasian female same-sex twin pairs sampled from the Virginia Twin Registry. The Virginia Twin Registry is a population-based register formed from a systematic review of all birth certificates in the Commonwealth of Virginia. Twins were eligible to participate in the study if they were born between 1934 and 1971 and if both members of the pair had previously responded to a mailed questionnaire, to which the individual response rate was approximately 64%. The cooperation rate was almost certainly higher than this, as an unknown number of twins did not receive their questionnaire due to faulty addresses, improper forwarding of mail, and so on. Of the total 1176 eligible pairs, neither twin was interviewed in 46, one twin was interviewed and the other refused in 97,

[6]Excessive contributions to the χ^2 by a small number of outliers could also be detected by fitting models directly to the raw data using Mx. Though a more powerful method of assessing the impact of outliers, it is beyond the scope of this volume.

and both twins were interviewed in 1033 pairs. Of the completed interviews, 89.3% were completed face to face, nearly all in the twins' home, and 10.7% (mostly twins living outside Virginia) were interviewed by telephone. The mean age (±SD) of the sample at interview was 30.1 (7.6) and ranged from 17 to 55.

Zygosity determination was based on a combination of review of responses to questions about physical similarity and frequency of confusion as children — which alone have proved capable of determining zygosity with over 95% accuracy (Eaves *et al.*, 1989b) — and, in over 80% of cases, photographs of both twins. From this information, twins were classified as either: definitely MZ, definitely DZ, probably MZ, probably DZ, or uncertain. For 118 of the 186 pairs in the final three categories, blood was taken and eight highly informative DNA polymorphisms were used to resolve zygosity. If all probes are identical then there is a .9997 probability that the pair is MZ (Spence *et al.*, 1988). Final zygosity determination, using blood samples where available, yielded 590 MZ pairs, 440 DZ pairs and 3 pairs classified as uncertain. The DNA methods validated the questionnaire- and photograph-based 'probable' diagnoses in 84 out of 104 pairs; all 26 of 26 pairs in the definite categories were confirmed as having an accurate diagnosis. The error rate in zygosity assignment is probably well under 2%.

Lifetime psychiatric illness was diagnosed using an adapted version of the Structured Clinical Interview for DSM-III-R Diagnosis (Spitzer *et al.*, 1987) an instrument with demonstrable reliability in the diagnosis of depression (Riskind *et al.*, 1987). Interviewers were initially trained for 80 hours and received bimonthly review sessions during the course of the study. Each member of a twin pair was invariably interviewed by a different interviewer. DSM-III-R criteria were applied by a blind review of the interview by K.S. Kendler, an experienced psychiatric diagnostician. Diagnosis of depression was not given when the symptoms were judged to be the result of uncomplicated bereavement, medical illness, or medication. Interrater reliability was assessed in 53 jointly conducted interviews. Chance corrected agreement (kappa) was .96, though this is likely to be a substantial overestimate of the value that would be obtained from independent assessments[7].

Contingency tables of MZ and DZ twin pair diagnoses are shown in Table 6.9. PRELIS estimates of the correlation in liability to depression are .435 for MZ and .186 for DZ pairs. Details of using PRELIS to derive these statistics and associated estimates of their asymptotic variances are given in Section 2.3. The `PMatrix` command is used to read in the tetrachoric correlation matrix, and the `ACov` command reads the asymptotic weight matrices. In both cases we use the `File=` keyword in order to read these data from files. Therefore our univariate Mx input script is unchanged from that shown in Appendix ?? on page ??, except for the title and the dat file used.

```
Major depressive disorder in adult female MZ twins
```

[7]Such independent assessments would risk retest effects if they were close together in time. Conversely, assessments separated by a long interval would risk actual phenotypic change from one occasion to the next. For a methodological review of this area, see Helzer (1977)

Table 6.9: Contingency tables of twin pair diagnosis of lifetime Major Depressive Disorder in Virginia adult female twins.

		MZ		DZ	
	Twin 1	Normal	Depressed	Normal	Depressed
Twin 2	Normal	329	83	201	94
	Depressed	95	83	82	63

Table 6.10: Major depressive disorder in Virginia adult female twins. Parameter estimates and goodness-of-fit statistics for models and submodels including additive genetic (A), common environment (C), random environment (E), and dominance genetic (D) effects.

	Parameter Estimates				Fit statistics		
Model	a	c	e	d	χ^2	df	p
E	—	—	1.00	—	56.40	2	.00
CE	—	0.58	0.81	—	6.40	1	.01
AE	0.65	—	0.76	—	.15	1	.70
ACE	0.65	—	0.76	—	.15	0	—
ADE	0.56	—	0.75	0.36	.00	0	—

```
Data NInput_vars=2 NObservations=590
#Include mzdepsum.dat
```

where the dat file reads

```
PMatrix File=MZdep.cov
ACov File=MZdep.asy
```

in the MZ group, with the same commands for the DZ group except for the number of observations (`NObs=440`) and a global replacement of DZ for MZ. For clarity, the comments at the beginning also should be changed.

Results of fitting the ACE and ADE models and submodels are summarized in Table 6.10. First, note that the degrees of freedom for fitting to correlation matrices are fewer than when fitting to covariance matrices. Although we provide Mx with two correlation matrices, each consisting of 1's on the diagonal and a correlation on the off-diagonal, the 1's on the diagonal cannot be considered unique. In fact, only one of them conveys information which effectively 'scales' the covariance. There is no information in the remaining three 1's on the diagonals of the MZ and DZ correlation matrices, *but Mx does not make this distinction.* Therefore, we must adjust the degrees of freedom by adding the option `Option DFreedom=-3`. Another way of looking at this is that the diagonal 1's convey no information whatsoever, but that we use one parameter to estimate the diagonal elements (e; it appears only in the expected variances, not the expected covariances). Thus, there are 4 imaginary variances and 1 parameter to estimate them — giving 3 statistics too many.

Second, the substantive interpretation of the results is that the model with just random environment fails, indicating significant familial aggregation for diagnoses of major depressive disorder. The environmental explanation of familial covariance also fails ($\chi^2_1 = 6.40$) but a model of additive genetic and random environment effects fits well ($\chi^2_1 = .15$). There is no possible room for significant improvement with the addition of any other parameter, since there are only .15 χ^2 units left. Nevertheless, we fitted both ACE and ADE models and found that dominance genetic effects could account for the remaining variability whereas shared environmental effects could not. This finding is in agreement with the observation that the MZ correlation is slightly greater than twice the DZ correlation. The heritability of liability to Major Depressive Disorder is moderate but significant at 42%, with the remaining variability associated with random environmental sources including error of measurement. These results are not compatible with the view that shared family experiences such as parental rearing, social class, or parental loss are key factors in the etiology of major depression. More modest effects of these factors may be detected by including them in multivariate model fitting (Kendler *et al.*, 1992a; Neale *et al.*, 1992).

Of course, every study has its limitations, and here the primary limitations are that: (i) the results only apply to females; (ii) the twin population is not likely to be perfectly representative of the general population, as it lacks twins who moved out of or into the state, or failed to respond to initial questionnaire surveys; (iii) a small number of the twins diagnosed as having major depression may have had bipolar disorder (manic depression), which may be etiologically distinct; (iv) the reliance on retrospective reporting of lifetime mental illness may be subject to bias by either currently well or currently ill subjects or both; (v) MZ twins may be treated more similarly as children than DZ twins; and (vi) not all twins were past the age at risk of first onset of major depression. Consideration of the first five of these factors is given in Kendler *et al.* (1992c). Of particular note is that a test of limitation (v), the 'equal environments' assumption, was performed by logistic regression of absolute pair difference of diagnosis (scored 0 for normal and 1 for affected) on a quasi-continuous measure of similarity of childhood treatment. Although MZ twins were on average treated more similarly than DZ twins, this regression was found to be non-significant. General methods to handle the effects of zygosity differences in environmental treatment form part of the class of data-specific models to be discussed in Section **??**. Overall there was no marked regression of age on liability to disease in these data, indicating that correction for the contribution of age to the common environment is not necessary (see the next section). Variable age at onset has been considered by Neale *et al.* (1989) but a full treatment of this problem is beyond the scope of this volume. Such methods incorporate not only censoring of the risk period, but also the genetic architecture of factors involved in age at onset and their relationship to factors relevant in the etiology of liability to disease. Note, however, that this problem, like the problem of measured shared environmental effects, may also be considered as part of the class of data-specific models.

6.4 Model for Age-Correction of Twin Data

We now turn to a slightly more elaborate example of univariate analysis, using data from the Australian twin sample that were used in the BMI example earlier, but in this case data on social attitudes. Factor analysis of the item responses revealed a major dimension with low scores indicating radical attitudes and high scores indicating attitudes commonly labelled as "conservative." Our *a priori* expectation is that variation in this dimension will be largely shaped by social environment and that genetic factors will be of little or no importance. This expectation is based on the differences between the MZ and DZ correlations; $r_{MZ} = 0.68$ and $r_{DZ} = 0.59$, indicating little, if any, genetic influence on social attitudes. We also might expect that conservatism scores are affected by age. We can use the Mx script in Appendix **??** to examine the age effects, reading in the age of each twin pair and the conservatism scores for twin 1 (`Cons_t1`) and twin 2 (`Cons_t2`). Since in this specification we have 3 indicator variables, we adjust `NInput_vars=3`. If we initially ignore age, as an exploratory analysis, we can select only the conservatism scores for analysis, using the `Select` command (note that the list of variables selected must end with a semicolon '`;`').

The script fits the ACE model. The results of this model are presented in the fourth line of the standardized results of Table 6.11, which shows that the squares of parameters estimated from the model sum to one, because these correspond to the proportions of variance associated with each source (A, C, and E).

The significance of common environmental contributions to variance in conservatism may be tested by dropping c (AE model) but this leads to a worsening of χ^2 by 53.8 for 1 d.f., confirming its importance. Similarly, the poor fit of the CE model confirms that genetic factors also contribute to individual differences (significance of a is $19.41 - 3.07 = 16.34$ for 1 df, which is highly significant). The e model, which hypothesizes that there is no family resemblance for conservatism, is overwhelmingly rejected, illustrating of the great power of this data set to discriminate between competing hypotheses. For interest, we also present the results of the ADE model. Since we have already noted that r_{DZ} is appreciably greater than half the MZ correlation, it is clear that this model is inappropriate. Symmetric with the results of fitting an ACE model to the BMI data (where $2r_{DZ}$ was still less than r_{MZ}, indicating dominance), we now find that the estimate of d gets "stuck" on its lower bound of zero. The BMI and conservatism examples illustrate

Table 6.11: Conservatism in Australian females: standardized parameter estimates for additive genotype (A), common environment (C), random environment (E) and dominance genotype (D).

	Parameter Estimates				Fit statistics		
Model	a	c	e	d	χ^2	df	p
E	—	—	1.000	—	823.76	5	.000
CE	—	0.804	0.595	—	19.41	4	.001
AE	0.836	—	0.549	—	56.87	4	.000
ACE	0.464	0.687	0.559	—	3.07	3	.380
ADE	0.836	—	0.549	0.000	56.87	3	.000

in a practical way the perfect reciprocal dependence of c and d in the classical twin design of which only one may be estimated. The issue of the reciprocal confounding of shared environment and genetic non-additivity (dominance or epistasis) in the classical twin design has been discussed in detail in papers by Martin *et al.,* (1978), Grayson (1989), and Hewitt (1989).

It is clear from the results above that there are major influences of the shared environment on conservatism. One aspect of the environment that is shared with perfect correlation by cotwins is their age. If a variable is strongly related to age and if a twin sample is drawn from a broad age range, as opposed to a cohort sample covering a narrow range of birth years, then differences between twin pairs in age will contribute to estimated common environmental variance. This is the case for the twins in the Australian sample, who range from 18 to 88 years old. It is clearly of interest to try to separate this variance due to age differences from genuine cultural differences contributing to the estimate of c.

Fortunately, structural equation modeling, which is based on linear regression, provides a very easy way of allowing for the effects of age regression while simultaneously estimating the genetic and environmental effects (Neale and Martin, 1989). Figure 6.2 illustrates the method with a path diagram, in which the regression of $Cons_{t1}$ and $Cons_{t2}$ on Age is s (for senescence), and this is specified in the script excerpt below.

We now work with the full 3×3 covariance matrices (so the `Select` statement is dropped from the previous job). We estimate simultaneously the contributions of additive genetic, shared and unique environmental factors on conservatism, the variance of age `V*V`, and the contribution of age to conservatism `S*V`.

```
Group 2: female MZ twin pairs
Data NInput_vars=3 NOberservations=941
Labels age cons_t1 cons_t2
CMatrix Symmetric File=ozconmzf.cov
```

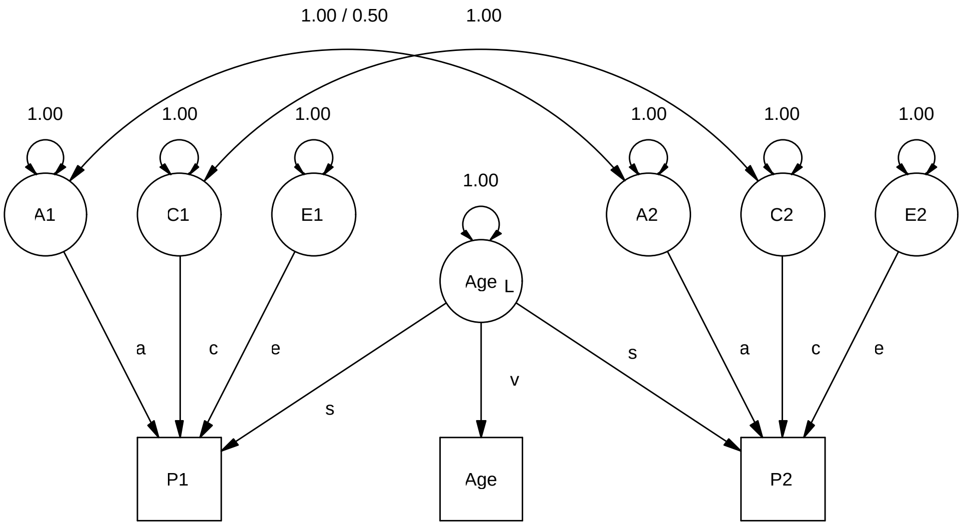


Figure 6.2: Path model for additive genetic (A), shared environment (C) and specific environment (E) effects on phenotypes (P) of pairs of twins ($T1$ and $T2$). α is fixed at 1 for MZ twins and at .5 for DZ twins. The effects of age are modelled as a standardized latent variable, Age_L, which is the sole cause of variance in observed Age.

```
Matrices= Group 1
Covariances V*V' | V*S'     | V*S'    _
            S*V' | A+C+E+G | A+C+G   _
            S*V' | A+C+G   | A+C+E+G;
```

The matrix algebra here is more complex than usual, and for univariate analysis it would be easier to draw the diagram with the GUI. However, the algebraic approach has the advantage that it is much easier to generalize to the multivariate case.

Results of fitting the ACE model with age correction are in the first row of Table 6.12. Standardized results are presented, from which we see that the standardized regression of conservatism on age (constrained equal in twins 1 and 2) is 0.422. In the unstandardized solution, the first loading on the age factor is the standard deviation of the sample for age, in this case 13.2 years. The latter is an estimated parameter, making five free parameters in total. In each group we have k(k+1)/2 statistics, where k is the number of observed variables, so there are 2× (k(k+1)/2-5 =7 degrees of freedom. Dropping either c or a still causes significant worsening of the fit, and it also is very clear that one cannot omit the age regression itself (final ACE model; $\chi^2_8 = 370.17, p = .000$).

It is interesting to compare the results of the ACE model in Table 6.11 with those of the ACES model in Table 6.12. We see that the estimates of e and a are

Table 6.12: Age correction of Conservatism in Australian females: standardized parameter estimates for models of additive genetic (A), common environment (C), random environment (E), and senescence or age (S).

	Parameter Estimates				Fit statistics		
Model	a	c	e	s	χ^2	df	p
ACES	0.474	0.534	0.558	0.422	7.41	7	.388
AES	0.720	—	0.547	0.426	31.56	8	.000
CES	—	0.685	0.595	0.421	25.49	8	.001
ACE	0.464	0.687	0.559	—	370.17	8	.000

identical in the two tables, accounting for $0.559^2 = 31\%$ and $0.464^2 = 22\%$ of the total variance, respectively. However, in the first table the estimate of $c = 0.687$, accounting for 47% of the variance. In the analysis with age however, $c = 0.534$ and accounts for 29% of variance, and age accounts for $0.422^2 = 18\%$. Thus, we have partitioned our original estimate of 47% due to shared environment into 18% due to age regression and the remaining 29% due to 'genuine' cultural differences. If we choose, we may recalculate the proportions of variance due to a, c, and e, as if we were estimating them from a sample of uniform age — assuming of course that the causes of variation do not vary with age (see Chapter 9). Thus, genetic variance now accounts for $22/(100 - 18) = 27\%$ and shared environment variance is estimated to be $29/82 = 35\%$.

Our analysis suggests that cultural differences are indeed important in determining individual differences in social attitudes. However, before accepting this result too readily, we should reflect that estimates of shared environment may not only be inflated by age regression, but also by the effects of assortative mating — the tendency of like to marry like. Since there is known to be considerable assortative mating for conservatism (spouse correlations are typically greater than 0.6), it is possible that a substantial part of our estimate of c^2 may arise from this source (Martin *et al.*, 1986). This issue will be discussed in greater detail in Chapter **??**.

Age is a somewhat unusual variable since it is perfectly correlated in both MZ and DZ twins (so long as we measure the members of a pair at the same time). There are relatively few variables that can be handled in the same way, partly because we have assumed a strong model that age *causes* variability in the observed phenotype. Thus, for example, it would be inappropriate to model length of time spent living together as a cause of cancer, even though cohabitation may lead to greater similarity between twins. In this case a more suitable model would be one in which the shared environment components are more highly correlated the longer the twins have been living together. Such a model would predict greater twin similarity, but would not predict correlation between cohabitation and cancer.

Some further discussion of this type of model is given in Section **??** in the context of data-specific models. One group of variables that may be treated in a similar way to the present treatment of age consists of maternal gestation factors. Vlietinck *et al.* (1989) fitted a model in which both gestational age and maternal age predicted birthweight in twins.

Finally we note that at a technical level, age and similar putative causal agents might most appropriately be treated as x-variables in a multiple regression model. Thus the observed covariance of the x-variables is incorporated directly into the expected matrix, so that the analysis of the remaining y-variables is conditional on the covariance of the x-variables. This type of approach is free of distributional assumptions for the x-variables, and is analogous to the analysis of covariance. However, when we fit a model that estimates a single parameter for the variance of age in each group, the estimated and observed variances are generally equal, so the same results are obtained.

Chapter 7

Power and Sample Size

7.1 Introduction

In this chapter we discuss the power of the twin study to detect variance components in behavioral characters. Our discussion is not in any way intended to be an exhaustive description of the power of the twin study under all possible combinations of causal factors and model parameters. Such a description is in large part available for the continuous case (Martin *et al.*, 1978) and the ordinal case (Neale *et al.*, 1994), and there is an extensive comparison of the power of various designs to detect cultural transmission (Heath *et al.*, 1985). As we move out of the framework of the univariate classical twin study to consider multivariate hypotheses and data from additional classes of relatives, a comprehensive treatment rapidly becomes unmanageably large. Fortunately, it seems rather unnecessary because the prospective researcher usually has certain specific aims of a study in mind, and often has a reasonable idea about the values of some of the parameters in the model. This information can be used to prune the prodigious tree of possible scenarios to manageable proportions. All that is required is an understanding of the factors contributing to power and the principles involved, which we aim to provide in (Section 7.2) and Section 7.3 respectively. We illustrate these methods with a limited range of examples for continuous (Section 7.4) and categorical (Section 7.5) twin data.

7.2 Factors Contributing to Power

One of the greatest advantages of the model-fitting approach is that it allows us to conduct tests of significance of alternative hypotheses. We can ask, for example, whether a given data set really supports our assertion that shared environmental effects contribute to variation in one trait or another (i.e., is $c^2 > 0$?).

Our ability to show that a specific effect is important obviously depends on a number of factors. These include:

1. The effect under consideration, for example, a^2 or c^2;

2. The actual size of the effect in the population being studied — larger values are detected more easily than small values;

3. The probability level adopted as the conventional criterion for rejection of the null-hypothesis that the effect is zero — rejection at higher significance levels will be less likely to occur for a given size of effect;

4. The actual size of the sample chosen for study — larger samples can detect smaller effects;

5. The actual composition of the sample with respect to the relative frequencies of the different biological and social relationships selected for study;

6. The level of measurement used — categorical, ordinal, or continuous.

All of these considerations lead us to the important question of *power*. If we are trying to get a sense of what we are likely to be able to infer from our own data set, or if we are considering a new study, we must ask either "What inferences can we hope to be able to make with our data set?" or "What kind of data set and sample sizes is it likely we will need to answer a particular set of questions?" In the next section we show how to answer these questions in relation to simple hypotheses with twin studies and suggest briefly how these issues may be explored for more complex designs and hypotheses.

7.3 Steps in Power Analysis

The basic approach to power analysis is to imagine that we are doing an identical study many times. For example, we pretend that we are trying to estimate a, c, and e for a given population by taking samples of a given number of MZ and DZ twins. Each sample would give somewhat different estimates of the parameters, depending on how many twins we study, and how big a, c, and e are in the study population. Suppose we did a very large number of studies and tabulated all the estimates of the shared environmental component, c^2. In some of the studies, even though there was some shared environment in the population, we would find estimates of c^2 that were not significant. In these cases we would commit "type II errors." That is, we would not find a significant effect of the shared environment even though the value of c^2 in the population was truly greater than zero. Assuming we were using a χ^2 test for 1 df to test the significance of the shared environment, and we had decided to use the conventional 5% significance level, the probability of Type II error would be the expected proportion of samples in which we mistakenly decided in favor of the null hypothesis that $c^2 = 0$. These cases would be those in which the observed value of χ^2 was less than 3.84, the 5% critical value for 1 df. The other samples in which χ^2 was greater than 3.84 are those in which we would decide, correctly, that there was a significant shared environmental effect in the population. The expected proportion of samples in which we decide correctly against the null hypothesis is the *power of the test.*

Designing a genetic study boils down to deciding on the numbers and types of relationships needed to achieve a given power for the test of potentially important genetic and environmental factors. There is no general solution to the problem of power. The answers will depend on the specific values we contemplate for all the factors listed above. Before doing any power study, therefore, we have to decide the following questions in each specific case:

1. What kinds of relationships are to be considered?
2. What significance level is to be used in hypothesis testing?
3. What values are we assuming for the various effects of interest in the population being studied?
4. What power do we want to strive for in designing the study?

When we have answered these questions exactly, then we can conduct a power analysis for the specified set of conditions by following some basic steps:

1. Obtain expected covariance matrices for each set of relationships by substituting the assumed values of the population parameters in the model for each relationship.
2. Assign some initial arbitrary sample sizes to each separate group of relatives.
3. Use Mx to analyze the expected covariance matrices just as we would to analyze real data and obtain the χ^2 value for testing the specific hypothesis of interest.
4. Find out (from statistical tables) how big that χ^2 has to be to guarantee the power we need.
5. Use a simple formula (given below) to multiply our assumed sample size and solve for the sample size we need.

It is essential to remember that the sample size we obtain in step five only applies to the particular effect, design, sample sizes, and even to the distribution of sample sizes among the different types of relationship assumed in a specific power calculation. To explore the question of power fully, it often will be necessary to consider a number, sometimes a large number, of different designs and population values for the relevant effects of genes and environment.

7.4 Power for the continuous case

A common question in genetic research concerns the ability of a study of twins reared together to detect the effects of the shared environment. Let us investigate this issue using Mx. Following the steps outlined above, we start by stipulating that we are going to explore the power of a classical twin study — that is, one in which we measure MZ and DZ twins reared together. We shall assume that 50% of the variation in the population is due to the unique environmental experiences of individuals ($e^2 = 0.5$). The expected MZ twin correlation is therefore 0.50. This intermediate value is chosen to be typical of many of the less-familial traits. Anthropometric traits, and many cognitive traits, tend to have higher MZ correlations than this, so the power calculations should be conservative as far as such variables are concerned. We assume further that the additive genetic component explains 30% of the total variation ($a^2 = 0.30$) and that the shared family environment accounts for the remaining 20% ($c^2 = 0.20$). We now substitute these parameter values into the algebraic expectations for the variances and covariances of MZ and DZ twins:

$$\begin{array}{llllllll}
\text{Total variance} & = & a^2 + c^2 + e^2 & = & 0.30 + 0.2 + 0.5 & = & 1.00 \\
\text{MZ covariance} & = & a^2 + c^2 & = & 0.30 + 0.2 & = & 0.50 \\
\text{DZ covariance} & = & .5a^2 + c^2 & = & 0.15 + 0.2 & = & 0.35
\end{array}$$

In Appendix **??** we show a version of the Mx code for fitting the ACE model to the simulated covariance matrices. In addition to the expected covariances we must assign an arbitrary sample size and structure. Initially, we shall assume the study involves equal numbers, 1000 each, of MZ and DZ pairs. In order to conduct the power calculations for the c^2 component, we can run the job for the full (ACE) model first and then the AE model, obtaining the expected difference in χ^2 under the full and reduced models just as we did earlier for testing the significance of the shared environment in real data.

