Preface

Before the beginning of years
There came to the making of man
Time with a gift of tears,
–Algernon Charles Swinburne

If we offend, it is with our good will.
That you should think, we come not to offend,
But with good will. To show our simple skill,
That is the true beginning of our end.
– William Shakespeare

Longitudinal data occurs when we repeatedly take the same type of measurement across time on the subjects in a study. My purpose in writing this textbook is to teach you how to think about and analyze longitudinal data.

As a graduate student, I joined the American Statistical Association and began to subscribe to professional journals. I was aware that most people did not read their journals, and in the natural exuberance of early graduate-student-hood I vowed to be different. I opened my first journal with the express intent to read it cover to cover; and quickly discovered not every article was interesting. However, I did read one article thoroughly. Ware (1985) had published an article titled “Linear Models for the Analysis of Longitudinal Studies.” I spent a lot of time trying to understand that article, and in a real sense I am still working on it today. This book is the outcome of my interest in longitudinal data, that began with that article.

Why This Book?

This is a textbook, not a monograph. Included material must be directly helpful when analyzing longitudinal data. Mathematical presentation is kept to a minimum although not eliminated, and statistical computing is not covered.

This book has several key features that other books on longitudinal data do not have. First of all, this book has chapter-length treatments of graphical methods, covariance modeling, and modeling the effects of covariates. These chapters are often only a small section in most other texts currently on the market. The effects of covariates requires at least one full chapter on top of what students have learned about covariates from their linear regression courses.

Many current texts are unbalanced in their coverage of this material. Many texts spend a lot of space on discrete data analysis–an entertaining and important topic. However, like courses on linear regression and generalized linear regression, students should cover linear regression in depth before moving on to logistic and Poisson regression. One book spends more than 25% of its space on missing data modeling. Understanding missing data and bias is an important part of statistical data analysis of longitudinal data. I do provide an introduction to missing data here, but first students need to know how to model regular longitudinal data before spending time learning about missing data.

Texts on longitudinal data from the 1980s and even 1990s are already out of date, usually concentrating on generalizations of analysis of variance rather than on generalizations of regression. The techniques they cover are often archaic. There are also several doctoral-level monographs on longitudinal data that cover multivariate analysis at a more advanced mathematical level, usually including substantial effort on computation and inference, but this is at the expense of not covering the nuts and bolts of data analysis, and those books cannot be read by master's-level students.

A number of texts treat longitudinal data as a special case of repeated measures or hierarchical or multi-level data. Those books emphasize the random effects approach to modeling to the detriment of other covariance models. Random effects models are powerful and flexible, and several sections of this text are devoted to random effects models. However, polynomial random effects models often do not provide the best fit to longitudinal data. Consequently, I treat random effects models as one of several covariance models to be considered when modeling the covariance matrix of longitudinal data.

Computation

I assume that computation will be handled by a software package. Statistical textbooks at the master's level typically do not cover statistical computation, and this book is no exception. My discussion of computation tries to aid the data analyst in understanding what the software does, why it may or may not work, and what implications this has for their own data analysis. I do not discuss code for particular packages because software changes too rapidly over time. It is altered, often improved, and eventually replaced. I am thankful to the vendors that supply programs for analyzing longitudinal data, and I wish them a long and successful run. Extensive software examples will be available on the book's Web site. A link to the book Web site will be located at `http://www.biostat.ucla.edu/books/mld`. You will find data sets, example code, example homework problem sets, computer labs, and useful longitudinal links.

Initially, example code for fitting these models in SAS® Proc Mixed® and Proc Nlmixed® will be available on the course Web site. Sets of computer labs will also be available for teaching longitudinal data analysis using SAS.

Mathematical Background

I have kept the mathematical level of the text as low as I could. Students really should be comfortable with the vector form of linear regression $Y = X\alpha+\delta$ where X is a matrix of known covariates with n rows and K columns, α is a K-vector of coefficients, and Y and δ are n-vectors of observations and residual errors, respectively. I use α rather than the more common β for the regression coefficients. Linear algebra beyond $X\alpha$ is rarely required, and those spots can be readily skipped. In chapters 5 and 6, I write down some likelihoods and the weighted least squares estimator for the regression coefficients in longitudinal data. This requires a few matrix inverses. This material is partly included to assuage my guilt had it been omitted and to provide hooks into future mathematical material should the reader cover more advanced material elsewhere. But this material is not central to the main theme. If the students do not swallow that material whole, it should not impede understanding elsewhere. I do review linear regression briefly, to remind the reader of what they learned before; one can't learn regression fresh from the review, but hopefully it will serve to exercise any neurons that need strengthening.

Multivariate Data and Multivariate Data Courses

Because longitudinal data is multivariate, you will learn something about multivariate data when you read this book. Longitudinal data is not the only type of multivariate data, although it is perhaps the most common type of multivariate data. One of the (dirty little?) secrets of statistical research in classical multivariate data methods is that many methods, while purporting to be multivariate, are actually illustrated on, and mainly useful for, longitudinal data.

Many statistics and biostatistics departments have courses in multivariate data analysis aimed at master's-level students and quantitative graduate students from other departments. These courses cover multivariate analysis of variance (MANOVA) and multivariate regression, among other things. I strongly recommend replacing such a course with a course in longitudinal data analysis using this book. The value of longitudinal data analysis to the student will be much greater than the value of MANOVA or multivariate regression. I often think of this course as a "money course." Take this course, earn a living. I hire many students to analyze data on different projects; it used to be that I required familiarity with regression analysis. Now familiarity with longitudinal data analysis is the most usual prerequisite.

Target Audience

Graduating master's students in statistics and biostatistics are very likely to be analyzing longitudinal data at least some of the time, particularly if they go into academia, the biotech/pharmaceutical industry, or other research environment. Doctoral students and researchers in many disciplines routinely collect longitudinal data. All of these people need to know about analyzing longitudinal data.

This book is aimed at master's and doctoral students in statistics and biostatistics and quantitative doctoral students from disciplines such as psychology, education, economics, sociology, business, epidemiology, sociology and engineering among many other disciplines. These are two different audiences. The common background must be a good course in linear regression. A course at the level of Kutner, Nachtsheim, and Neter (2004), Fox (1997), Weisberg (2004) or Cook and Weisberg (1999) is a necessary prerequisite to reading this book. The seasoning provided by an additional statistics or biostatistics course at this level will be exceedingly helpful. I have taught this material to students from other disciplines whose mathematical background was not up to this level. They found this course rewarding but challenging.

The statistics and biostatistics students bring a deeper knowledge of mathematics and statistics to the course, but often little knowledge of longitudinal data other than perhaps knowledge that longitudinal data is likely to be in their future or on their comprehensive exam. Students from outside stat/biostat tend to have much less mathematical and statistical background. Instead, they bring with them the motivation that comes from having data in hand and needing to analyze it, often for their dissertation. The two different backgrounds can both lead to success in learning this material.

Applied researchers with a good regression course under their belt and some added statistical sophistication should be able to read this book as well. For anyone reading this book, the single best supplemental activity when reading the text would be to have your own data set and to draw all the relevant plots and fit all the relevant models you read about to your own data.

An Overview

This overview is for anyone; but I'm writing it as if I were talking to another teacher.

Chapter 1, *Introduction*, introduces longitudinal data, gives examples, talks about time, discusses how longitudinal data is different from linear regression data, why analyzing longitudinal data is more difficult than analyzing linear regression data and defines notation.

Chapter 2, *Plots*, discusses the plotting of longitudinal data. Intertwined with the plots are ways of thinking about longitudinal data, issues that are naturally part of longitudinal data analysis. Even if you do not wish to cover every last piece of this material in a course, I recommend that the students read the whole chapter.

Chapter 3, *Simple Analyses*, discusses things like paired t-tests and two-sample t-tests and the two-sample t-test on paired differences, called the difference of differences, (DoD) design. These simple analyses are done on various subsets of the data or on summaries of the data. The ideas are re-used in the chapter on specifying covariates. Chapter 4, *Critiques of Simple Analyses*, complains about these analyses and explains some of the problems. Perhaps the real cost of simple analyses is the loss of the richness of multivariate data.

