# Chapter 1

# Introduction

This book is intended to be a practical guide for the analysis of longitudinal behavioral data. Longitudinal data consist of repeated measures collected on the same subjects over time. Such data is collected by researchers in psychology, education, organization studies, public policy, and related fields. A variety of substantive research questions are addressed with longitudinal data, including how student achievement changes over time, how psychopathology develops, and how intra-group conflict evolves.

There are practical issues associated with longitudinal data collection that impact analysis. Longitudinal data are a special case of repeated measures data, with duration or time constituting the dimension over which the measurements are collected. In contrast to repeated measurements across experimental conditions, the time dimension does not allow the order of presentation to be randomized or counterbalanced. Consequently, longitudinal data have characteristics that require a flexible statistical model for analysis. Some of the traditional methods, such as repeated measures analysis of variance (RM-ANOVA), are not as flexible as the more modern methods discussed in this book.

A common feature of longitudinal data is that the variance of the response variable tends to change over time. It is common in behavioral studies for individuals to change at different rates, leading to increasing or decreasing dispersion as the study progresses. It seems important then, to choose a statistical model that allows for fanning-out or squeezing-in of the data points over time.

Another feature of longitudinal data is the tendency of observations more closely spaced to have a higher correlation than observations more distantly spaced. One reason for this phenomenon is the presence of intervening factors that affect responses, such as life events of the subjects. When the repeated measurements are closely spaced, intervening factors are expected to change little and have a relatively small influence. This results in a relatively high correlation for response variable scores at adjacent time points. When the repeated measurement are not closely spaced, the intervening factors have a greater chance of influencing responses, leading to decreased correlations.

The timing of the observations, then, has an influence on the correlations. Measurement times are usually determined by the content under study. However, available resources or happenstance might also play a role. The frequency of observations is usually set by the researcher to be sensitive to "real" change in the response variable. Observations too closely spaced might not allow for the observation of phenomena that take time to manifest. On the other hand, observations too distantly spaced may miss important intervening landmarks of change or development. It follows that the researcher should think carefully about the intervals of measurement and whether they should be a matter of seconds, minutes, days, months, or years.

A complication in having a relatively long duration between repeated measurements is an increased potential for missing data. As time elapses, subjects can grow weary of being measured, or they can have significant life events, such as changes in employment status or marital status. Such events can precipitate dropout from the study. When a subject drops out, they might not re-enter the study and they might not be replaced. Study resources are usually not allocated for replacement subjects or for severely staggered start times. This is especially so when the dropout is late in a study that spans a relatively long period of time, such as several years.

From an applied data analysis perspective, it is desirable to have a statistical method that can accommodate missing data and adequately account for the typical pattern of variances and correlations among the

repeated measures. One method that addresses both issues is *linear mixed effects regression* (LMER), which is the primary analysis tool discussed in this book. (LMER is pronounced as "el-mer".)

In addition to providing an appropriate statistical model for longitudinal data, LMER allows for the examination of predictors of change, sometimes referred to as covariates of change. In behavioral science analysis, the term *covariate* is closely associated with a control variable used in the analysis of covariance (ANCOVA). Here the term is used more broadly and refers to any type of regression predictor.

Predictors of change are vital when a researcher wants to know about conditional change, or change that varies according to the values of one or more covariates. For example, a researcher might want to know if change in reading achievement is conditional on gender. Such an analysis involves an examination of the reading achievement trajectory of males to determine if it is substantially different than the trajectory of females. In another example, a researcher might want to examine if trajectories of externalizing behavior vary based on the quality of parenting. Individuals with relatively low quality of parenting might have higher levels and higher rates of growth than individuals with relatively high quality. Research questions such as these are common in behavioral research and LMER is well-suited for addressing them.

## 1.1 Statistical Computing

Statistical computing is broadly defined as any use of a computer program for manipulating and analyzing data. Consistent with the practical goals of this book, statistical computing is discussed in detail. Rather than relegate computer examples to the end of chapters or a remote website, the approach is to interweave the examples within the narrative of the book. In future chapters, statistical concepts will be presented immediately followed by illustrations of how the related analysis can be carried out.

A single data set is used throughout much of the book, presented in table form later in this chapter and also available on the book website. The intention is that the reader can duplicate what is seen in the book, perhaps in a dynamic fashion by having the book and computer side-by-side. This will help the reader to gain insights into both the statistical methods and statistical computing.

True to the title of this book, the freely available `R` program (`R` Development Core Team, 2010) is used for all computer examples. Among other features, `R` has the advantage of accessibility. Anyone with an Internet connection has access to this very powerful program for statistical computing and graphics. There are free add-on packages for manipulating, graphing, and analyzing longitudinal data that are among the best available in any statistical computing program.

Though `R` is widely used in the statistical sciences, it is not as widespread in the behavioral sciences. For this reason, it is assumed the reader has never used the program, and a primer on `R` is provided in Chapter 2. Rather than attempt to pack all the relevant concepts into the primer, many statistical computing concepts and methods are presented as the occasion warrants in later chapters.

## 1.2 Preliminary Issues

Every discipline provides a specific context for data analysis. In the analysis of longitudinal behavioral data, there are particular preliminary issues that are important to ponder prior to the data analysis, or even prior to data collection. Consideration of these preliminary issues can help a researcher determine what type of statistical analysis is appropriate for their data. A survey of these issues also helps provide an indication of the extent to which the data lend themselves to drawing valid inferences. If certain data conditions are met, then this can increase the likelihood of drawing sound conclusions from a LMER analysis.

### 1.2.1 Means versus Correlations

The primary focus of this book is examining change in the quantity or level of a response variable repeatedly measured over time. It is assumed the researcher has obtained repeated measurements on several individuals, and the primary goal of the analysis is to examine aggregate change. Throughout this text, aggregate change will be indexed by the arithmetic mean. Thus, mean change will be the focus of analysis.

In addition to means, correlations between pairs of time points are informative. Correlations indicate the strength of dependency of the response over time. The correlation matrix among all the time points is

useful as an indication of the strength of dependency, and how it is related to the spacing of observations. In many longitudinal data sets, observations more closely spaced have a higher correlation than observations more distantly spaced. When considering longitudinal data analysis, it is advantageous to use a model that accounts for such a pattern. As will be explained later, LMER is one such model.

Correlations are standardized covariances, and the diagonal of a correlation matrix consists of 1s. Sometimes in longitudinal analysis, covariances among the time points rather than correlations are discussed. Similar information about strength of dependency can be obtained with the covariance matrix, but it has the advantage that the diagonal elements are variances. Thus, the covariance matrix allows examination of change in variability over time in addition to between-time dependency.

Mean change might not be an obvious touchstone for some researchers, as there are many instances in the behavioral sciences when correlations rather than means are used in longitudinal analysis (e.g., Burt, Obradović, Long, & Masten, 2008; Masten et al., 2005). It seems valuable then, to highlight the types of information represented by means and correlations, and how this relates to the study of change.

A correlation cannot be used to make inferences about absolute – as opposed to relative – change in quantity. Correlations are based on standard scores or $z$-scores that have a mean value of 0. When computed for each time point, $z$-scores guarantee a constant mean of 0 over time, precluding meaningful examination of change in quantity or level.

When an analysis uses correlations and excludes means, the focus is not on the change in quantity, but on the change in relative position of subjects over time. Between-time correlations largely index changes in rank order of individuals. For this reason, the between-time correlation is known as a *stability coefficient* (McDonald, 1999, chap. 5). The extent of stability is reflected in the value of the coefficient, with higher values indicating greater persistence of rank-order over time.

The mean and correlation difference is illustrated with the made-up achievement data that appear in Figure 1.1. Each graph depicts achievement scores ($y$) measured at two time points for the same cohort of individuals. For each subject, the two repeated measures scores are connected by a line. The lines are called *growth curves* or *change curves*. In this book, these terms will be used to refer either to lines connecting observed scores, as in Figure 1.1, or lines connecting fitted or predicted scores, as discussed in later chapters.

The extent of the crossing lines is related to the strength of the stability coefficient, which is Pearson's correlation coefficient (cor) computed between the time points. The graphs in the first row show relatively few crossing lines, which indicates little rank order change and high stability. This is reflected in a high correlation, cor = 0.94. The graphs in the second row show extensive crossing of lines, which indicates low stability and a correlation that is close to 0, cor = 0.06.

A comparison of the left and right columns in Figure 1.1 illustrate that mean change is indifferent to the between-time correlation. For the graphs in a row of the figure, the correlations are identical, but the mean difference, mean diff $= \bar{y}_2 - \bar{y}_1$, is quite different. Figure 1.1a in the upper left has mean diff = 0.74, whereas Figure 1.1b in the upper right has mean diff = 10.74. The graphs in the bottom row, Figures 1.1c and 1.1d, show a greater contrast in mean difference for the same between-time correlation.

The right-hand graphs (Figures 1.1b and 1.1d) were created by starting with the left-hand graphs (Figures 1.1a and 1.1c) and adding a constant to each score at the second time point. A value of 10 was added for the top graphs and a value of 15 for the bottom graphs. Any value can be added, showing that any size of mean difference can be induce without changing the correlation between time points.

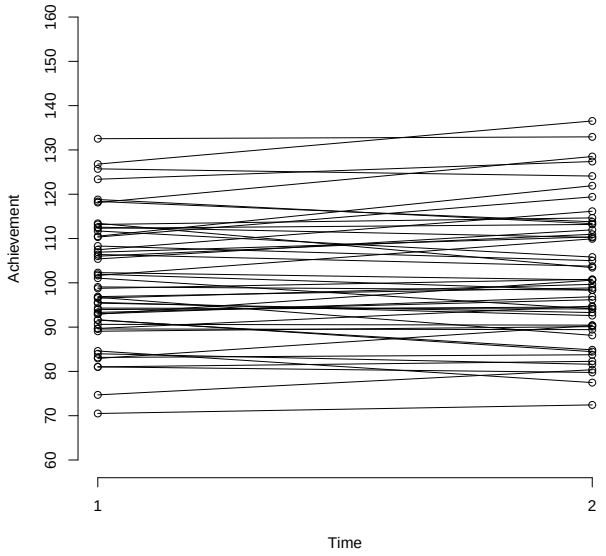
In addition to leaving the correlation undisturbed, adding a constant does not change the variance or the covariance among the time points. This shows that, like the correlation, the covariance is not informative about mean change. Though the covariances among the time points cannot be used to directly study mean change, it is shown later that they are important in an indirect sense. The covariances are used in computing the standard errors (SEs) of parameter estimates that directly index mean change.