Notice that fitting the full ACE model yields a goodness-of-fit χ^2 of zero. This should always be the case when we use Mx to solve for all the parameters of the model we used to generate the expected covariance matrices because, since there is no sampling error attached to the simulated covariance matrices, there is perfect agreement between the matrices supplied as "data" and the expected values under the model. In addition, the parameter estimates obtained should agree precisely with those used to simulate the data; if they are not, but the fit is still perfect, it suggests that the model is not identified (see Section 5.7) Therefore, as long as we are confident that we have specified the structural model correctly and that the full model is identified, there is really no need to fit the full model to the simulated covariances matrices since we know in advance that the "χ^2" is expected to be zero. In practice it is often helpful to recover this known result to increase our confidence that both we and the software are doing the right thing.

For our specific case, with samples of 1000 MZ and DZ pairs, we obtain a

Table 7.1: Non-centrality parameter, λ, of non-central χ^2 distribution for 1 df required to give selected values of the power of the test at the 5% significance level (selected from Pearson and Hartley, 1972).

Desired Power	λ
0.25	1.65
0.50	3.84
0.75	6.94
0.80	7.85
0.90	10.51
0.95	13.00

goodness-of-fit χ^2_4 of 11.35 for the AE model. Since the full model yields a perfect fit ($\chi^2_3 = 0$), the expected difference in χ^2 for 1 df — testing for the effect of the shared environment — is 11.35. Such a value is well in excess of the 3.84 necessary to conclude that c^2 is significant at the 5% level. However, this is only the value expected in the ideal situation. With real data, individual χ^2 values will vary greatly as a function of sampling variance. We need to choose the sample sizes to give an expected value of χ^2 such that observed values exceed 3.84 in a specified proportion of cases corresponding to the desired power of the test.

It turns out that such problems are very familiar to statisticians and that the expected values of χ^2 needed to give different values of the power at specified significance levels for a given df have been tabulated extensively (see Pearson and Hartley, 1972). The expected χ^2 is known as the *centrality parameter* (λ) of the non-central χ^2 distribution (i.e., when the null-hypothesis is false). Selected values of the non-centrality parameter are given in Table 7.1 for a χ^2 test with 1 df and a significance level of 0.05.

With 1000 pairs of MZ and DZ twins, we find a non-centrality parameter of 11.35 when we use the χ^2 test to detect c^2 which explains 20% of the variation in our hypothetical population. This corresponds to a power somewhere between 90% ($\lambda = 10.51$) and 95% ($\lambda = 13.00$). That is, 1000 pairs each of MZ and DZ twins would allow us to detect, at the 5% significance level, a significant shared environmental effect when the true value of c^2 was 0.20 in about 90-95% of all possible samples of this size and composition. Conversely, we would only fail to detect this much shared environment in about 5-10% of all possible studies.

Suppose now that we want to figure out the sample size needed to give a power of 80%. Let this sample size be N^*. Let N_0 be the sample size assumed in the initial power analysis (2000 pairs, in our case). Let the expected χ^2 for the particular test being explored with this sample size be χ^2_E (11.35, in this example). From Table 7.1, we see that the non-centrality parameter, λ, needs to be 7.85 to give a power of 0.80. Since the value of χ^2 is expected to increase linearly as a function of sample size we can obtain the sample size necessary to give 80% power by solving:

$$\begin{aligned} N^* &= \frac{\lambda}{\chi^2_E} N_0 \qquad (7.1) \\ &= \frac{7.85}{11.35} \times 2000 \\ &= 1383 \end{aligned}$$

That is, in a sample comprising 50% MZ and 50% DZ pairs reared together, we would require 1,383 pairs in total, or approximately 692 pairs of each type to be 80% certain of detecting a shared environmental effect explaining 20% of the total variance, when a further 30% is due to additive genetic factors.

It must be emphasized again that this particular sample size is specific to the study design, sample structure, parameter values and significance level assumed in the simulation. Smaller samples will be needed to detect larger effects. Greater power requires larger samples. Larger studies can detect smaller effects, and finally, some parameters of the model may be easier to detect than others.

7.5 Loss of Power with Ordinal Data

An important factor which affects power but is often overlooked is the form of measurement used. So far we have considered only continuous, normally distributed variables, but of course, these are not always available in the biosocial sciences. An exhaustive treatment of the power of the ordinal classical twin study is beyond the scope of this text, but we shall simply illustrate the loss of power incurred when we use more crude scales of measurement (Neale *et al.*, 1994). Consider the example above, but suppose this time that we wish to detect the presence of additive genetic effects, a^2, in the data. For the continuous case this is a trivial modification of the input file to fit a model with just c and e parameters. The chi-squared from running this program is 19.91, and following the algebra above (equation 7.1) we see that we would require $2000 \times 7.85/19.91 = 788$ pairs in total to be 80% certain of rejecting the hypothesis that additive genes do not affect variation when in the true world they account for 30%, with shared environment accounting for a further 20%. Suppose now that rather than measuring on a continuous scale, we have a dichotomous scale which bisects the population; for example, an item on which 50% say 'yes' and 50% say no. The data for this case may be summarized as a contingency table, and we wish to generate tables that: (i) have a total sample size of 1000; (ii) reflect a correlation in liability of .5 for MZ and .35 for DZ twins; and (iii) reflect our threshold value of 0 to give 50% either side of the threshold. Any routine that will compute the bivariate normal integral for given thresholds and correlation is suitable to generate the expected proportions in each cell. In this case we use a short Mx script (Neale, 1991) to generate the data for PRELIS. We can use the weight option in PRELIS to indicate the cell counts for our contingency tables. Thus, the PRELIS script might be:

```
Power calculation MZ twins
DA NI=3 NO=0
LA; SIM1 SIM2 FREQ
RA FI=expectmz.frq
WE FREQ
OR sim1 sim2
OU MA=PM SM=SIMMZ.COV SA=SIMMZ.ASY PA
```

with the file `expectmz.frq` looking like this:

```
0 0 333.333
0 1 166.667
1 0 166.667
1 1 333.333
```

A similar approach with the DZ correlation and thresholds gives expected frequencies which can be used to compute the asymptotic variance of the tetrachoric correlation for this second group. The simulated DZ frequency data might appear as

```
0 0 306.9092
0 1 193.0908
1 0 193.0908
1 1 306.9092
```

The cells display considerable symmetry — there are as many concordant 'no' pairs as there are concordant 'yes' pairs because the threshold is at zero. Running PRELIS generates output files, and we can see immediately that the correlations for MZ and DZ twins remain the desired .5 and .35 assumed in the population. The next step is to feed the correlation matrix and the weight matrix (which only contains one element, the asymptotic variance of the correlation between twins) into Mx, in place of the covariance matrix that we supplied for the continuous case. This can be achieved by changing just three lines in each group of our Mx power script:

```
#NGroups 2
Data NInput_vars=2 NObservations=1000
PMatrix File=SIMMZ.COV
ACov File=SIMMZ.ASY
```

with corresponding filenames for the DZ group, of course. When we fit the model to these summary statistics we observe a much smaller χ^2 than we did for the continuous case; the χ^2 is only 6.08, which corresponds to a requirement of 2,582 pairs in total for 80% power at the .05 level. That is, *we need more than three times as many pairs to get the same information about a binary item than we need for a continuous variable.* The situation further deteriorates as we move the threshold to one side of the distribution. Simulating contingency tables, computing tetrachorics and weight matrices, and fitting the false model when the threshold is one standard deviation (SD) to the right (giving 15.9% in one category and 84.1% in the other), the χ^2 is a mere 3.29, corresponding a total sample size of 4,772 total pairs. More extreme thresholds further reduce power, so that for an item (or a disease) with a 95:5% split we would require 13,534 total pairs. Only in the largest studies could such sample sizes be attained, and they are quite unrealistic for data that could be collected by personal interview or laboratory measurement. On the positive side, it seems unlikely that given the advantages of the clinical interview or laboratory setting, our only measure could be made at the crude 'yes or no' binary response level. If we are able to order our data into more than two categories, some of the lost power can be regained. Following the procedure outlined above, and assuming that there are two thresholds, one at -1 SD and one at $+1$ SD, then the χ^2 obtained is 8.16, corresponding to 'only' 1,924 pairs for 80% chance of finding additive variance significant at the .05 level. If one threshold is 0 and the other at 1 SD then the χ^2 rises slightly to 9.07, or 1,730 pairs. Further improvement can be made if we increase the measurements to comprise four categories. For example, with thresholds at -1, 0, and 1 SD the χ^2 is 12.46, corresponding to a sample size of 1,240 twin pairs.

While estimating tetrachoric correlations from a random sample of the population has considerable advantages, it is not always the method of choice for studies focused on a single outcome, such as schizophrenia. In cases where the base rates are so low (e.g., 1%) then it becomes inefficient to sample randomly, and an ascertainment scheme in which we select cases and examine their relatives is a practical and powerful alternative, if we have good information on the base rate in the population studied. The necessary power calculations can be performed using the computer packages LISCOMP (Muthén, 1987) or Mx (Neale, 1997).

7.6 Exercises

1. Change the example program to obtain the expected χ^2 for the test for additive genetic effects. Find out how many pairs are needed to obtain significant estimates of a^2 in 80% of all possible samples.

2. Explore the effect of power of a particular test of altering the proportion of MZ and DZ twins in the sample.

3. Show that the change in expected χ^2 is proportional to the change in sample size.

4. Obtain and tabulate the sample sizes necessary to detect a significant a^2 when the population parameter values are as follows:

a^2	c^2
0.10	0.00
0.30	0.00
0.60	0.00
0.90	0.00

In what way do these values change if there are shared environmental effects?

5. Show that with small sample sizes for the number of pairs in each group, some bias in the chi-squared is introduced. Consider whether or not this may be due to the $n-1$ part of the maximum likelihood loss function (Equation **??** on page **??**).

Chapter 8

Social Interaction

8.1 Introduction

This chapter introduces a technique for specifying and estimating paths between dependent variables, so called non-recursive models. Uses of this technique include: modeling social interactions, for example, sibling competition and cooperation; testing for direction of causation in bivariate data, e.g., whether life events cause depression or vice versa; and developmental models for longitudinal or repeated measurements.

Models for sibling interaction have been popular in genetics for some time (Eaves, 1976b), and the reader should see Carey (1986b) for a thorough treatment of the problem in the context of variable family size. Here we provide an introductory outline and application for the restricted case of pairs of twins, and we assume no effects of other siblings in the family. We further confine our treatment to sibling interactions *within variables.* Although multivariate sibling interactions (such as aggression in one twin causing depression in the cotwin) may in the long run prove to be more important than those within variables, they are beyond the scope of this introductory text. Section 8.2 provides a summary of the basic univariate genetic model without interaction. The extension to include sibling interaction is described in Section 8.3. Details on the consequences of sibling interaction on the variation and covariation are discussed in Section 8.4

8.2 Basic Univariate Model without Interaction

Up to this point, we have been concerned primarily with decomposing observed phenotypic variation into its genetic and environmental components. This has been accomplished by estimating the paths from latent or independent variables to dependent variables. A basic univariate path diagram is set out in Figure 8.1. This

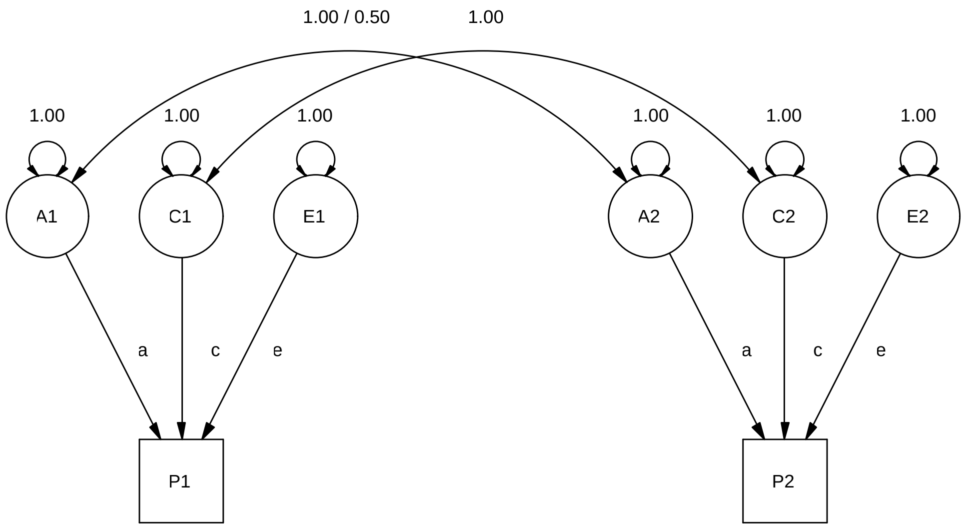


Figure 8.1: Basic path diagram for univariate twin data.

path diagram shows the deviation phenotypes P_1 and P_2, of a pair of twins. Here we refer to the phenotypes as *deviation phenotypes* to emphasize the point that the model assumes variables to be measured as deviations from the means, which is the case whenever we fit models to covariance matrices and do not include means. The deviation phenotypes P_1 and P_2 result from their respective additive genetic deviations, A_1 and A_2, their shared environment deviations, C_1 and C_2, and their non-shared environmental deviations, E_1 and E_2. The linear model corresponding to the path diagram is:

$$\begin{aligned} P_1 &= aA_1 + cC_1 + eE_1 \\ P_2 &= aA_2 + cC_2 + eE_2 \end{aligned}$$

In matrix form we can write:

$$\begin{pmatrix} P_1 \\ P_2 \end{pmatrix} = \begin{pmatrix} a & c & e & 0 & 0 & 0 \\ 0 & 0 & 0 & a & c & e \end{pmatrix} \begin{pmatrix} A_1 \\ C_1 \\ E_1 \\ A_2 \\ C_2 \\ E_2 \end{pmatrix}$$

or as a matrix expression

$$\mathbf{y} = \mathbf{G}\mathbf{x}$$

Details of specifying and estimating this basic univariate model are given in Chapter 6. One of the interesting assumptions of this basic ACE model is that the siblings' or twins' phenotypes have no influence on each other. This assumption may well be true of height or finger print ridge count, but is it necessarily true for a behavior like smoking, a psychiatric condition like depression, delinquent behavior in children or even an anthropometric measure like the body mass index? We should not, in general, assume *a priori* that a source of variation is absent, especially when an empirical test of the assumption may be readily performed. However, we may as well recognize from the onset that evidence for social interactions or sibling effects is pretty scarce. The fact is that usually one form or another of the basic univariate model adequately describes a twin or family data set, within the power of the study. This tells us that there will not be evidence of significant social interactions since, were such effects substantial, they would lead to failure of basic univariate models. Nevertheless, this extension of the basic models is of considerable theoretical interest and studying its outcome on the expectations derived from the models can provide insight into the nature and results of social influences. The applications to bivariate and multivariate causal modeling are perhaps even more intriguing and will be taken up in chapter **??**.

8.3 Sibling Interaction Model

Suppose that we are considering a phenotype like number of cigarettes smoked. For the sake of exposition we will set aside questions about the appropriate scale of measurement, what to do about non-smokers and so on, and assume that there is a well-behaved quantitative variable, which we can call 'smoking' for short. What we want to specify is the influence of one sibling's (twin's) smoking on the other sibling's (cotwin's) smoking. Figure 8.2 shows a path diagram which extends the basic univariate model for twins to include a path of magnitude s from each twin's smoking to the cotwin. If the path s is positive then the sibling interaction is essentially cooperative, i.e., the more (less) one twin smokes the more (less) the cotwin will smoke as a consequence of this direct influence. We can easily conceive of a highly plausible mechanism for this kind of influence when twins are cohabiting; as a twin lights up she offers her cotwin a cigarette. If the path s is negative then the sibling interaction is essentially competitive. The more (less) one twin smokes the less (more) the cotwin smokes. Although such competition contributes negatively to the covariance between twins, it may well not override the positive covariance resulting from shared familial factors. Thus, even in the presence of competition the observed phenotypic covariation may still be positive. If interactions are cooperative in some situations and competitive in others, our analyses will reveal the predominant mode. But before considering the detail of our expectations, let us look more closely at how the model is specified. The linear model is now:

$$P_1 = sP_2 + aA_1 + cC_1 + eE_1 \qquad (8.1)$$

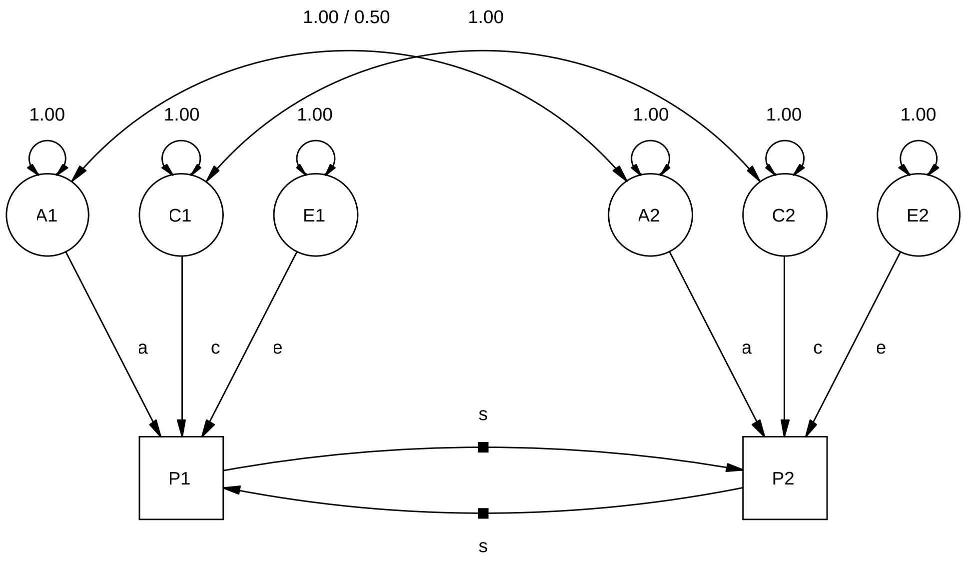


Figure 8.2: Path diagram for univariate twin data, incorporating sibling interaction.

$$P_2 = sP_1 + aA_2 + cC_2 + eE_2 \tag{8.2}$$

In matrix form we have

$$\begin{pmatrix} P_1 \\ P_2 \end{pmatrix} = \begin{pmatrix} 0 & s \\ s & 0 \end{pmatrix} \begin{pmatrix} P_1 \\ P_2 \end{pmatrix} + \begin{pmatrix} a & c & e & 0 & 0 & 0 \\ 0 & 0 & 0 & a & c & e \end{pmatrix} \begin{pmatrix} A_1 \\ C_1 \\ E_1 \\ A_2 \\ C_2 \\ E_2 \end{pmatrix}$$

or

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \mathbf{G}\mathbf{x}$$

In this form the **B** matrix is a square matrix with the number of rows and columns equal to the number of dependent variables. The leading diagonal of the **B** matrix contains zeros. The element in row i and column j represents the path from the j^{th} dependent variable to the i^{th} dependent variable. From this equation we can deduce, as shown in more detail below, that:

$$\mathbf{y}(\mathbf{I} - \mathbf{B}) = \mathbf{G}\mathbf{x}$$

$$\mathbf{y} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{G}\mathbf{x}$$

8.3.1 Application to CBC Data

By way of illustration we shall analyze data collected using the Achenbach Child Behavior Checklist (CBC; Achenbach & Edelbrock, 1983) on juvenile twins aged 8 through 16 years living in Virginia. Mothers were asked the extent to which a series of problem behaviors were characteristic of each of their twin children over the last six months. The 118 problem behaviors that were rated can be categorized, on the basis of empirical clustering, into two broad dimensions of *internalizing* and *externalizing* problems. The former are typified by fears, psychosomatic complaints, and symptoms of anxiety and depression. Externalizing behaviors are characterized by "acting out" — delinquent and aggressive behaviors. The factor patterns vary somewhat with the age and sex of the child but there are core items which load on the broad factors in both boys and girls at younger (6-11 years) and older (12-16 year) ages. The 24 core items for the externalizing dimension analyzed by Silberg *et al.* (1992) and Hewitt *et al.* (1992) include among other things: arguing a lot, destructive behavior, disobedience, fighting, hanging around with children who get into trouble, running away from home, stealing, and bad language. For such behaviors we might suspect that siblings will influence each other in a cooperative manner through imitation or mutual reinforcement. The Mx script in Appendix **??** specifies the model for sibling interactions shown in Figure 8.2.

By varying the script, the standard E, AE, CE, and ACE models may be fitted to the data to obtain the results shown in Table 8.1. Clearly the variation and

Table 8.1: Preliminary results of model fitting to externalizing behavior problems in Virginia boys from larger families.

	Fit statistics			Parameter Estimates		
Model	df	χ^2	AIC	a	c	e
AE	4	32.57	24.6	.78	—	.33
CE	4	29.80	21.8	—	.78	.43
ACE	3	4.95	-1.0	.50	.64	.34

co-aggregation of boys' behaviors problems cannot be explained either by a model which allows only for additive genetic effects (along with non-shared environmental influences), nor by a model which excludes genetic influences altogether. The ACE model fits very well ($p = .18$) and suggests a heritability of 33% with shared environmental factors accounting for 52% of the variance[1]. But is the ACE model the best in this case? We observe that the pooled individual phenotypic variances of

[1]The reader might like to consider what the components of this shared variance might include in these data obtained from the mothers of the twins and think forward to our treatment of rating data in Chapter 11.

the MZ twins (0.915) are greater than those of the DZ twins (0.689) and, although this discrepancy is apparently not statistically significant with our sample sizes (171 MZ pairs and 194 DZ pairs), we might be motivated to consider sibling interactions.

Fitting the model shown in Figure 8.2 yields results given in Table 8.2. Our general conclusion is that while the evidence for social interactions is not unequivocal, a model including additive genetic effects, non-shared environments, and reciprocal sibling cooperation provides the best account of these data.

Table 8.2: Parameter estimates and goodness of fit statistics from fitting models of sibling interaction to CBC data.

	Fit statistics			Parameter estimates			
Model	df	χ^2	AIC	a	c	e	s
E+s	4	29.80	21.8	—	—	*	*
AE+s	3	1.80	-4.2	.611	—	.419	.230
CE+s	3	29.80	21.8	—	.882	.282	-.101
ACE+s	2	1.80	-2.2	.611	.000[1]	.419	.230

* Indicates parameters out of bounds.
[1]This parameter is fixed on the lower bound (0.0) by Mx

8.4 Consequences for Variation and Covariation

In this section we will work through the matrix algebra to derive expected variance and covariance components for a simplified model of sibling interaction. We then show how this model can be adapted to handle the specific cases of additive and dominant genetic, and shared and non-shared environmental effects. Numerical examples of strong competition and cooperation will be used to illustrate their effects on the variances and covariances of twins and unrelated individuals reared in the same home.

8.4.1 Derivation of Expected Covariances

To understand what it is about the observed statistics that suggests sibling interactions in our twin data we must follow through a little algebra. We shall try to keep this as simple as possible by considering the path model in Figure 8.3, which depicts the influence of an arbitrary latent variable, X, on the phenotype P. As long as our latent variables — A, C, E, etc. — are independent of each other, their effects can be considered one at a time and then summed, even in the presence of social interactions. The linear model corresponding to this path diagram is

$$P_1 = sP_2 + xX_1 \tag{8.3}$$

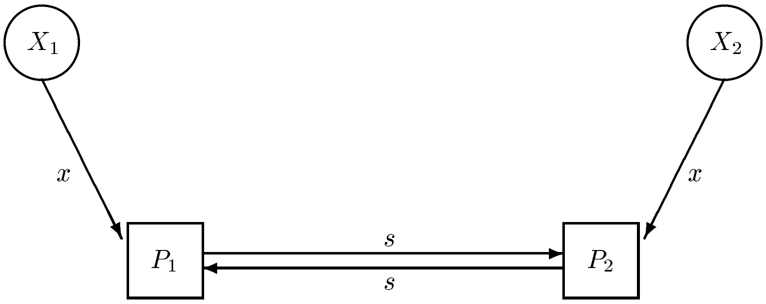


Figure 8.3: Path diagram showing influence of arbitrary exogenous variable X on phenotype P in a pair of relatives (for univariate twin data, incorporating sibling interaction).

$$P_2 = sP_1 + xX_2 \tag{8.4}$$

Or, in matrices:

$$\begin{pmatrix} P_1 \\ P_2 \end{pmatrix} = \begin{pmatrix} 0 & s \\ s & 0 \end{pmatrix} \begin{pmatrix} P_1 \\ P_2 \end{pmatrix} + \begin{pmatrix} x & 0 \\ 0 & x \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

which in turn we can write more economically as

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \mathbf{G}\mathbf{x}$$

Following the rules for matrix algebra set out in Chapters 4 and **??**, we can rearrange this equation, as before:

$$\mathbf{y} - \mathbf{B}\mathbf{y} = \mathbf{G}\mathbf{x} \tag{8.5}$$
$$\mathbf{I}\mathbf{y} - \mathbf{B}\mathbf{y} = \mathbf{G}\mathbf{x} \tag{8.6}$$
$$(\mathbf{I} - \mathbf{B})\mathbf{y} = \mathbf{G}\mathbf{x}\ , \tag{8.7}$$

and then, multiplying both sides of this equation by the inverse of (**I** - **B**), we have

$$\mathbf{y} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{G}\mathbf{x}\ . \tag{8.8}$$

In this case, the matrix (**I** - **B**) is simply

$$\begin{pmatrix} 1 & -s \\ -s & 1 \end{pmatrix} ,$$

which has determinant $1 - s^2$, so $(\mathbf{I} - \mathbf{B})^{-1}$ is

$$\frac{1}{1-s^2} \otimes \begin{pmatrix} 1 & s \\ s & 1 \end{pmatrix} .$$

The symbol $\otimes$ is used to represent the Kronecker product, which in this case simply means that each element in the matrix is to be multiplied by the constant $\frac{1}{1-s^2}$.