Chapter 5, the *Multivariate Normal Linear Model*, starts with the iid multivariate normal model for data, then introduces parameterized covariance matrices and covariates and the basic aspects of and techniques for drawing conclusions.

Chapter 6, *Tools and Concepts*, contains a grab-bag of useful tools (likelihood ratio tests, model selection, maximum likelihood and restricted maximum likelihood, back-transforming a transformed response, an introduction to design) and discussions about issues with longitudinal data analysis (assuming normality, computation). These tools may be skipped at first reading. However, my suspicion is that those readers who only read a section or two out of the entire book are most likely to dip into this chapter or into one of the topics chapters at the end. Many readers will come back to the various sections of chapter 6 when needed or interested. Most readers will continue on to chapters 7 and 8, coming back to pick up material on model selection, computation, inference as needed.

Chapters 7 and 8, *Specifying Covariates* and *Modeling the Covariance Matrix*, respectively, are the chapters that allow the flavor and beauty of longitudinal data analysis to come to full bloom. As best as possible, I have tried to write these chapters so they could be read in either order. I have tried both orders; my preference is to study covariates first. Covariate specification in longitudinal data analysis requires additional modeling skills beyond what is taught in linear regression and is where the science usually comes in when analyzing longitudinal data. I prefer to have that as early as possible so students can start thinking about their own longitudinal data problems and how to specify the scientific questions. Another reason is that otherwise we are well past the mid-quarter mark before having talked about covariates and that is too long in the quarter to put off talking about covariates. Because there are many short references to covariance matrix specification in chapters 5 and 7, it allows for a softer introduction to the material on covariance models. The downside of this order is that students tend to ask a lot of questions about covariance models before you are ready to discuss them.

Chapter 9, *Random Effects Models*, discusses the random effects model as a hierarchical model, with discussions of random effects estimation and shrinkage. Longitudinal data sets frequently have subjects nested inside larger groups, for example students in classrooms or children in families. We explain how to model this data as well.

Chapter 10, *Residuals and Case Diagnostics*, presents current knowledge about residuals and case diagnostics with emphasis on residuals in random effects models as more is known (by me at any rate) about residuals there than in the general multivariate linear regression model.

Chapter 11, *Discrete Longitudinal Data* introduces discrete longitudinal data models. I discuss the random intercept model for binary data and for count data.

Chapter 12, *Missing Data*, is an introduction to issues surrounding missing data in longitudinal data. We talk about intermittently observed data and dropout and missing at random and variants.

Finally, chapter 13, *Analyzing Two Longitudinal Variables*, introduces bivariate longitudinal data, when you measure two variables repeatedly over time on subjects and wish to understand the interrelationship of the two variables over time.

Teaching from This Book

I teach this book as a quarter course, covering essentially the entire text. Lectures are supplemented with a computer lab that covers the use of a computer program for analyzing longitudinal data.

I have also taught precursors of this material as a subset of a quarter course on multivariate analysis for biostatistics doctoral students. In this course, I cover material from chapters 1, 2, 7, 8, and 9 in three to four weeks, concentrating on the mathematical presentation. I replace chapter 5 with a substantially higher level of mathematical rigor. Chapters 1 and 2 are shortened and the material tightly compacted. Next time I teach that course, I plan to require that students read the entire book and may add parts from chapter 11 and 13 to lectures as well.

A number of homework problems are included. That is how you can tell this is a textbook and not a monograph. The most important homework problems should lead students through a complete analysis of a simple data set. I use the Dental data for this first set of homework problems, which is why it does not appear in the text. Students should first plot and summarize the data, then explore the fixed effects, model the covariance matrix, look at the residuals and finally put their results all together in a report. This can be over a set of three homework assignments. The next assignment(s) can either be a report on the complete analysis of a somewhat more complicated longitudinal data set or another three homework assignments analyzing a data set with unbalanced or random times and more covariates. The last project should be the analysis of a still more complex data set supplied by the teacher or a data set supplied by the student. I do not give exams when I teach this material as a stand-alone course. Report writing supplies a useful form of training that was often historically lacking in statistical training. Ironically, the initial motivation for chapter 7 came from observing the difficulty that many very smart biostatistics doctoral students had in setting up even simple covariate matrices for longitudinal data during comprehensive exams. The Web site has homework assignments that I have used.

Feedback

Comments are actively solicited; especially comments that will help me make the reading and learning experience more helpful for future readers.

Acknowledgments

Many people have provided assistance in the writing of this book, in ways large and small. A number of colleagues have helped indirectly by talking with me about longitudinal data and directly with information for and comments on various drafts of the book. My apologies for omitting way too many of them from these acknowledgments. My thanks to my colleagues at UCLA both inside the Department of Biostatistics and outside for putting up with, encouraging, ignoring, and abetting. Particular thanks to Robert Elashoff, Bill Cumberland, and Abdelmonem Afifi for early encouragement.

Students in the courses Biostat 236 and Biostat 251 at UCLA have sat through many presentations of this material over a number of years and have contributed much through their questions, enthusiasms, homework answers, report writing, yawns, laughter, and typo reports. I'd like to thank them for being willing participants in reading early versions as they were written and in particular for letting me read the notes to them and for helping me catch the typos on the fly. A number of students have written master's papers and doctoral dissertations with me on longitudinal data analysis; every one has helped me understand the subject better.

I have had tons of assistance in data management and in writing code using SAS, R©/Splus®, and ARC©/xlispstat©. Thanks to Charlie Zhang, Zhishen Ye, Yunda Huang, Susan Alber, Leanne Streja, Lijung Liang, Luohua Jiang, Jim Sayre, John Boscardin, Scott Comulada, Zhen Qian, Wenhua Hu, and others who I have unfortunately omitted.

I'd like to thank Sandy Weisberg for LaTeX® help and for always answering my questions about almost anything; John Boscardin for programming, LaTeX, and longitudinal help; Marc Suchard for detailed comments, and many many discussions, breakfasts, bagels, and pushes; Steve West; Eric Bradlow; Bill Rosenberger; Billy Crystal in Throw Momma from the Train: "A writer writes: always!"; Lynn Eberly for particularly helpful comments and for encouragement; several anonymous reviewers for comments both general and detailed; Susan Alber for comments on the writing, help with SAS, and teaching me about longitudinal data analysis. A big thanks to John Kimmel for his patience, encouragement, stewardship, and for finding the right answers when it mattered.

I have gotten data sets from a number of places. I'd like to thank Dr. Lonnie Zeltzer for the Pediatric Pain and Vagal Tone data; Dr. Mary Jane Rotheram for the BSI and other data sets; Dr. Charlotte Neumann for the Kenya data; Robert Elashoff for the Weight Loss data.

Finally, I would like to thank my multi-generational family for putting up with me while I worked on this. You have often asked when this book would be done. If you are reading this for the first time, check your watch and you will have your answer.

Robert Weiss
Los Angeles
2005

2
Plots

"Do you see the big picture, [Meehan]?"
"Never have, Your Honor," Meehan told her. "I'm lucky if I make sense of the inset."
– From Donald E. Westlake's *Put a Lid on It.*

There I shall see mine own figure.
Which I take to be either a fool or a cipher.
– William Shakespeare

Overview

In this chapter, we cover the following

- Plotting longitudinal data
  - What we want from our graphics
  - Defining profile plots
  - Interpreting profile plots
  - Variations on profile plots
- Empirical residuals
- Correlation
  - Correlation matrix
  - Scatterplot matrices
  - Correlograms

- Empirical summary plots
- How much data do we have?

This chapter presents exploratory data graphics for longitudinal data. Most graphics in research reports are used to present inferences. Long before we draw conclusions, we must understand our data and determine the models we will use. In drawing exploratory graphics, we plot the data in ways that shed light on modeling decisions. Longitudinal data are more complicated than cross-sectional data, and our plots will be more complex as well.

Our discussions will assume balanced or balanced with missing data. Most ideas apply equally well to random time data. The problem with explicitly including random time data in the discussion is that there is a significant increase in notational complexity without much corresponding benefit.