What accounts for the crossing lines in Figure 1.1? One factor is variability in individual rate of change; people increase or decrease at different rates. Another factor is random measurement error. The scores from the instrument are not reliable, which causes inconsistent measurement over time. A statistical model is required to sort out the impact of these two influences. The LMER models considered in this book can be used to estimate the magnitude of the variability in rate of change and measurement unreliability from longitudinal sample data.

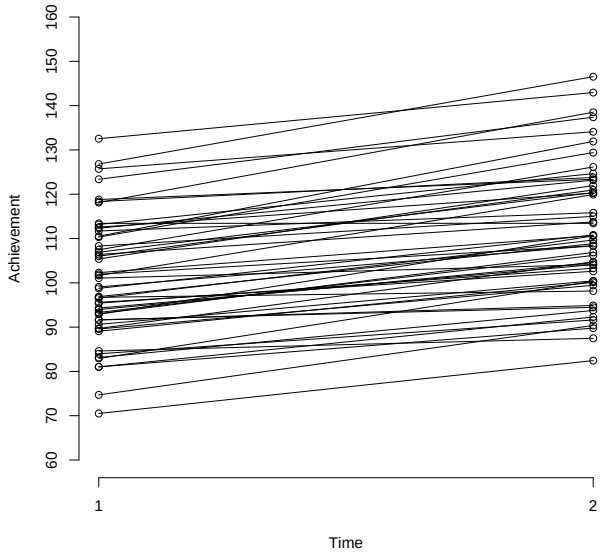
It is informative to examine both mean change and correlations (covariances) in a longitudinal analysis. As mentioned, some longitudinal analyses in the behavioral sciences involve only correlations. The focus

Figure 1.1: Illustration of longitudinal correlation and mean change.

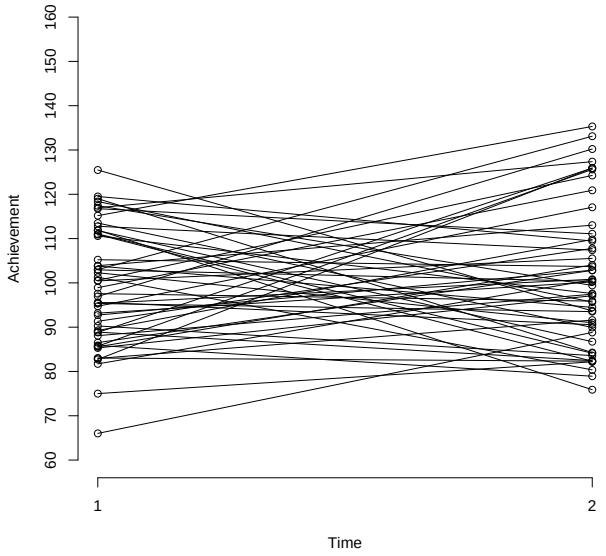
(a) Cor = 0.94, mean diff = 0.74



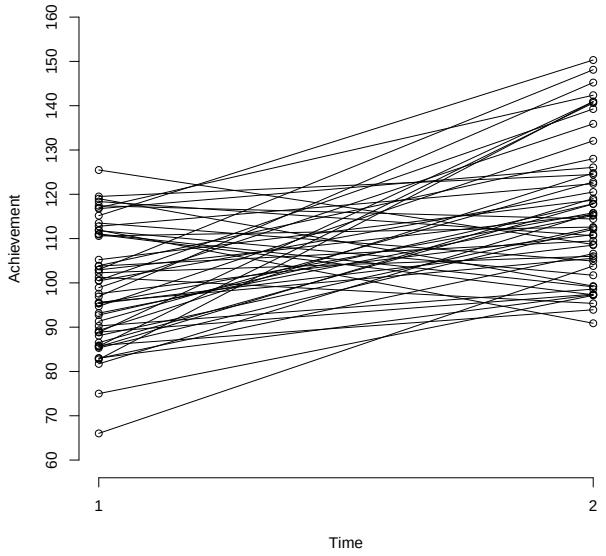

(b) Cor = 0.94, mean diff = 10.74




(c) Cor = 0.06, mean diff = 2.26




(d) Cor = 0.06, mean diff = 17.26




on correlations may be due to particular characteristics of the data, or the research design. One reason for considering correlations rather than means is to accommodate the switching of measures over the course of data collection. Measures are typically switched because subjects outgrow them. If the span of the data collection crosses landmark developmental thresholds, this can render the use of the same instrument developmentally inappropriate.

As an example of development inappropriateness, consider the longitudinal measurement of externalizing

behavior. Externalizing behavior is an outward display of a person's negative reaction to their environment, such as aggression or hyperactivity. The tracking of externalizing behaviors over time can be important in examining the development of psychopathology (Klimes-Dougan et al., 2010).

Suppose the goal is to assess the aggression of subjects who are tracked from 5 to 30 years old. For 5 year olds, aggression might be defined as the frequency at which the subject bites others while interacting on the playground. This seems to be a perfectly valid measure of aggression for young children, but it may be completely invalid for 30 year olds. Assuming a normative sample, biting and playgrounds seem to have no relevance for indexing externalizing behavior in adults. A different item should probably be used at some point, such as the ability of the subjects to get along with their boss at work. This idea generalizes to instruments made up of several items. The point is that researchers should ponder the appropriateness of their measures for indexing the same construct over time. This is especially pertinent if the observation period spans several years.

Switching instruments precludes a valid analysis of change in level, unless considerable ancillary work is accomplished prior to the analysis. The problem is that the instrument switch is confounded with change, making it difficult to draw unambiguous conclusions. Suppose a childhood externalizing behaviors measure is switched for an adult version over the course of the longitudinal study, when the subjects turn 18. Imagine that the means of the childhood instrument are relatively small in value, but the means of the adult instrument are relatively large. How is one to assess if this is an instance of true development or an artifact of switching instruments?

When instruments are switched over time, it is possible to study mean change, provided some type of vertical equating is used (Kolen & Brennan, 2004). In contrast to $z$-scores, vertical equating sets different instruments on a common scale, but the mean can change over time. Therefore, under certain assumptions, scores on the common scale can be validly used to track changes in the quantity of the construct. A potential complication is that proper equating typically requires a large sample size and knowledge of test theory, especially item response theory (IRT; McDonald, 1999, chap. 12). These requirements can be onerous for applied researchers. In addition, there is the drawback that substantial resources may be expended on an issue that is tangential to the major thrust of the research.

### 1.2.2 Measurement Issues

When the same instrument is used over the course of a longitudinal study, there still may be questions as to whether change in quantity can be validly studied. As mentioned, items of instruments can change their meaning as subjects age. For this reason, it cannot always be taken for granted that the same instrument will measure the same construct(s) as a study progresses (Long, Herring, Brekke, Test, & Greenberg, 2007).

It is possible to conduct empirical checks on the consistent measurement of a construct over time. These checks are typically based on longitudinal IRT models (Muthén & Christoffersson, 1981). Evidence of longitudinal validity is provided in the form of invariance of measurement parameters over time (e.g., difficulty and discrimination). Similar to equating, longitudinal validity analysis typically constitutes a preliminary or stand-alone analysis based on a large number of subjects. For examples of longitudinal validity analysis, see Long et al. (2007), Obradović, Pardini, Long, and Loeber (2007), Pitts, West, and Tein (1996), and Vandenberg and Lance (2000).

In addition to consistency in measurement, the analysis of mean change requires that the construct in question be measured in an absolute sense. Certain physical constructs, such as height, meet this criterion. The height of a single subject is determined in reference to an objective distance scale, for example, a meter stick. An individual's height is not influenced by the measurement of other subjects, at least not explicitly.

In contrast to physical constructs, there are certain behavioral constructs that have only a relative determination. This occurs when a construct is measured by assigning ratings for a subject based on their relative standing to other subjects. An example is when a teacher assigns a Likert rating to a student for social cooperation, sometimes called citizenship, relative to other students in a classroom.

Such ratings are difficult to interpret in an absolute sense, as they are affected by changes in the composition of the class. If students are added or removed from the class, it is plausible the assigned social cooperation for a given student could change, even though the student's true (unknown) absolute level remains unchanged. A student initially rated as relatively uncooperative might improve in ranking if uncooperative students happen to be added to the class. A student rated as relatively affable might be downgraded if some

apple-polishers come into the class. Due to this problem, relative scores appear to be an inadequate basis for studying mean change.

### 1.2.3 Response Variable Assumptions

In light of the above discussion, it is assumed throughout that the repeatedly measured response variable constitutes a valid foundation for making inferences about mean change. The researcher should carefully consider this assumption before commencing with an analysis. In the material to follow, variances, covariances, and correlations will be considered, but they will not have as high a priority as means. A proper model for the covariance matrix, for example, is required for proper inferences about mean change, but ultimate interest is in the nature of the change over time.

It is also assumed throughout that the response variable is quantitative or continuous. A quantitative variable has numbers that represent amounts of things. This is in contrast to categorical variables whose values denote different categories of membership, or count variables that record the frequency of an event. A generalized type of LMER can be used for the case of categorical and count response variables; for an introduction, see Fitzmaurice et al. (2004), chap. 10. Though the response is always considered to be quantitative in this book, the predictors can be quantitative, categorical, or a combination. Additional detail is provided below and in future chapters.

## 1.3 Conceptual Overview of Linear Mixed Effects Regression

The method of analysis featured in this book is LMER. The focus of LMER is mean change over time, but it also allows for the examination of individual variability in change. Furthermore, LMER incorporates a random error term that accounts for measurement unreliability. Components of the LMER model for longitudinal data are presented in Table 1.1. The symbols in the last column are used in the algebraic formulas presented in future chapters.