We have a vector of phenotypes on the left hand side of equation 8.8. In the chapter on matrix algebra (p. 84) we showed how the covariance matrix could be computed from the raw data matrix $\mathbf{T}$ by expressing the observed data as deviations from the mean to form matrix $\mathbf{U}$, and computing the matrix product $\mathbf{UU}'$. The same principle is applied here to the vector of phenotypes, which has an expected mean of $\mathbf{0}$ and is thus already expressed in mean deviate form. So to find the expected variance-covariance matrix of the phenotypes P_1 and P_2, we multiply by the transpose:

$$\begin{aligned} \mathcal{E}\left\{\mathbf{yy}'\right\} &= \left\{(\mathbf{I}-\mathbf{B})^{-1}\mathbf{Gx}\right\}\left\{(\mathbf{I}-\mathbf{B})^{-1}\mathbf{Gx}\right\}' && (8.9) \\ &= (\mathbf{I}-\mathbf{B})^{-1}\mathbf{G}\mathcal{E}\left\{\mathbf{xx}'\right\}\mathbf{G}'(\mathbf{I}-\mathbf{B})^{-1'} . && (8.10) \end{aligned}$$

Now in the middle of this equation we have the matrix product $\mathcal{E}\{\mathbf{xx}'\}$. This is the covariance matrix of the $\mathbf{x}$ variables. For our particular example, we want two standardized variables, X_1 and X_2 to have unit variance and correlation r so the matrix is:

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} .$$

We now have all the pieces required to compute the covariance matrix, recalling that for this case,

$$\begin{aligned} \mathbf{G} &= \begin{pmatrix} x & 0 \\ 0 & x \end{pmatrix} && (8.11) \\ (\mathbf{I}-\mathbf{B})^{-1} &= \frac{1}{1-s^2} \otimes \begin{pmatrix} 1 & s \\ s & 1 \end{pmatrix} && (8.12) \\ \mathcal{E}\left\{\mathbf{xx}'\right\} &= \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} . && (8.13) \end{aligned}$$

The reader may wish to show as an exercise that by substituting the right hand sides of equations 8.11 to 8.13 into equation 8.10, and carrying out the multiplication, we obtain:

$$\mathcal{E}\left\{\mathbf{yy}'\right\} = \frac{x^2}{(1-s^2)^2} \otimes \begin{pmatrix} 1+2sr+s^2 & r+2s+rs^2 \\ r+2s+rs^2 & 1+2sr+s^2 \end{pmatrix} \qquad (8.14)$$

We can use this result to derive the effects of sibling interaction on the variance and covariance due to a variety of sources of individual differences. For example, when considering:

1. additive genetic influences, $x^2 = a^2$ and $r = \alpha$, where α is 1.0 for MZ twins and 0.5 for DZ twins;

2. shared environment influences, $x^2 = c^2$ and $r = 1$;

3. non-shared environmental influences, $x^2 = e^2$ and $r = 0$;

4. genetic dominance, $x^2 = d^2$ and $r = \delta$, where $\delta = 1.0$ for MZ twins and $\delta = 0.25$ for DZ twins.

These results are summarized in Table 8.3.

Table 8.3: Effects of sibling interaction(s) on variance and covariance components between pairs of relatives.

Source	Variance	Covariance
Additive genetic	$\omega(1+2s\alpha+s^2)a^2$	$\omega(\alpha+2s+\alpha s^2)a^2$
Dominance genetic	$\omega(1+2s\delta+s^2)d^2$	$\omega(\delta+2s+\delta s^2)d^2$
Shared environment	$\omega(1+2s+s^2)c^2$	$\omega(1+2s+s^2)c^2$
Non-shared environment	$\omega(1+s^2)e^2$	$\omega 2se^2$

ω represents the scalar $\frac{1}{(1-s^2)^2}$ obtained from equation 8.14.

8.4.2 Numerical Illustration

To illustrate these effects numerically, let us consider a simplified situation in which $a^2 = .5, d^2 = 0, c^2 = 0, e^2 = .5$ in the absence of social interaction (i.e., $s = 0$); in the presence of strong cooperation, $s = .5$; and in the presence of strong competition, $s = -.5$. Table 8.4 gives the numerical values for MZ and DZ twins and unrelated pairs of individuals reared together (e.g., adoptive siblings). In terms of correlations, phenotypic cooperation mimics the effects of shared environment while phenotypic competition may mimic the effects of non-additive genetic variance. However, the effects can be distinguished because social interactions result in different total phenotypic variances for differently related pairs of individuals. All of the other kinds of models we have considered predict that the population variance of individuals is not affected by the presence or absence of relatives. However, cooperative interactions increase the variance of more closely related individuals the most, while competitive interactions increase them the least and under some circumstances may decrease them. Thus, in twin data, cooperation is distinguished from shared environmental effects because cooperation results in greater total phenotypic variance in MZ than in DZ twins. Competition is distinguished from non-additive genetic effects because it results in lower total phenotypic variance in MZ

Table 8.4: Effects of strong sibling interaction on the variance and covariance between MZ, DZ, and unrelated individuals reared together. The interaction parameter s takes the values 0, .5, and $-.5$ for no sibling interaction, cooperation, and competition, respectively.

	MZ twins			DZ twins			Unrelated		
Interaction	Var	Cov	r	Var	Cov	r	Var	Cov	r
None	1.00	.50	.50	1.00	.25	.25	1.00	.00	.00
Cooperation	3.11	2.89	.93	2.67	2.33	.88	2.22	1.78	.80
Competition	1.33	.44	.33	1.78	-.67	-.38	2.22	-1.78	-.80

than in DZ twins. This is the bottom line: social interactions cause the variance of a phenotype to depend on the degree of relationship of the social actors.

There are three observations we should make about this result. First, a test of the contrary assumption, i.e., that the total observed variance is independent of zygosity in twins, was set out by Jinks and Fulker (1970) as a preliminary requirement of their analyses and, as has been noted, is implicitly provided whenever we fit models without social interactions to covariance matrices. For I.Q., educational attainment, psychometric assessments of personality, social attitudes, body mass index, heart rate reactivity, and so on, the behavior genetic literature is replete with evidence for the *absence* of the effects of social interaction. Second, analyses of family correlations (rather than variances and covariances) effectively standardize the variances of different groups of individuals and throw away the very information we need to distinguish social interactions from other influences. Third, if we are working with categorical data and adopting a threshold model (see Chapter 2), we can make predictions about the standardized thresholds in different groups. Higher quantitative variances lead to smaller (i.e., less deviant) thresholds and therefore higher prevalence for the extreme categories. Thus, for example, if abstinence vs. drinking status is influenced by sibling cooperation on a latent underlying phenotype, and abstinence has a frequency of 10% in DZ twins, we should expect a higher frequency of abstinence in MZ twins. These models are relatively simple to implement in Mx (Neale, 1997).

Chapter 9

Sex-limitation and G $\times$ E Interaction

9.1 Introduction

As described in Chapter 6, the basic univariate ACE model allows us to estimate genetic and environmental components of phenotypic variance from like-sex MZ and DZ twin data. When data are available from both male and female twin pairs, an investigator may be interested in asking whether the variance profile of a trait is similar across the sexes or whether the magnitude of genetic and environmental influences are sex-dependent. To address this issue, the ACE model may be fitted independently to data from male and female twins, and the parameter estimates compared by inspection. This approach, however, has three severe limitations: (1) it does not test whether the heterogeneity observed across the sexes is significant; (2) it does not attempt to explain the sex differences by fitting a particular sex-limitation model; and (3) it discards potentially useful information by excluding dizygotic opposite-sex twin pairs from the analysis. In the first part of this chapter (Section 9.2), we outline three models for exploring sex differences in genetic and environmental effects (i.e., models for sex-limitation) and provide an example of each by analyzing twin data on body mass index (BMI) (Section 9.2.4).

Just as the magnitude of genetic and environmental influences may differ according to sex, they also may vary under disparate environmental conditions. If differences in genetic variance across environmental exposure groups result in differential heritability estimates for these groups, a genotype $\times$ environment interaction is said to exist. Historically, genotype $\times$ environment (G $\times$ E) interactions have been noted in plant and animal species (Mather and Jinks, 1982); however, there is increasing evidence that they play an important role in human variability as well (Heath and Martin, 1986; Heath *et al.*, 1989b). A simple method for detecting

G × E interactions is to estimate components of phenotypic variance conditional on environmental exposure (Eaves, 1982). In the second part of this chapter (Section 9.3), we illustrate how this method may be employed by suitably modifying models for sex-limitation. We then apply the models to depression scores of female twins and estimate components of variance conditional on a putative buffering environment, marital status (Section 9.3.2).

9.2 Sex-limitation Models

9.2.1 General Model for Sex-limitation

The general sex-limitation model allows us to (1) estimate the *magnitude* of genetic and environmental effects on male and female phenotypes and (2) determine whether or not it is the *same set* of genes or shared environmental experiences that influence a trait in males and females. Although the first task may be achieved with data from like-sex twin pairs only, the second task requires that we have data from opposite-sex pairs (Eaves *et al.*, 1978). Thus, the Mx script we describe will include model specifications for all 5 zygosity groups (MZ–male, MZ–female, DZ–male, DZ–female, DZ–opposite-sex).

To introduce the general sex-limitation model, we consider a path diagram for opposite-sex pairs, shown in Figure 9.1. Included among the ultimate variables in the diagram are female and male additive genetic (A_f and A_m), dominant genetic (D_f and D_m), and unique environmental (E_f and E_m) effects, which influence the latent phenotype of the female (P_f) or male (P_m) twin. The additive and dominant genetic effects are correlated within twin pairs ($\alpha = 0.50$ for additive effects, and $\beta = 0.25$ for dominant effects) as they are for DZ like-sex pairs in the simple univariate ACE model. This correlational structure implies that the genetic effects represent *common* sets of genes which influence the trait in both males and females; however, since a_m and a_f or d_m and d_f are not constrained to be equal, the common effects need not have the same *magnitude* across the sexes. Figure 9.1 also includes ultimate variables for the male (or female) member of the opposite-sex twin pair (A'_m and D'_m) which do not correlate with genetic effects on the female phenotype. For this reason, we refer to A'_m and D'_m as sex-specific variables. Significant estimates of their effects indicate that the set of genes which influences a trait in males is not identical to that which influences a trait in females. To determine the extent of male-female genetic similarity, one can calculate the male-female genetic correlation (r_g). As usual (see Chapter 2) the correlation is computed as the covariance of the two variables divided by the product of their respective standard deviations. Thus, for additive genetic effects we have

$$r_g = \frac{a_m a_f}{\sqrt{a_f^2(a_m^2 + a_m'^2)}}$$

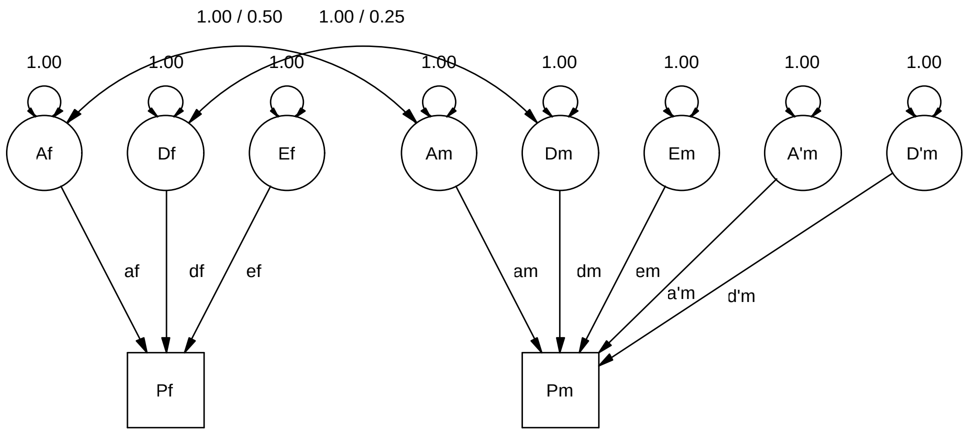


Figure 9.1: The general genotype × sex interaction model for twin data. Path diagram is shown for DZ opposite-sex twin pairs. $\alpha = 0.5$ and $\beta = 0.25$.

Alternatively, a similar estimate may be obtained for dominant genetic effects. However, the information available from twin pairs reared together precludes the estimation of *both* sex-specific parameters, a'_m and d'_m and, consequently, both additive and dominance genetic correlations. Instead, models including A'_m *or* D'_m may be fit to the data, and their fits compared using appropriate goodness-of-fit indices, such as Akaike's Information Criteria (AIC; Akaike, 1987; see Section **??**). This criterion may be used to compare the fit of an ACE model to the fit of an ADE model. AIC is one member of a class of indices that reflect both the goodness of fit of a model and its parsimony, or ability to account for the observed data with few parameters.

To generalize the model specified in Figure 9.1 to other zygosity groups, the parameters associated with the female phenotype are equated to similar effects on the phenotypes of female same-sex MZ and DZ twin pairs. In the same manner, all parameters associated with the male phenotype (reflecting effects which are common to both sexes as well as those specific to males) are equated to effects on both members of male same-sex MZ and DZ pairs. As a result, the model predicts that variances will be equal for all female twins, and all male twins, regardless of zygosity group or twin status (i.e., twin 1 vs. twin 2). The model does not necessarily predict equality of variances *across* the sexes.

9.2.2 General Sex-limitation Model Mx Script

The full Mx specification for the general sex-limitation model is provided in Appendix **??**. In theory, the same approach that was used to specify the simple univariate ACE model (Chapter 6) in Mx could be used for the general sex-limitation model. That is, genetic and environmental parameters can be specified in calculation groups and the matrices can be included in the data groups to specify the expected covariance matrices. The female and male parameters are declared in separate groups which simplifies the data groups. The only differences between the male and female data groups are the details about the data and the number of the group from which matrices are being imported. Note that for the general sex limitation model, one extra matrix (`N`) is declared in the male group to account for the male-specific additive genetic effects.

While the specification of the same-sex groups is a straightforward extension of the univariate model, the opposite-sex group requires some special attention. First, the matrices for both male and female *variance components* are read in to formulate the expected variance for females (twin 1) and males (twin 2). Second, the expected covariance between male and female twins can be specified by multiplying the male and female *path coefficient matrices*. Although not a problem in the univariate analysis, note that the male-female expected covariance matrix is not necessarily symmetric.

Without boundary constraints on the parameters, this specification may lead to *negative* parameter estimates for one sex, especially when the DZ opposite-sex correlation is low, as compared to DZ like-sex correlations. Such negative parameter estimates result in a negative genetic (or common environmental) covariation between the sexes. Although a negative covariation is plausible, it seems quite unlikely that the *same* genes or common environmental influences would have *opposite* effects across the sexes. With the availability of linear and non-linear constraints in Mx, we can parameterize the general sex-limitation model so that the male-female covariance components are constrained to be non-negative by using a boundary statement:

```
Bound .000 10 X 1 1 1 Z 1 1 1 W 1 1 1
Bound .000 10 X 2 1 1 Z 2 1 1 W 2 1 1 N 2 1 1
```

where `.000` is the lower boundary, `10` is the upper boundary followed by matrix elements.

In this example, we estimate sex-specific additive genetic effects (and fix the sex-specific dominance effects to zero). The data are log-transformed indices of body mass index (BMI) obtained from twins belonging to the Virginia and American Association of Retired Persons twin registries. A detailed description of these data will be provided in section 9.2.4, in the discussion of the model-fitting results.

9.2.3 Restricted Models for Sex-limitation

In this section, we describe two restricted models for sex-limitation. The first we refer to as the *common effects sex-limitation model*, and the second, the *scalar sex-limitation model*. Both are sub-models of the general sex-limitation model and therefore can be compared to the more general model using likelihood-ratio χ^2 difference tests.

Common Effects Sex-limitation Model

The common effects sex-limitation model is simply one in which the sex-specific pathways in Figures 9.1 (a'_m or d'_m) are fixed to zero or the additve or dominant genetic correlation between males and females is fixed to .50. As a result, only the genetic effects which are *common* to both males and females account for phenotype variance and covariance. Although the genes may be the same, the magnitude of their effect is still allowed to differ across the sexes. This restricted model may be compared to the general sex-limitation model using a χ^2 difference test with a single degree of freedom.

Information to discern between the general sex-limitation model and the common effects model comes from the covariance of DZ opposite-sex twin pairs. Specifically, if this covariance is significantly less than that predicted from genetic effects which are *common* to both sexes (i.e., less than $[(a_m \times a_f) + (d_m \times d_f)]$), then there is evidence for sex-specific effects. Otherwise, the restricted model without these effects should not fit significantly worse than the general model. Mere inspection of the *correlations* from DZ like-sex and opposite-sex pairs may alert one to the fact that sex-specific effects are playing a role in trait variation, if it is found that the opposite sex-correlation is markedly less than the like-sex DZ correlations.

Scalar Effects Sex-limitation Model

The scalar sex-limitation model is a sub-model of both the general model and the common effects model. In the scalar model, not only are the sex-specific effects removed, but the variance components for females are all constrained to be equal to a *scalar* multiple (k^2) of the male variance components, such that $a_f^2 = k^2 a_m^2$, $d_f^2 = k^2 d_m^2$, and $e_f^2 = k^2 e_m^2$. As a result, the standardized variance components (e.g., heritability estimates) are equal across sexes, even though the unstandardized components differ.

Figure 9.2 shows a path diagram for DZ opposite-sex under the scalar sex-limitation model, and Appendix **??** provides the Mx specification. Unlike the model in Figure **??**, the scalar model does not include separate parameters for genetic and environmental effects on males and females — instead, these effects are equated across the sexes. Because of this equality, negative estimates of male-female genetic covariance cannot result. To introduce a scaling factor for the male (or female) variance components, we can pre and postmultiply the expected variances by a scalar.

The full scalar sex-limitation model may be compared to the full common effects model using a χ^2 difference test with 2 degrees of freedom. Similarly, the scalar sex-limitation model may be compared to the model with no sex differences (that is, one which fixes k to 1.0) using a χ^2 difference test with a single degree of freedom.

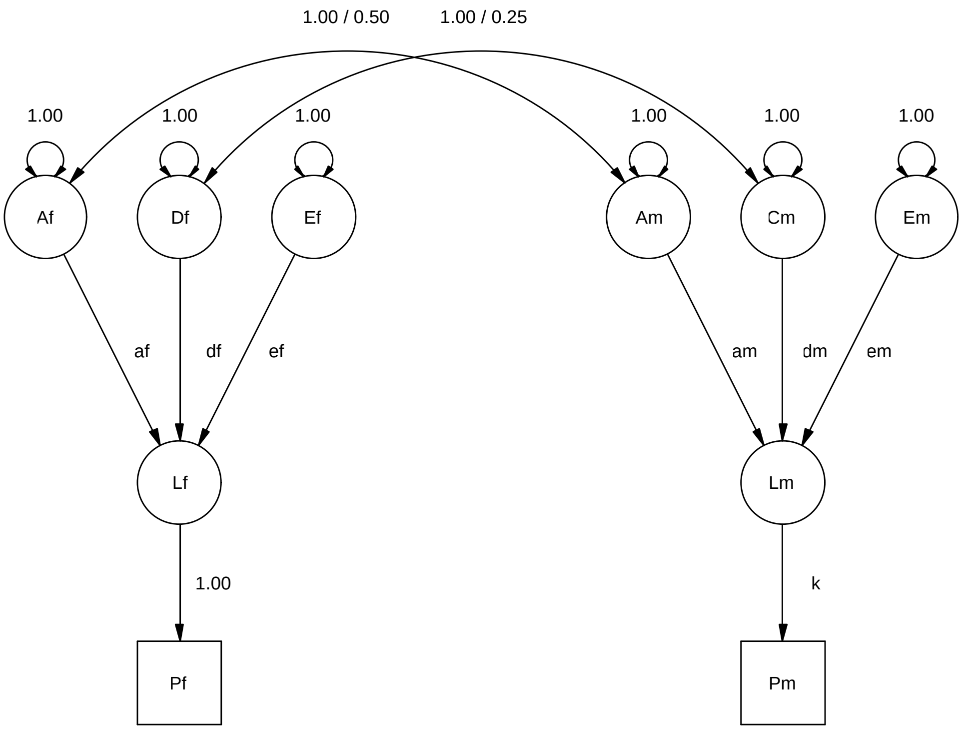


Figure 9.2: The scalar genotype × sex interaction model for twin data. Path diagram is shown for DZ opposite-sex twin pairs. The $\alpha = 0.5$ and $\beta = 0.25$.

The restricted sex-limitation models described in this section are not an exhaustive list of the sub-models of the general sex-limitation model. Within either of these restricted models (as within the general model), one can test hypotheses regarding the significance of genetic or environmental effects. Also, within the common effects sex-limitation model, one may test whether *specific* components of variance are equal across the sexes (e.g., a_m may be equated to a_f, or e_m to e_f). Again, sub-models may be compared to more saturated ones through χ^2 difference tests, or to models with the same number of parameters with Akaike's Information Criteria.

9.2.4 Application to Body Mass Index

In this section, we apply sex-limitation models to data on body mass index collected from twins in the Virginia Twin Registry and twins ascertained through the American Association of Retired Persons (AARP). Details of the membership of these two twin cohorts are provided in Eaves *et al.* (1991), in their analysis of BMI in extended twin-family pedigrees. In brief, the Virginia twins are members of a population based registry comprised of 7,458 individuals (Corey *et al.*, 1986), while the AARP twins are members of a volunteer registry of 12,118 individuals responding to advertisements in publications of the AARP. The Virginia twins' mean age is 39.7 years (SD = 14.3), compared to 54.5 years (SD = 16.8) for the AARP twins. Between 1985 and 1987, Health and Lifestyle questionnaires were mailed to twins from both of these cohorts. Among the items on the questionnaire were those pertaining to physical similarity and confusion in recognition by others (used to diagnose zygosity) and those asking about current height and weight (used to compute body mass index). Questionnaires with no missing values for any of these items were returned by 5,465 Virginia and AARP twin pairs.

From height and weight data, body mass index (BMI) was calculated for the twins, using the formula:

$$\text{BMI} = wt(kg)/ht(m)^2$$

The natural logarithm of BMI was then taken to normalize the data. Before calculating covariance matrices of log BMI, the data from the two cohorts were combined, and the effects of age, age squared, sample (AARP vs. Virginia), sex, and their interactions were removed. The resulting covariance matrices are provided in the Mx scripts in Appendices **??** and **??**, while the correlations and sample sizes appear in Table 9.1 below.

Table 9.1: Sample sizes and correlations for BMI data in Virginia and AARP twins.

Zygosity Group	N	r
MZF	1802	0.744
DZF	1142	0.352
MZM	750	0.700
DZM	553	0.309
DZO	1341	0.251

We note that both like-sex MZ correlations are greater than twice the respective DZ correlations; thus, models with dominant genetic effects, rather than common environmental effects, were fit to the data.

In Table 9.2, we provide selected results from fitting the following models: general sex-limitation (I); common effects sex-limitation (II-IV); and scalar sex-limitation (V). We first note that the general sex-limitation model provides a good fit to the data, with $p = 0.32$. The estimate of a'_m under this model is fairly small, and when set to zero in model II, found to be non-significant ($\chi^2_1 = 2.54$, $p > 0.05$). Thus, there is no evidence for sex-specific additive genetic effects, and the common effects sex-limitation model (model II) is favored over the general model. As an exercise, the reader may wish to verify that the same conclusion is reached if the general sex-limitation model with sex-specific dominant genetic effects is compared to the common effects model with d'_m removed.

Note that under model II the dominant genetic parameter for females is quite small; thus, when this parameter is fixed to zero in model III, there is not a significant worsening of fit, and model III becomes the most favored model. In model IV, we consider whether the dominant genetic effect for males can also be fixed to zero. The goodness-of-fit statistics indicate that this model fits the data poorly ($p < 0.01$) and provides a significantly worse fit than model III ($\chi^2_1 = 26.73$, $p < 0.01$). Model IV is therefore rejected and model III remains the favored one.

Finally, we consider the scalar sex-limitation model. Since there is evidence for dominant genetic effects in males and not in females, it seems unlikely that this model, which constrains the variance components of females to be scalar multiples of the male variance components, will provide a good fit to the data, unless the additive genetic variance in females is also much smaller than the male additive genetic variance. The model-fitting results support this contention: the model provides a marginal fit to the data ($p = 0.05$), and is significantly worse than model II ($\chi^2_2 = 7.82$, $p < 0.05$). We thus conclude from Table 9.2 that III is the best fitting model. This conclusion would also be reached if AIC was used to assess goodness-of-fit.

Table 9.2: Parameter estimates from fitting genotype × sex interaction models to BMI.