2.1 Graphics and Longitudinal Data

General multivariate data are difficult to plot in a way that provides insight into the data. Much ingenuity has been expended on creating such plots. In contrast, a number of useful plots for longitudinal data exist. The reason for the difference is that the units of measurement for longitudinal observations Y_{i1}, Y_{i2}, $\ldots$, are all identical. Within-subject comparisons make no immediate sense for general multivariate data. It is hard to compare heart rate and blood pressure as part of a multivariate observation; the units are not directly comparable. Only between-subject comparisons of corresponding measurements such as $Y_{i1} - Y_{l1}$ or $Y_{i2} - Y_{l2}$ make sense. In contrast, with longitudinal data we can take differences between observations within subjects. The difference $Y_{i2} - Y_{i1}$ is the increase in the response from the first to the second observation, and we want plots to show that change. We may take differences of similarly timed measures between subjects. Suppose that $t_{ij} = t_{lj}$ for subjects i and l, then $Y_{ij} - Y_{lj}$ is the amount that subject i is higher than subject l at time j. Even if times t_{ij} and t_{lk} are different, the difference $Y_{ij} - Y_{lk}$ is still interpretable.

What are the quantities we want our graphics to show? Some basic quantities we are interested in are the value of a particular observation Y_{ij} and the average response from subject i

$$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$

We want to compare observations within a subject. For example, we want to evaluate the difference $Y_{ij} - Y_{i(j-1)}$, and we need to compare observations across subjects at a particular time as in $Y_{ij} - Y_{lj}$. We want to answer questions such as which observations Y_{ij} are highest or lowest and which subjects are highest or lowest on average. We would like to assess the average response across subjects at a single time j

$$\bar{Y}_{.j} = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}$$

and the sample standard deviations

$$s_{jj} = \left[\frac{1}{n-1}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_{.j})^2\right]^{1/2}$$

of the Y_{ij} at a specific time j. We want to know if these means and standard deviations are increasing, constant, or decreasing over time. The ratio

$$\gamma_{ij} = \frac{Y_{ij} - Y_{i(j-1)}}{t_{ij} - t_{i(j-1)}} \tag{2.1}$$

is the slope of the line segment between observation $j-1$ and j for subject i. We will want to compare these slopes for different j within subject and also across different subjects at similar or different times. Are the slopes increasing over time or decreasing? Is the typical subject's slope increasing or decreasing at time t? What is the average of γ_{ij} over subjects i? We would like to see which subjects have similar profiles on average $\bar{Y}_i = \bar{Y}_l$ or have profiles with similar patterns over time such as $Y_{ij} - \bar{Y}_i = Y_{lj} - \bar{Y}_l$ even though their averages may not be the same.

So far we have mentioned basic features of our data; observation level Y_{ij}, subject average level $\bar{Y}_i$, across subject within time average level $\bar{Y}_{.j}$; differences $Y_{ij} - Y_{ik}$, $Y_{ij} - Y_{lj}$, standard deviations s_{jj}, and slopes $(Y_{ij} - Y_{i(j-1)})/(t_{ij} - t_{i(j-1)})$. Thinking now not about features of the data, but features of the models we will be creating, what are the basic components of our models? We want our plots to help us with specification of these components. The basic features of our models will be

- the population mean response at a particular time,
- the population variance or standard deviation of the responses at a particular time,
- the correlations between observations within subjects, and
- the effects of covariates on these quantities.

We want our plots to show us information so we can specify our models appropriately. Mainly we want to make qualitative judgments from our plots; quantitative judgments are reserved for the output of our models. We do not need to learn that the mean of the observations at day 2 is 200 and that at day 20 it is 950. Rather, we need to learn from our plots if the mean response is increasing over time or not. If the mean response is increasing, is the increase linear or something more complicated?

2.2 Responses Over Time

Time permeates all longitudinal data analyses. The first graphic we make plots the longitudinal response against time. The obvious first plot we might consider plots all responses Y_{ij} against time t_{ij}. Figure 2.1 shows this plot for the *Big Mice* data. The response is the weight in milligrams for $n = 35$ mice with each mouse contributing observations from various days starting at birth, day 0, through day 20. Thirty-three of the mice were weighed every three days for a total of seven observations each. Eleven mice in group 1 were weighed beginning on day 0, ending on day 18; group 2 has 10 mice weighed beginning on day 1 ending on day 19; and group 3 has 12 mice weighed beginning on day 2 ending on day 20. The last two mice are in group 4 and were weighed daily from day 0 to day 20. A subset of the Big Mice data forms the *Small Mice* consisting of the group three mice plus the group four observations on the same days. The Small Mice form a balanced subset of the data, whereas the Big Mice data are balanced with lots of data missing. Each mouse comes from a separate litter, so it is reasonable to treat mice as independent. All weighings were performed by a single person using a single scale.

See the data set appendix for details about any given data set. We will discuss data sets in the text as we need the information. To make it easy to find this information at a later time, data set descriptions are kept in the data set appendix. The mice data set description is in section A.2.

2.2.1 Scatterplots

Figure 2.1 is a scatterplot of weight against time with all $33 \times 7 + 2 \times 21 = 273$ observations plotted. The weights start out low and grow rapidly. On days 0 and 1, the weights are all less than 200 milligrams (mg), by day 5 the average weight has more than doubled, and the weights more than double again by the end of the study at day 20. Somewhere around day 10 the daily increase in weight appears to slow down although the exact pattern of increase is unclear.

2.2.2 Box Plots

There is a fair amount of over-plotting of circles in figure 2.1, and with smaller page size, poorer quality graphics or larger data sets, over-plotting can be even worse. One solution that people have used is to plot repeated box plots over time. Rather than attempting to plot all of the observations, the box plot summarizes the observations at each time point and does a careful job of presenting the summary.

Figure 2.2 shows 21 repeated box plots of the mice data. Each box plot summarizes the observations taken on one particular day. The central divided rectangular box plots the lower quartile (lowest line), the median

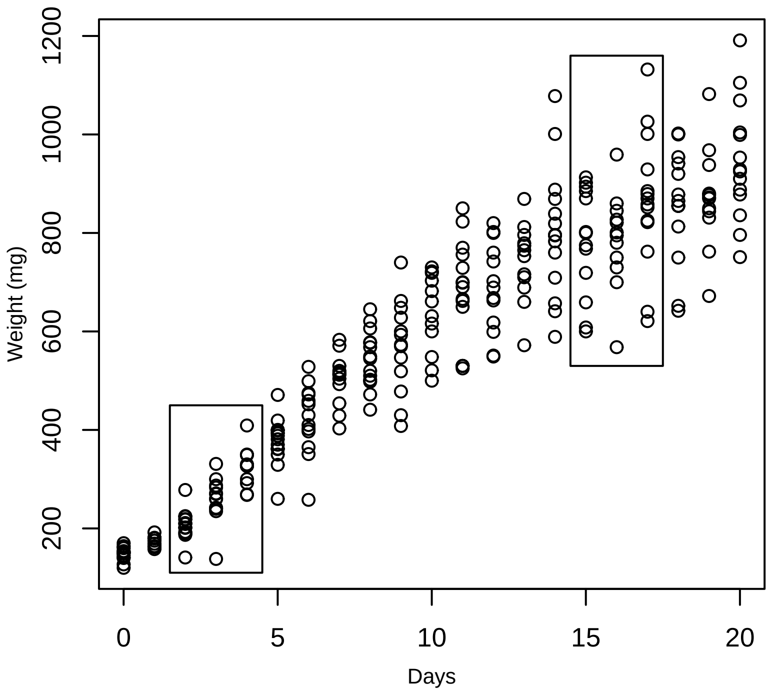

Figure 2.1. Scatterplot of Big Mice weights in milligrams against time in days. Thirty-three of the mice contribute 7 observations each and 2 mice contribute 21 observations each. The boxes are explained in section 2.3.1.

(middle line), and the upper quartile (upper line) of the observations measured on that day. The lower (upper) quartile is the observation with at least 25% (75%) of the observations at or below it and at least 75% (25%) of the observations at or above it. The *whiskers* are dashed lines extending from the lower and upper quartiles to the minimum and maximum values indicated by the short horizontal lines. The box plot shows the interquartile range, the upper quartile minus the lower quartile, and it shows the range, the maximum minus the minimum. The box plot is less crowded than 2.1. Figure 2.1 tries to show every data point, while the box plot displays five summary statistics of the data at each time point.