Table 1.1: Components of linear mixed effects regression.

| Component | Description | Level of Effect | Symbol |
|---|---|---|---|
| Fixed effects | Regression coefficients | Group | $\beta$ |
| Random effects | Individual deviations from fixed effects | Individual | $b$ |
| Random error | Regression error term | Individual | $\varepsilon$ |

LMER can be viewed as an extension of traditional multiple regression. The extension is made by introducing individual-level terms to the regression model that index variability among the subjects. These terms are known as *random effects*. The random effects are in contrast to the *fixed effects*, which are similar to the traditional regression coefficients. Fixed effects are constant among individuals and index group-level change. It is the mix of fixed and random effects that is the genesis of the term *linear mixed effects regression*. Random effects vary among individuals and index deviations from the group. In the context of longitudinal models, the random effects reflect variation among individual change curves.

The basic ideas of LMER can be illustrated with the small data set of Table 1.2. The four subjects depicted in the table are from a larger data set to be presented later in this chapter. The data are in *long format*, meaning the blocks of data for the subjects are stacked one atop another and separated by horizontal lines in the table. Chapter 3 provides a discussion of how traditionally structured data can be converted into long format. For the moment, the data are simply presented in this format.

There are five columns in the table, with the first being the subject identifier or subject variable, `subid`.[1] The response variable measured over time is reading achievement (`read`) in the second column. The third column has the grade at which reading achievement is measured (`grade`), and the predictor variables, gender (`gen`) and attendance proportion (`att`), are in the last two columns. The variable `grade` is considered to be a *time predictor*, as it is an index of duration. That is, `grade` is a proxy for annual measurement.

[1]The convention used throughout this book is that a variable name in an actual data set appears in `typewriter font`.

Table 1.2: Data of four subjects in long format.

| subid | read | grade | gen | att |
|---|---|---|---|---|
| 1 | 172 | 5.00 | F | 0.94 |
| 1 | 185 | 6.00 | F | 0.94 |
| 1 | 179 | 7.00 | F | 0.94 |
| 1 | 194 | 8.00 | F | 0.94 |
| 3 | 191 | 5.00 | M | 0.97 |
| 3 | 199 | 6.00 | M | 0.97 |
| 3 | 203 | 7.00 | M | 0.97 |
| 3 | 215 | 8.00 | M | 0.97 |
| 5 | 207 | 5.00 | F | 0.85 |
| 5 | 213 | 6.00 | F | 0.85 |
| 5 | 212 | 7.00 | F | 0.85 |
| 5 | 213 | 8.00 | F | 0.85 |
| 7 | 199 | 5.00 | M | 0.97 |
| 7 | 208 | 6.00 | M | 0.97 |
| 7 | 213 | 7.00 | M | 0.97 |
| 7 | 218 | 8.00 | M | 0.97 |

### 1.3.1 Goals of Inference

The general goal of regression analysis is to account for the variability of the response variable. In the case of longitudinal data there are two types of response variability, *within-subjects* and *between-subjects*. Within-subjects variability is represented by row-to-row change for individual subjects in Table 1.2. Between-subjects variability is represented by block-to-block change, with the blocks being demarcated by the horizontal lines in the table.

Within-subjects variability is due to changes in the response variable over time, that is, variability among the repeated measures. Within each block of Table 1.2, it can be seen that `read` changes over time for each subject. This row-to-row change indicates that within-subjects variability does exit for the response variable.

Between-subject variability is variation due to individual differences. This variability is evident when subjects differ on some characteristic that is not measured over time. For example, Table 1.2 shows that subjects vary in terms of their gender and attendance proportion. This block-to-block change indicates between-subjects variability does exist for the predictor variables.

Accounting for the types of variability constitutes two goals of inference in LMER. These goals are accomplished by the use of different types of predictors. The predictor variables in Table 1.2 account for the different types of variability based on whether their scores change over time. Any predictor that changes over time accounts for within-subjects variability, and any predictor that is constant over time – but not constant among subjects – accounts for between-subjects variability.

From Table 1.2, it can be seen that `grade` accounts for within-subject variability, whereas `gen` and `att` account for between-subject variability. Within-subject predictors are known as time-varying predictors or *dynamic predictors*. A special case of a dynamic predictor is a time predictor mentioned above. Dynamic predictors need not be time predictors as when, for example, math scores are collected over time and used to predict reading scores. Models for dynamic predictors other than time predictors are discussed in Chapter 13.

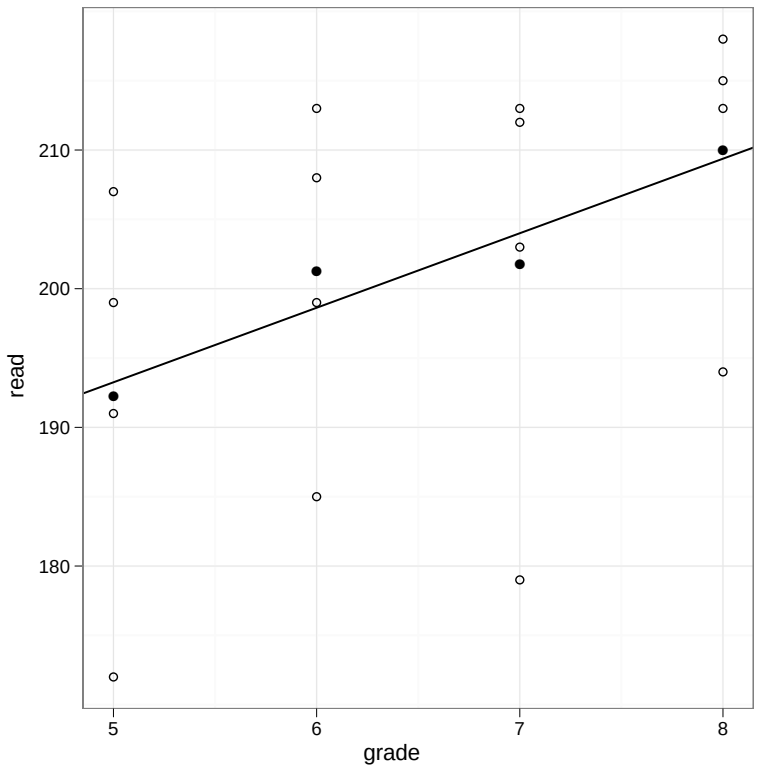
Between-subject predictors are known as time-invariant predictors or *static predictors*. Static predictors have their scores repeated for the duration of time, which is illustrated in Table 1.2.

As explained in Chapter 5, when the dynamic predictor is a time predictor like `grade`, the regression analysis involves the fitting of a curve for the response variable over time. Such curves are traditionally known as growth curves, but the term change curves is preferred as not all responses accumulate over time. In many instances, the response goes up or down over time, as when depression scores increase and then

decrease or vice versa.

The values of `grade` are consecutive integers, indicating the regression of `read` on `grade` results in fitting a linear change curve. A straight line is fit to the scatter of points defined by the two variables. The regression line consists of the fitted values of `read`, which can be interpreted as predicted mean values for the fixed values of `grade`. Therefore, the regression of a longitudinal response variable on a time predictor focuses on the trend of the means, or the *mean trend* over time.

Figure 1.2: Regression of reading scores on grade. Open circles are the individual data points, filled circles are the means, and the solid line is the regression line.




A graph of `read` by `grade` for the Table 1.2 data is shown in Figure 1.2. The open circles are the individual `read` scores for the fixed values of `grade`, and the filled circles are the means. The regression line shows that the means tend to increase over time.

For purposes of illustration, the regression line in Figure 1.2 was produced using traditional regression methods. Traditional regression is wrong for this situation because there are repeated measures. In a moment, a more appropriate analysis using LMER is presented. Additional details of the contrast between traditional regression and LMER are provided in Chapter 5.

For a number of data sets, a linear time variable does not adequately account for the within-subjects variability. In such cases, transformations of the original time variable can be substituted and/or added to the regression equation to account for nonlinear change. It is common to include power transformations of the time variable, for example, $\texttt{grade}^2$, in order to add bends to the change curve. Transformations of the time predictor for modeling nonlinear change are discussed beginning in Chapter 3, but the most extensive material appears in Chapter 12.

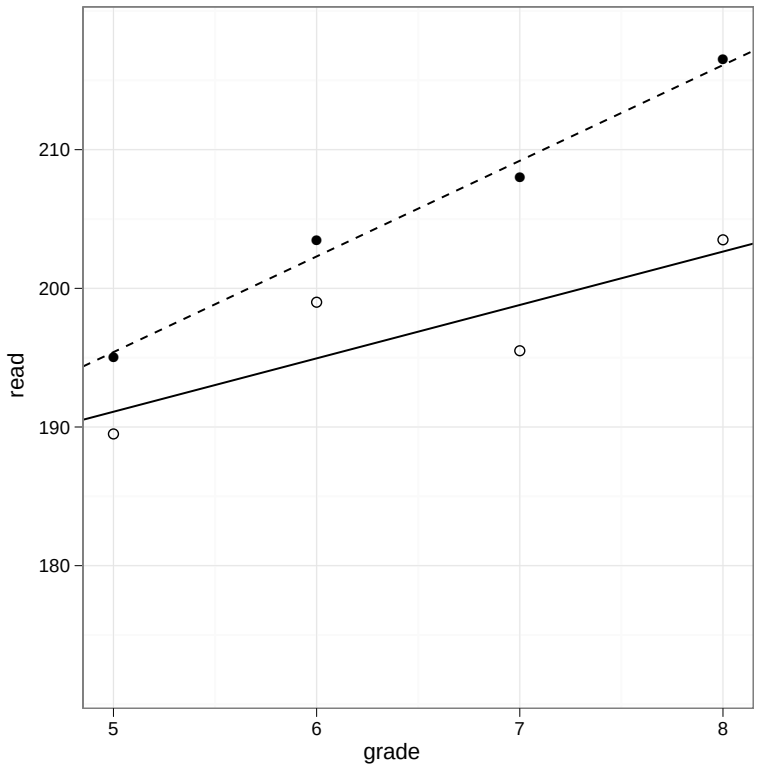
In some cases, the only variables collected by the researcher are the response variable (e.g., `read`) and the time predictor (e.g., `grade`). The emphasis is on accounting for the within-subjects variability, or the change over time. The analysis typically involves trying different time predictor transformations to account for nonlinear trends. Or, if a linear trend is apparent, the focus is on the estimation of the change curve parameters and statistical inference for the fixed effects. Between-subjects variability can also be examined

in this case. This provides an assessment of the correspondence of individual subjects' regression lines with the overall regression line, as discussed below.