	MODEL				
Parameter	I	II	III	IV	V
a_f	0.449	0.454	0.454	0.454	0.346
d_f	0.172	0.000	–	–	0.288
e_f	0.264	0.265	0.265	0.267	0.267
a_m	0.210	0.240	0.240	0.342	–
d_m	0.184	0.245	0.245	–	–
e_m	0.213	0.213	0.213	0.220	–
a'_m	0.198	–	–	–	–
k	–	–	–	–	0.778
χ^2	9.26	11.80	11.80	38.53	19.62
$d.f.$	8	9	10	11	11
p	0.32	0.23	0.30	0.00	0.05
AIC	-6.74	-6.20	-8.20	16.53	-2.38

Using the parameter estimates under model III, the expected variance of log

BMI (residuals) in males and females can be calculated. A little arithmetic reveals that the phenotypic variance of males is markedly lower than that of females (0.17 vs. 0.28). Inspection of the parameter estimates indicates that the sex difference in phenotypic variance is due to increased *genetic* and *environmental* variance in females. However, the increase in genetic variance in females is proportionately greater than the increase in environmental variance, and this difference results in a somewhat larger broad sense (i.e., $a^2 + d^2$) heritability estimate for females (75%) than for males (69%).

The detection of sex-differences in environmental and genetic effects on BMI leads to questions regarding the nature of these differences. Speculation might suggest that the somewhat lower male heritability estimate may be due to the fact that males are less accurate in their self-report of height and weight than are females. With additional information, such as test-retest data, this hypothesis could be rigorously tested. The sex-dependency of genetic dominance is similarly curious. It may be that the common environment in females exerts a greater influence on BMI than in males, and, consequently, masks a genetic dominance effect. Alternatively, the genetic architecture may indeed be different across the sexes, resulting from sex differences in selective pressures during human evolution. Again, additional data, such as that from reared together adopted siblings, could be used to explore these alternative hypotheses.

One sex-limitation model that we have not considered, but which is biologically reasonable, is that the across-sex correlation between additive genetic effects is the same as the across-sex correlation between the dominance genetic effects[1]. Fitting a model of this type involves a non-linear constraint which can easily be specified in Mx.

9.3 Genotype × Environment Interaction

As stated in the introduction of this chapter, genotype × environment (G × E) interactions can be detected by estimating components of phenotypic variance *conditional* on environmental exposures. To do so, MZ and DZ covariance matrices are computed for twins concordant for exposure, concordant for non-exposure , and discordant for exposure, and structural equation models are fitted to the resulting six zygosity groups. The Mx specifications for alternative G × E interaction models are quite similar to those used in a sex-limitation analysis; however, there are important differences between the two. In a G × E interaction analysis, the presence of a sixth group provides the information for an additional parameter to be estimated. Further, the nature of alternative hypotheses used to explain heterogeneity across groups differs from those invoked in a sex-limitation analysis.

[1] The reasoning goes like this: (e.g.) males have a elevated level of a chemical that prevents *any* gene expression from certain loci, at random with respect to the phenotype under study. Thus, both additive and dominant genetic effects would be reduced in males vs females, and hence the same genetic correlation between the sexes would apply to both.

In section 9.3.1 we detail these differences, and in section 9.3.2 we illustrate the method with an application to data on marital status and depression.

9.3.1 Models for G $\times$ E Interactions

The models described in this section are appropriate for analyzing G $\times$ E interaction when genes and environment are acting independently. However, if there is genotype – environment correlation, then more sophisticated statistical procedures are necessary for the analysis. One way of detecting a G – E correlation is to compute the cross-correlations between one twin's environment and the trait of interest in the cotwin (Heath *et al.*, 1989b). If the cross-correlation is not significant, there is no evidence for a G – E correlation, and the G $\times$ E analysis may proceed using the methods described below.

General G $\times$ E Interaction Model

First we consider the general G $\times$ E interaction model, similar to the general sex-limitation model discussed in section 9.2.3. This model not only allows the magnitude of genetic and environmental effects to vary across environmental conditions, but also, by using information from twin pairs discordant for environmental exposure, enables us to determine whether it is the same set of genes or environmental features that are expressed in the two environments. Just as we used twins who were discordant for sex (i.e., DZO pairs) to illustrate the sex-limitation model, we use twins discordant for environmental exposure to portray the general G $\times$ E interaction model. Before modeling genetic and environmental effects on these individuals, one must order the twins so that the first of the pair has not been exposed to the putative modifying environment, while the second has (or *vice versa,* as long as the order is consistent across families and across groups). The path model for the discordant DZ pairs is then identical to that used for the dizygotic opposite-sex pairs in the sex-limitation model; for the discordant MZ pairs, it differs only from the DZ model in the correlation structure of the ultimate genetic variables (see Figure 9.3).

Among the ultimate variables in Figure 9.3 are genetic effects that are correlated between the unexposed and exposed twins and those that influence only the latter (i.e., environment-specific effects). For the concordant unexposed and concordant exposed MZ and DZ pairs, path models are comparable to those used for female-female and male-male MZ and DZ pairs in the sex-limitation analysis, with environment-specific effects (instead of sex-specific effects) operating on the exposed twins (instead of the male twins). As a result, the model predicts equal variances *within* an exposure class, across zygosity groups.

In specifying the general G $\times$ E interaction model in Mx, one must again use boundary constraints, in order to avoid negative covariance estimates for the pairs discordant for exposure (Appendix **??**).

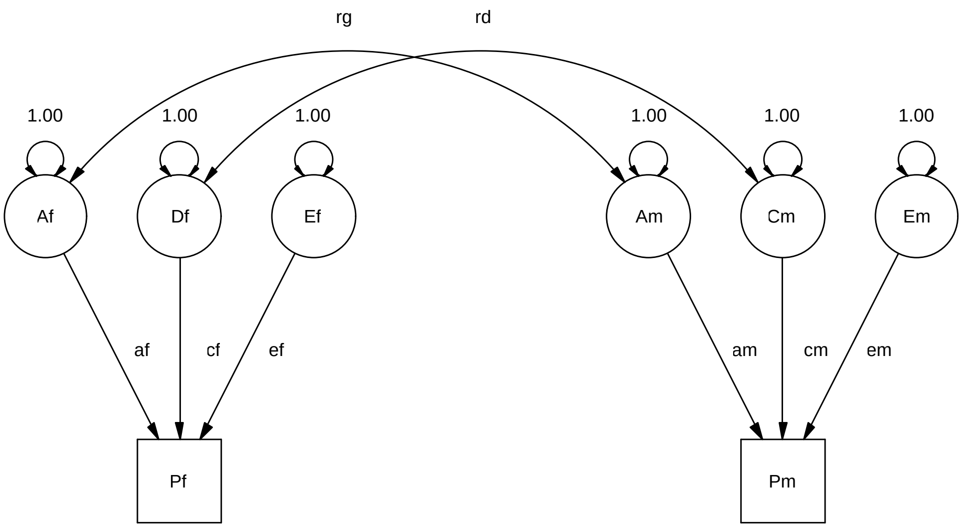


Figure 9.3: The general genotype × environment interaction model for twin data. Path diagram is for MZ and DZ twins discordant for environmental exposure. For MZ pairs, $\alpha = 1.0$ and $\beta = 1.0$; for DZ pairs, $\alpha = 0.5$ and $\beta = 0.25$. The subscripts u and e identify variables and parameters and unexposed and exposed twins, respectively.

Unlike the general sex-limitation analysis, there is enough information in a G × E analysis to estimate two environment-specific effects. Thus, the magnitude of environment-specific additive and dominant genetic *or* additive genetic and common environmental effects can be determined. It still is not possible to simultaneously estimate the magnitude of common environmental and dominant genetic effects.

Common Effects G × E Interaction Model

A *common effects G × E model* can also be fitted to covariance matrices computed conditionally on environmental exposure by simply fixing the environment-specific effects of the general model to zero, and comparing the two using a χ^2 difference test. The information from pairs discordant for environmental exposure allows for this comparison.

A critical sub-model of the common effects G × E model is one which tests the hypothesis that exposure group heterogeneity is *solely due to heteroscedasticity*, or *group differences in random environmental variance*, rather than *group differences in genetic variance*. To fit this model, the genetic parameters are simply equated across groups, while allowing the random environmental effects to take on different values. If this model does not fit worse than the full common effects model, then there is evidence for heteroscedasticity.

A second sub-model of the common effects G $\times$ E interaction model is one which constrains the *environmental* parameters to be equal across exposure groups, while allowing the genetic variance components to differ. If this model is not significantly worse than the full common effects model, then there is evidence to suggest that the environmental interaction only involves a differential expression of genetic, but not environmental, influences.

Scalar Effects G $\times$ E Interaction Model

As with the scalar sex-limitation model, the *scalar G $\times$ E interaction model* equates genetic and environmental effects on exposed twins to be a scalar multiple of similar effects on twins who have not been exposed to a modifying environment. As a consequence, the heritability of a trait remains constant across exposure groups, and there is *no* evidence for a genotype $\times$ environment interaction. This situation may arise if there is a mean-variance relationship, and an increase in trait mean under a particular environmental condition is accompanied by an increase in phenotypic variation. When this is the case, the ratio of the genetic variance component and environmental variance component is expected to remain the same in different environments.

The Mx specification for the scalar G $\times$ E interaction model is identical to that used for the scalar sex-limitation model, except for the addition of MZ discordant pairs. The Mx script in Appendix **??** illustrates how these pairs may be included.

9.3.2 Application to Marital Status and Depression

In this section, we determine whether the heritability of self-report depression scores varies according to the marital status of female twins. Our hypothesis is that marriage, or a marriage-type relationship, serves as a buffer to decrease an individual's inherited liability to depression, consequently decreasing the heritability of the trait.

The data were collected from twins enrolled in the Australian National Health and Medical Research Council Twin register. In this sample, mailed questionnaires were sent to the 5,967 pairs of twins on the register between November 1980 and March 1982 (see also Chapter 10). Among the items on the questionnaire were those from the state depression scale of the Delusions-Symptoms States Inventory (DSSI; Bedford *et al.*, 1976) and a single item regarding marital status. The analyses performed here focus on the like-sex MZ and DZ female pairs who returned completed questionnaires. The ages of the respondents ranged from 18 to 88 years; however, due to possible differences in variance components across age cohorts, we have limited our analysis to those twins who were age 30 or less at the time of their response. There were 570 female MZ pairs in this young cohort, with mean age 23.77 years (SD=3.65); and 349 DZ pairs, with mean age 23.66 years (SD=3.93).

Using responses to the marital status item, pairs were subdivided into those who were concordant for being married (or living in a marriage type relationship); those who were concordant for being unmarried; and those who were discordant for marital status. In the discordant pairs, the data were reordered so that the first twin was always unmarried. Depression scores were derived by summing the 7 DSSI item scores, and then taking a log-transformation of the data [$x' = log_{10}(x+1)$] to reduce heteroscedasticity. Covariance matrices of depression scores were computed for the six zygosity groups after linear and quadratic effects of age were removed. The matrices are provided in the Mx scripts in Appendices **??** and **??**, while the correlations and sample sizes are shown in Table 9.3. We note (i) that in all cases, MZ correlations are greater than the corresponding DZ correlations; and (ii) that for concordant married and discordant pairs, the MZ:DZ ratio is greater than 2:1, suggesting the presence of genetic dominance.

Table 9.3: Sample sizes and correlations for depression data in Australian female twins.

Zygosity Group	N	r
MZ - Concordant single	254	0.409
DZ - Concordant single	155	0.221
MZ - Concordant married	177	0.382
DZ - Concordant married	107	0.098
MZ - Discordant	139	0.324
DZ - Discordant	87	0.059

Before proceeding with the G × E interaction analyses, we tested whether there was a G – E correlation involving marital status and depression. To do so, cross-correlations between twins' marital status and cotwins' depression score were computed. In all but one case (DZ twin 1's depression with cotwin's marital status; $r = -0.156$, $p < 0.01$), the correlations were not significant. This near absence of significant correlations implies that a genetic predisposition to depression does not lead to an increased probability of remaining single, and indicates that a G – E correlation need not be modeled.

Table 9.4 shows the results of fitting several models: general G×E (I); full common-effects G × E (II); three common-effects sub-models (III-V); scalar G x E (VI); and no G × E interaction (VII). Parameter estimates subscripted s and m refer respectively to single (unexposed) and married twins. Models including genetic dominance parameters, rather than common environmental effects, were fitted to the data. The reader may wish to show that the overall conclusions concerning

G × E interaction do not differ if shared environment parameters are substituted for genetic dominance.

Table 9.4: Parameter estimates from fitting genotype × marriage interaction models to depression scores.

	MODEL						
Parameter	I	II	III	IV	V	VI	VII
a_s	0.187	0.187	0.207	0.209	0.186	0.206	0.188
d_s	0.106	0.105	–	–	–	–	–
e_s	0.240	0.240	0.246	0.245	0.257	0.247	0.246
a_m	0.048	0.048	0.163	0.162	0.186	0.206	0.188
d_m	0.171	0.173	–	–	–	–	–
e_m	0.232	0.232	0.243	0.245	0.232	0.247	0.246
a'_m	0.008	–	–	–	–	–	–
k	–	–	–	–	–	0.916	–
χ^2	15.44	15.48	18.88	18.91	22.32	20.08	27.19
$d.f.$	11	12	14	15	15	15	16
p	0.16	0.22	0.17	0.22	0.10	0.17	0.04
AIC	-6.56	-9.52	-9.12	-11.09	-7.68	-9.92	-4.81

Model I is a general G × E model with environment-specific additive genetic effects. It provides a reasonable fit to the data ($p = 0.16$), with all parameters of moderate size, except a'_m. Under model II, the parameter a'_m is set to zero, and the fit is not significantly worse than model I ($\chi^2_1 = 0.04$, $p = 0.84$). Thus, there is no evidence for environment-specific additive genetic effects. As an exercise, the reader may verify that the same conclusion can be made for environment-specific dominant genetic effects.

Under model III, we test whether the dominance effects on single and married individuals are significant. A χ^2 difference of 3.40 ($p = 0.183$, 2 df.) between models III and II indicates that they are not. Consequently, model III, which excludes common dominance effects while retaining common additive genetic and specific environmental effects, is favored.

Models IV - VII are all sub-models of III: the first specifies no differences in environmental variance components across exposure groups; the second specifies no differences in genetic variance components across groups; the third constrains the genetic and environmental variance components of single twins to be scalar multiples of those of married twins; and the fourth specifies no genetic or environmental differences between the groups. When each of these is compared to model III using a χ^2 difference test, only model VII (specifying complete homogeneity across groups) is significantly worse than the fuller model ($\chi^2_2 = 8.28$, p = 0.004).

In order to select the best sub-models from IV, V and VI, Akaike's Information Criteria were used. These criteria indicate that model IV — which allows for group differences in genetic, but not environmental, effects — gives the most parsimonious explanation for the data. Under model IV, the heritability of depression is 42% for single, and 30% for married twins. This finding supports our hypothesis that marriage or marriage type relationships act as a buffer against the expression of inherited liability to depression.

Chapter 10

Multivariate Analysis

10.1 Introduction

Until this point we have been concerned primarily with methods for analyzing single variables obtained from twin pairs; that is, with estimation of the relevant sources of genetic and environmental *variation* in each variable separately. Most studies, however, are not designed to consider single variables, but are trying to understand what factors make sets of variables correlate, or *co-vary*, to a greater or lesser extent. Just as we can partition variation into its genetic and environmental components, so too we can try to determine how far the covariation between multiple measures is due to genetic and environmental factors. This partitioning of covariation is one of the first tasks of multivariate genetic analysis, and it is one for which the classical twin study, with its simple and regular structure, is especially well-suited.

In Chapter 1 we described three of the main issues in the genetic analysis of multiple variables. These issues include

1. contribution of genes and environment to the correlation between variables

2. direction of causation between observed variables

3. genetic and environmental contributions to developmental change.

Each of these questions presumes either a different data collection strategy or a different model or both; for example, analysis of measurements of correlated traits taken at the same time (question 1) requires somewhat different methods than assessments of the same trait taken longitudinally (question 3). However, all of the multivariate issues share the requirement of multiple measurements from the same subjects. In this chapter we direct our attention to the first issue: genetic and environmental contributions to observed correlations among variables. We describe twin methods for the other two questions in Chapters **??** – **??**.

The treatment of multivariate models presented here is intended to be introductory. There are many specific topics within the broad domain of multivariate genetic analysis, some of which we address in subsequent chapters. Here we exclude treatment of observed and latent variable means and analysis of singleton twins.

10.2 Phenotypic Factor Analysis

Factor analysis is one of the most widely used multivariate methods. The general idea is to explain variation within and covariation between a large number of observed variables with a smaller number of latent factors. Here we give a brief outline of the method — those seeking more thorough treatments are referred to e.g., Gorsuch (1983), Harman (1976), Lawley and Maxwell (1971). Typically the free parameters of primary interest in factor models are the *factor loadings* and *factor correlations*. Factor loadings indicate the degree of relationship between a latent factor and an observed variable, while factor correlations represent the relationships between the hypothesized latent factors. An observed variable that is a good indicator of a latent factor is said to "load highly" on that factor. For example, in intelligence research, where factor theory has its origins (Spearman, 1904), it may be noted that a vocabulary test loads highly on a hypothesized (latent) verbal ability factor, but loads to a much lesser extent on a latent spatial ability factor; i.e., the vocabulary test relates strongly to verbal ability, but less so to spatial ability. Normally a factor loading is identical to a path coefficient of the type described in Chapter 5.

In this section we describe factor analytic models and present some illustrative applications to observed measurements without reference to genetic and environmental causality. We turn to genetic factor models in Section 10.3.

10.2.1 Exploratory and Confirmatory Factor Models

There are two general classes of factor models: exploratory and confirmatory. In exploratory factor analysis one does not postulate an *a priori* factor structure; that is, the number of latent factors, correlations among them, and the *factor loading pattern* (the pattern of relative weights of the observed variables on the latent factors) is calculated from the data in some manner which maximizes the amount of variance/covariance explained by the latent factors. More formally, in exploratory factor analysis:

1. There are no hypotheses about factor loadings (all variables load on all factors, and factor loadings cannot be constrained to be equal to other loadings)

2. There are no hypotheses about interfactor correlations (either all correlations are zero — orthogonal factors, or all may correlate — oblique factors)

3. Only one group is analyzed

4. Unique factors (those that relate only to one variable) are uncorrelated,

5. All observed variables need to have specific variances.

These models often are fitted using a statistical package such as SPSS or SAS, in which one may *explore* the relationships among observed variables in a latent variable framework.

In contrast, confirmatory factor analysis requires one to formulate a hypothesis about the number of latent factors, the relationships between the observed and latent factors (the factor pattern), and the correlations among the factors. Thus, a possible model of the data is formulated in advance as a *factor structure,* and the factor loadings and correlations are estimated from the data[1]. As usual, this model-fitting process allows one to *test* the ability of the hypothesized factor structure to account for the observed covariances by examining the overall fit of the model. Typically the model involves certain constraints, such as equalities among certain factor loadings or equalities of some of the factor correlations. If the model fails then we may relax certain constraints or add more factors, test for significant improvement in fit using the chi-squared difference test, and examine the overall goodness of fit to see if the new model adequately accounts for the observed covariation. Likewise, some or all of the correlations between latent factors may be set to zero or estimated. Then we can test if these constraints are consistent with the data. Confirmatory factor models are the type we are concerned with using Mx.

10.2.2 Building a Phenotypic Factor Model Mx Script

The factor model may be written as

$$Y_{ij} = b_i X_j + E_{ij}$$

with

$$\begin{aligned} i &= 1, \cdots, p \text{ (variables)} \\ j &= 1, \cdots, n \text{ (subjects)} \end{aligned}$$

and where the measured variables Y are a function of a subject's value on the underlying factor X (henceforth the j subscript indicating subjects in Y will be omitted). These subject values are called *factor scores.* Although the use of factor scores is always implicit in the application of factor analysis, they cannot be determined precisely but must be estimated, since the number of common and unique factors always exceeds the number of observed variables. In addition, there is a specific part (E) to each variable. The b's are the p–variate factor loadings of measured variables on the latent factors. To estimate these loadings we do not need

[1] In exploratory factor analysis the term "factor structure" is used to describe the correlations between variables and factors, but in confirmatory analysis, as described here, the term often describes the characteristics of a hypothesized factor model.

to know the individual factor scores, as the expectation for the $p \times p$ covariance matrix ($\Sigma_{Y,Y}$) consists only of a $p \times m$ matrix of factor loadings (**B**) **(**m **equals the number of latent factors), a** $m \times m$ **correlation matrix of factor scores (P), and a** $p \times p$ **diagonal matrix of specific variances (E) :**

$$\Sigma_{Y,Y} = \mathbf{BPB}' + \mathbf{E}. \tag{10.1}$$

In problems with uncorrelated latent factors, P is an identity matrix, so equation 10.1 reduces to

$$\Sigma_{Y,Y} = \mathbf{BB}' + \mathbf{E}. \tag{10.2}$$

Thus, the parameters in the model consist of factor loadings and specific variances (sometimes also referred to as error variances).

10.2.3 Fitting a Phenotypic Factor Model

Martin *et al.* **(1985) obtained data on arithmetic computation from male and female twins who were measured once before and three times after drinking a standard dose of alcohol. To illustrate the use of a confirmatory factor analysis model in Mx, we analyze data from MZ females (first born twin only). The observed variances and correlations are shown in Table 10.1. The confirmatory model is one in which a single latent factor is hypothesized to account for all the covariances among the four variables. The Mx script below shows the model specifications and the** 4×4 **input matrix.**

Table 10.1: Observed correlations (with variances on the diagonal) for arithmetic computation variables from female MZ twins before (time 0) and after (times 1 – 3) standard doses of alcohol.

	Time 0	Time 1	Time 2	Time 3
Time 0	259.66			
Time 1	.81	259.94		
Time 2	.83	.87	245.24	
Time 3	.87	.87	.90	249.30

Mx Script for Phenotypic Factor Analysis of Four Variables

```
Single factor phenotypic model: 4 arithmetic computation variables
 Data NGroups=1 NInput_vars=4 NObservations=42
 CMatrix
  259.664
  209.325 259.939
  209.532 220.755 245.235
  221.610 221.491 221.317 249.298
 Labels Time1 Time2 Time3 Time4
 Begin Matrices;
  B Full 1 4 Free
  P Symm 1 1
  E Diag 4 4 Free
 End Matrices;
  Value 1 P 1 1
  Start 9 All
 Covariances B*P*B'+E;
 Option RSiduals
End
```

The parameters in the group type statement indicate that we have only `NGroup=1` group (consisting of `NObservations=42` subjects) and there are `NInput_vars=4` input variables. The loadings of the four variables on the single common factor are estimated in matrix B and their specific variances are estimated on the diagonal of matrix E. In this phenotypic factor model, we have sufficient information to estimate factor loadings and specific variances for the four variables, but we cannot simultaneously estimate the variance of the common factor because the model would then be underidentified. We therefore fix the variance of the latent factor to an arbitrary non-zero constant, which we choose to be unity in order to keep the factor loadings and specific variances in the original scale of measurement (`Value 1 P 1 1`).

The Mx output (after editing) from this common factor model is shown below. The `PARAMETER SPECIFICATIONS` section illustrates the assignment of parameter numbers to matrices declared `Free` in the matrices section. Consecutive parameter numbers are given to free elements in matrices in the order in which they appear. It is always advisable to check the parameter specifications for the correct assignment of free and constrained parameters. The output depicts the single common factor structure of the model: there are free factor loadings for each of the four variables on the common factor, and specific variance parameters for each of the observed variables. Thus, the model has a total of 8 parameters to explain the $4(4+1)/2 = 10$ free statistics.

The results - from the `MX PARAMETER ESTIMATES` section of the Mx output - are summarized in Table 10.2.3. The chi-squared goodness-of-fit value of 1.46 for 2 degrees of freedom suggests that this single factor model adequately explains the observed covariances (p = .483). This also may be seen by comparing the elements of the fitted covariance matrix and the observed covariance matrix, which are seen to be very similar. The fitted covariance matrix is printed by Mx when the `RSiduals` option is added. The fitted covariance matrix is calculated by Mx using expression 10.2 with the final estimated parameter values.

Mx Output from Phenotypic Factor Model

```
----------------------------------------------------------
PARAMETER SPECIFICATIONS
MATRIX B
                  1
   TIME1          1
   TIME2          2
   TIME3          3
   TIME4          4

MATRIX E
          TIME1       TIME2       TIME3       TIME4
   1          5           6           7           8
```

Table 10.2: Parameter estimates and expected covariance matrix from the phenotypic factor model

	B	E		$Time1$	$Time2$	$Time3$	$Time4$
Time 1	**14.431**	**51.422**	**Time 1**	**259.670**			
Time 2	**14.745**	**42.509**	**Time 2**	**212.784**	**259.927**		
Time 3	**14.699**	**29.174**	**Time 3**	**212.115**	**216.736**	**245.229**	
Time 4	**15.119**	**20.709**	**Time 4**	**218.181**	**222.933**	**222.233**	**249.297**
$\chi^2 = .46$, **2 df, p=.483**							

10.3 Simple Genetic Factor Models

The factor analytic approach outlined above can be readily applied to multivariate genetic problems. This was first suggested by Martin and Eaves (1977) for the analysis of twin data (although in their original publication they use matrices of mean squares and cross-products between and within twin pairs). As in the phenotypic example above, a single common factor is proposed to account for correlations among the variables, but now one such factor is hypothesized for each of the components of variation, genetic, shared environmental, and non-shared environmental. Data from genetically related individuals are used to estimate loadings of variables on common genetic and environmental factors, so that variances and covariances may be explained in terms of these factors.