We again see the sharp rise in the weights over time. The increase accelerates around days 2–6. The medians, for example, increase rapidly each day until around day 11, when they grow less quickly. Around days 12–14 the rise in the medians continues, but perhaps at not such a sharp rate. Thereafter the increases are uneven. At days 14–16 the median weight is nearly constant and again for days 17–19.

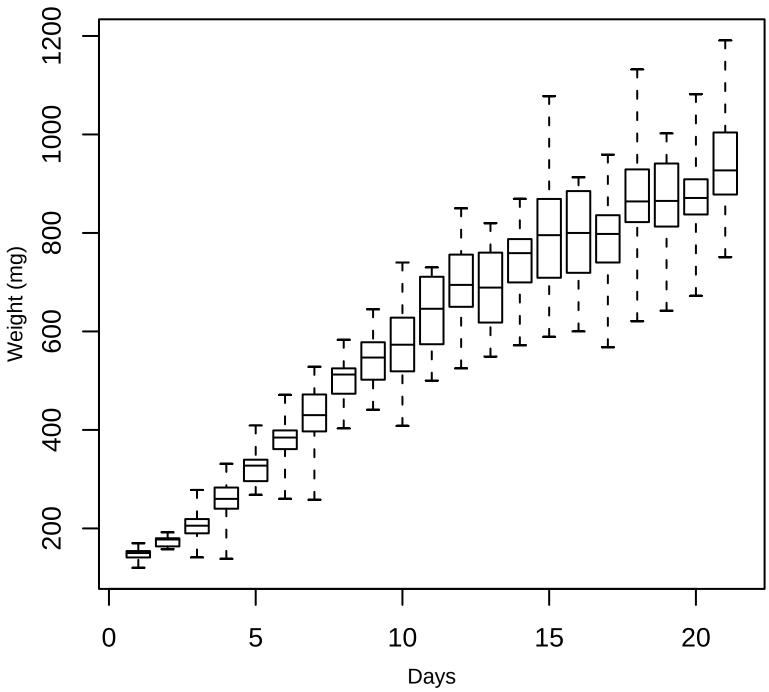

Figure 2.2. Repeated box plots of the Big Mice weights over time. Each box plot summarizes the distribution (minimum, lower quartile, median, upper quartile, maximum) of weights observed on that day.

The variability of observations on a given day increases as the mice get older. As the mice age, we see in figures 2.1 and 2.2 that the variability in weights increases up till perhaps around day 9, then at some point, possibly day 14, it increases again. In figure 2.2, the range and the interquartile range appear to increase over time as well. As the mice grow, the range increases and then appears to stabilize around day 9 or so.

2.2.3 Flaws in Scatterplots and Box Plots for Longitudinal Data

The data points and data summaries in figures 2.1 and 2.2 are not independent. One may have experience looking at a scatterplot of a response y versus a predictor x and deciding whether there is a significant or important differences in the response as functions of time. Because our observations come from the same mice at different time points, our intuition based on independent observations may not apply. For neighboring days, the observations are almost independent except for the two mice in group 4 who contribute data to all days. If we wish to compare the data from two days that are multiples of three days apart, we have correlated data. If we were to compare the means between days 4 and 7 for example, we would do it with a paired t-test, not a two-sample t-test. Figures 2.1 and 2.2 do not show the connection between observations from the same mouse.

It is a flaw of these figures that we cannot tell which observations come from the same mouse. Three of the four largest weights are the largest weights at days 14, 17 and 20. Because most mice are weighed every three days, we suspect, but cannot tell, that these observations belong to the same mouse. The largest observations on days 14, 17, and 20 weigh much more than the largest observations at days 15, 16, 18. Day 14's maximum is even slightly higher than the maximum at day 19. Similarly, days 2 and 5 have measurements distinctly lower than the other observations, and we suspect, but cannot tell, that we are looking at two measurements of one mouse. There is a similar low pair of observations at days 3 and 6.

Additional features of our data that we cannot identify include the differences $Y_{ij} - Y_{i(j-1)}$ or the slope between consecutive observations within a mouse, nor can we identify whether a particular mouse is high or low compared to the remaining mice. Longitudinal data have a natural hierarchical structure that should be reflected in our plots. Observations within subject are correlated and the nesting or clustering of observations within subject should be encoded in our plots.

2.2.4 Profile Plots

A *profile* is the set of points (t_{ij}, Y_{ij}), $j = 1, \ldots, n_i$. A *profile plot* improves on the basic scatterplot 2.1 by using line segments to connect consecutive observations (t_{ij}, Y_{ij}) and $(t_{i(j+1)}, Y_{i(j+1)})$ within a subject. No lines are drawn between observations from different subjects. Profile plots are useful because the clustering of observations within subject is directly visible. In a profile plot, the basic plotting unit is not the observation (t_{ij}, Y_{ij}), rather it is the entire profile (t_i, Y_i). The profile plot in figure 2.3 displays the mice data. We can see that a single mouse is heaviest at days 14, 17, and 20. It was not the heaviest mouse at time 11 or earlier. The second heaviest mouse at days 14 and 17 is outweighed by yet another mouse at day 20. We see that the mouse that was heaviest at days 3–6 was one of the two mice that were measured daily. It ends up among the heaviest mice but is not the heaviest.

Generally we see that the mice all grow in parallel; if mouse A is heavier than mouse B at an earlier time, it has a tendency to be heavier at a later time. This is particularly clear after day 9 or 10; the plot is cluttered before day 9 and it is not so easy to see if this is true for observations from before day 9. Mice that are close in weight may change rank from day to day, but if mouse A is more than 100 milligrams greater than mouse B after day 9, it is unlikely to ever be lighter than mouse B.

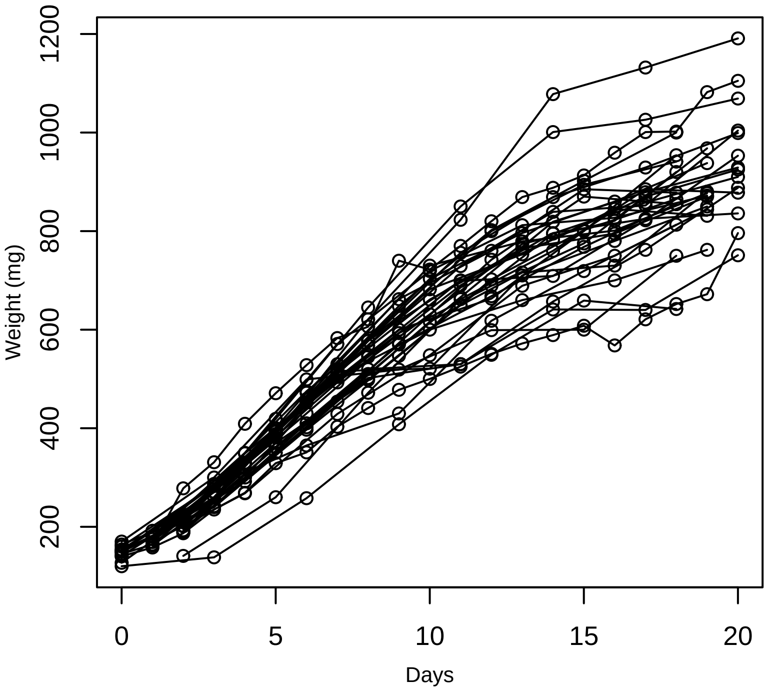

Figure 2.3. Profile plot of mice weights against time. Consecutive observations within a mouse are connected by line segments.

In figure 2.3, we also notice that on a few occasions, mice actually decrease in weight from one observation to the next, something we could only infer from the previous two plots and only for a few special circumstances. Examples include the heaviest mouse at day 9, which lost weight at day 10, and the second lightest mouse at day 15, which is the lightest mouse on days 16 and 17.

2.2.5 The Need to Connect-the-Dots in Profiles

The need for line segments in profile plots is illustrated in figure 2.4. Five fictional subjects contribute six observations each to figure 2.4(a). From this plot we do not know which observations belong to which subjects. We can only learn about the *marginal distribution* of the data at any given time point. Pick any point t on the time axis, and look at the collection of observations above t, and perhaps within a window slightly to either side, say observations with times in the range $t - \Delta, t + \Delta$ for some modest value of Δ. Average the responses in the window, and look at a number of windows centered at different times t. We learn that the average value appears to be fairly constant across time. To the right of the middle of the time axis, there appears to be possibly less variability in the response values or possibly there are merely fewer observations at that time. We do not know the reason for this lower variability. From this plot we do not learn about the *joint distribution* of observations within a subject. In 2.4(a) we do not know, for example, if the largest half a dozen observations across all time points belong to the same or different subjects.