In other analyses (almost all?), the researcher collects data on static predictors, in addition to the response and the time predictor. The focus is on both within-subjects change and between-subjects differences in that change. As shown in Chapter 5, when the static predictors are considered in combination with the time predictors, between-subjects differences in change curves can be examined in a regression equation.

For example, one can study whether reading achievement change curves are different for gender groups, or whether change curves differ by levels of attendance proportion. This is accomplished by including time predictors and static predictors as single predictors and as interactions. Interactions in this context are product terms among the predictors. The mathematical details are deferred until Chapter 5.

Figure 1.3: Regression lines and means of gender groups. Open circles are for females and filled circles are for males.




Consider Figure 1.3 that shows the means and regression lines by gender group. Males start a bit higher than females and increase at a faster rate. In terms of the sample regression lines, there is a difference in intercepts meaning a difference in starting points,[2] and a difference in slopes. Here the effects are assessed based on a graph of the sample data; inferences can also be based on LMER as discussed below.

When the static predictor is a categorical variable like gender, then the change curves in question are mean trends for the groups. This concept generalizes to the case of quantitative static predictors like attendance proportion, for which the levels of the predictor may be considered as strata. In such a case, mean trends are considered for each stratum, at least in principle.

### 1.3.2 Random Effects

The regression lines in the examples above were computed using traditional regression methods, and this is pause for concern. Traditional regression does not correctly associate a subject with their repeated measures,

[2]It is assumed here that the intercept is anchored at the first time point; see Chapter 5 for details.

and no distinction is made between dynamic and static predictors. Thus, when traditional regression is used with the Table 1.2 data, the variance of the response is not properly partitioned into its within-subjects and between-subjects portions. This is important, as valid SEs, among other things, are predicated on such partitioning.

Moreover, since the blocks of Table 1.2 are ignored, every row is treated as a unique subject. This leads to incorrect bookkeeping, meaning that important quantities, such as degrees of freedom ($df$), cannot possibly be correct. The typical assumption of independence of observations is certainly violated, which casts doubt on the validity of any inferential results. Clearly, traditional regression is inappropriate for analyzing longitudinal data.

In contrast to traditional regression, LMER does associate subjects with their repeated measures. This is accomplished by introducing random effects in the regression equation. Random effects summarize individual change. The proper bookkeeping provided by the random effects means that LMER is a better basis for inference with longitudinal data than traditional regression. LMER recognizes the distinction between dynamic and static predictors, so that the response variance is partitioned into within-subjects and between-subjects portions. Such partitioning provides more accurate SEs, among other things.

In Figures 1.2 and 1.3, group-level curves were considered that summarized the reading/grade relationship for all the subjects. To provide a conceptual understanding of random effects, it is useful to consider individual change curves. Figure 1.4a shows the observed points for the four individuals in the example data set, along with their fitted regression lines. There is one panel per subject, and the subject identifier is in the panel title (`subid`). For simplicity, linear change is considered, but the concepts to be discussed also generalize to nonlinear change.

Random effects are useful for modeling and examining individual variation in change. As the graph in Figure 1.4a reveals, there is variability among the subjects in terms of their intercepts, or where the regression lines begin. Subject 1 starts relatively low, whereas subject 5 starts relatively high. There is also variability among their slopes, or variability in the slant of their regression lines. Subject 5 has a relatively shallow slope, being the closest to a horizontal line of the bunch. The other individuals have lines with greater slants rising from lower left to upper right.

The first key to understanding random effects is that each subject in Figure 1.4a has their own intercept and slope. Though these intercepts and slopes are unobserved random variables in LMER, they can be computed based on sample data as discussed in Chapter 10. A crude method inferior to those discussed in Chapter 10, is to compute the intercepts and slopes using ordinary least squares (OLS) of traditional regression. Since OLS is assumed to be familiar to the reader, it is used here for purposes of illustration. The OLS equations for the sample intercept and slope are the following,

$$\begin{aligned} \text{slope} = \hat{\beta}_1^* &= \frac{\sum(y-\bar{y})(x-\bar{x})}{\sum(x-\bar{x})^2}, \\ \text{inter} = \hat{\beta}_0^* &= \bar{y} - (\hat{\beta}_1^*)(\bar{x}). \end{aligned} \tag{1.3.1}$$

The asterisk (*) is used to denote that the formulas are for a **single subject**. Additional details about the notation are provided in Chapter 5. Since the formulas of Equation 1.3.1 apply to an individual, the summation ($\sum$) is over the repeated measures for the individual (not over subjects).

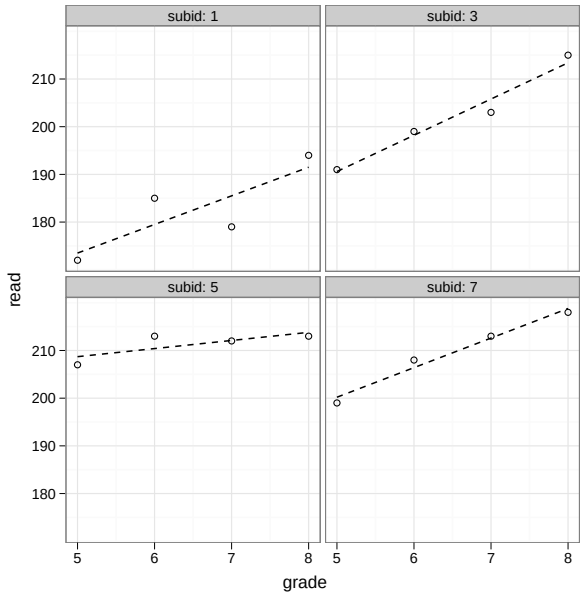
Table 1.3: Slope and intercept calculations for subject 1.

| time | $y$ | $x$ | $(y-\bar{y})$ | $(x-\bar{x})$ | $(y-\bar{y})(x-\bar{x})$ | $(x-\bar{x})^2$ |
|---|---|---|---|---|---|---|
| 1 | 172 | 5 | −10.50 | −1.50 | 15.75 | 2.25 |
| 2 | 185 | 6 | 2.50 | −0.50 | −1.25 | 0.25 |
| 3 | 179 | 7 | −3.50 | 0.50 | −1.75 | 0.25 |
| 4 | 194 | 8 | 11.50 | 1.50 | 17.25 | 2.25 |
| Sum | 730 | 26 | 0 | 0 | 30 | 5 |

As an example of the calculations, consider Table 1.3 that shows the relevant quantities for subject 1. The data in the $x$ and $y$ columns are from the first block in Table 1.2. From Table 1.3, it should be clear that $\bar{y} = \frac{730}{4} = 182.50$, $\bar{x} = \frac{26}{4} = 6.50$, slope $= \hat{\beta}_1^* = \frac{30}{5} = 6$, and inter $= \hat{\beta}_0^* = 182.50 - (6)(6.50) = 143.50$.

Figure 1.4: Illustration of random effects based on four subjects.

(a) Observed points (circles) and fitted regression lines (dashed).




(b) Individual fitted regression lines (dashed) and group fitted line (solid).

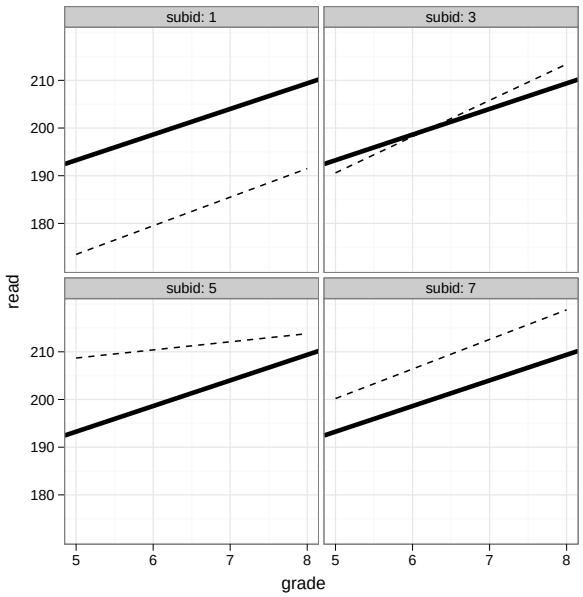



Table 1.4 lists the intercept (`inter`) and slope for each subject computed with the OLS formulas of Equation 1.3.1. The last row of the table lists the group-level intercept and slope, which is computed by regressing `read` on `grade` for the entire sample. The regression using all the subjects need not be performed, as the group-level quantities are the mean of the individual values.

The second key to understanding random effects is that an individual's intercept and slope are expressed in relation to the group intercept and slope. The concept is illustrated in Figure 1.4b that depicts the

individual regression lines along with the group regression line. The solid group regression line is the same in each panel illustrating that a group-level effect is constant among individuals. The dashed individual lines are particular to the subjects illustrating individual variability. **Random effects are represented by the deviation of the individual regression lines from the group regression line.**

In the case of Figure 1.4b, there are two components of the regression lines that are of interest: the intercept and the slope. As will be see in future chapters, it is not compulsory to include every possible random effect in the model. The first random effect, which is a *random intercept*, is the difference between an individual's intercept and the group intercept. As can be seen in Figure 1.4b, subject 1 has a negative value for their random intercept, as the person's curve starts (and continues) at a lower point than the group curve. Subject 3 also has a negative value, but it is much closer to 0 than subject 1. Subjects 5 and 7 have positive random intercept values, as their curves start at a higher point than the group curve.

The second random effect is a *random slope*, which indicates the discrepancy of an individual's slope from the group slope. Subject 1 has an individual curve that is almost parallel with the group curve. This results in a random slope that is near to 0, as the person's rate of change is almost the same as the group rate of change – there is almost no deviation from the group trend. A similar situation applies for subject 7.

The slope for subject 3 is greater than the slope for the group. This means the random slope for subject 3 will be a positive value. In contrast, the slope for subject 5 is lower than the group slope indicating a negative random slope value.