10.3.1 Multivariate Genetic Factor Model

Using genetic notation, the genetic factor model can be represented as

$$P_{ij} = a_i A_j + c_i C_j + e_i E_j + U_{ij}$$

with

$$\begin{aligned} i &= 1, \cdots, p \text{ (variables)} \\ j &= 1, \cdots, n \text{ (subjects)} \end{aligned}$$

The measured phenotype (P) (again, omitting the j subscript) consists of multiple variables that are a function of a subject's underlying additive genetic deviate (A), common (between-families) environment (C), and non-shared (within-families) environment (E). In addition, each variable P_j has a specific component U_j that itself may consist of a genetic and a non-genetic part. In this initial application, we assume that U_j is entirely random environmental in origin, an assumption we relax later. Parameters a, c, and e are the p–variate factor loadings of measured variables on the latent factors. A path diagram of this model is shown in Figure ??.

In Mx, there are a number of alternative ways to specify the model. One approach is to specify the factor structure for the genetic, shared and specific environmental factors in one matrix, e.g. B with twice the number of variables (for both twins) as rows and the number of factors for each twin as columns. If we assume one genetic, one shared environmental and one specific environmental common factor per twin $(A_1, A_2, C_1, C_2, E_1, E_2)$ for our four-variate arithmetic computation example (shown as T0 – T3 to represent administration times 0–3 before

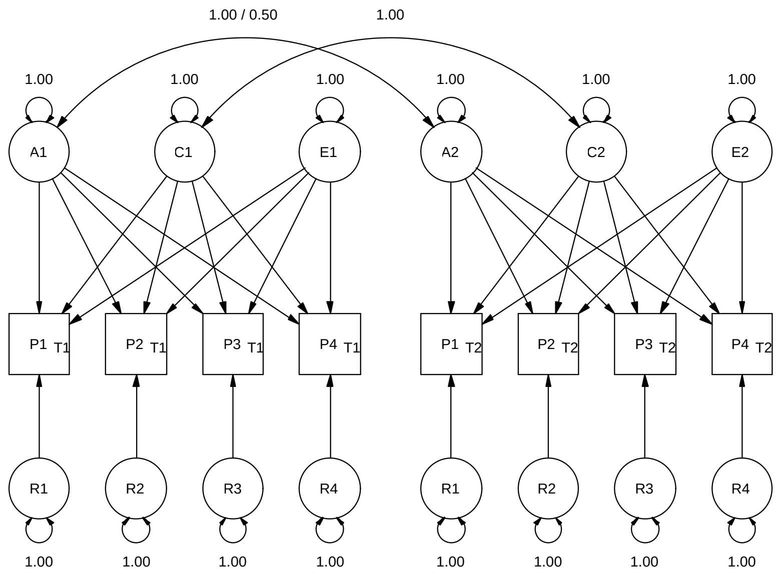


Figure 10.1: Multivariate Genetic Factor model for four variables.

and after standard doses of alcohol for twin 1 (Tw1) and twin 2 (Tw2) respectively), the B matrix would look like

	A_1	C_1	E_1	A_2	C_2	E_2
Tw1-T0	1	5	9	0	0	0
Tw1-T1	2	6	10	0	0	0
Tw1-T2	3	7	11	0	0	0
Tw1-T3	4	8	12	0	0	0
Tw2-T0	0	0	0	1	5	9
Tw2-T1	0	0	0	2	6	10
Tw2-T2	0	0	0	3	7	11
Tw2-T3	0	0	0	4	8	12

In this case with $m = 6$ factors and four observed variables for each twin (p=8), B would be a $p \times m$ (8×6) matrix of the factor loadings, P the $m \times m$ correlation matrix of factor scores, and E a $p \times p$ diagonal matrix of unique variances. The expected covariance may then be calculated as in equation 10.1:

$$\Sigma_{Y,Y} = \mathbf{BPB}' + \mathbf{E}. \tag{10.3}$$

In a multivariate analysis of twin data according to this factor model, Σ is a $2p \times 2p$ predicted covariance matrix of observations on twin 1 and

twin 2 and B is a $2p \times 2m$ matrix of loadings of these observations on latent genotypes and non-shared and common environments of twin 1 and twin 2. The factor loadings between A_1 and A_2, E_1 and E_2, and C_1 and C_2 are constrained to be equal for twin 1 and twin 2, similar to the path coefficients of the univariate models discussed in previous chapters. The equality constraints on the parameters are obtained in Mx by using the same non-zero parameter number in a `Specification` statement for the free parameters. The unique variances also are equal for both members of a twin pair. These may be estimated on the diagonal of the $2p \times 2p$ E matrix (e.g., Heath *et al.*, 1989c). To fit this model, B and E are estimated from the data and P $(2m \times 2m)$ must be fixed *a priori* (for example, the correlation between A_1 for twin 1 and A_2 for twin 2 is 1.0 for MZ and 0.5 for DZ twins; the correlation between the C variables of twin 1 and twin 2 is 1.0).

One alternative specification of this model is to include the unique variances in matrix B and fix E to zero. The factor patterns for A and E of twin 1 and twin 2 are identical to that in Section 10.2.3. The main difference lies in the treatment of the unique variances. In the earlier example these were estimated as variances on the diagonal of E, but now they are modeled as the *square roots of the variances*. These quantities are now square roots because the unique variances are calculated as the product BPB$'$ in the expected covariance expression whereas in the previous example the quantities were estimated as the unproducted quantity E. One might expect that this subtle change would have no effect on the model (as indeed it does not in this example), but on occasion these alternative residual specifications may produce different outcomes. The situation of residual variances < 0.0 makes little sense in genetic analyses because it implies an impossible negative variance component. Consequently, although it may be possible to make alternative representations like this in Mx, we recommend this model, as it constrains unique variances to be ≥ 0.0. Nevertheless, both methods give identical solutions when fitted to the data used in these examples.

10.3.2 Alternate Representation of the Multivariate Genetic Factor Model

One of the features of Mx is its flexibility for specifying the same or very similar models in different ways. Frequently the choice of model specification is simply a matter of individual preference, convenience, or familiarity with Mx notation, particularly when a model can be written in several different ways with no change in the substantive or numerical outcome. However, at other times very subtle changes in the Mx formulation of a model translate into a completely different substantive question. While it may be true that flexibility imparts confusion, it is important to recognize and distinguish alternative representations of genetic models in Mx.

While the approach discussed above may be fairly intuitive, the B matrix may become relatively big, therefore increasing the chance of errors in editing. An alternative approach is to specify the common factors and residual variances for genetic, shared and specific environmental factors in separate matrices. One advantage of this approach is that the model can be easily adapted for a different number of common factors or observed variables. For example, if we use a 4×1 matrix **X** for the genetic common factor, a 4×1 matrix **Y** for the shared environmental common factor, a 4×1 matrix **Z** for the specific environmental common factor and a 4×4 diagonal matrix **F** for the unique variances, the matrices section in Mx would be

```
X Full 4 1 Free    ! genetic common factor
Y Full 4 1 Free    ! shared environmental common factor
Z Full 4 1 Free    ! specific environmental common factor
F Diag 4 4 Free    ! specific environmental unique variances
```

We can then pre-calculate the genetic, shared and specific environmental variance components in the algebra section:

```
A= X*X';
C= Y*Y';
E= Z*Z' +F*F';
```

and these matrices can be used to specify the expected covariance matrices for MZ and DZ twins in a similar fashion as the univariate models. Note that by using a Kronecker product for the genetic variance component in DZ twins (`H@A`) every element of the **A** matrix is multiplied by one half. One additional feature in Mx that allows for flexible model specification is the `#define` statement. One possible use is to define the number of variables up front, e.g.

```
#define nvar 4
```

and use the 'defined' variables in the matrices section:

```
X Full nvar 1 Free       ! genetic common factor
Y Full nvar 1 Free       ! shared environmental common factor
Z Full nvar 1 Free       ! specific environmental common factor
F Diag nvar nvar Free    ! specific environmental unique variances
```

If we wanted to do an analysis with just three variables, the only change to be made, besides the `NInput_vars` and `Select` statements, is the `#define` statement.

10.3.3 Fitting the Multivariate Genetic Model

To illustrate the genetic common factor model we fit it to the arithmetic computation data, but now using both members of the female twin pairs and specifying two groups for the MZ and DZ twins. The observed variances and correlations examined in this analysis are presented in Table 10.3. Appendix ?? shows the full Mx script for this model.

Table 10.3: Observed female MZ (above diagonal) and DZ (below diagonal) correlations and variances for arithmetic computation variables.

		Twin 1				Twin 2			
		T0	T1	T2	T3	T0	T1	T2	T3
T1	T0	1.0	.81	.83	.87	.78	.65	.71	.68
	T1	.89	1.0	.87	.87	.74	.74	.74	.71
	T2	.85	.90	1.0	.90	.73	.66	.72	.70
	T3	.83	.86	.86	1.0	.74	.71	.74	.75
T2	T0	.23	.31	.36	.34	1.0	.73	.78	.79
	T1	.22	.32	.34	.38	.81	1.0	.86	.87
	T2	.16	.23	.27	.35	.79	.86	1.0	.87
	T3	.23	.31	.34	.37	.81	.86	.87	1.0
	MZ	297.9	229.4	247.4	274.9	281.9	359.7	326.9	281.1
	DZ	259.7	259.9	245.2	249.3	283.8	249.5	262.1	270.9

The results from this common factor model are shown in Table 10.3.3 The parameter estimates in the `MX PARAMETER ESTIMATES` section indicate a substantial genetic basis for the observed arithmetic covariances, as the genetic loadings are much higher than either the shared and non-shared environmental effects. The unique variances in `F` also appear substantial but these do not contribute to covariances among the measures, only to the variance of each observed variable. The χ^2_{56} value of 46.77 suggests that this single factor model provides a reasonable explanation of the data. (Note that the 56 degrees of freedom are obtained from $2 \times 8(8+1)/2$ free statistics minus 16 estimated parameters).

Earlier in this chapter we alluded to the fact that confirmatory factor models allow one to statistically test the significance of model parameters. We can perform such a test on the present multivariate genetic model. The Mx output above shows that the shared environment factor loadings are much smaller than either the genetic or non-shared environment loadings. We can test whether these loadings are significantly different from zero by modifying slightly the Mx script to fix these pa-

Table 10.4: Parameter estimates from the full genetic common factor model

	A_C	C_C	E_C	E_S
Time 1	15.088	1.189	4.142	46.208
Time 2	13.416	5.119	6.250	39.171
Time 3	13.293	4.546	7.146	31.522
Time 4	13.553	5.230	5.765	34.684
$\chi^2 = 46.77$, 56 df, p=.806				

rameters and then re-estimating the other model parameters. There are several possible ways in which one might modify the script to accomplish this task, but one of the easiest methods is simply to change the Y to have no free elements.

Performing this modification in the first group effectively drops all C loadings from all groups because the `Matrices= Group 1` statement in the second and third group equates its loadings to those in the first. Thus, the modified script represents a model in which common factors are hypothesized for genetic and non-shared environmental effects to account for covariances among the observed variables, and unique effects are allowed to contribute to measurement variances. All shared environmental effects are omitted from the model.

Since the modified multivariate model is a sub- or nested model of the full common factor specification, comparison of the goodness-of-fit chi-squared values provides a test of the significance of the deleted C factor loadings (see Chapter ??). The full model has 56 degrees of freedom and the reduced one: $2 \times 8(8+1)/2 - 12 = 60$ d.f. Thus, the difference chi-squared statistic for the test of C loadings has $60 - 56 = 4$ degrees of freedom. As may be seen in the output fragment below, the χ^2_{60} of the reduced model is 51.08, and, therefore, the difference χ^2_4 is $51.08 - 46.77 = 4.31$, which is non-significant at the .05 level. This non-significant chi-squared indicates that the shared environmental loadings can be dropped from the multivariate genetic model without significant loss of fit; that is, the arithmetic data are not influenced by environmental effects shared by twins. Parameter estimates from this reduced model are given below in Table 10.3.3

The estimates for the genetic and non-shared environment parameters differ somewhat between the reduced model and those estimated in the full common factor model. Such differences often appear when fitting nested models, and are not necessarily indicative of misspecification (of course, one would not expect the estimates to change in the

Table 10.5: Parameter estimates from the reduced genetic common factor model

	A_C	C_C	E_C	E_S
Time 1	**14.756**	–	**3.559**	**59.502**
Time 2	**14.274**	–	**6.331**	**39.433**
Time 3	**14.081**	–	**7.047**	**30.843**
Time 4	**14.405**	–	**5.845**	**36.057**
$\chi^2 = 51.08$, **60 df, p=.787**				

case where parameters to be omitted are estimated as 0.0 in the full model). The fitting functions used in Mx (see Chapter ??) are designed to produce parameter estimates that yield the closest match between the observed and estimated covariance matrices. Omission of selected parameters, for example, the C loadings in the present model, generates a different model Σ and thus may be expected to yield slightly different parameter estimates in order to best approximate the observed matrix.

10.3.4 Fitting a Second Genetic Factor

The genetic common factor model we introduced in Sections 10.3.3 and 10.3.2 may be extended to address more specific questions about the data. In the arithmetic computation measures, for example, it is reasonable to hypothesize two genetic factors: one general factor contributing to all measurements of arithmetic computation, and a second "alcohol" factor which influences the measures taken after the challenge dose of alcohol. The most parsimonious extension of our common factor model may involve the addition of only 1 free parameter which represents each of the factor loadings on the alcohol factor (that is, the alcohol loadings may be equated for all alcohol measurements).

The Mx script corresponds very closely to that used in section 10.3.2, using the X for the genetic common factors We add the latent alcohol factors for twins 1 and 2 as a second column with the following specification statement:

```
Specify X
1 0
2 5
3 5
4 5
```

The addition of the single parameter for all alcohol loadings reflects

a model having 13 parameters and $2 \times 8(8+1)/2 - 13 = 59$ degrees of freedom. We can, therefore, test the significance of the alcohol factor by comparing the goodness-of-fit chi-squared value for this model with that obtained from the model of Section 10.3.2 for a $60 - 59 = 1$ d.f. test. Table 10.3.4 shows the results of the two-factor multivariate genetic model.

Table 10.6: Parameter estimates from the two genetic factors model

	A_C1	A_C2	E_C	E_S
Time 1	**15.067**	**0.000**	**4.408**	**6.674**
Time 2	**13.701**	**4.270**	**6.091**	**6.277**
Time 3	**13.518**	**4.270**	**6.800**	**5.644**
Time 4	**13.832**	**4.270**	**5.695**	**5.928**
$\chi^2 = 47.52$, **59 df, p=.858**				

The estimated genetic factor loading for the alcohol variables (4.27) is reasonably large, but much smaller than the loadings on the general genetic factor. This difference is more apparent when we consider proportions of genetic variance accounted for by these two factors, being $4.27^2/(13.70^2 + 4.27)$ or 9% for the alcohol factor, and $100 - 9 = 91\%$ for the general genetic factor. The model yields a $\chi^2_{59} = 47.52$ ($p =$ .86), indicating a good fit to the data. The chi-squared test for the significance of the alcohol factor loadings is $51.08 - 47.52 = 3.56$, which is not quite significant at the .05 level. Thus, while the hypothesis of there being genetic effects on the alcohol measures additional to those influencing arithmetic skills fits the observed data better, the increase in fit obtained by adding the alcohol factor does not reach statistical significance.

10.4 Multiple Genetic Factor Models

10.4.1 Genetic and Environmental Correlations

We now turn from the one- and two-factor multivariate genetic models described above and consider more general multivariate formulations which may encompass many genetic and environmental factors. These more general approaches subsume the simpler techniques described above.

Consider a simple extension of the one- and two-factor AE models for multiple variables (sections 10.3.2–10.3.4). The total phenotypic covariance matrix in a population, $\mathbf{C_p}$, can be decomposed into an additive

genetic component, A, and a random environmental component, E:

$$\mathbf{C_p} = \mathbf{A} + \mathbf{E} \ , \tag{10.4}$$

We are leaving out the shared environment in this example just for simplicity. More complex expectations for 10.4 may be written without affecting the basic idea. "A" is called the *additive genetic covariance matrix* and "E" the *random environmental covariance matrix.* If A is diagonal, then the traits comprising A are genetically independent; that is, there is no "additive genetic covariance" between them. One interpretation of this is that different genes affect each of the traits. Similarly, if the environmental covariance matrix, E, is diagonal, we would conclude that each trait is affected by quite different environmental factors.

On the other hand, suppose A were to have significant off-diagonal elements. What would that mean? Although there are many reasons why this might happen, one possibility is that at least some genes are having effects on more than one variable. This is known as *pleiotropy* in the classical genetic literature (see Chapter 3). Similarly, significant off-diagonal elements in E (or C, if it were included in the model) would indicate that some environmental factors influence more than one trait at a time.

The extent to which the same genes or environmental factors contribute to the observed phenotypic correlation between two variables is often measured by the *genetic* or *environmental correlation* between the variables. If we have estimates of the genetic and environmental covariance matrices, A and E, the genetic correlation (r_g) between variables i and j is

$$r_{g_{ij}} = \frac{a_{ij}}{\sqrt{(a_{ii} \times a_{jj})}} \tag{10.5}$$

and the environmental correlation, similarly, is

$$r_{e_{ij}} = \frac{e_{ij}}{\sqrt{(e_{ii} \times e_{jj})}} \ . \tag{10.6}$$

The analogy with the familiar formula for the correlation coefficient is clear. The genetic covariance between two phenotypes is quite distinct from the genetic correlation. It is possible for two traits to have a very high genetic correlation yet have little genetic covariance. Low genetic covariance could arise if either trait had low genetic variance. Vogler (1982) and Carey (1988) discuss these issues in greater depth.

10.4.2 Cholesky Decomposition

Clearly, we cannot resolve the genetic and environmental components of covariance without genetically informative data such as those from twins. Under our simple AE model we can write, for MZ and DZ pairs, the expected covariances between the multiple measures of first and second members very simply:

$$\begin{aligned} \mathbf{C}_{\mathbf{MZ}} &= \mathbf{A} \\ \mathbf{C}_{\mathbf{DZ}} &= \alpha \mathbf{A} \end{aligned}$$

with the total phenotypic covariance matrix being defined as in expression 10.4. The coefficient α in DZ twins is the familiar additive genetic correlation between siblings in randomly mating populations (i.e., 0.5).

The method of maximum likelihood, implemented in Mx, can be used to estimate A and E. However, there is an important restriction on the form of these matrices which follows from the fact that they are covariance matrices: they *must* be positive definite. It turns out that if we try to estimate A and E without imposing this constraint they will very often not be positive definite and thus give nonsense values (greater than or less than unity) for the genetic and environmental correlations. It is very simple to impose this constraint in Mx by recognizing that any positive definite matrix, F, can be decomposed into the product of a *triangular matrix* and its transpose:

$$\mathbf{F} = \mathbf{T}\mathbf{T}' , \tag{10.7}$$

where T is a triangular matrix (i.e., one having fixed zeros in all elements above the diagonal and free parameters on the diagonal and below). This is sometimes known as a *triangular decomposition* or a *Cholesky factorization* of F. Figure 10.2 shows this type of model as a path diagram for four variables. In our case, we represent the genetic and environmental covariance matrices in Mx by their respective Cholesky factorizations:

$$\mathbf{A} = \mathbf{X}\mathbf{X}' \tag{10.8}$$

and

$$\mathbf{E} = \mathbf{Z}\mathbf{Z}' , \tag{10.9}$$

where X and Z are triangular matrices of additive genetic and within-family environment factor loadings.

A triangular matrix such as T, X, or Z is square, having the same number of rows and columns as there are variables. The first column has non-zero entries in every element; the second has a zero in the first element and free, non-zero elements everywhere else, and so on. Thus, the Cholesky factors of F, when F is a 3×3 matrix of the product $\mathbf{T}\mathbf{T}'$, will have the form:

$$\mathbf{T} = \begin{pmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{pmatrix} .$$

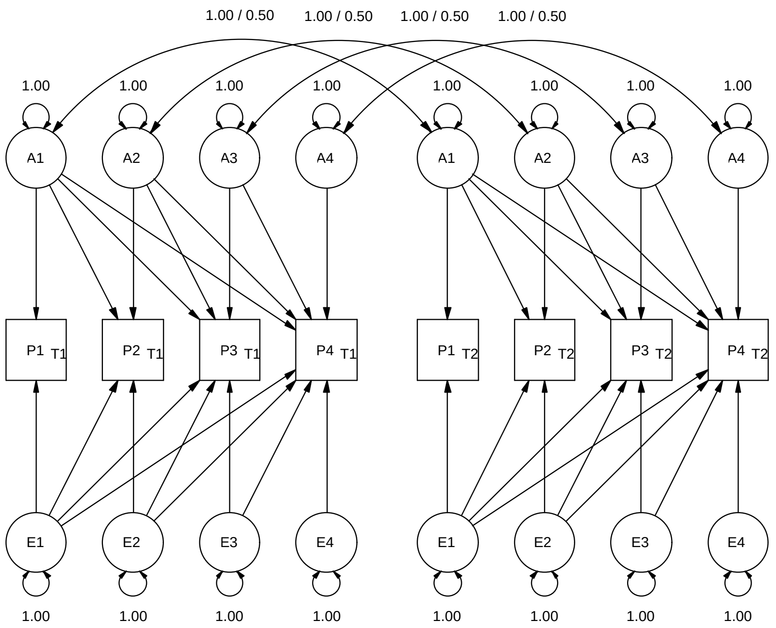


Figure 10.2: Phenotypic Cholesky decomposition model for four variables.

It is important to recognize that common factor models such as the one described in Section 10.3 are simply reduced Cholesky models with the first column of parameters estimated and all others fixed at zero.

10.4.3 Analyzing Genetic and Environmental Correlations

We illustrate the estimation of the genetic and environmental covariance matrices for a simple case of skinfold measures made on 11 year-old male twins from the Medical College of Virginia Twin Study (Schieken *et al.*, 1989)[2]. Our skinfold assessments include four different measures which were obtained using standard anthropometric techniques. The measures were obtained for biceps (BIC), subscapular (SSC), suprailiac (SUP), and triceps (TRI)skinfolds. The raw data were averaged for the left and right sides and subjected to a logarithmic transformation prior to analysis in order to remove the correlation between error variance and skinfold measure. The 8×8 covariance matrices for the male MZ and DZ twins are given in Table 10.7.

[2]We are grateful to Dr. Richard Schieken for making these data, gathered as part of a project supported by NHLBI award HL-31010, available prior to publication.

Table 10.7: Covariance matrices for skinfold measures in adolescent Virginian male twins.

	BIC1	SSC1	SUP1	TRI1	BIC2	SSC2	SUP2	TRI2
Dizygotic Male Pairs (N=33)								
BIC1	**.1538**							
SSC1	**.1999**	**.3007**						
SUP1	**.2266**	**.3298**	**.3795**					
TRI1	**.1285**	**.1739**	**.2007**	**.1271**				
BIC2	**.0435**	**.0336**	**.0354**	**.0376**	**.1782**			
SSC2	**.0646**	**.0817**	**.0741**	**.0543**	**.2095**	**.3081**		
SUP2	**.0812**	**.0901**	**.0972**	**.0666**	**.2334**	**.3241**	**.3899**	
TRI2	**.0431**	**.0388**	**.0376**	**.0373**	**.1437**	**.1842**	**.2108**	**.1415**
Monozygotic Male Pairs (N=84)								
	BIC1	**SSC1**	**SUP1**	**TRI1**	**BIC2**	**SSC2**	**SUP2**	**TRI2**
BIC1	**.1285**							
SSC1	**.1270**	**.1759**						
SUP1	**.1704**	**.2156**	**.3031**					
TRI1	**.1035**	**.1101**	**.1469**	**.1041**				
BIC2	**.0982**	**.1069**	**.1491**	**.0824**	**.1233**			
SSC2	**.0999**	**.1411**	**.1848**	**.0880**	**.1295**	**.1894**		
SUP2	**.1256**	**.1654**	**.2417**	**.1095**	**.1616**	**.2185**	**.2842**	
TRI2	**.0836**	**.0907**	**.1341**	**.0836**	**.1010**	**.1134**	**.1436**	**.1068**

Variable Labels: BIC=Biceps; SSC=Subscapular; SUP=Suprailiac; TRI=Triceps. "1" and "2" refer to measures on first and second twins

An example Mx program for estimating the Cholesky factors of the additive genetic and within-family environmental covariance matrices is given in Appendix ??. The matrices X and Z are now declared as free lower triangular matrices.

When this program is run with the data from male twins, we obtain a goodness-of-fit chi-squared of 68.92 for 52 d.f. ($p = .058$) suggesting that the AE model gives a reasonable fit to these data. Setting the off-diagonal elements of the genetic factors to zero yields a chi-squared that may be compared using the difference test to see whether the measures can be regarded as genetically independent. This chi-squared turns out to be 110.96 for 6 d.f. which is highly significant. Therefore, the genetic correlations between these skinfold measures cannot be ignored. Similarly, setting the environmental covariances to zero yields a significant increase in chi-squared of 356.98, also for 6 d.f. Clearly, there are also highly significant environmental covariances among the four variables.

Table 10.8 gives the estimates of the Cholesky factors of the genetic and environmental covariance matrices produced by Mx. Carrying

Table 10.8: Parameter estimates of the cholesky factors in the genetic and environmental covariance matrices.