Figure 2.4(b) presents a possible profile plot for the data in 2.4(a) with observations within subjects connected by consecutive line segments. Profiles are labeled by subject id from 1 to 5 at the left of each profile. Subject 1 has the highest response values at all times, and subject 5 has the lowest responses. Generally, observations within subject at different times have similar responses Y_{ij} across time. If subject A begins higher than subject B at the left side of the plot (early times), then A's observations are higher in the middle (middle times) and again at the right side at the latest times.

In contrast, figure 2.4(c) represents a different assignment of observations to subjects. At the earliest times, the subject profiles are the same as in figure 2.4(b). However, somewhere in the middle of time, subjects who start low tend to rise, while subjects who start high tend to fall, and at the late times, subject 5 who started lowest is highest, while subject 2 for example, who started second highest ends as second lowest. Subject 3 has a flat profile throughout: subject 3 had an average response in the beginning is still average at the end.

Figures 2.4(b) and 2.4(c) suggest different explanations for the reduced variance of the responses in the late middle time region. In figure 2.4(b), it appears that the reason for the gap is that there were few observations on subjects 4 and 5 around that time, whereas there were plenty of observations for subjects 1, 2, and 3. If we had observations on subjects 4 and 5 to the right of the middle time we would expect them to be low; we expect the variability of the responses across subjects to be roughly constant across time. Figure 2.4(c) is different; we see that each subject's responses appear to be following their own line as a function of time. Each subject has a different slope and intercept. Subjects 1 and 2 have negative slopes, 3 is flat, and subjects 4 and 5 have positive slopes. In figure 2.4(c), it appears that no matter how many observations we collected from these 5 subjects, we would see the same decrease in variance to the right of the middle; observations around the point where the lines cross will always be tightly clustered, and thus the variability of the responses will be lower to the right of the middle time.

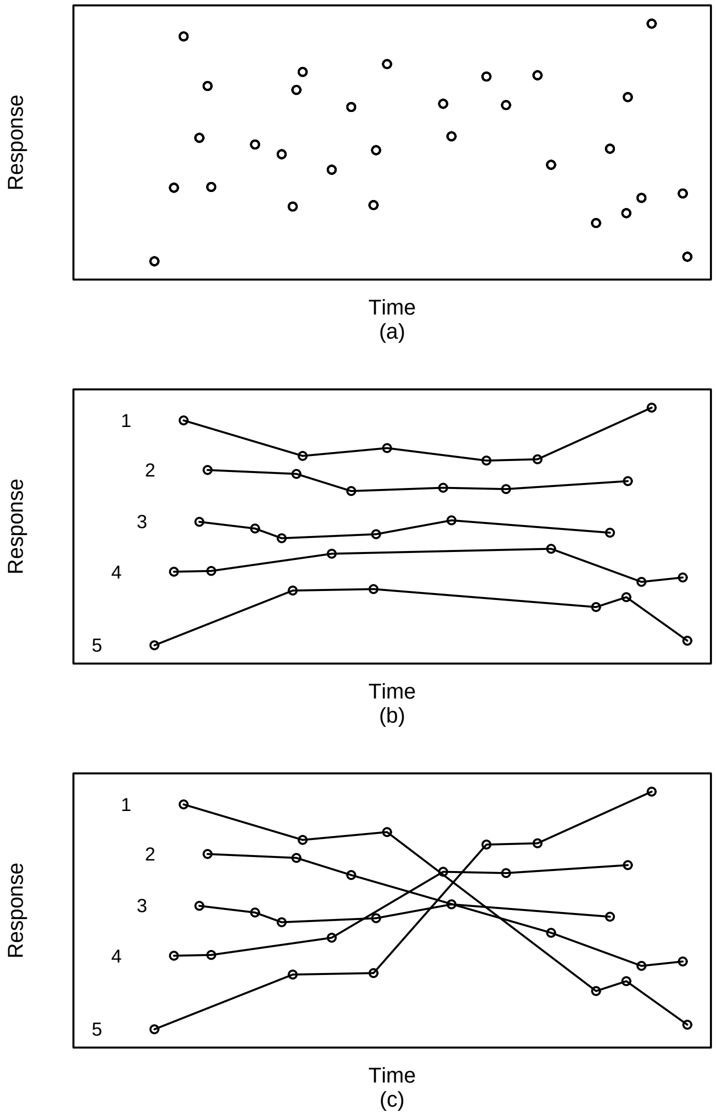

Figure 2.4. Plotting longitudinal data. (a) Scatterplot of responses Y_{ij} against t_{ij} for 5 fictional subjects of 6 observations each. (b) Possible profile plot based on the observations in plot (a). (c) Alternative profile plot based on the observations in plot (a). Subjects are labeled 1–5 to the left of their earliest observation in (b) and (c).

2.3 Interpreting Profile Plots

We can read basic information about subjects' response patterns from a profile plot. Figure 2.5 illustrates four different situations. Each subfigure displays data from eight hypothetical subjects. Subjects are measured at random times, usually with 5 observations per subject. Individual observations are plotted using a circle and, as before, line segments connect consecutive observations within a person. Later on we will drop the circles and only use the connected line segments. Figure 2.5(a) shows a very common situation not unlike that in 2.4(b). We see that each profile is roughly flat, with observations near a subject-specific mean. Individual profiles generally do not cross other profiles, that is, they are roughly parallel. There are but a few exceptions in the middle of the data where one profile crosses another.

If we extrapolate each subject's profile back in time to the time $t = 0$ axis, each profile would intersect the axis at a subject-specific intercept. If subjects are a random sample from the population of interest, then any subject-specific characteristic is also a sample from the population of possible values of that characteristic. In particular, the intercepts are a sample from the population of intercepts. We say that the data in 2.5(a) has a *random intercept*. The term random is used in the same way as when we said that the subjects in our study are a random sample from the population under study. Another way to say random intercept is to say that each subject has their own subject-specific mean response and that observations vary around the mean response.

In figure 2.5(b), the profiles are again parallel, but this time each has a linear time trend with a positive slope. If we extrapolate by eye back to the origin, the profiles all appear to have different intercepts, and again we conclude that the data has a random intercept. When we look at the slopes of the profiles, all of the slopes appear to be about the same. Here, we have a *fixed slope*, a slope that does not vary by subject. We conclude that the population also has a fixed slope; each subject's responses increase at the same rate over time.

The data in figure 2.5(c) illustrate a different pattern of responses. Most of the profiles start low at time $t = 1$, and grow larger as time progresses. There is one unusual profile that starts high and does not grow over time. We identify that subject as an outlier, and would strongly consider removing it from the data set before fitting models to this data. The remaining profiles are linear with similar initial values at the earliest measurement but they increase over time at different rates. We conclude that we have a random slope and a fixed intercept in the population. The unusual subject's earliest observation is a univariate outlier; we can identify univariate outliers on a profile plot when a single observation (t_{ij}, Y_{ij}) is the most extreme Y-value, either highest or lowest at time t_{ij}, or, for random times,