For this simple example, crude types of random effects are computed by subtracting out the respective group values from the individual scores; a better method is described in Chapter 10. The random intercept of subject 1, for example, is $143.50 - 166.38 = -22.88$, and the random slope is $6.00 - 5.38 = 0.63$. The random effects scores are listed in the `inter-M` and `slope-M` columns. The notation of the variable labels indicates that the individual scores are mean-corrected.

In this example, the group intercept and group slope are the means of the respective individual estimates. Thus, subtracting out the group component is equivalent to mean-correcting the individual intercept and slope scores. In LMER, the random effects are the mean-corrected scores, `inter-M` and `slope-M`. However, interest is usually in the variances and covariances of the random effects, which are the same whether the mean-corrected or mean-uncorrected versions are used.

As discussed in future chapters, a key assumption for inference in LMER is that the subjects have been randomly sampled from a population. Under this assumption, the scores in the rows of Table 1.4 are independent. The random sampling assumption is the genesis of the term, random effects. In addition to independence, inference in LMER is often predicated on the joint normality of the random effects, an assumption made throughout this book.

Table 1.4: Intercepts and slopes of four subjects, mean uncorrected and mean corrected version, and two static predictors.

| subid | inter | slope | inter-M | slope-M | gen | att |
|---|---|---|---|---|---|---|
| 1 | 143.50 | 6.00 | -22.88 | 0.63 | F | 0.94 |
| 3 | 152.60 | 7.60 | -13.77 | 2.22 | M | 0.97 |
| 5 | 200.20 | 1.70 | 33.83 | -3.68 | F | 0.85 |
| 7 | 169.20 | 6.20 | 2.82 | 0.83 | M | 0.97 |
| Mean | 166.38 | 5.38 | 0 | 0 | | |

Many of the conceptual aspects of LMER can be gleaned from a comparison of Table 1.4 and Table 1.2. The random effects act to reduce the dimension of the original regression problem by summarizing change over time. The random effects scores represent a collapsing over the time dimension that yields the traditional one-row-per-subject data set of Table 1.4. Assuming random sampling, the rows of Table 1.4 are independent, which is in contrast to the dependency among the rows in Table 1.2.

In collapsing over the time dimension, there is an inevitable loss of information about the response scores. As depicted in Figure 1.4a, there is scatter of the observed points about an individual's regression line. This scatter is *within-subject error*. The error is assumed to be random, so it is also known as the random

error previously mentioned. The error variance computed among all the subjects is the *within-subjects error variance*. The within-subjects error variance is considered to be a result of measurement unreliability. Typical of regression models, the assumption is made that the within-subjects error is normally distributed with a mean equal to 0 and a constant variance; see Chapter 5.

Having collapsed over the within-subjects time dimension, attention now turns to differences between individuals, or put another way, between-subjects variability. Table 1.4 shows there is variability in random intercepts and in random slopes. Subjects start at different levels, and they change over time at different rates. The variability can be quantified by computing the variance of the intercepts and the variance of the slopes. The variance does not change under subtraction of a constant, so it does not matter if `inter` and `slope` are used, or `inter-M` and `slope-M`.

One might also want to compute the covariance or correlation between the random effects, as this indicates how the two are related. In this example, the correlation between the random effects is cor $= -0.87$, indicating that those who start lower have a faster rate of change than those who start higher. This is illustrated in Figure 1.4b. Subject 3 starts at a lower point than the group curve, but increases at a faster rate. Conversely, subject 5 starts at a higher point than the group curve, but increases at a slower rate.

When there are no static predictors, between-subjects effects are represented by the variance-covariance matrix among the random effects. When there are static predictors, as in the example of gender and attendance, then an additional set of between-subjects effects can be addressed. The random intercepts and random slopes can be regressed on the static predictors to learn about between-subject differences in change. Gender and/or attendance might account for variance in the random intercepts and/or slopes. To the extent this is the case, either the starting point (intercept), or rate of change (slope), or both, can be predicted from the static predictors. These types of static predictor effects are often of prime interest in LMER analysis.

As a conceptual example, consider the correlation between `att` and `inter`, and then `att` and `slope`. Using information from the table, it can be shown that cor(`att`, `inter`) $= -0.79$, and cor(`att`, `slope`) $= 0.97$. The negative correlation with the intercept indicates that those with lower attendance proportions tend to have higher starting points (intercepts), and those with higher attendance proportions tend to have lower starting points. The positive correlation with the slopes indicates that those with lower attendance proportions tend to have lower rates of linear change, and those with higher attendance proportions tend to have higher rates of linear change. In later chapters, methods of inferences for such effects are discussed.

### 1.3.3 How Important are Random Effects?

Having introduced random effects as a crucial feature for the proper analysis of longitudinal data, it may seem strange to ask about their importance. Random effects are certainly important for analyzing longitudinal data, as they provide a proper accounting of the repeated measures. Furthermore, the random effects provide a model for the response correlation between different time points, as mentioned in Section 1.2. That is, the random effects account for the dependency due to repeated measures, which is crucial for a proper longitudinal model. The demonstration of the random effects in this role is relatively complicated, requiring the development of some preliminary material. Details are provided in Chapters 5 and 6.

Despite the importance of random effects, the primary objects of inference in the majority of behavioral science applications are the fixed effects. Recall the fixed effects are the group-level regression coefficients, which are in the last row of Table 1.4. Similar to traditional regression, applied researchers are usually interested in summarizing among subjects, meaning the group-level fixed effects are of prime interest.

With the above in mind, one view of LMER is as a fixed effects model with an error term appropriate for longitudinal data; or more simply, a regression model with a correlated error term (Verbeke & Molenberghs, 2000, chap. 3). As shown in Chapter 5, the random effects can be considered as part of a composite error term. In this view, the variances and covariances of the random effects are largely treated as "nuisance" parameters. They are a necessary consequence of including random effects, but they are the handmaiden of the fixed effects used primarily to account for the dependency due to repeated measures.

Though the random effects per se are rarely a concern, their variances and covariances (or correlations) are of interest, especially when one wants to know about individual variability in change. Output from LMER analysis provides estimates of the variances and correlations among the random effects. The interpretation of these estimates are discussed in future chapters when they are encountered.

Based on what appears to be the typical priority of applied researchers, LMER will mainly be treated

as a fixed effects model with correlated errors induced by including random effects. Not all readers will appreciate this emphasis, but this approach reflects the primary interest of most applied researchers.

There are practical advantages of focusing on fixed effects. As discussed in Chapter 7, model building is much simpler when the number of random effects is determined before the fixed effects are considered. Things can be quite involved when both random and fixed effects are simultaneously selected in the same analysis. Among other problems, statistical tests for inclusion of random effects are not standard and require special consideration; see Chapter 10.

Such problems hold the possibility that undue time and resources will be expended on effects that are not of primary interest. In many instances, graphs of the data can be used to suggest a reasonable number of random effects. Graphing procedures are covered in Chapters 3, 9, and 10. Informal methods are stressed for the selection of the random effects, but formal inferential methods are also discussed.

## 1.4 Traditional Approaches

The use of LMER in behavioral science analysis is relatively new. Until recently, longitudinal data was analyzed with traditional approaches such as RM-ANOVA, and repeated measures multivariate analysis of variance (RM-MANOVA). In this book, the traditional approaches are not discussed in detail because they have particular features that make them less attractive for applied longitudinal analysis than LMER. It is worthwhile to highlight some of these features so as to be clear why a growing number of applied researchers are turning to LMER.

From an applied research perspective, the Achilles heal of the traditional methods is the inability to accommodate missing data. Traditional methods require the data set to be complete, and subjects who have any amount of missing data are mercilessly omitted. For the situation in which there is a small amount of missing data per subject, but a moderate number of subjects in this predicament, *df* can be small relative to the number of original subjects. Many applied researchers are not comfortable with the potentially large loss of information. The result is a turn to alternative methods that allow subjects with incomplete data to be included in the analysis.

In most computer realizations of RM-ANOVA and RM-MANOVA, the statistical program fits a series of orthogonal polynomials to the data; see Chapter 12. The number of polynomials fit is one less than the number of time points. In most cases, there is no user control over this fitting. When the number of time points is large, say greater than four, high order polynomials are fit regardless if they are useful or interpretable. For the applied researcher analyzing 10 time points, an all-too-often occurrence is to find a single statistically significant high order polynomial, such as the 8th. Rarely is such an effect hypothesized, and there seems to be scant direction as to how such a finding should be interpreted. A statistical consultant might comment that an 8th order polynomial is "wavy", but this is probably not much help to the researcher. It would be better for the researcher to specify only the effects of interest, which is possible with LMER.

Another issue with RM-ANOVA and RM-MANOVA is they represent two extremes. On the one hand is RM-ANOVA, whose basic inferential methods rely on the assumption that the rate of change for all the subjects is constant. This scenario seems overly restrictive for much longitudinal data, as individual variability in rate of change appears to be the rule rather than the exception; see Figure 1.1. A constant rate of change is plausible when the presentation of repeated measures is counterbalanced. Thus, RM-ANOVA seems most appropriate for experimental designs, in which the repeated measures are conditions that subjects endure, rather than time points at which behavior is observed. This is why the history of the traditional approaches is closely associated with randomized experimental designs (Keppel, 1991, chap. 15).

At the other extreme is RM-MANOVA, which relaxes the requirement of constant rate of change for all subjects. However, the model in effect, provides a unique correlation or covariance parameter for each pair of time points. This means the number of covariance components expands with the number of time points, potentially resulting in a large number of parameters. When there are more than two time points, the number of covariance components can be greater than the number of change curve parameters. As discussed above, the correlations among the repeated measures are important, but they are of relatively minor interest in comparison to mean change over time. Therefore, RM-MANOVA devotes an undue number of parameters to an aspect of the model that is largely ignored in the results. In the LMER models considered later, adequate accounting of the correlation among the repeated measures is accomplished with many fewer parameters.

The advantages of LMER over the traditional methods are summarized in the following list.

- **LMER can accommodate missing data.** Subjects who have response data for at least one time point can be included in the analysis. However, valid inference under these conditions is predicated on assumptions regarding the mechanism that caused the missingness; see Chapter 3.
- **LMER provides control over the number and nature of terms used to model change over time.** Regardless the number of time points, the researcher can fit a low order polynomial or fit other types of curves that directly address research questions; see Chapter 12.
- **LMER is very flexible regarding the structure of the data.** Timing of observations can vary among subjects, and the distance between adjacent time points need not be equal.
- **Finally, LMER allows for various types of predictors.** Predictors can be dynamic and/or static, quantitative and/or categorical.