	Genetic Factor				Environmental Factor			
Variable	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
BIC	**0.340**	**0.000**	**0.000**	**0.000**	**0.170**	**0.000**	**0.000**	**0.000**
SSC	**0.396**	**0.182**	**0.000**	**0.000**	**0.160**	**0.138**	**0.000**	**0.000**
SUP	**0.487**	**0.159**	**0.148**	**0.000**	**0.180**	**0.117**	**0.093**	**0.000**
TRI	**0.288**	**0.016**	**0.036**	**0.110**	**0.117**	**0.039**	**-0.004**	**0.085**

out the pre- and post-multiplication of the Cholesky factors (see equations 10.8 and 10.9) gives the maximum-likelihood estimates of the genetic and environmental covariance matrices, which we present in the upper part of Table 10.9. The lower part of Table 10.9 gives the matrices of genetic and environmental correlations derived from these covariances (see 10.5 and 10.6).

We see that the genetic correlations between the four skinfold measures are indeed very large, suggesting that the amount of fat at different sites of the body is almost entirely under the control of the same genetic factors. However, in this example, the environmental correlations also are quite large, suggesting that environmental factors which affect the amount of fat at one site also have a generalized effect over all sites.

Table 10.9: Maximum-likelihood estimates of genetic and environmental covariance (above the diagnoals) and correlation (below the diagonals) matrices for skinfold measures.

	Genetic				Environmental			
Variable	**BIC**	**SSC**	**SUP**	**TRI**	**BIC**	**SSC**	**SUP**	**TRI**
BIC	**0.116**	**0.135**	**0.166**	**0.098**	**0.029**	**0.027**	**0.030**	**0.020**
SSC	**0.909**	**0.190**	**0.222**	**0.117**	**0.759**	**0.044**	**0.045**	**0.024**
SUP	**0.914**	**0.955**	**0.284**	**0.148**	**0.769**	**0.908**	**0.054**	**0.025**
TRI	**0.927**	**0.863**	**0.894**	**0.097**	**0.778**	**0.757**	**0.716**	**0.023**

Note: The variances are given on the diagonals of the two matrices

10.5 Common vs. Independent Pathway Genetic Models

As another example of multivariate analysis we consider four atopic symptoms reported by female twins in a mailed questionnaire study (Duffy *et al.*, 1990; 1992). Twins reported whether they had ever (versus never) suffered from asthma, hayfever, dust allergy and eczema. Tetrachoric correlation matrices were calculated with PRELIS and are shown in the Mx script in Appendix ?? and in Table 10.10. Tetrachoric or polychoric matrices and their corresponding asymptotic covariance matrices are read in with the `PMatrix` and `ACov` statements. The script shows that asymptotic covariance matrices are stored in files named `ahdemzf.acv` and `ahdedzf.acv` respectively for MZ and DZ twins. Reading polychoric matrices flags Mx that the weighted least squares (WLS) fit function is required, rather than maximum likelihood. Maximum-likelihood estimation is not appropriate when there are glaring departures from normality; the dichotomous items used in this example are inevitably non-normal.

10.5.1 Independent Pathway Model for Atopy

Inspection of the correlation matrices in Table 10.10 reveals that the presence of any one of the symptoms is associated with an increased risk of the others within an individual (hence the concept of "atopy"). All four symptoms show higher MZ correlations (0.592, 0.593, 0.518, 0.589) than DZ correlations in liability (0.262, 0.318, 0.214, 0.313) and there is a hint of genetic dominance (or epistasis) for asthma and dust allergy (DZ

Table 10.10: Tetrachoric correlations for female MZ (above diagonal) and DZ (below diagonal) twins for asthma (A), hayfever (H), dust allergy (D), and eczema (E).

		Twin 1				Twin 2			
		A	H	D	E	A	H	D	E
Twin 1	**Asthma**		.56	.57	.27	.59	.41	.43	.09
	Hayfever	.52		.76	.26	.37	.59	.42	.20
	Dust Allergy	.59	.75		.31	.40	.45	.52	.19
	Eczema	.29	.31	.28		.23	.15	.19	.59
Twin 2	**Asthma**	.26	.17	.04	.14		.55	.64	.15
	Hayfever	.13	.32	.26	.09	.40		.77	.12
	Dust Allergy	.08	.17	.21	.02	.68	.72		.22
	Eczema	.22	.11	.09	.31	.25	.22	.28	

correlations less than half their MZ counterparts). Preliminary multivariate analysis suggests that dominance is acting at the level of a common factor influencing all symptoms, rather than as specific dominance contributions to individual symptoms. Our first model for covariation of these symptoms is shown in the path diagram of Figure 10.3

Because each of the three common factors (A, D, E) has its own paths to each of the four variables, this has been called the *independent pathway model* (Kendler *et al.*, 1987) or the *biometric factors model* (McArdle and Goldsmith, 1990). This is translated into Mx in the Appendix ?? script. The specification of this example is very similar to the multivariate genetic factor model described earlier in this chapter. The three common factors are specified in `nvar`×1 matrices `X`, `W` and `Z`, where `nvar` is defined as 4, representing the four atopy measures. The genetic and environmental specifics are estimated in `nvar`×`nvar` matrices `G` and `F`. The genetic, dominance and specific environmental covariance matrices are then calculated in the algebra section. The rest of the script is virtually identical to that for the univariate model.

One important new feature of the model shown in Figure 10.3 is the treatment of variance specific to each variable. Such residual variance does not generally receive much attention in regular non-genetic factor analysis, for at least two reasons. First, the primary goal of factor analysis (and of many multivariate methods) is to understand the covariance between variables in terms of reduced number of factors. Thus the residual, variable specific, components are not the focus. A second reason is that with phenotypic factor analysis, there is simply no information very

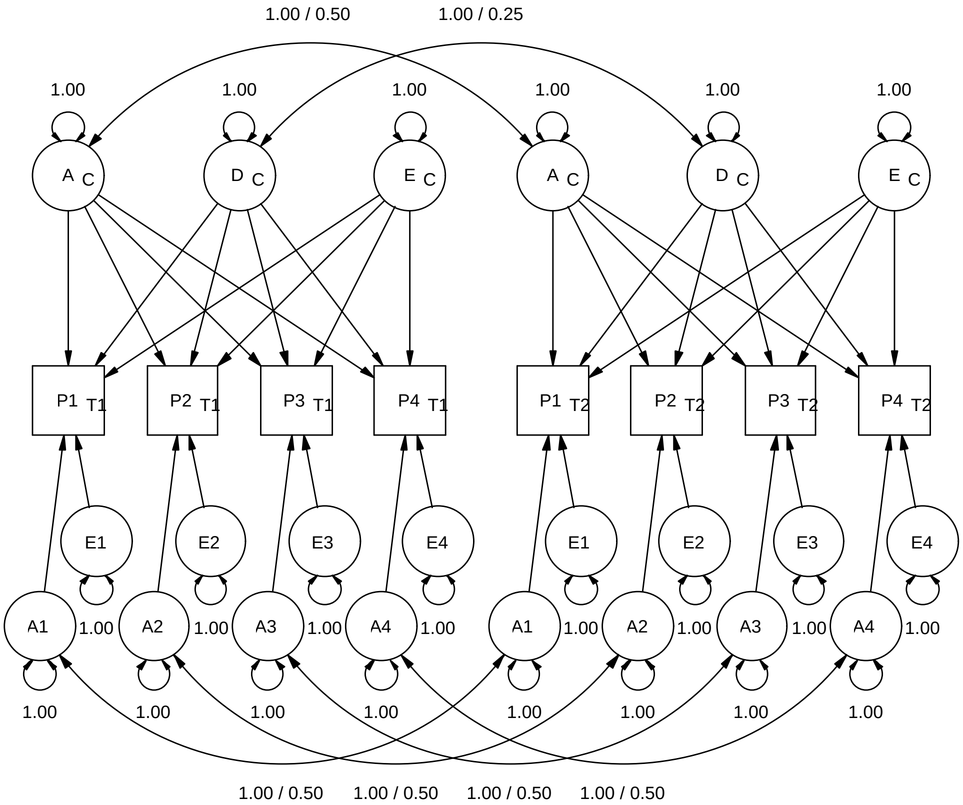


Figure 10.3: Independent pathway model for four variables. All labels for path-coefficients have been omitted. All four correlations at the bottom of the figure are fixed at 1 for MZ and .5 for DZ twins.

similar to further decompose the variable specific variance. However, in the case of data on groups of relatives, we have two parallel goals of understanding not only the within-person covariance for different variables, but also the across-relatives covariance structure both within and across variables. The genetic and environmental factor structure at the top of Figure 10.2 addresses the genetic and environmental components of variance common to the different variables. However, there remains information to discriminate between genetic and environmental components of the residuals, which in essence answers the question of whether family members correlate for the variable specific portions of variance.

A second important difference in this example — using correlation matrices in which diagonal variance elements are standardized to one — is that the degrees of freedom available for model testing are different from the case of fitting to covariance matrices in which all $k(k+1)/2$ elements are available, where k is the number of input variables. We encountered this difference in the univariate case in Section 6.3.1, but it is slightly more complex in multivariate analysis. For correlation matrices, since the k diagonal elements are fixed to one, we apparently have $g \times k$ fewer degrees of freedom than if we were fitting to covariances, where g is the number of data groups. However, since for a given variable the sum of squared estimates always equals unity (within rounding error), it is apparent that not all the parameters are free, and we may conceptualize the unique environment specific standard deviations (i.e., the e_i's) as being obtained as the square roots of one minus the sum of squares of all the other estimates. Since there are v (number of variables) such constrained estimates, we actually have v more degrees of freedom than the above discussion indicates, the correct adjustment to the degrees of freedom when fitting multivariate genetic models to correlation matrices is $-(g \times k - v)$. Since in most applications $k = 2v$, the adjustment is usually $-3v$. In our example $v = 4$ and the adjustment is indicated by the option `DFreedom=-12`. (Note that the `DFreedom` adjustment applies for the goodness-of-fit chi-squared for the whole problem, not just the adjustment for that group).

Edited highlights of the Mx output are shown below and the goodness-of-fit chi-squared indicates an acceptable fit to the data. The adjustment of -12 to the degrees of freedom which would be available were we working with covariance matrices (72) leaves 60 statistics. We have to estimate 3×4 factor loadings and 2×4 specific loadings (20 parameters in all), so there are $60 - 20 = 40$ d.f. It is a wise precaution always to go through this calculation of degrees of freedom — not because Mx is likely to get them wrong, but as a further check that the model has been specified correctly.

Table 10.11: Parameter estimates from the independent pathway model for atopy

	E_{Atopy}	H_{Atopy}	D_{Atopy}	H_{spec}	E_{Spec}
Asthma	.320	.431	.466	.441	.548
Hayfever	.494	.772	.095	.000	.388
Dust Allergy	.660	.516	.431	.297	-.159
Eczema	.092	.221	.260	.712	.606
$\chi^2 = 38.44$, 40 df, p=.540					

We can test variations of the above model by dropping the common factors one at a time, or by setting additive genetic specifics to zero. This is easily done by dropping the appropriate elements. Note that fixing E specifics to zero usually results in model failure since it generates singular expected covariance matrices (Σ)[3]. Neither does it make biological sense since it is tantamount to saying that a variable can be measured without error; it is hard to think of a single example of this in nature! We could also elaborate the model by specifying a third source of specific variance components, or by substituting shared environment for dominance, either as a general factor or as specific variance components.

10.5.2 Common Pathway Model for Atopy

In this section we focus on a much more stringent model which hypothesizes that the covariation between symptoms is determined by a single 'phenotypic' latent variable called "atopy." Atopy itself is determined by additive, dominance and individual environmental sources of variance. As in the independent pathway model, there are still specific genetic and environmental effects on each symptom. The path diagram for this model is shown in Figure 10.4. Because there is now a latent variable ATOPY which has direct phenotypic paths to each of the symptoms, this has been called the *common pathway model* (Kendler *et al.*, 1987) or the *psychometric factors model* (McArdle and Goldsmith, 1990).

The Mx script corresponding to this path diagram, given in Appendix ??, contains several new features. Again, there are a number of alternative ways to specify this model in Mx. We use the same approach as in previous models and specify the genetic and environmental covariance matrices in a calculation group up front. In this example, matrices `X`, `W` and `Z` represent the additive and dominance genetic and specific environmental loadings on the latent phenotype. The factor loadings on the observed variables are estimated in 4×1 matrix `S`. The residual variances are decomposed in genetic and environmental diagonal matrices `G` and `F`. The data groups are identical to those of the independent pathway model.

One final feature of the model is that since ATOPY is a latent variable whose scale (and hence variance) is not indexed to any measured variable, we must fix its residual variance term (EATOPY) to unity to make the model identified. This inevitably means that the estimates for the loadings contributing to ATOPY are arbitrary and hence so are the paths leading from ATOPY to the symptoms. It is thus particularly im-

[3]This problem is extreme when maximum likelihood is the fit function, because the inverse of Σ is required (see Chapter **??**).

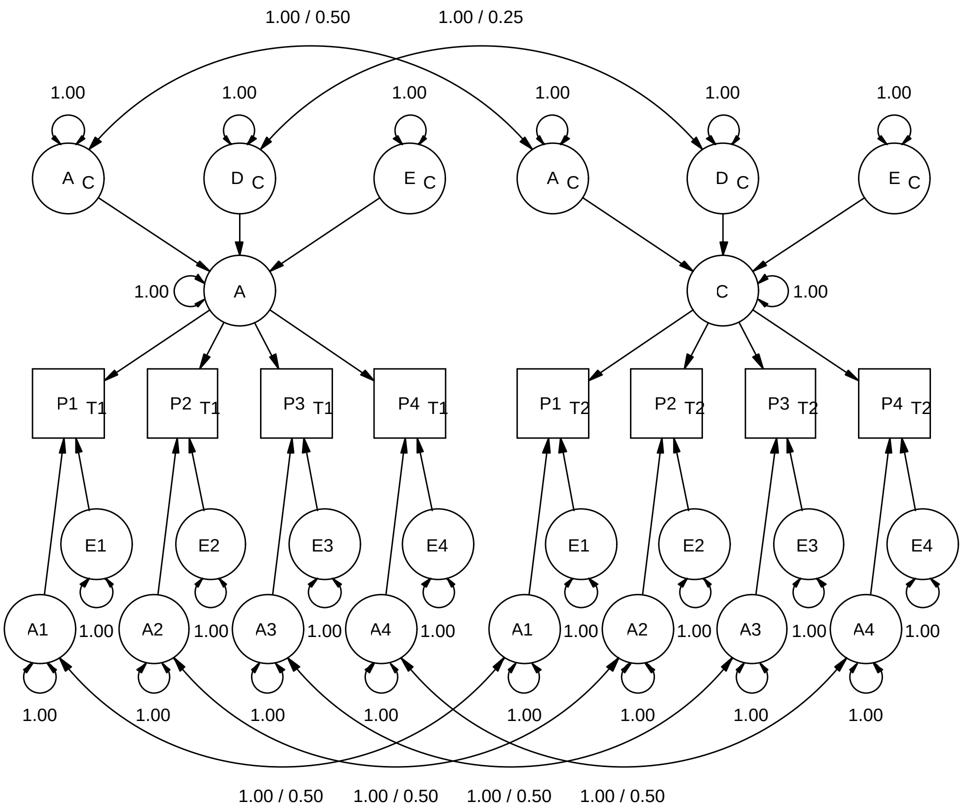


Figure 10.4: Common pathway model for four variables.

portant to standardize the solution so that the total variance explained for each symptom is unity. The fixing of the loading on EATOPY clearly has implications for the calculation of degrees of freedom, as we shall see below.

The condensed output for this model is presented below, showing the completely standardized estimates which give unit variance for each variable.

Note that here `NInput_vars=8` so there are 56 $(2 \times \mathtt{NI}(\mathtt{NI}-1)/2)$ unique correlations. From the above table it appears that 15 parameters have been estimated, but in fact EATOPY was fixed and the four E specifics are obtained by difference, so there are only 10 free parameters in the model, hence 46 degrees of freedom.

The latent variable ATOPY has a broad heritability of over 0.6

Table 10.12: Parameter estimates from the common pathway model for atopy

	Atopy	E_{Atopy}	H_{Atopy}	D_{Atopy}	H_{spec}	E_{Spec}
Asthma	.671	–	–	–	.531	.517
Hayfever	.814	–	–	–	.456	.358
Dust Allergy	.941	–	–	–	-.059	.334
Eczema	.301	–	–	–	.735	.608
Atopy	–	.686	.397	.610	–	–
$\chi^2 = 51.37$, 46 df, p=.238						

$(1 - .610^2 = .686^2 + .397^2)$ of which approximately a quarter is due to dominance, and this factor has an important phenotypic influence on all symptoms, particularly dust allergy (0.941) and hayfever (0.814). There are still sizeable specific genetic influences not accounted for by the ATOPY factor on all symptoms except dust allergy $(.059^2)$. However, despite the appeal of this model, it does not fit as well as the independent pathway model and the imposition of constraints that covariation between symptoms arises purely from their phenotypic relation with the latent variable ATOPY has worsened fit by $\chi^2 = 12.93$ for 6 degrees of freedom, which is significant at the 5% level.

We conclude that while there are common environmental, additive, and nonadditive genetic factors which influence all four symptoms of atopy, these have *differential* effects on the symptoms; the additive and non-additive factors, for example, having respectively greater and lesser proportional influence on hayfever than the other symptoms. While it is tempting to interpret this as evidence for at least two genes, or sets of genes, being responsible for the aggregation of symptoms we call atopy, this is simplistic as in fact such patterns could be consistent with the action of a single gene — or indeed with polygenic effects. For a full discussion of this important point see Mather and Jinks (1982) and Carey (1988).

Chapter 11

Observer Ratings

11.1 Introduction

Rather than measuring an individual's phenotype directly, we often have to rely on ratings of the individual made by an observer. An important example is the assessment of children via ratings from parents and teachers. In this chapter we consider in some detail the assessment of children by their parents. Since the ratings obtained in this case are a function of both parent and child, disentangling the child's phenotype from that of the rater becomes an important methodological problem. For the analysis of genetic and environmental contributions to children's behavior, solutions to this are available when multiple raters, e.g., two parents, rate multiple children, e.g., twins. This chapter describes and illustrates simple Mx models for the analysis of parental ratings of children's behavior (Section 11.2. We show how the assumption that mothers and fathers are rating the same behavior in children can be contrasted with the weaker alternative that parents are rating correlated behaviors. Given the stronger assumption, which appears adequate for ratings of some children's behavior problems, the contribution of rater bias and unreliability may be separated from the shared and non-shared environmental components of variation of the true phenotype of the child. The models are illustrated with an application to CBC data (Section 11.2.5

11.2 Models for Multiple Rating Data

A primary source of information about a child's behavior is the description of that behavior by his or her parents. In the study of child and adolescent psychopathology for example, parental reports are fundamental to the widely used assessment system developed by Achenbach and Edelbrock (1981). However, different informants do not generally agree in detail about a given child's behavior (Achenbach *et al.*, 1987; Loeber *et al.,* 1989) and, of course, there are very good reasons why this should be so (Cox and Rutter, 1985). Different informants, such as the child, parents, teachers or peers, have different situational exposure, different degrees of insight, and different perceptions, evaluations and normative standards that may create rater differences of various kinds in reporting problem behaviors. How we analyze parental ratings of children's behavior, and the models we employ in the course of our analyses, will depend on the assumptions we make. In this chapter we discuss the application of three classes of models — biometric, psychometric, and bias models.

First, suppose we took an agnostic view of the relationship between the ratings by different informants by thinking of them as assessing different phenotypes of the child. The phenotypes may be correlated but for unspecified reasons. This view may be appropriate if mothers and fathers reported on behaviors observed in distinct situations, or if they did not share a common understanding of the behavioral descriptions. In such a case it would be appropriate to treat the analysis of mothers' and fathers' ratings as a standard bivariate genetic and environmental analysis where the two variables are the mothers' ratings and fathers' ratings. We shall refer to the class of standard bivariate factor model as *biometric models* (see Chapter 10 for examples).

Second, suppose we made the more restrictive assumption that there is (i) a common phenotype of the children which is assessed both by mothers and by fathers, and (ii) a component of each parent's ratings which results from an assessment of an independent aspect of the child. Mothers' ratings and fathers' ratings would correlate because they are indeed making assessments based on shared observations and have a shared understanding of the behavioral descriptions used in the assessments. In this case, we approach the analysis of parental ratings through a special form of model for bivariate data which we will refer to as *psychometric models* (see Chapter 10 for examples).

Third, we consider a model of *rater bias.* Bias in this context is considered to be the tendency of an individual rater to overestimate or underestimate scores consistently. This tendency is a deviation from the mean of all possible raters in the rater group; no reference is made here to any external criterion such as a clinician's judgement. Neale and Stevenson (1989) considered the general problem of rater bias and the particular issues of parental biases in ratings of children. They presented a model in which the rating of a child's phenotype is considered to be a function both of the child's phenotype and of the bias introduced by the rater. In this way it is possible, when two parents rate each of their twin children, to conduct a behavior genetic analysis of the variation in the latent phenotype while allowing for variation due to rating biases. If the rater bias model adopted by Neale and Stevenson (1989) provides an adequate account of the ratings of children by their parents, it becomes possible to partition the variance in these parental ratings into their components due to reliable trait variance, due to parental bias , and due to unreliability or error in the particular rating of a particular child. The reliable trait variance can then be decomposed into its components due to genetic influences, shared environments, and individual environments. Since rater bias models represent restricted special cases for the parental ratings of more general biometric and psychometric models of the kind discussed by Heath *et al.*, (1989) and McArdle and Goldsmith (1990) and in Chapter 10 of this volume, it is possible to compare the adequacy of bias models with the alternative bivariate psychometric and biometric models. Further, comparison of the biometric and psychometric models indicates how reasonable it is to assume that two raters are assessing the same phenotype in a child. As we move from the biometric to the psychometric to the bias models, our assumptions become more restrictive but, if appropriate, our analyses become more directly informative psychologically. Here we outline how an analysis of parental ratings using the bias model can be implemented simply using Mx. We discuss the properties of the alternative models and illustrate their application with data from a twin study of child and adolescent behavior problems.

11.2.1 Rater Bias Model

Figure 11.1 shows a path model for the ratings of twins by their parents, in which the phenotypes of a pair of twins (PT_1 and PT_2) are functions of additive genetic influence (A), shared environments (C) and non-shared environments (E). The ratings by the mother (MRT) and father (FRT) are functions of the twin's phenotype, the maternal (B_M) or paternal (B_F) rater bias, and residual errors (R_{1MRT}, etc).

If this model is correct, the following discriminations may be made:

1. the structural analysis of the latent phenotypes of the children can be considered independently of the rater biases and unreliability of the ratings;

2. the extent of rater biases and unreliability of ratings can be estimated;

3. the relative accuracy of maternal and paternal ratings can be assessed.

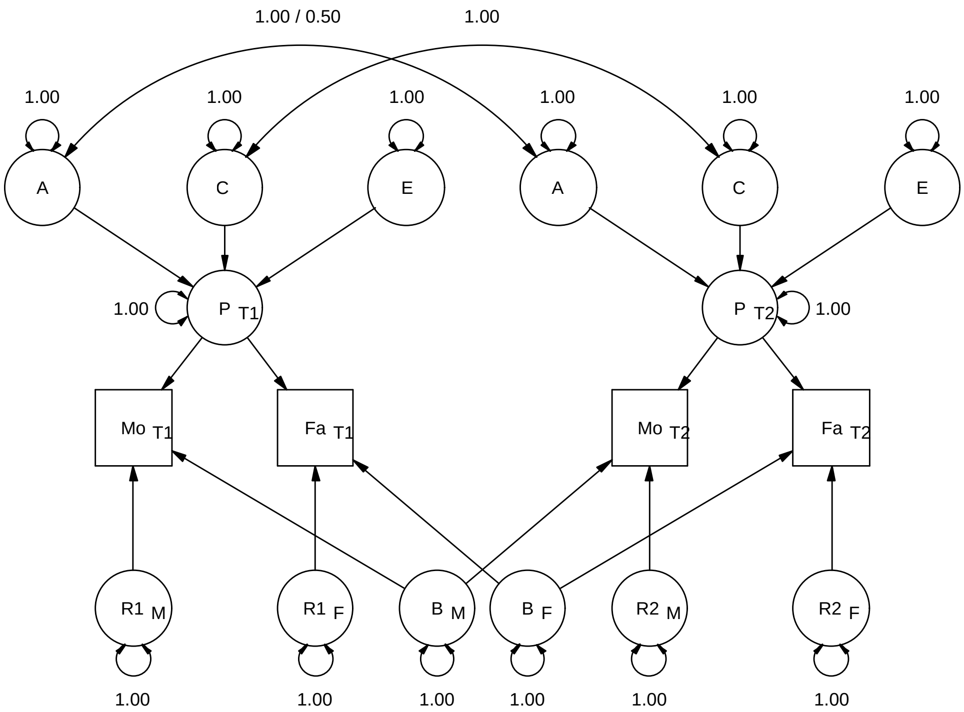


Figure 11.1: Model for ratings of a pair of twins (1 and 2) by their parents. Maternal and paternal observed ratings (MRT and FRT) are linear functions of the true phenotypes of the twins (PT), maternal and paternal rater bias (B_M and B_F), and residual error (R_{MRT} and R_{FRT}).