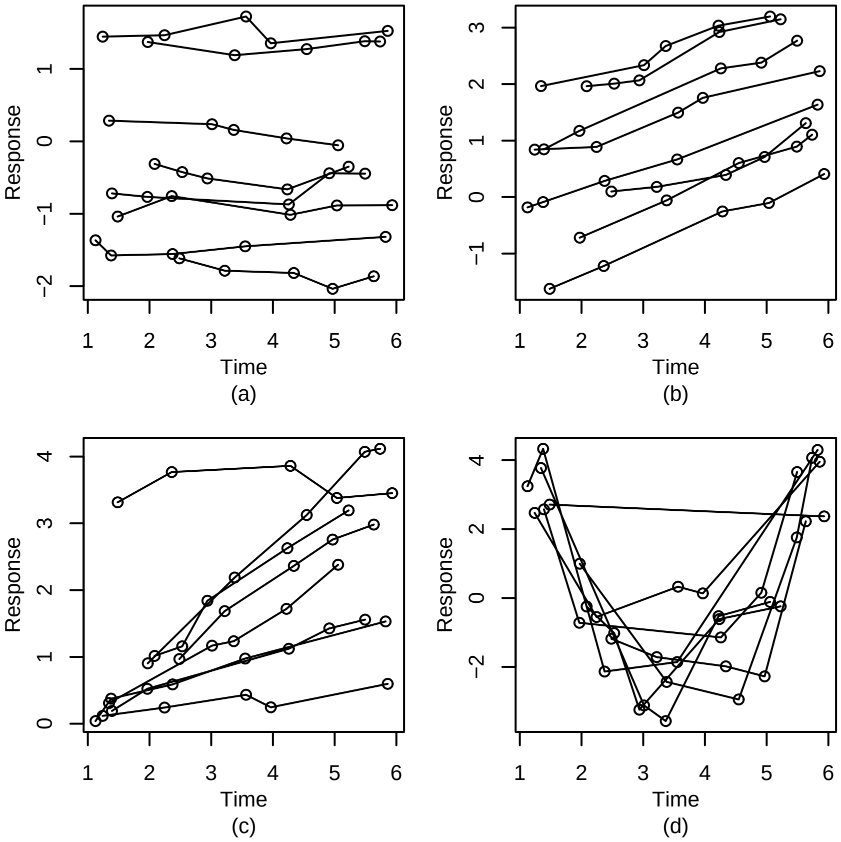

Figure 2.5. Example profile plots. (a) Random intercept, constant population mean; (b) random intercept, positive fixed population time trend; (c) random slope, fixed intercept, with one outlying profile; (d) fixed quadratic.

if Y_{ij} is the most extreme Y-value for all observations within a narrow window of time centered on t_{ij}.

Bivariate outliers (t_{ij}, Y_{ij}), (t_{ij+1}, Y_{ij+1}) can be identified if the line segment connecting them is unlike all the other line segments in the same region of time. The unusual subject in figure 2.5(c) also begins with a bivariate outlier, as no other subject has a high followed by a high first two observations. Bivariate outliers not necessarily need be univariate outliers. Imagine in figure 2.5(a) a subject with points at (t, y) equal to $(2, -1)$ followed by $(3, +1)$. Neither y-value of -1 or $+1$ is unusual, but the line segment connecting them would have the largest slope of any other line segment in the plot, by a substantial margin, indicating that this was an unusual pair of observations.

The final figure 2.5(d) shows subjects following a quadratic trend in time. The shape of the quadratic appears to be the same for each subject, and we conclude that the data follows a fixed quadratic path. At first glance there appears to be a single outlying subject. Closer inspection reveals that that subject has but two observations, one early at $t \approx 1.5$ and one late at $t \approx 6$. The $\approx$ is read *approximately equal to.* The impression of an outlier is given because we use a line segment to interpolate between the two observations, while the bulk of the observations follow a distinct non-linear trend between those times.

Many other possible patterns of profiles exist. We can imagine a plot where every profile is approximately linear, each with a different slope and intercept. The data would have random intercepts and slopes. We can imagine many forms of curvature over time too innumerable to even begin to discuss. Problem 9 presents a few possibilities.

2.3.1 Sample Means and Standard Deviations

Key features we can estimate informally from profile plots are population quantities such as the population mean or population standard deviation of the data as a function of time. Suppose we have a balanced data set with no missing data, and subjects are a random sample from our population. We might take a mean of all observations at each time point where we have data. These means estimate the *population mean* as a function of time. If we hypothetically had observed all subjects, then the mean of all observations at a given time is the population mean at that time! Depending on need, we may plot these sample means over time, or we might roughly eyeball them merely by viewing the profile plot. Inspection of the means will indicate to us whether the population mean is constant over time or if it is increasing or decreasing and whether the population trend is linear or not. If the linear trend is modest, we may need to resort to a formal statistical test to determine significance, and when presenting our conclusions to others we almost always supplement our informal judgments with statistical tests.

The population standard deviation at a given time is the standard deviation of a set of observations, one per subject, if we had observed all subjects at a single time. Given a sample of subjects, we have an estimate of the population standard deviation. The population standard deviation measures the within-time across-subject variability.

When we have random times or balanced with many missing data, or just sparse data, we may not have enough data to calculate a mean or standard deviation at a given time or there may be too few observations to get a reliable estimate. Instead, we may pool responses taken from observations with similar times to calculate our mean or standard deviation (sd). In particular, we might take all observations $Y(t)$ within a *window* along the time axis. The window has a midpoint at time t_M, a width w, a left endpoint $t_L = t_M - w/2$, and a right endpoint $t_R = t_M + w/2$. We collect all the

Y_{ij} values from observations whose t_{ij} are in the window, $t_L \leq t_{ij} \leq t_R$; and we perform some statistical operation, for example, mean, sd, min, max, or median on those observations. We plot the resultant mean, sd, or other quantity against the midpoint t_M. Next we move the window along the t axis, moving t_M from one end of the data set to the other. For each window, we calculate the same statistical operation, and we plot the result against the window midpoint, connecting the resulting points by line segments. When our operation is the mean or a quantile, we may plot the summary on the same plot as the data. For a range or standard deviation, we would plot the ranges or sd's against the window midpoint on another plot because these values lie on a scale different from the original data. We typically pick the window width just large enough to give us enough data to make a decent estimate but not so wide that the estimate becomes meaningless.

Figure 2.1 illustrates two windows of width 3 days, one from 1.5 to 4.5 and one from 14.5 to 17.5. The two boxes in the plot enclose all of the observations in the two windows. The mean of the observations in the left window is 260 mg while the mean of the observations in the right window is 810 mg. The standard deviations are left window 61 mg and right window 120 mg. We reasonably conclude that both the population mean and sd are larger around time $t = 16$ than around $t = 3$. An issue is how big the window should be. Around time $t = 3$, the means are increasing rapidly, and taking a wide window may cause us to overestimate the standard deviation. With the Big Mice data, we have enough data to keep the window width down to a width less than 1. We would then take the mean and sd at each time point, and plot them against that time point. These two plots are illustrated in figure 2.6. We see in figure 2.6(b) that the sd at time $t = 3$ is between 40 and 50, and because our earlier window was wider than necessary, it did indeed overestimate the standard deviation.

Inspecting the two plots, we conclude that the mice means increase smoothly over time. The increase is not quite linear, with a slight acceleration in the beginning, and then a slight deceleration after day 10. The sd's also increase in a smooth but somewhat curvilinear pattern. The sd's bounce around more from day to day than do the means. In general, standard deviations are harder to estimate than means, and this is reflected in the greater variability of the standard deviations over time.

Often we do not formally draw figures such as 2.6 or decide on a specific window width. For example, in figure 2.5(a) we see that the minimum response, the maximum response, the average observed response, and the range of the responses seem to be nearly constant over time. We identify these statistics as a function of time by, for example, looking at the subset of observations in the window between the times $t = 1$ and $t = 2$ and comparing that set of observations to the set of observations between for example $t = 5$ and $t = 6$. The maximum value in these two time intervals is nearly identical and come from the same subject. The minimum values

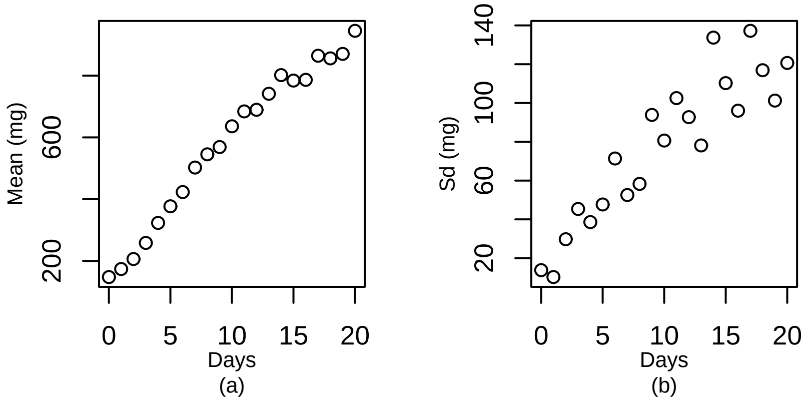

Figure 2.6. Big mice data. (a) Sample means by day. (b) Sample standard deviations by day.

are slightly different and come from different subjects. But this difference is small and is attributable to sampling variability and the fact that the lowest valued subject did not provide any observations before $t \approx 2.5$. Because the time trend within subjects seems flat, and the overall impression of the sample average time trend seems flat, we reasonably hypothesize that the trend of the population mean over time is flat.