The above features are attractive for the analysis of longitudinal behavioral data. LMER provides the flexibility to cover a multitude of empirical data structures. The remainder of the book focuses on LMER, though passing reference is sometimes made to RM-ANOVA/MANOVA to point out similarities of the models.

## 1.5 MPLS Data Set

For the majority of the examples in this book, a single data set will be used. The data appear in Table 1.6 and are also available on the book website. A missing value is coded as $-99$. A relatively small data set of $N = 22$ is considered because it is manageable. The reader can print the data set to the computer screen or a single sheet of paper. For the reader just becoming familiar with the `R` computer program, the ability to see the entire data set is beneficial. Results of data manipulations can be readily seen, which facilitates understanding of the workings of particular functions. There is benefit in using the same data set for different illustrations, such as graphing and data analysis. The continuity reduces the need to comprehend new substantive issues along with new statistical issues that can arise when data sets are switched.

The data in Table 1.6 are taken from a much larger data set collected by the Minneapolis Public School District (MPLS) in Minnesota (USA), and deidentified for the purposes of secondary data analysis. The data were collected in part because of district efforts to comply with recent federal accountability requirements, namely Title X of the No Child Left Behind Act of 2001 (2002). Data collection began with the 2004-2005 school year, and the larger data set has $N = 16561$ students.

The variables in Table 1.6 are reading achievement scores from grades 5 through 8 (`read.5`-`read.8`), risk group (`risk`), gender (`gen`), ethnicity (`eth`), English language learner status (`ell`), special education services (`sped`), and attendance proportion (`att`). The reading scores constitute the repeatedly measured response variable that is referred to as `read`. All variables other than `read` and `grade` are static predictors. A limited description of the variables is provided below; additional details are found in Obradović et al. (2009).

The reading achievement scores are from the reading portion of the Northwest Achievement Levels Tests (NALT; Northwest Evaluation Association, 2003). The NALT is a multiple-choice adaptive assessment of academic achievement. NALT raw scores are converted to vertically-equated scaled scores based on IRT models (see Kolen & Brennan, 2004). In principle, the scaling provides a valid basis for using NALT scores to study change in level over time. NALT administration was in the fall for all students, so that `grade` represents a yearly measurement.

The risk group variable constitutes classification of individuals based on the MPLS assessment of homeless and highly mobile (HHM) status and poverty status (POV). HHM status is defined by a family living in inadequate and/or transitional conditions, such as on the street, in a shelter, or doubled up with relatives. Individuals are classified as HHM if they meet the district definition at any time over the course of the study.

Poverty status (POV) is defined by student eligibility for free or reduced-price meals based on federal standards. A person is classified in the POV group if they qualify for free or reduced lunch at least once over the course of the study, but at no time qualify as HHM. All those in the HHM group qualify for free or reduced lunch and thus, are also a special case of POV. Individuals are classified as advantaged (ADV) if

they do not qualify for classification in the other two categories. As indicated below and in future chapters, the HHM and POV groups will often be combined and designated as the disadvantaged (DADV) group.

As for the other static predictors, gender is coded as F = female and M = male. English language learner status is coded as 0 = is not a learner and 1 = is a learner, and special education services is coded as N = not receiving and Y = receiving. Attendance proportion is the average number of days attended by a student divided by the number of days enrolled for the duration of the data collection. Ethnicity is coded as Afr = African American, His = Hispanic, Whi = White (W). For some examples, the first two ethnicity groups will be combined and designated as the non-White (NW) group.

The static predictors in Table 1.6 can be classified as categorical or quantitative variables based on their level of measurement (Long, Feng, & Cliff, 2003). Categorical variables denote membership in different groups or classes and can use numbers, letters, or names. When numbers are used, they do not represent amounts of things, but different categories of membership. All the static predictors, except for attendance proportion, are categorical. Quantitative variables use only numbers and the numbers stand for amounts of things. Attendance proportion is the sole quantitative predictor. It is possible to have categories with a natural ordering, such as 1 = low, 2 = medium, 3 = high. A variable like this might be classified as ordinal or ordered categorical. Ordinal variables are sometimes treated as quantitative if there are many categories (e.g., ten), or as categorical if there are few categories (e.g., three).

A point of clarification should be mentioned regarding dynamic and static variables. It is possible for a dynamic variable to have constant values over time, such as when a subject has the same reading score at each time point. Such an event is unlikely, but the possibility nonetheless does exist. The distinction between static and dynamic variables is that the researcher fixes the values of static variables, but does not fix the values of dynamic variables. Dynamic variables can be treated as static if the researcher chooses, for example, to repeat the baseline value over all the repeated measures, or average over all the time points and repeat this value. The latter approach was taken in computing attendance proportion in Table 1.6. A related discussion is found in Section 5.5.3.

The decision to treat a predictor as dynamic or static depends on such things as the research questions, the nature of the variable, and the study design. In the MPLS study, the yearly variation in attendance was judged by district officials to not be particularly important due to unreliability of reporting. Thus, the researchers chose to represent an overall level for each subject.

It must be emphasized that though analysis with the larger data set is no trifling matter, the Table 1.6 data set should be considered a toy. Serious issues regarding academic achievement are discussed below and in future chapters. **Analysis with the Table 1.6 data set should not be taken as a serious attempt to address these deep issues.** It is perhaps a bitter reality of applied research that small data sets have limited information, and this precludes addressing certain types of research questions. For example, perusal of the ethnicity variable (`eth`) reveals that only one subject is Hispanic (`His`). We certainly would not want to make ethnic group comparisons based on a single case.

On a related note, LMER models can have relatively many parameters when there are numerous static predictors. Each static predictor can have an intercept and slope effect, and possibly more, when the change curve is nonlinear (see Chapters 5 and 12). Though relatively complex models will be considered for pedagogical purposes, such models can be unstable when the sample size is small. Results based on the sample data set of Table 1.6 should not be used to make definitive assessments regarding the plausibility of the estimated models.

## 1.6 Statistical Strategy

Applied data analysis consists of more than statistical concepts and formulas. There is always an underlying philosophy of the data analysis, referred to as the *statistical strategy* (Hand, 1996; Mallows, 1998). Unfortunately, in much applied data analysis, statistical strategy for an analysis is not explicitly stated. All too often, data analysis is guided by convention and intuition. The justification for using particular approaches and/or analysis steps is left vague, or swept under the rug altogether.

Statistical strategy should be made explicit so that the consumer of one's results will be clear about the justification, rationale, and viewpoint of the statistical methods used, even if there is disagreement. To this end, the statistical strategy of this book will emphasize what is known as an *information-theoretic* approach

to statistical analysis. For model evaluation, focus will be on the use of a global statistical fit measure known as *Akaike's Information Criterion* (AIC), due to the statistician Hirotugu Akaike (Akaike, 1973, 1974, 1981). (Akaike is roughly pronounced as "ah-kah-ee-kay".)

Why is the AIC emphasized? The evaluation of statistical evidence obtained from sample data is considered to be a primary goal of applied analysis. The AIC provides a clear measure of such evidence. Once AIC values are determined for a set of fitted models, clear evidential statements can be made regarding the plausibility of the models and the scientific hypotheses they represent. A limited example is provided below with a more extensive example provided in Chapter 7.

The AIC can be use to rank-order models in terms of their plausibility, and the probability of a particular model can be determined, given some conditions. Meaningful statements regarding differences in the plausibility of the models can also be made. The best fitting model, for example, might be slightly more plausible than the next best model, or much more plausible.

An emphasis on the AIC means a deemphasis on traditional methods, such as null hypothesis statistical testing (NHST), with its reliance on the $p$-value. This does not imply that NHST is wrong. NHST is a sound and perfectly appropriate method for certain purposes, such as evaluating evidence from randomized controlled experiments. However, for the case of regression models with observational data – the primary domain of this book – the AIC provides a clearer representation and evaluation of statistical evidence. In the dominant traditional interpretation, $p$-values do not indicate effect size (Goodman, 1999). Thus, they cannot be used to rank-order a set of regression models in terms of their plausibility. Rank-ordering and effects size are fundamental to the scientific enterprise. Thus, the AIC is preferred to NHST in most applications considered in this book. Additional points are discussed in Chapters 7 and 8.

The use of the AIC is closely associated with an approach to data analysis know as *multimodel inference* (Burnham & Anderson, 2004, 2002; Anderson, 2008). Multimodel inference has three main aspects.

1. A set containing two or more candidate models (e.g., regression models) is formulated prior to the data analysis.

2. The AIC is used to evaluate the models in terms of relative fit and plausibility.

3. Limited results are presented for all the models in the set, such as the AIC values, and detailed results are presented only for the best fitting model(s).

An advantage of multimodel inference is that it stresses careful and deliberate thought prior to the analysis. This is probably a reaffirmation of what many researcher already do, but it speaks to the merits of planned analysis rather than data-driven analysis. It has long been documented that data-driven analysis, sometimes called *exploratory analysis*, is highly vulnerable to errors of inference (Chatfield, 1995; Faraway, 1992; Freedman, 1983; Lovell, 1983).

Sample data contain *information* about the phenomenon under study, and its compliment, *non-information* (Anderson, 2008, chap.1). Researchers are concerned with correctly winnowing information from non-information, as the former is the enduring part of the sample data. The enduring part of the sample data is instructive about the mechanisms underlying the content of interest.

It has long been argued that the best strategy for data analysis is to carefully formulate several scientific hypotheses prior to the analysis (Chamberlain, 1890; Feyerabend, 1968; Peirce, 1903; Platt, 1964). These hypotheses, also known as *working hypotheses*, are formulated based on theory, expert knowledge, the extent literature, and interaction with colleagues. The working hypotheses are the embodiment of the underlying processes that give rise to the information in the sample data. Statistical models, such as regression models, represent the hypotheses and are estimated based on the sample data. The estimation is sometimes referred to as *fitting the model to the data*.