A simple implementation of the model in Mx is achieved by defining the model by the following matrix equations:

$$\begin{pmatrix} MRT_1 \\ MRT_2 \\ FRT_1 \\ FRT_2 \end{pmatrix} = \begin{pmatrix} b_m & 0 \\ b_m & 0 \\ 0 & b_f \\ 0 & b_f \end{pmatrix} \begin{pmatrix} B_M \\ B_F \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \alpha & 0 \\ 0 & \alpha \end{pmatrix} \begin{pmatrix} PT_1 \\ PT_2 \end{pmatrix}$$

$$+ \begin{pmatrix} r_{m1} & 0 & 0 & 0 \\ 0 & r_{m2} & 0 & 0 \\ 0 & 0 & r_{f1} & 0 \\ 0 & 0 & 0 & r_{f2} \end{pmatrix} \begin{pmatrix} R_{1MRT} \\ R_{2MRT} \\ R_{1FRT} \\ R_{2FRT} \end{pmatrix} \tag{11.1}$$

or

$$\mathbf{y} = \mathbf{Bb} + \mathbf{Ll} + \mathbf{Rr}$$

and

$$\begin{pmatrix} PT_1 \\ PT_2 \end{pmatrix} = \begin{pmatrix} a & c & e & 0 & 0 & 0 \\ 0 & 0 & 0 & a & c & e \end{pmatrix} \begin{pmatrix} A_1 \\ C_1 \\ E_1 \\ A_2 \\ C_2 \\ E_2 \end{pmatrix} \tag{11.2}$$

$$\tag{11.3}$$

or

$$\mathbf{l} = \mathbf{Gx}$$

Thus

$$\mathbf{y} = \mathbf{Bb} + \mathbf{LGx} + \mathbf{Rr}$$

Then, the covariance matrix of the ratings is given by

$$\begin{aligned} \mathcal{E}\{\mathbf{yy}'\} &= \mathcal{E}\{\mathbf{Bb} + \mathbf{LGx} + \mathbf{Rr}\}\{\mathbf{Bb} + \mathbf{LGx} + \mathbf{Rr}\}' & (11.4) \\ &= \mathbf{BB}' + \mathbf{RR}' + \mathbf{LG}\mathcal{E}\{\mathbf{xx}'\}\mathbf{G}'\mathbf{L}' & (11.5) \end{aligned}$$

The term $\mathbf{G}\mathcal{E}\{\mathbf{xx}'\}\mathbf{G}'$ generates the usual expectations for the ACE model. The expectations are filtered to the observed ratings through the factor structure **L** and are augmented by the contributions from rater bias (**B**) and residual influences (**R**). An Mx script for this model is listed in Appendix ??. In considering the rater bias model, and the other models discussed below, we should note that parameters need not be constrained to be equal when rating boys and girls and, as Neale and Stevenson (1988) pointed out, we need not necessarily assume that parental biases are equal for MZ and DZ twins' ratings. This latter relaxation of the parameter constraints allows us to consider the possibility that twin correlations differ across zygosities for reasons related to differential parental biases based on beliefs about their twins' zygosity.

11.2.2 Psychometric Model

Figure 11.2 shows a bivariate psychometric or 'common pathway' model. Implementation of this model in Mx can be achieved by the approaches illustrated in Chapter 10. The psychometric model estimates, for each source of influence (A, C, and E) the variance for mothers' ratings, the variance for fathers' ratings and the covariance between these ratings. These estimates are subject to the constraints that the covariances are positive and neither individual rating variance can be less than the

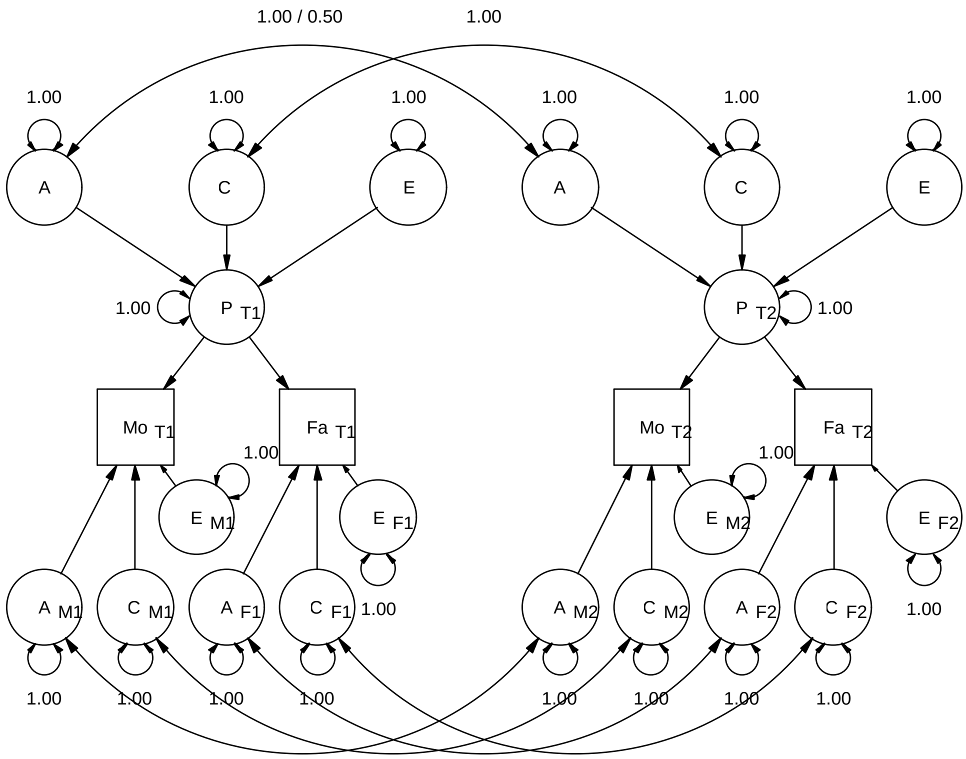


Figure 11.2: Psychometric or common pathway model for ratings of a pair of twins (1 and 2) by their parents. Maternal and paternal observed ratings (MRT and FRT) are linear functions of the latent phenotypes of the twins (PT), and rater specific variance (e.g., A_M, C_M and E_M).

covariance between the ratings. The psychological implication of this psychometric model is that the mothers' and fathers' ratings are composed of consistent assessments of reliable trait variance, together with assessments of specific phenotypes uncorrelated between the parents.

There are some technical points to note with this model. First, bivariate data for MZ and DZ twins (of a given sex) yield 20 observed variances and covariances. However, only 9 of these have unique expectations under the classes of model we are considering, the remaining 11 being replicate estimates of particular expectations (e.g., the variance of maternal ratings of MZ twin 1, of MZ twin 2, of DZ twin 1 and of DZ twin 2 are four replicate estimates of the variance of maternal ratings in

the population). Given this, we might expect our 9 parameter psychometric model to fit as well as any other 9 parameter model for bivariate twin data. However, there are some implicit constraints in our psychometric model. For example, the phenotypic covariance of mothers' and fathers' ratings cannot be greater than the variance of either type of rating. Such constraints may cause the model to fail in some circumstances even though the 9 parameter biometric model discussed below (Figure 11.3) may fit adequately[1]. The second technical point is that if we do not constrain the loadings of the common factor to be equal on the mothers' ratings and on the fathers' ratings, and assume that there is no specific genetic variance for either mothers' ratings or for fathers' ratings, then this variant of the psychometric model is formally equivalent to our version in the Neale and Stevenson bias model described above. In this case the "shared environmental" specific variances for the mothers' and fathers' ratings are formally equivalent to the maternal and paternal biases in the earlier model, while the "non-shared" specific variances are equal to the unreliability variance of the earlier parameterization. Thus, although the 9 parameter psychometric model and the bias model do not form a nested pair (Mulaik *et al.*, 1989), they represent alternative sets of constraints on a more general 10 parameter model (which is not identified with two-rater twin data) and these constrained models may be compared in terms of parsimony and goodness of fit. Furthermore, we may consider a restricted bias model in which the scaling factor in Figure 11.1 is set to unity and which, therefore, has 7 free parameters and is nested within both the psychometric model and the unrestricted bias model. This restricted bias model may therefore be tested directly against either the psychometric or the unrestricted bias models by a likelihood ratio chi-square.

11.2.3 Biometric Model

The final model to be considered is the biometric model shown in Figure 11.3, and again may be readily implemented using the procedure described in Chapter 10. In this model there are two factors for each source of variance (A, C, and E). One factor is subscripted M, e.g., A_M, and loads on the maternal rating (MRT) and on the paternal rating (FRT). The other factor subscripted F, e.g., A_F, loads only on the paternal rating. Thus, for each source of influence we estimate three factor loadings which enable us to reconstruct estimates of the contribution of this influence to the variance of maternal ratings, the variance of pa-

[1] There are in fact some other special cases such as scalar sex-limitation – where identical genetic or environmental factors may have different factor loadings for males and females — when the psychometric model may fit as well or better than the biometric model.

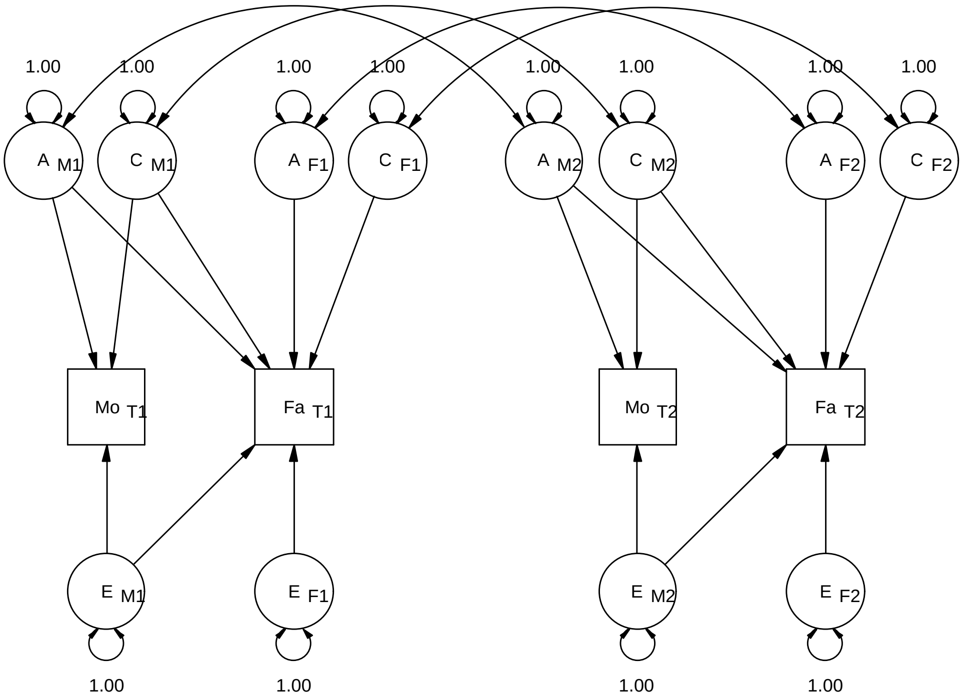


Figure 11.3: Biometric or independent pathway model for ratings of a pair of twins (1 and 2) by their parents. Maternal and paternal observed ratings (MRT and FRT) are linear functions of general (subscript M) and restricted (subscript F) genetic and environmental factors.

ternal ratings and the covariance between them. Which factor loads on both types of rating and which on only one is arbitrary. This type of model is referred to as a Cholesky model or decomposition or a triangular model and provides a standard general approach to multivariate biometrical analysis (see Chapter 10). This biometric model is a saturated unconstrained model for the nine unique expected variances and covariances (in the absence of sibling interactions or other influences giving rise to heterogeneity of variances across zygosities, cf. Heath *et al.*, 1989) and provides the most general approach to estimating the genetic, shared environmental and non-shared environmental components of variance and covariance. However, the absence of theoretically motivated constraints lessens the psychological informativeness of the model for the analysis of parental ratings. In this context, we may use the bio-

metric model first to test the adequacy of the assumption that of the 20 observed variances and covariances for bivariate twin data of a given sex, 11 represent replicate estimates of the 9 unique structural expectations. Once again, sex differences in factor loadings (scalar sex limitation) may in principle lead to model failure for opposite sex data even though the biometric model is adequate for a given sex. In this case the non-scalar sex limitation model described in Heath *et al.* (1989) and Chapter 9 would be required. The bivariate biometric model provides a baseline for comparison of the adequacy of the psychometric and bias models. This comparison alerts us to the important possibility that mothers and fathers are assessing different (but possibly correlated) phenotypes as, for example, they might be if mothers and fathers were reporting on behaviors observed in different situations or without a common understanding of the behavioral descriptions used in the assessment protocol.

11.2.4 Comparison of Models

We have considered four alternative models for parental ratings of children's behavior. Each model is for bivariate twin data where the two variables are the special case of mothers' ratings and fathers' ratings of the children's behavior. The least restrictive model, the biometric model, provides a baseline for comparison with the psychologically more informative psychometric and bias models. The most restricted bias model may be formally tested by likelihood ratio chi-square against either the psychometric or the unrestricted bias models. However, these latter two are not themselves nested. The relationships between these models, without taking into account sex limitations, are summarized in Figure 11.4. In this figure the solid arrows represent the process of constraining a more general model to yield a more restrictive model; the model at the arrow head is nested within the model at the tail of the arrow and may be tested against it by a likelihood ratio chi square. The dashed arrows represent rotational constraints on the biometric model. The nine parameter psychometric model requires, for example, that the covariance between maternal and paternal ratings be no greater than the variance of either type of rating; in factor analytic terms this would require a constrained rotation of the biometric model solution. The ten parameter psychometric model, allowing α not equal to unity, still imposes the constraints that the contributions of the common influences to the variance of maternal ratings, the variance of paternal ratings, and the covariance between them be in the ratio $1 : \alpha^2 : \alpha$ for each source of influence. Thus, even though this model has 10 parameters (and hence is not identified for bivariate twin data) any of its solutions, arrived at by fixing one of the parameters to an arbitrary value, will again rep-

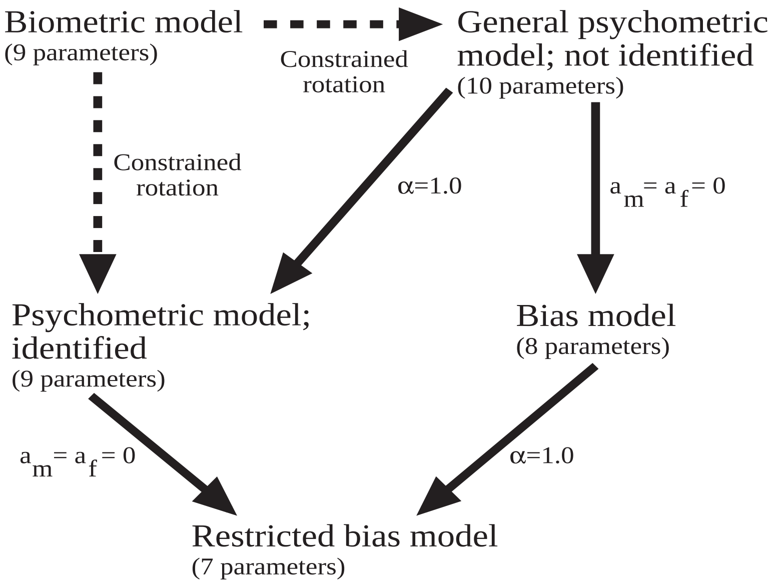


Figure 11.4: Diagram of nesting of biometric, psychometric, and rater bias models.

resent in factor analytic terms a constrained rotation of the biometric model.

11.2.5 Application to Data from the Child Behavior Checklist

To illustrate the application of these models we consider an updated set of data first presented by Hewitt, *et al.,* (1990) and now based on 983 families where both parents rated each of their twin children using Achenbach's Child Behavior Checklist (CBC; Achenbach and Edelbrock, 1983). For the full analysis, published in Hewitt *et al.* (1992), data from a population-based sample of 500 MZ twin pairs and 483 DZ twin pairs were considered and ratings were included irrespective of the biological or social relationship of the parent to the child. The children were Caucasian and ranged in age from 8 to 16 years. Ratings on 23 core items assessing children's internalizing behavior in both younger and older children and in either boys or girls were totalled to obtain an internalizing scale score for each child. The items contributing to this scale are listed in Appendix ??.

For illustrative purposes in this chapter we just consider the "prepu-

bertal" subsample of younger children aged 8-11 years. More detailed analyses, including older children, may be found in Hewitt *et al.* (1992). The scale scores were log-transformed to approximate normality and adjusted for linear regression on age and sex within age cohorts. The observed variances, covariances, and correlations of the resulting scores are given in Table 11.1 by zygosity and sex group.

Table 11.1: Observed variance-covariance matrices (lower triangle) and twin correlations (above the diagonal) for parental ratings (mother (Mo); father (Fa)) of internalizing behavior problems in five zygosity-sex groups (MZ female, N=96; MZ male, N=102; DZ female, N=102; DZ male, N=97; DZ male-female, N=103). All twins were between 8 and 11 years at assessment.

Zygosity/sex		Male					Female			
		Twin 1		Twin 2			Twin 1		Twin 2	
		Mo	Fa	Mo	Fa		Mo	Fa	Mo	Fa
MZ	MoT1	.675	.40	.74	.43	MoT1	.694	.47	.84	.46
	FaT1	.265	.652	.35	.77	FaT1	.312	.638	.37	.72
	MoT2	.513	.237	.714	.51	MoT2	.569	.238	.666	.45
	FaT2	.292	.513	.354	.676	FaT2	.308	.461	.293	.647
DZ	MoT1	.621	.47	.70	.34	MoT1	.565	.41	.55	.29
	FaT1	.315	.719	.35	.73	FaT1	.241	.604	.25	.57
	MoT2	.434	.236	.623	.37	MoT2	.291	.137	.488	.52
	FaT2	.233	.531	.251	.743	FaT2	.171	.347	.285	.604
DZMF	MoT1	.538	.26	.49	.18					
	FaT1	.162	.730	.17	.56					
	MoT2	.243	.102	.465	.37					
	FaT2	.103	.372	.191	.574					

A summary of the adequacy of the models fitted to these data on younger children's internalizing problems is shown in Table 11.2. The illustrative program in Appendix ?? runs the analysis for the bias model with 34 degrees of freedom.

As can be seen from Table 11.2, all three types of model give excellent fits to the data for younger children, with the psychometric model being preferred by Akaike's Information Criterion. Thus, our first conclusion would be that to a very good approximation, mothers and fathers can be assumed to be rating the same phenotype in their children when using the Child Behavior Checklist, at least as far as these internalizing behaviors are concerned. This may not be so for other behaviors or assessment instruments and in each particular case the assumption ought

Table 11.2: Model comparisons for internalizing problems analysis.

	Fit statistics		
Model*	**df**	χ^2	**AIC**
Restricted bias	**36**	**30.07**	**-41.9**
Bias	**34**	**25.78**	**-42.2**
Psychometric	**32**	**20.71**	**-43.3**
Biometric	**32**	**20.95**	**-43.1**

to be tested by a comparison of models of the kind we have described. Although there are numerous submodels or alternative models that may be considered, (for example: no sex limitation; non-scalar sex-limitation; and setting non-significant parameters to zero), only a subset will be presented here for illustration.

Table 11.3 shows the parameter estimates for the full bias and psycho-

Table 11.3: Parameter estimates from fitting bias, psychometric, and biometric models for parental ratings of internalizing behaviors.

Bias model			Psychometric model			Biometric model		
Path	**Boys**	**Girls**	**Path**	**Boys**	**Girls**	**Path**	**Boys**	**Girls**
a	**.519**	**.163**	a	**.370**	**.145**	a_m	**.513**	**.134**
c	**.277**	**.363**	a_m	**.338**	**-.027**	a_{fm}	**.261**	**.132**
e	**.189**	**.156**	a_f	**-.069**	**.281**	a_f	**.265**	**.286**
a	**.671**	**1.416**						
b_m	**.320**	**.545**	c	**.308**	**.449**	c_m	**.440**	**.659**
b_f	**.509**	**.473**	c_m	**.332**	**.479**	c_{fm}	**.225**	**.308**
r_m^2	**.074**	**.154**	c_f	**.437**	**.507**	c_f	**.490**	**.603**
r_f^2	**.175**	**.115**						
			e	**.176**	**.200**	e_m	**.328**	**.423**
			e_m	**.278**	**.372**	e_{fm}	**.096**	**.097**
			e_f	**.386**	**.333**	e_f	**.414**	**.377**

metric models allowing for scalar sex limitation and, in the case of the biometric model, we have allowed for non-scalar sex-limitation[2] of the shared environmental influences specific to fathers' ratings ($\chi^2_{31} = 20.76$ for the model presented with the correlation between boys' and girls' effects of this kind estimated at 0.86 rather than unity). To show the

[2]This is to avoid estimated loadings of opposite sign in boys and girls – see Chapter 9.

relationship between the more parsimonious bias model and the full parameterization of the biometric model, in Table 11.4 we present the

Table 11.4: Contributions to the phenotypic variances and covariance of mothers' and fathers' ratings of young boys' internalizing behavior.

	Biometric model			Bias model		
	Ratings		Cov (r)	Ratings		Cov (r)
Source	**Mother**	**Father**	**M-F**	**Mother**	**Father**	**M-F**
A	**.268**	**.138**	**.134 (.70)**	**.269**	**.121**	**.181 (*1.0*)**
C	**.194**	**.291**	**.099 (.42)**	**.077**	**.035**	**.051 (*1.0*)**
Bias	—	—	—	**.102**	**.259**	**.000 (*.00*)**
C + Bias	**.194**	**.291**	**.099 (.42)**	**.179**	**.294**	**.051 (.22)**
E	**.108**	**.181**	**.031 (.22)**	**.036**	**.016**	**.024 (*1.0*)**
Residual	—	—	—	**.074**	**.175**	**.000 (.00)**
E + Residual	**.108**	**.181**	**.031 (.22)**	**.110**	**.191**	**.024 (.17)**
Phenotypic Total	**.564**	**.609**	**.264 (.45)**	**.558**	**.606**	**.256 (.44)**

Italicized numbers indicate parameters are fixed *ex hypothesi* in the rater bias model.

expected contributions of A, C, and E to the variance of mothers' ratings, fathers' ratings, and the covariances between mothers' and fathers' ratings. What Table 11.4 shows is that, providing the rater bias model is adequate, we can partition the environmental variance of mothers' and fathers' ratings into variance attributable to those effects consistently rated by both parents and those effects which either represent rater bias or residual unreliable environmental variance. In this particular case, while a univariate consideration of maternal ratings would suggest a heritability of 47% $[= .263/(.263 + .194 + .108)]$**, a shared environmental influence of 34%, and a non-shared environmental influence of 19%, it is clear that more than half of the shared environmental influence can be attributed to rater bias, and the major portion of the non-shared environmental influence to unreliability or inconsistency between ratings. The heritability of internalizing behaviors in young boys rated consistently by both parents may be as high as 70%** $[= .269/(.269 + .077 + .036)]$**.**

11.2.6 Discussion of CBC Application

The data we have analyzed are restricted to parental checklist reports of their twin children's behavior problems, without the benefit of self re-

ports, teachers' reports or clinical interviews. As such, they are limited by the ability of parents to provide reliable and valid integrative assessments of their children, using cursorily defined concepts like 'Sulks a lot,' 'Worrying,' or 'Fears going to school.' It is clear from meta-analyses of intercorrelations of ratings of children by different types informants that while the level of agreement between mothers and fathers is often moderate (e.g., yielding correlations around .5 to .6) the level of agreement between parents and other informants (e.g., parent with child or parent with teacher) is modest and generally yields a correlation around 0.2 to 0.3 (Achenbach *et al.*, 1987). Thus, parental consistency in evaluating their children does not guarantee cross situational validity, although it does provide evidence that ratings of behavior observable by parents are not simply reflecting individual rater biases. In assessing the importance of the home environment on children's behavior this becomes a critical issue since studies of children's behavior based on ratings by a single individual in each family, e.g., the mother, confound the rater bias with the influence of the home environment. This may have the dual effect of inflating global estimates of the home environment's influence while at the same time either attenuating the relationship between objective indices of the environment and children's behavior (which is being assessed by a biased observer) or spuriously augmenting apparent relationships which are in fact relationships between environmental indices and maternal or paternal rating biases.

An issue distinguishable from that of bias is that of behavior sampling or situational specificity. Thus maternal and paternal ratings of children may differ not because of the tendency of individual parents to rate children in general as more or less problematic (bias), but because they are exposed to different samples of behavior. If this is so, then treating informants' ratings as if they were assessing a common phenotype, albeit in a biased or unreliable way, will be misleading. It is of considerable psychological importance to know whether different observers are being presented with different behaviors. The approach outlined in this chapter first enables us to examine the adequacy of the assumption that different informants are assessing the same behaviors and then, if that assumption is deemed adequate, to separate the contributions of rater bias and unreliability from the genetic and environmental contribution to the common behavioral phenotype. For our particular example, all the models fit our data adequately and the bias model, even in its restricted version, does not fit significantly worse than the psychometric or biometric models.