Both the minimum and the maximum in figure 2.5(b) appear to be increasing linearly over time. However, the range = max − min is roughly constant over time. The distribution of observations between the max and min is fairly uniform and we conclude that the *population variance* of the responses is constant over time.

The population sd over time of the responses increases in figure 2.5(c). Ignoring the outlier, the range of the 5 responses taken at around $t = 1$ is less than 1/2 of a unit, from just above zero to less than .5. At time $t = 6$, the observations range from a minimum of around .5 to a maximum near 4. We conclude that the population mean, sd, and range are increasing over time. A rough estimate of the standard deviation is range/4; the range appears to be linearly increasing with time, and we conclude that the range and the sd are increasing in an approximately linear fashion.

We do these sample mean and sd calculations as steps to a further end: the development of a model for the responses as a function of time. For the Big Mice data, we have learned that any model must allow for a population mean and sd that are increasing smoothly with time.

2.3.2 Skewness and the Pediatric Pain Data

Pain is a difficult subject to study because it is hard to design formal experiments if one does not wish to inflict pain on humans; rats cannot tell

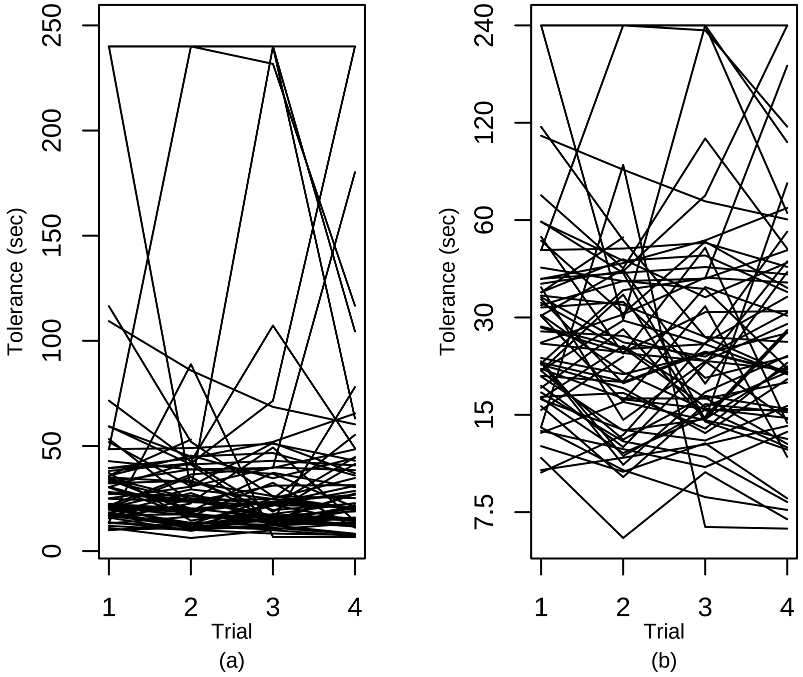

Figure 2.7. Profile plot of the Pediatric Pain data. (a) Original scale. (b) Log scale.

us where it hurts. Most human pain studies are observational. The Pediatric Pain data are unusual in being the result of a designed experiment. The data consist of up to four observations on 64 children aged 8 to 10. The response is the length of time in seconds that the child can tolerate keeping his or her arm in very cold water, a proxy measure of pain tolerance. After the cold becomes intolerable, the child removes his or her arm. The arm is toweled off and no harm is caused. The procedure was considered enjoyable by the children. No one declined to participate in the experiment initially, and no one dropped out for reasons related to the experiment although there is some missing data due to absences and broken arms, common occurrences in children and unrelated to the experiment. Two measures were taken during a first visit followed by two more measures during a second visit after a two-week gap.

During the first visit, the children were classified into one of two groups, *attenders* (A) or *distracters* (D) according to their style of coping (CS) with the pain. The children were asked what they were thinking about during the trials. Those who were thinking about the experiment, the experimental apparatus, the feelings from their arms and so on were classified as attenders. Those who thought about other things, such as the wall, homework from school, going to the amusement park, all unrelated to the experiment, were classified as distracters.

A treatment (TMT) was administered prior to the fourth trial. The treatment consisted of a ten-minute counseling intervention to either attend (A), distract (D), or no advice (N). The N treatment consisted of a discussion without advice regarding any coping strategy. Interest lies in the main effects of TMT and CS and interactions between TMT and CS. Interactions between TMT and CS were anticipated.

The data are plotted in figure 2.7(a), circles for individual observations are omitted. Time is the trial number, ranging from 1 to 4. We see a large mass of observations at times under 50 seconds and relatively sparse data at larger times. This suggests that the data are skewed and that we should consider a transformation. Figure 2.7(b) is the same data on the log scale, labeled with a logarithmic scale on the y axis. The profiles are evenly distributed throughout the range of the data. The data are perhaps slightly more sparse near the top than the bottom, and while we could consider a slightly stronger transformation, the amount of skewness seems minor.

In linear regression, we often use a histogram of our responses to determine if we should transform the data. For longitudinal data, it is not correct to pool all observations into a single histogram. One could draw histograms of the data at a single time point. If we have random times, then we could plot a data set consisting of but one observation per subject, all taken from a narrow window of time. Would it make sense to plot all of the Big Mice data in a single histogram? Two mice would contribute 21 observations and the rest would contribute 7. Consider data like in 2.5(c), with fixed intercept and with the random slopes ranging from 0 up to some positive value. If we had no outlier, and if the study had continued on longer, a histogram of the entire data set would look skewed, yet it would not be correct to transform the data. Histograms of the data in a reasonably narrow window about any time would correctly indicate no need to transform the data.

In figure 2.7(a), there seemed to be a lot of univariate outliers; all the observations above approximately 75 seconds or so. In figure 2.7(b), these high observations seem much less troublesome. Still, a few outliers are visible. The subject with the lowest times at trials 3 and 4 has an unusually high pain tolerance at trial 2. This high trial 2 value causes the line segment between trials 1 and 2 and also between trials 2 and 3 to travel in directions very different from the other line segments between these times. This indicates that the (Y_{i1}, Y_{i2}) pair and the (Y_{i2}, Y_{i3}) pairs for this subject are bivariate outliers. We identify this subject as an overall outlier.

In figure 2.7(a) and also (b), we also see that there appears to be a fixed maximum above which no child scores. Inspection of the data shows that these values are 240.00 seconds, or 4 minutes exactly. Sometimes the investigator may not mention this to the statistician; graphics help us discover these features of the data. The investigators felt that if immersion lasted

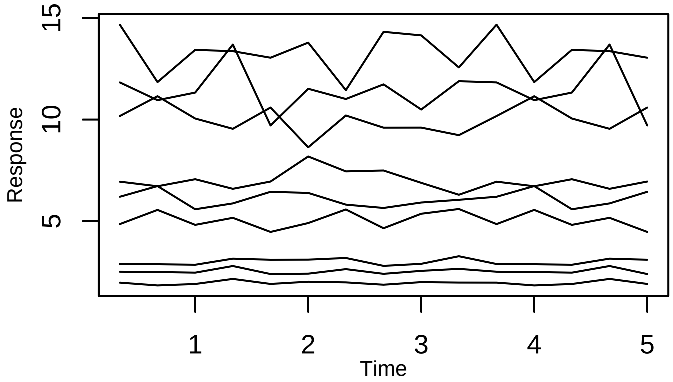

Figure 2.8. Example illustrating different within subject variability. Observations from a single subject have a random intercept and no time trend. Bottom three subjects have observations with the least amount of within-subject variability. Middle three subjects have a middle amount of variability within subject, and the top three subjects have the most within-subject variability.

past 4 minutes, there was no extra information to be gained from allowing the child to keep their arms in longer. This happened in 11 observations out of the 245 total number of observations. The investigators recorded 240.00 seconds as the response in these trials. This *censoring* should be taken account of in the modeling although we do not do this in the analyses presented here.