When scientific hypotheses are carefully and pain-stakingly formulated, it is reasonable that the information from the sample – as opposed to the non-information – is being drawn upon to validate the associated models. If one model has better fit than another, it is thought that the associated hypothesis is a better representation of the processes that generated the information.

Exploratory analysis involves little or no extra-data knowledge to guide it. As a result, exploratory analysis is extremely prone to conflating information and non-information. The statistical models are suggested by the data, meaning they are suggested by information **and** non-information. This phenomenon is referred to as *overfitting* the data. Because non-information is a potentially blatant influence in exploratory analysis,

the hypotheses generated from data-driven models are misleading. They are potentially poor representations of the true underlying mechanisms that generated the information in the sample.

Multimodel inference is desirable because it emphasizes confirmatory rather than exploratory research. Though it is acknowledged that both confirmatory and exploratory analysis are needed, the former is to be preferred. This partisan stance is motivated by the fact that exploratory analysis is vulnerable to misuse. Given the speed and ease of modern computers, there is a temptation to devote considerable time to sifting the data with statistical procedures. In many cases, time is probably better spent sifting the available literature on the substantive topic of interest. Statistical computing should never be substituted for careful thought about the variables in the analysis, and formulating sensible scientific hypotheses and statistical models.

The purpose of exploratory analysis is to generate scientific hypotheses for future research. For this reason, exploratory analysis should be based only on graphical and descriptive methods. Exploratory analysis should not be based on inferential methods like $p$-values, hypothesis tests, or confidence intervals (Tukey, 1977). To this end, an entire chapter is devoted to graphing longitudinal data (Chapter 3), and descriptive methods are discussed throughout the book, especially in Chapters 9 and 10.

Output from LMER analysis qualifies as descriptive information, and its use can be important in exploratory analysis. However, when LMER is used for exploratory analysis, the output should not be used to generate $p$-values for hypothesis testing because of the problem of overfitting. As discussed in Chapter 6, the `R` package that will be used for longitudinal data analysis is very desirable in this regard, as only descriptive statistics are printed in much of the default output. It has been repeatedly demonstrated that given a sufficiently thorough search through the data, a seemingly meaningful pattern can always be found, even when the variables are randomly generated with a computer (Miller, 2002, chap.1). Iterative procedures, in which multiple models are fit with different numbers of variables based on statistical tests, are especially prone to spurious results (Freedman, 1983).

There is a certain sympathy for researchers who want to "let the data speak", but it is important to remember that the data always speak a combination of truth and lies, or in statistics speak, information and non-information. The first line of defense against letting non-information drive the selection of models is to bring theory, previous research, experience, and expert opinion to bare on the problem prior to the analysis. The second line of defense is to be extremely cautious with analyses that are not pre-planned.

Another advantage of multimodel inference is that the results are more realistic than traditional methods. A single final model is not presented as if it is the true model or even an adequate model. Rather, a "bouquet of models" (Tukey, 1995) is considered in any one analysis, with some models – the better fitting ones – more sweet smelling than others. The bouquet provides an indication of the relative plausibility of the scientific hypotheses the models represent. Based on sample evidence, some models are expected to die off and be discarded, whereas other models are born and added to the bouquet for further consideration. Details of multimodel inference are discussed in Chapter 7, but a thumbnail sketch is provided later in this chapter.

Lest some readers have become nervous by the last few paragraphs, rest assured that NHST is also throughly treated in Chapters 6 and 8, emphasizing a test procedure known as the *likelihood ratio test* (LRT). Though multimodel inference is preferred, it is acknowledged that NHST is compulsory in a number of quarters. In the hope that researchers will transition from NHST to multimodel inference, the LRT is related to the AIC in Chapter 8. However, the LRT may be used for traditional testing without reference to information theory. In addition to the analytic LRT, parametric bootstrap approaches are also discussed that are applicable for a wider number of testing situations.

Finally, the statistical inference discussed in this book is based on classical or frequentist principles, as opposed to other schools of thought, such as Bayesian statistics. The frequentist approach is based on the idea of repeatable research procedures (Fisher, 1991, p.14). Sample data are assumed to come from a hypothetical infinite population and are generated by the repeated research procedures in question (Sprott, 2000, chap.1). This concept forms the basis of inference using the predictive accuracy underlying the AIC (see Chapter 7), and the method of maximum likelihood (ML) underlying the LRT (see Chapters 6 and 8).

## 1.7 LMER and Multimodel Inference

Having introduced LMER, the AIC, and multimodel inference, a limited analysis example is provided using the MPLS data set. A more extensive analysis is provided in Chapter 7. To keep things simple, examination is restricted to questions regarding risk group differences in reading achievement over time. More complex research questions are examined in the remainder of the book.

Details of LMER, parameter estimation, and multimodel inference are provided in future chapters. Here a conceptual thumbnail sketch of the important issues is provided. The hope is that the example will motivate exploration of the remaining chapters.

### 1.7.1 Statistical Hypotheses

Several longitudinal studies have reported initial and on-going differences in reading achievement among socioeconomically advantaged and disadvantaged groups (Arnold & Doctoroff, 2003; Chatterji, 2006; Eamon, 2002; McLoyd, 1998; Pungello, Kupersmidt, Burchinal, & Patterson, 1996). Advantaged students generally have higher achievement scores than disadvantaged students, with an important issue being the extent and duration of the achievement gap. There is conflicting evidence regarding the persistence of an early achievement gap into latter grades. Depending on the historical epoch, and relevant factors like school-based interventions, early differences might remain stable over time (Applebee & Langer, 2006), or narrow or widen (Caro, McDonald, & Willms, 2009; Dearing, Kreider, Simkins, & Weiss, 2006).

The research findings suggest several LMER models that might be fit to the sample data. For this example, the single static predictor of risk is considered. Due to the small sample size of the MPLS data set, the dichotomous version of risk will be used, consisting of the disadvantaged group (DADV) and the advantaged group (ADV). Recall the disadvantaged students are classified as such if they meet eligibility for free or reduced-price meals based on federal standards. All HHM students are eligible for this aid, and thus, are absorbed into the disadvantaged group. The advantaged students have high enough family incomes to disqualify them for disadvantaged status. There is considerable controversy regarding whether free and reduced lunch is a valid indicator of disadvantage due to poverty, but this consideration is beyond the scope of this example (see e.g., Harwell & LeBeau, 2010).

The single time predictor is grade, which is considered to be an essential part of every model. Based on a voluminous research literature (Shin, Davison, & Long, 2009), there is a strong expectation that reading achievement will increase over time regardless of risk. Given the relatively short grade span (5-8), the increase is expected to be linear, as this is consistent with similar studies (Ding, Davison, & Petersen, 2005). Therefore, all models considered will be linear growth models and include grade as the sole time predictor.

Assuming linear change and two risk groups, there are four main models one might consider for LMER analysis. The mathematical details of LMER are delayed until Chapter 5. Here a heuristic approach is presented using graphs to convey the essence of the models.

Graphical illustrations of the four LMER models are shown in Figure 1.5. Model A in Figure 1.5a depicts the situation in which the DADV and ADV change curves are identical. There is no intercept difference or slope difference among the groups. The advantaged and disadvantaged lines are atop one another, giving the appearance of only one curve in the graph. It follows that this is a model with no achievement gap.

Model B in Figure 1.5b depicts an intercept difference. This is a stable achievement gap model. The change curves increase over time, but they are parallel. The DADV curve starts at a lower point on the horizontal axis than the ADV curve, and the difference endures over time.

Figure 1.5c shows Model C with a slope difference, but not an intercept difference. This is a model that has no initial achievement gap, but one develops over time. The groups begin at the same value, but the ADV group increases at a faster rate than the DADV group.

Finally, Model D in Figure 1.5d depicts both an intercept difference and a slope difference. This is an achievement gap model, with the gap narrowing over time. The ADV group starts at a higher value than the DADV group, but the DAVD group increases at a faster rate. Another possibility for a changing-gap model is that the gap widens over time, that is, the group curves increasingly diverge as time elapses.

Multimodel inference begins with a set of working hypotheses. The statistical models discussed in future chapters are considered as representatives of these working hypotheses. The distinction between a working hypothesis and its representative statistical model is usually blurred. The consequence is that the represen-

tative is often referred to as a *statistical hypothesis*. **One tenet of multimodel inference is that each statistical hypothesis should have a sound scientific justification.** The intention here is to reinforce the good practice of carefully thinking about theoretical issues before examining the data. Consequently, based on the brief sketch of the extent literature provided above, two of the models in Figure 1.5 are not plausible, Model A and Model C. As mentioned, an initial achievement gap has been consistently found over many studies spanning many years of research. Model A and C do not have an early achievement gap, and thus, are not scientifically justified.

The omission of unjustifiable hypotheses is in contrast to the traditional approach of NHST. NHST often involves an unrealistic *null hypothesis*, such as Model A or C. Usually a justifiable hypothesis or *alternative hypothesis*, like Model B or D, is tested against the null hypothesis. When the data are used to evaluate the hypotheses, it is no surprise that the justifiable hypothesis is usually chosen over the unjustifiable hypothesis. For this reason, the use of NHST has been criticized when used with a meaningless null hypothesis, as the results are not helpful in evaluating the alternative hypothesis of interest (Cohen, 1994; Hubbard & Lindsay, 2008; Meehl, 1978, 1997; Tukey, 1991).

**In contrast, multimodel inference considers only alternative hypotheses.** In the case of the models represented in Figure 1.5, only Model B and D are considered justifiable and thus, are the two alternative models of interest. The set of models for the analysis then, consists of Model B and Model D. Model B represents an enduring achievement gap model with an intercept difference. Model D represents a changing achievement gap model with intercept and slope differences. In an actual analysis, it is common to consider many more than two alternative models.

The graphs of Figure 1.5b and Figure 1.5d illustrate the fixed effects portion of the models. The details of the random effects portion will be discussed in future chapters. Suffice it to say, the two models will both have two random effects: random intercepts and random slopes. Recall this means the models allow for individual variation in starting points and rates of change over time. These two types of variation are typically modeled in an actual analysis (Cnaan, Laird, & Slasor, 1997; Ware, 1985).

In the results below, focus is on the fixed effects, as these are usually the object of interest for applied researchers. Information about the variability of the random effects is also provided, but such information is usually not emphasized. Furthermore, the focus is on the static risk predictor effects rather than grade, as both models (B and D) have linear change curves.