Although not presented here, there is some evidence that for externalizing behavior mothers and fathers cannot be assumed to be simply assessing the same phenotype with bias. In this context it is worth noting, however, that the adequacy of the assumption that parents are assessing the same phenotype in their children does not imply a high parental correlation (which may be lowered by bias and unreliability) and, conversely, even though parents may be shown to be assessing different phenotypes in their children to a significant degree the parental correlation in assessments may predominate over variance specific to a given parent. Our comparison of the bias with the psychometric and biometric models provides important evidence of the equivalence of the internalizing behaviors assessed by mothers and fathers using this instrument. This equivalence does not preclude bias or unreliability and the evidence presented in Table 11.4 provides a striking illustration of the impact of these sources of variation on maternal or paternal assessments. A shared environmental component which might be estimated to account for 34% of variance if mothers' ratings alone were considered, may correspond to only 20% of the variance when maternal biases have been removed. Similarly, a non-shared environmental variance component of 19% of variance may correspond to 9% of variance in individual differences between children that can be consistently rated by both parents. Finally, once allowance has been made for bias and inconsistency or unreliability, the estimated heritability rises from 47% to 70% in this case.

We have not been concerned here to seek the most parsimonious submodel within each of the model types. We should be aware that although we have, for the younger children, presented the full models with sex limitation, differences between boys and girls are not necessarily significant (for example, although the biometric model without sex limitation fit our data significantly worse than the corresponding model allowing for sex limitation ($\chi^2_9 = 21.31$, $p < .05$), the overall fit without sex limitation is still adequate, $\chi^2_{41} = 42.26$). Furthermore, individual parameter estimates reported for our full models may not depart significantly from zero. Other limitations of the method are that it does not allow for interaction effects between parents and children[3] and, in our application, assumed the independence of maternal and paternal biases. The analysis of parental bias under this model requires that both parents rate each of two children. Distinguishing between correlated parental biases and shared environmental influences would require a third, independent, rater (e.g., a teacher); thus we cannot rule out a contribution of correlated biases to our estimates of the remaining shared family environmental influence.

The final caveat against overinterpretation of particular parameter

[3]However, if these effects were substantial and if MZ twins correlated more highly than DZ twins in their interactional style, the variance of parents' ratings should differ (Neale *et al.*, 1992). Given sufficient sample size, these effects would lead to failure of these models.

estimates is that we have reported analyses for families in which both parents have returned a questionnaire and we have made no distinction between different biological or social parental statuses. Clearly, we anticipate that the inclusion and exclusion criteria are not neutral with respect to children's behavior problems and their perception by parents. However, we have illustrated that behavior genetic analyses are possible even when we have to rely on ratings by observers, providing that we have at least two degrees of relatedness among those being rated (e.g., MZ and DZ twins). Without an approach of this sort we have no way of establishing whether parents are assessing the same behaviors in their children and whether analyses will spuriously inflate estimates of the shared environment as much as parental biases inflate the correlations for pairs of twins independent of zygosity. Extension of the model to include other raters, for example, teachers, is straightforward.

Bibliography

Achenbach, T. M. & Edelbrock, C. S. (1981). *Behavior problems and competencies reported by parents of normal and disturbed children age four through sixteen. Monographs of the Society for Research in Child Development*. Number 188 in 46.

Achenbach, T. M. & Edelbrock, C. S. (1983). *Manual for the Child Gehavior Checklist and Revised Child Behavior Profile*. Burlington, VT: University of Vermont Dept. of Psychiatry.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213–232.

Aitken, A. C. (1934). Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society B*, *4*, 106–110.

Akaike, H. (1987). Factor analysis and aic. *Psychometrika*, *52*, 317–332.

Australian Bureau of Statistics (1977). *Alcohol and tobacco consumption patterns: February 1977* (catalogue no. 4312.0. ed.). Australian Bureau of Statistics.

Bedford, A., Foulds, G. A., & Sheffield, B. F. (1976). A new personal disturbance scale (DSSI/SAD). *British Journal of Social Clinical Psychology*, *15*, 387–394.

Bock, R. D. & Vandenberg, S. G. (1968). Components of heritable variation in mental test scores. In S. G. Vandenberg (Ed.), *Progress in human behavior genetics*, (pp. 233–260). The Johns Hopkins Press: Baltimore.

Bodmer, W. F. (1987). HLA, immune response, and disease. In F. Vogel & K. Sperling (Eds.), *Human Genetics: Proceedings of the 7th international congress*, (pp. 107–113). Springer-Verlag: New York.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* New York: John Wiley.

Boomsma, D. I. & Molenaar, P. C. M. (1986). Using lisrel to analyze genetic and environmental covariance structure. *Behavior Genetics*, *16*, 237–250.

Boomsma, D. I. & Molenaar, P. C. M. (1987). The genetic analysis of repeated measures. *Behavior Genetics*, *17*, 111–123.

Boyer, C. B. (1985). *A history of mathematics.* Princeton, New Jersey: Princeton University Press.

Bray, J. A. (1976). *The Obese Patient.* Philadelphia: W. B. Saunders.

Cantor, R. M. (1983). A multivariate genetic analysis of ridge count data from the offspring of monozygotic twins. *Acta Geneticae Medicae et Gemellologiae*, *32*, 161–208.

Cardon, L. R. & Fulker, D. W. (1992). Genetic influences on body fat from birth to age 9. *Genetic Epidemiology.* (in press).

Carey, G. (1986a). A general multivariate approach to linear modeling in human genetics. *American Journal of Human Genetics*, *39*, 775–786.

Carey, G. (1986b). Sibling imitation and contrast effects. *Behavior Genetics*, (pp. 319–341).

Carey, G. (1988). Inference about genetic correlations. *Behavior Genetics*, *18*, 329–338.

Castle, W. E. (1903). The law of heredity of Galton and Mendel and some laws governing race improvement by selection. *Proceedings of the American Academy of Sciences*, *39*, 233–242.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*, 1–22.

Cavalli-Sforza, L. L. & Feldman, M. (1981). *Cultural Transmission and Evolution: A Quantitative Approach.* Princeton: Princeton University.

Cloninger, C. R. (1980). Interpretation of intrinsic and extrinsic structural relations by path analysis: Theory and application to assortative mating. *Genetic Research*, *36*, 133–145.

Cloninger, C. R., Rice, J., & Reich, T. (1979a). Multifactorial inheritance with cultural transmission and assortative mating. II. A general model of combined polygenic and cultural inheritance. *American Journal of Human Genetics*, *31*, 176–198.

Cloninger, C. R., Rice, J., & Reich, T. (1979b). Multifactorial inheritance with cultural transmission and assortative mating. III. Family structure and the analysis of separation experiments. *American Journal of Human Genetics*, *31*, **366–388.**

Corey, L. A., Eaves, L. J., Mellen, B. G., & Nance, W. E. (1986). Testing for developmental changes in gene expression on resemblance for quantitative traits in kinships of monozygotic twins. *Genetic Epidemiology*, *3*, **73–83.**

Cox, A. & Rutter, M. (1985). Diagnostic appraisal and interviewing. In M. Rutter & L. Hersor (Eds.), *Child and adolescent psychiatry*. **Blackwell: Oxford, (2nd ed.).**

Crow, J. F. & Kimura, M. (1970). *Introduction to Population Genetics Theory*. **New York: Harper and Row.**

Darlington, C. D. (1971). Axiom and process in genetics. *Nature*, *234*, **131–133.**

Dawkins, R. (1982). *The Extended Phenotype: The Gene as the Unit of Selection*. **Oxford: Oxford University Press.**

Duffy, D. L. & Martin, N. G. (1992). Inferring the direction of causation in cross-sectional twin data: theoretical and empirical considerations. *Genetic Epidemiology*, *In press*.

Duffy, D. L., Martin, N. G., Battistutta, D., Hopper, J. L., & Mathews, J. D. (1990). Genetics of asthma and hayfever in australian twins. *American Review of Respiratory Disease*, *142*, **1351–1358.**

Eaves, L. J. (1976a). The effect of cultural transmission on continuous variation. *Heredity*, *37*, **41–57.**

Eaves, L. J. (1976b). A model for sibling effects in man. *Heredity*, *36*, **205–214.**

Eaves, L. J. (1982). The utility of twins. In V. Anderson, et al (Ed.), *Genetic Bases of the Epilepsies*. **New York: Raven Press.**

Eaves, L. J., Fulker, D. W., & Heath, A. C. (1989). The effects of social homogamy and cultural inheritance on the covariances of twins and their parents. *Behavior Genetics*, *19*, **113–122.**

Eaves, L. J., Heath, A. C., Neale, M. C., Hewitt, J. K., & Martin, N. G. (1992). Sex differences and non-additivity in the effects of genes on personality. *Psychological Science*. **(in press).**

Eaves, L. J., Hewitt, J. K., Meyer, J. M., & Neale, M. C. (1990). Approaches to quantitative genetic modeling of development and age-related changes. In M. E. Hahn, J. K. Hewitt, N. D. Henderson, & R. Benno (Eds.), *Developmental Behavior Genetics. Neural, Biometrical and Evolutionary Approaches*, (pp. 266–277). Oxford University Press: Oxford.

Eaves, L. J., Last, K. A., Martin, N. G., & Jinks, J. L. (1977). A progressive approach to non-additivity and genotype-environmental covariance in the analysis of human differences. *British Journal of Mathematical and Statistical Psychology*, *30*, 1–42.

Eaves, L. J., Last, K. A., Young, P. A., & Martin, N. G. (1978). Model-fitting approaches to the analysis of human behavior. *Heredity*, *41*, 249–320.

Eaves, L. J., Long, J., & Heath, A. C. (1986). A theory of developmental change in quantitative phenotypes applied to cognitive development. *Behavior Genetics*, *16*, 143–162.

Eaves, L. J., Neale, M. C., & Meyer, J. M. (1991). A model for comparative ratings in studies of within-family differences. *Behavior Genetics*, *21*, 531–536.

Falconer, D. S. (1960). *Quantitative Genetics*. Edinburgh: Oliver and Boyd.

Falconer, D. S. (1990). *Introduction to Quantitative Genetics* (3rd ed.). New York: Longman Group Ltd.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Translations of the Royal Society, Edinburgh*, *52*, 399–433.

Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Notices of the Royal Astronomical Society*, *80*, 758–770.

Fulker, D. W. (1982). Extensions of the classical twin method. In *Human genetics, part A: The unfolding genome* (pp. 395–406). New York: Alan R. Liss.

Fulker, D. W. (1988). Genetic and cultural transmission in human behavior. In B. S. Weir, E. J. Eisen, M. M. Goodman, & G. Namkoong (Eds.), *Proceedings of the second international conference on quantitative genetics* (pp. 318–340). Sunderland, MA: Sinauer.

Fulker, D. W., Baker, L. A., & Bock, R. D. (1983). Estimating components of covariance using LISREL. *Data Analyst*, *1*, 5–8.

Fuller, J. L. & Thompson, W. R. (1978). *Foundations of Behavior Genetics*. St. Louis: C. V. Mosby.

Gill, P. E., Murray, W., , & Wright, M. H. (1981). *Practical Optimization*. New York: Academic Press.

Gorsuch, R. L. (1983). *Factor Analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Graybill, F. A. (1969). *Introduction to Matrices with Applications in Statistics*. Belmont, CA: Wadsworth Publishing Company.

Grayson, D. A. (1989). Twins reared together: Minimizing shared environmental effects. *Behavior Genetics*, *19*, 593–603.

Grilo, C. M. & Pogue-Guile, M. F. (1991). The nature of environmental influences on weight and obesity: A behavior genetic analysis. *Psychological Bulletin*, *110*, 520–537.

Haley, C. S., Jinks, J. L., & Last, K. (1981). The monozygotic twin half-sib method for analyzing maternal effects and sex-linkage in humans. *Heredity*, *46*, 227–238.

Harman, H. H. (1976). *Modern Factor Analysis*. Chicago: University of Chicago Press.

Heath, A. C. (1983). *Human Quantitative Genetics: Some Issues and Applications*. Unpublished doctoral dissertation, University of Oxford, Oxford, England.

Heath, A. C. (1987). The analysis of marital interaction in cross-sectional twin data. *Acta Geneticae Medicae et Gemellologiae*, *36*, 41–49.

Heath, A. C., Jardine, R., & Martin, N. G. (1989a). Interactive effects of genotype and social environment on alcohol consumption in female twins. *Journal of Studies on Alcohol*, *50*, 38–48.

Heath, A. C., Kendler, K. S., Eaves, L. J., & Markell, D. (1985). The resolution of cultural and biological inheritance: Informativeness of different relationships. *Behavior Genetics*, *15*, 439–465.

Heath, A. C. & Martin, N. G. (1986). Detecting the effects of genotype $\times$ environment interaction on personality and symptoms of anxiety and depression. *Behavior Genetics*, *16*, 622.

Heath, A. C., Neale, M. C., Hewitt, J. K., Eaves, L. J., & Fulker, D. W. (1989b). Testing structural equation models for twin data using LISREL. *Behavior Genetics*, *19*, 9–36.

Heise, D. R. (1975). *Causal Analysis*. New York: Wiley-Interscience.

Helzer, J. E., Robins, L. N., Taibleson, M., Woodruff, R. A., Reich, T., & Wish, E. D. (1977). Reliability of psychiatric diagnosis. *Archives of General Psychiatry*, *34*, 129–133.

Hewitt, J. K. (1989). Of biases and more in the study of twins reared together: A reply to Grayson. *Behavior Genetics*, *19*, 605–610.

Hewitt, J. K., Silberg, J. L., & Erickson, M. (1990). Genetic and environmental influences on internalizing and externalizing behavior problems in childhood and adolescence. *Behavior Genetics*, *20*, 725. (abstract).

Hewitt, J. K., Silberg, J. L., Neale, M. C., Eaves, L. J., & Erickson, M. (1992). The analysis of parental ratings of children's behavior using LISREL. *Behavior Genetics*. (in press).

IMSL (1987). *IMSL User's Manual. Version 1.0*. Houston, Texas: IMSL, Inc.

Jardine, R. (1985). *A twin study of personality, social attitudes, and drinking behavior*. Unpublished doctoral dissertation, Australian National University, Australia.

Jeffrey, D. B. & Knauss, M. R. (1981). The etiologies, treatments, and assessments of obesity. In S. N. Haynes & L. Gannon (Eds.), *Psychosomatic disorders: A psychophysiological apprach to etiology and treatment*. New York: Praeger.

Jencks, C. (1972). *Inequality: A Reassessment of the Effect of Family and Schooling in America*. New York: Basic Books.

Jinks, J. L. & Fulker, D. W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychological Bulletin*, *73*, 311–349.

Jöreskog, K. G. & Sörbom, D. (1989). *LISREL 7: A Guide to the Program and Applications* (2nd ed.). Chicago: SPSS, Inc.

Kempthorne, O. (1960). *Biometrical Genetics*. New York: Pergammon Press.

Kendler, K. S., Heath, A. C., Martin, N. G., & Eaves, L. J. (1986). Symptoms of anxiety and depression in a volunteer twin population: The etiologic role of genetic and environmental factors. *Archives General Psychiatry*, *43*, 213–221.

Kendler, K. S., Heath, A. C., Martin, N. G., & Eaves, L. J. (1987). Symptoms of anxiety and symptoms of depression: Same genes, different environments? *Archives General Psychiatry*, *44*, 451–457.

Kendler, K. S. & Kidd, K. K. (1986). Recurrence risks in an oligogenic threshold model: The effect of alterations in allele frequency. *Annals Human Genetics*, *50*, 83–91.

Kendler, K. S., Neale, M. C., Heath, A. C., Kessler, R. C., & Eaves, L. J. (1991). Life events and depressive symptoms: A twin study perspective. In P. McGuffin & R. Murray (Eds.), *The New Genetics of Mental Illness* (pp. 144–162). London: Butterworth-Heinemann.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C., & Eaves, L. J. (1992a). Childhood parental loss and adult psychopathology in women: A twin study perspective. *Archives General Psychiatry*, *49*, 109–116.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C., & Eaves, L. J. (1992b). Generalized anxiety disorder in women: A population based twin study. *Archives General Psychiatry*. (in press).

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C., & Eaves, L. J. (1992c). Major depression and generalized anxiety disorder: Same genes, (partly) different environments? *Archives General Psychiatry*. (in press).

Kenny, D. A. (1979). *Correlation and Causality*. New York: Wiley-Interscience.

Lawley, D. N. & Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*. London: Butterworths.

Li, C. C. (1975). *Path Analysis: A Primer*. Pacific Grove, CA: Boxwood Press.

Loeber, R., Green, S. M., Lahey, B., & Stouthamer-Loeber, M. (1989). Optimal informants on childhood disruptive behaviors. *Developmental Psychopathology*, *1*, 317–337.

Loehlin, J. C. (1987). *Latent Variable Models*. Baltimore: Lawrence Erlbaum.

Loehlin, J. C. & Vandenberg, S. G. (1968). Genetic and environmental components in the covariation of cognitive abilities: An additive model. In S. G. Vandenberg (Ed.), *Progress in human behavior genetics*, (pp. 261–285). Johns Hopkins University Press: Baltimore.

Lykken, D. T., McGue, M., & Tellegen, A. (1987). Recruitment bias in twin research: the rule of two-thirds reconsidered. *Behavior Genetics*, *17*, 343–362.

Lytton, H. (1977). Do parents create, or respond to differences in twins? *Developmental Psychology*, *13*, 456–459.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis*. New York: Academic Press.

Martin, N. G. & Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity*, *38*, 79–95.

Martin, N. G., Eaves, L. J., Heath, A. C., Jardine, R., Feindgold, L. M., & Eysenck, H. J. (1986). Transmission of social attitudes. *Proceedings of the National Academy of Science*, *83*, 4364–4368.

Martin, N. G., Eaves, L. J., Kearsey, M. J., & Davies, P. (1978). The power of the classical twin study. *Heredity*, *40*, 97–116.

Martin, N. G., Eaves, L. J., & Loesch, D. Z. (1982). A genetical analysis of covariation between finger ridge counts. *Annals Human Biology*, *9*, 539–552.

Martin, N. G. & Jardine, R. (1986). Eysenck's contribution to behavior genetics. In S. Modgil & C. Modgil (Eds.), *Hans Eysenck: Consensus and Controversy*. Falmer Press: Lewes, Sussex.

Martin, N. G., Oakeshott, J. G., Gibson, J. B., Starmer, G. A., Perl, J., & Wilks, A. V. (1985). A twin study of psychomotor and physiological responses to an acute dose of alcohol. *Behavior Genetics*, *15*, 305–347.

Mather, K. & Jinks, J. L. (1971). *Biometrical Genetics*. London: Chapman and Hall.

Mather, K. & Jinks, J. L. (1977). *Introduction to Biometrical Genetics*. Ithaca, New York: Cornell University Press.

Mather, K. & Jinks, J. L. (1982). *Biometrical genetics: The Study of Continuous Variation* (3rd ed.). London: Chapman and Hall.

Maxwell, A. E. (1977). *Multivariate Analysis in Behavioral Research*. New York: John Wiley.

McArdle, J. J., Connell, J. P., & Goldsmith, H. H. (1980). Structural modeling of stability and genetic influences: Some results from a longitudinal study of behavioral style. *Behavior Genetics*, *10*, 487.

McArdle, J. J. & Goldsmith, H. H. (1990). Alternative common-factor models for multivariate biometric analyses. *Behavior Genetics*, *20*, 569–608.

Molenaar, P. C. M. & Boomsma, D. I. (1987). Application of nonlinear factor analysis to genotype-environment interaction. *Behavior Genetics*, *17*, 71–80.

Mood, A. M. & Graybill, F. A. (1963). *Introduction to the Theory of Statistics* (2nd ed.). New York: McGraw-Hill.

Mulaik, S. A., James, L. R., VanAlstine, J., Bennett, N. Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equations models. *Psychological Bulletin*, *105*, 430–445.

Muthen, B. O. (1987). *LISCOMP: Analysis of Linear Structural Equations with a Comprehensive Measurement Model*. Mooresville, IN: Scientific Software, Inc.

NAG (1990). *The NAG Fortran Library Manual, Mark 14*. Oxford: Numerical Algorithms Group.

Nance, W. E. & Corey, L. A. (1976). Genetic models for the analysis of data from the families of identical twins. *Genetics*, (pp. 811–825).

Neale, M. C. (1988). Handedness in a sample of volunteer twins. *Behavior Genetics*, *18*, 69–79.

Neale, M. C. (1991). *Mx: Statistical Modeling*. Box 3 MCV, Richmond, VA 23298: Department of Human Genetics.

Neale, M. C. & Eaves, L. J. (1992). Estimating and controlling for the effects of volunteer bias with pairs of relatives. *Behavior Genetics*, *in press*.

Neale, M. C., Heath, A. C., Hewitt, J. K., Eaves, L. J., & Fulker, D. W. (1989). Fitting genetic models with LISREL: Hypothesis testing. *Behavior Genetics*, *19*, 37–69.

Neale, M. C. & Martin, N. G. (1989). The effects of age, sex and genotype on self-report drunkenness following a challenge dose of alcohol. *Behavior Genetics*, *19*, 63–78.

Neale, M. C., Rushton, J. P., & Fulker, D. W. (1986). The heritability of items from the eysenck personality questionnaire. *Personality and Individual Differences*, *7*, 771–779.

Neale, M. C. & Stevenson, J. (1989). Rater bias in the EASI temperament scales: A twin study. *Journal of Personality and Social Psychology*, (pp. 446–455).

Ott, J. (1985). *Analysis of Human Genetic Linkage*. Baltimore, MD: Johns Hopkins University Press.

Pearson, E. S. & Hartley, H. O. (1972). *Biometrika Tables for Statisticians, volume* 2. Cambridge: Cambridge University Press.

Pearson, K. (1904). On a generalized theory of alternative inheritance, with special references to Mendel's laws. *Phil. Trans. Royal Society A*, *203*, 53–86.

Plomin, R. & Bergeman, C. S. (1991). The nature of nurture: Genetic influence on environmental measures. *Behavior and Brain Sciences*, *14*, 373–397.

Rao, D. C., Morton, N. E., & Yee, S. (1974). Analysis of family resemblance II. A linear model for familial correlation. *American Journal of Human Genetics*, *26*, 331–359.

Rice, J., Cloninger, C. R., & Reich, T. (1978). Multifactorial inheritance with cultural transmission and assortative mating. I. Description and basic properties of the unitary models. *American Journal of Human Genetics*, *30*, 618–643.

Riskind, J. H., Beck, A. T., Berchick, R. J., Brown, G., & Steer, R. A. (1987). Reliability of DSM-III diagnoses for major depression and generalized anxiety disorder using the Structured Clinical Interview for DSM-III. *Archives General Psychiatry*, *44*, 817–820.

SAS (1985). *SAS/IML User's Guide, Version 5 edition*. Cary, NC: SAS Institute.

SAS (1988). *SAS/STAT User's guide: Release 6.03*. Cary, NC: SAS Institute, Inc.

Schieken, R. M., Eaves, L. J., Hewitt, J. K., Mosteller, M., Bodurtha, J. M., Moskowitz, W. B., & Nance, W. E. (1989). Univariate genetic analysis of blood pressure in children: the mcv twin study. *American Journal of Cardiology*, *64*, 1333–1337.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics.* New York: John Wiley.

Silberg, J. L., Erickson, M. T., Eaves, L. J., & Hewitt, J. K. (1992). The contribution of environmental factors to maternal ratings of behavioral and emotional problems in children and adolescents. (manuscript in preparation).

Spence, J. E., Corey, L. A., Nance, W. E., Marazita, M. L., Kendler, K. S., & Schieken, R. M. (1988). Molecular analysis of twin zygosity using VNTR DNA probes. *American Journal of Human Genetics*, *43(3)*, A159 (Abstract).

Spitzer, R. L., Williams, J. B., & Gibbon, M. (1987). *Structured Clinical Interview for DSM-III-R.* New York: Biometrics Research Dept. and New York State Psychiatric Institute.

SPSS (1988). *SPSS-X User's Guide* (3rd ed.). Chicago: SPSS Inc.

Stunkard, A. J., Foch, T. T., & Hrubec, Z. (1986). The body-mass index of twins who have been reared apart. *New England Journal of Medicine*, *314*, 193–198.

van Eerdewegh, P. (1982). *Statistical selection in multivariate systems with applications in quantitative genetics.* Unpublished doctoral dissertation, Washington University, St. Louis.

Vandenberg, S. G. (1965). Multivariate analysis of twin differences. In S. G. Vandenberg (Ed.), *Methods and goals in human behavior genetics*, (pp. 29–43). Academic Press: New York.

Vlietinck, R., Derom, R., Neale, M. C., Maes, H., Van Loon, H., Van Maele, G., Derom, C., & Thiery, M. (1989). Genetic and environmental variation in the birthweight of twins. *Behavior Genetics*, *19*, 151–161.

Vogler, G. P. (1982). Multivariate behavior genetic analyses of correlations vs. phenotypically standardized covariances. *Behavior Genetics*, *12*, 473–478.

Vogler, G. P. (1985). Multivariate path analysis of familial resemblance. *Genetic Epidemiology*, *2*, 35–53.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*, 557–585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, *5*, 161–215.

Wright, S. (1960). The treatment of reciprocal interaction, with or without lag, in path analysis. *Biometrics*, *16*, **189–202.**

Wright, S. (1968). *Evolution and the Genetics of Populations. Volume 1. Genetic and Biometric Foundations.* **Chicago: University of Chicago Press.**

Yule, G. U. (1902). Mendel's laws and their probable relation to intra-racial heredity. *New Phytology*, *1*, **192–207.**