Table 1.5 summarizes the two alternative models and provides statistical evidence based on the fit to the MPLS data set. In future chapters, parameter estimation and the computation of the AIC will be discussed in detail. For the moment, the AIC values and related information are simply listed, having been obtained from fitting the models to the data. Given the small sample size of this example, a variant of the AIC should be used that is appropriate for small samples. This variant is presented in Chapter 7.

Table 1.5: Set of two models for multimodel inference.

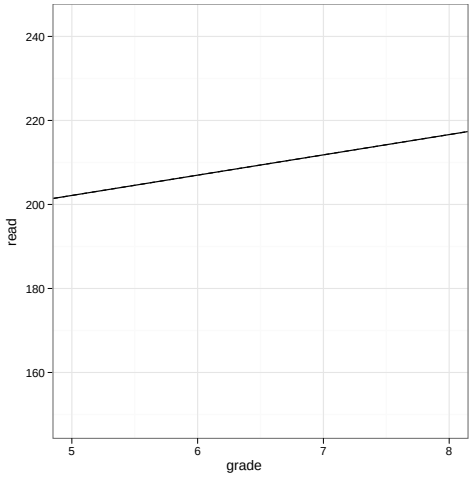
| Model | Description | AIC | Weight |
|---|---|---|---|
| B | Risk group intercept difference | 563.89 | 0.71 |
| D | Risk group intercept and slope differences | 565.70 | 0.29 |

A smaller AIC value indicates better fit, and argues for greater plausibility of the model in question. As seen in Table 1.5, Model B has a smaller AIC than Model D, indicating that Model B is the more plausible of the two. Plausibility is a relative concept, and much more will be said about this in Chapter 7. For the moment, it is stressed that there is no statistical testing with the AIC. All that can be said is that a model with the smaller value is better fitting and more plausible than a model with a larger value.

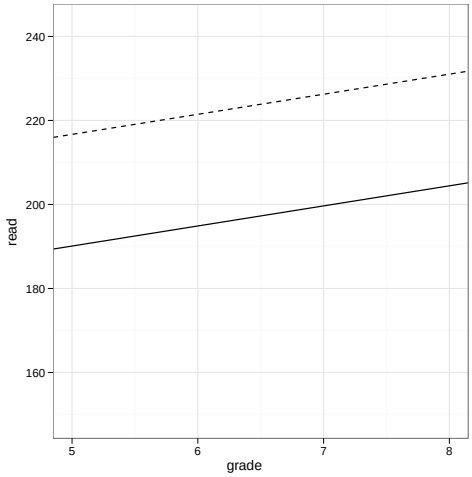
An advantage of the AIC is that much can be said regarding relative fit and plausibility. It is possible to quantify the extent of the plausibility using probability statements. Along with the AIC, Table 1.5 lists the weight of evidence, which is a probability scaling of the AIC. The weights fall between 0 and 1 inclusive, and their sum is always 1. The weight of evidence is the probability of the model given the data, the model set, and inability to know the true model; see Chapter 7. In this case, the weight of Model B is approximately $2\frac{1}{2}$ times larger than the weight of Model D. The interpretation is that Model B is more plausible than Model D,

Figure 1.5: Graphical representations of four models of risk group intercept and slope differences. The dashed line is for the advantaged group and the solid line is for the disadvantaged group.

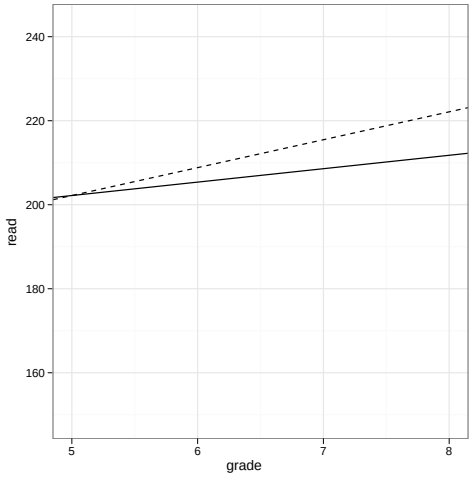
(a) Model A: No differences (lines are atop one another).



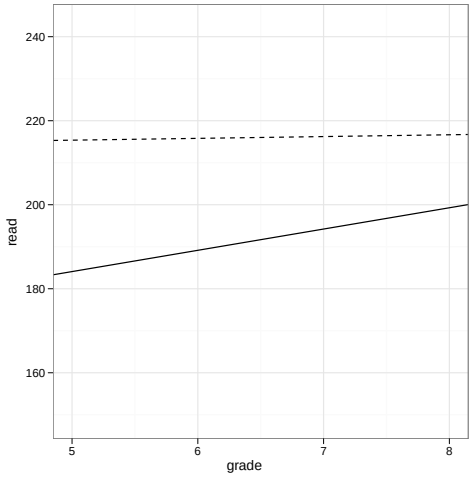

(b) Model B: Intercept difference.




(c) Model C: Slope difference.




(d) Model D: Intercept and slope differences.




but Model D probably cannot be completely ruled out. The basis for such statemens is provided in Chapter 7.

In a write-up of the results, Table 1.5 would be included along with the interpretative statements of the last paragraphs. In addition, parameter estimates of the most plausible Model B, would also be presented. This is to provide additional detail important for comprehending the results, such as the estimated values of the intercepts and the estimated rate of growth over time.

Details of obtaining parameter estimates are discussed in Chapters 5 and 6. For now, the parameter estimates obtained with the data are presented in the narrative, but usually a table of estimates is constructed. Beginning with the fixed effects estimates, the results indicate the ADV group had an estimated intercept of 192.78, with an estimated SE of 6.51. The intercept estimate indicates the estimated mean reading achievement for the ADV group at the 5th grade was 192.78. The DADV group had an estimated intercept 26.58 less than the ADV group, with a SE of 4.83 for the difference. The estimated rate of change for the two groups was 4.78, with a SE of 0.73. The slope estimate indicates the groups increased in their reading

achievement at a mean annual rate of 4.78.

In terms of individual variability, the estimated variance of the random intercepts was 579.5, or a standard deviation (SD) of 24.07. Using the vertical axes in Figure 1.5 as a reference, this indicates relatively extensive individual variability in starting points. The estimated variance of the random slopes was 6.75 (SD = 2.6), illustrating there were individual differences in rate of change over time. The estimated correlation between the random intercepts and slopes was cor = −0.9. The correlation indicates those who started lower tended to have a faster increase over time, and those who started higher tended to have a slower increase over time. Estimates of SEs for the variances and correlation are not presented for reasons discussed in Chapter 5.

To summarize, Model B with a constant achievement gap had superior fit to Model D with a changing gap. The fit of Model B was approximately $2\frac{1}{2}$ times better. Estimated parameters for Model B were provided to illustrate the differences between the groups. The difference was such that the advantaged group started higher, but both groups had a constant rate of increase over time. In terms of the bouquet of models, Model B was definitely more fragrant than Model D. The weights of evidence indicate, however, that Model D was not entirely malodorous and should not be completely discounted in future research.

Table 1.6: MPLS data set in wide format.

| subid | read.5 | read.6 | read.7 | read.8 | risk | gen | eth | ell | sped | att |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 172 | 185 | 179 | 194 | HHM | F | Afr | 0 | N | 0.94 |
| 2 | 200 | 210 | 209 | -99 | HHM | F | Afr | 0 | N | 0.91 |
| 3 | 191 | 199 | 203 | 215 | HHM | M | Afr | 0 | N | 0.97 |
| 4 | 200 | 195 | 194 | -99 | HHM | F | Afr | 0 | N | 0.88 |
| 5 | 207 | 213 | 212 | 213 | HHM | F | Afr | 0 | N | 0.85 |
| 6 | 191 | 189 | 206 | 195 | HHM | M | Afr | 0 | N | 0.90 |
| 7 | 199 | 208 | 213 | 218 | POV | M | Afr | 1 | N | 0.97 |
| 8 | 191 | 194 | 194 | -99 | POV | F | His | 1 | Y | 0.97 |
| 9 | 149 | 154 | 174 | 177 | POV | F | Afr | 0 | Y | 0.97 |
| 10 | 200 | 212 | 213 | -99 | POV | F | Afr | 0 | N | 0.96 |
| 11 | 218 | 231 | 233 | 239 | POV | F | Whi | 0 | N | 0.98 |
| 12 | 228 | 232 | 248 | 246 | POV | F | Whi | 0 | Y | 0.96 |
| 13 | 228 | 236 | 228 | 239 | ADV | F | Whi | 0 | N | 0.99 |
| 14 | 199 | 210 | 225 | 235 | ADV | M | Afr | 1 | N | 0.99 |
| 15 | 218 | 223 | 236 | -99 | ADV | F | Whi | 0 | N | 1.00 |
| 16 | 228 | 226 | 234 | 227 | ADV | M | Whi | 0 | N | 0.97 |
| 17 | 201 | 210 | 208 | 219 | ADV | M | Whi | 0 | N | 0.98 |
| 18 | 218 | 220 | 217 | 221 | ADV | M | Whi | 0 | N | 1.00 |
| 19 | 215 | 216 | 221 | -99 | ADV | F | Whi | 0 | N | 0.96 |
| 20 | 204 | 215 | 219 | 214 | ADV | F | Afr | 0 | N | 0.95 |
| 21 | 237 | 241 | 243 | -99 | ADV | M | Whi | 0 | N | 0.98 |
| 22 | 219 | 233 | 236 | -99 | ADV | F | Afr | 0 | N | 0.96 |

## 1.8 Overview of the Remainder of the Book

The example of LMER and multimodel inference in the last section was conceptual. The intent of the remaining chapters is to fill in the details that were glossed over. This includes details of LMER, such as the algebraic formulas and how the parameter estimates are obtained from sample data using `R`. Particulars of multimodel inference and NHST will be presented, illustrating among other things, how the values in Table 1.5 are obtained and interpreted.

An additional goal of the remaining chapters is to expand the repertoire, so that more complex analyses can also be handled. Examples of such analyses include models with multiple static predictors, and models

with nonlinear change curves. For researchers who anticipate using exploratory analysis, descriptive and graphical methods will be covered.