Figure 2.7(a), is not a beautiful plot; one would never publish it in a medical journal as part of reporting an analysis of this data. Still, this is potentially the single most important step in the analysis of this data! We learned (1) that we should transform the response, (2) that there was possible non-constant variance, (3) that there were some outliers, and (4) that there was a maximum value imposed on the data by the investigators.

2.3.3 Within-Subject Variability

In the Big Mice data, we saw different marginal variances at different times. This was summarized in figure 2.6(b), which plotted standard deviations across-subjects within-time. We also can think about *within-subject, across-time* variability. Figure 2.8 illustrates. The nine subjects each have their own (random) intercept and profiles that have no trend over time. The three subjects with the smallest responses have observations that vary around their means in a tight pattern without much variance. The three subjects in the middle have greater variability around their means, and the three subjects with the highest means have observations with the highest variability around their individual means. The range of the observations within person is low for the subjects with the lowest values; it is middling for the

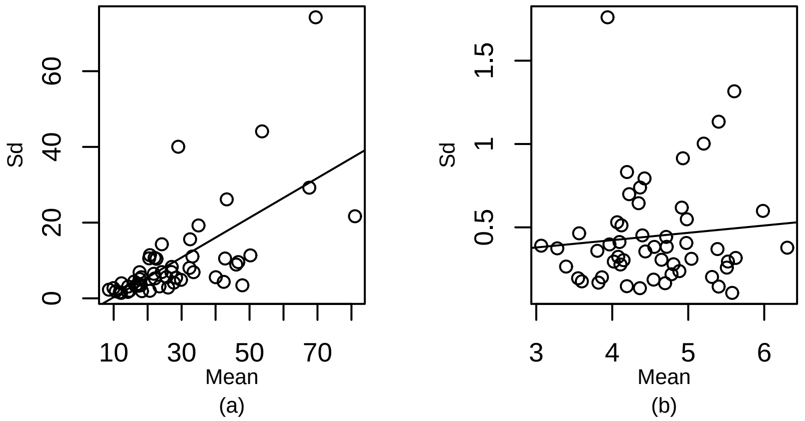

Figure 2.9. Pain data. Plot of standard deviations (y) against means (x) for subjects with 4 observations and no values of 240 seconds. (a) original data, (b) log base 2 data.

subjects with the middle means, and it is highest for the subjects with the highest means.

The Pediatric Pain data in figure 2.7(a) appear to have higher within-subject variability for subjects with higher means, and subjects with lower means appear to have smaller variability. It is somewhat hard to be absolutely certain. Inspection of figure 2.7(b) suggests that the log transformation was useful in eliminating the non-constant within-subject variance along with the skewness. Skewness of the response and non-constant variance are often associated, and it is not surprising that the one transformation does a good job of reducing both.

When profiles are flat or linear, that is, they exhibit a random intercept and do not have a time trend, there is a plot that can help clarify whether within-subject variability increases with the subject-specific mean. For each subject, we can calculate the mean of the n_i observations and the standard deviation of the n_i observations. Then we plot the n standard deviations against the means. Figure 2.9(a) plots, for the Pediatric Pain raw data, the within-subject standard deviations of the four observations on the vertical axis versus the means of the four observations on the horizontal axis. The line is a least squares line drawn through the points without particular regard to assumptions. We see that the standard deviations do definitely increase with the mean. Figure 2.9(b) shows the same plot for the log base 2 data. It shows little if any correlation between the within-subject standard deviation and the subject-specific mean. For both plots, subjects with less than four observations or with a measurement of 240 are not included.

For our Pediatric Pain analyses, we take a log base two transformation of the responses before analyzing the data. The base two log transformation

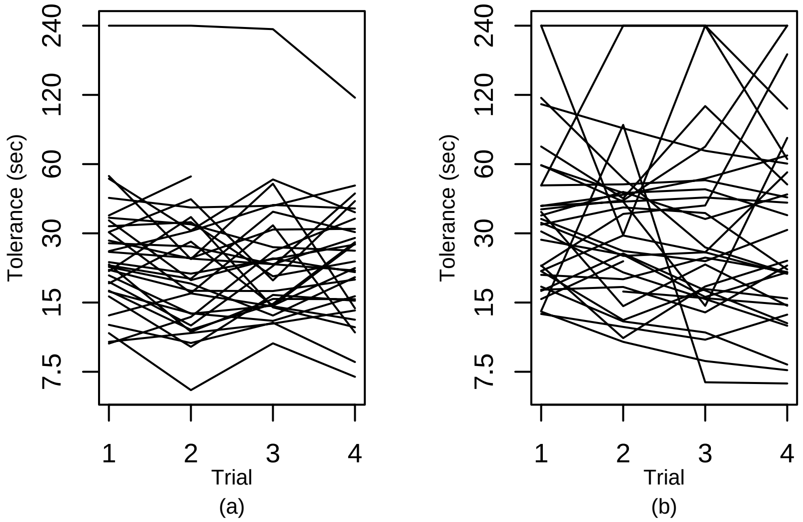

Figure 2.10. Profile plots of the Pediatric Pain data on the log scale separately by coping style: (a) attenders; (b) distracters.

is no different from a base 10 or base e transformation in terms of getting rid of skewness. However, powers of two are much easier to do in our head than powers of e to invert the log transformation and allow us to interpret what a particular log measurement signifies on the original seconds scale. Powers of 10 are also easy to do, but are only useful for data with a wide dynamic range covering several powers of 10.

2.4 Elaborations of Profile Plots

The profile plot is a useful all-purpose tool for understanding longitudinal data. Any truly useful tool develops many variations that are helpful in different circumstances. In this section, we discuss some modifications to the basic profile plot for (i) data sets with covariates, (ii) data sets where the range of the response across subjects obscures the trend of the profiles within subject, and (iii) two kinds of empirical residuals that can be helpful for understanding our data.

2.4.1 Covariates

So far we have plotted entire data sets in our profile plots. It is not required that we include the entire data set in a single plot, and we may use our creativity to determine when it might be helpful to look at a subset of the data in a single plot, or to look at all of the data but spread across several different profile plots.

When we have a discrete covariate, we may plot subjects with different values of the covariate in separate plots. Figure 2.10 illustrates this for the Pain data and the coping style covariate. The left-hand plot shows subjects who are attenders and the right-hand plot shows distracters. Compared with figure 2.7(b), there are fewer subjects plotted on the same plot, and we can more easily distinguish individual profiles of subjects.

Figure 2.10 reveals several interesting features of the data. In figure 2.10(a), most of the subjects range uniformly on the log scale from approximately 7.5 seconds up to slightly below 60 seconds. There is one exceptional subject that we identify as a high outlier. In figure 2.10(b), the distracters range uniformly from around 7.5 seconds up to 240 seconds. A number of distracters have observations over 60 seconds. The average of the attenders appears to be approximately half way between 15 and 30 seconds. Because this is a logarithmic scale, that point is at $\sqrt(2) \times 15 \approx 1.4 \times 15 = 24$ seconds. The average of the distracters appears to be around one quarter of the way between 30 and 60, which puts it at around $2^{1/4} \times 30 \approx 36$ seconds. We conclude that distracters have greater average pain tolerance than attenders.

We could have used different line types for the two groups and plotted all subjects on the same plot. This works reasonably well when the groups are well separated in their responses or if there are very few subjects. The Pain data have a bit too many subjects for separate line types to be helpful.

With a continuous covariate, we might slice the covariate into a small number of intervals and create separate profile plots for subjects with covariate values that fall into each interval. A common way of slicing continuous variables is called a *median split*. Subjects with covariate values above the median form one group, and those with values below the median form a second group. All subjects who fall at the median, if they exist, may go all together into either group. We do a median split when there is no particular scientific rationale for splitting the covariate at some other value.

2.4.2 Ozone Data

Ozone is an invisible pollutant that irritates the lungs and throat and causes or exacerbates health problems in humans. Crops may grow less if exposed to excess ozone, and chemical products such as paint may degrade when exposed to ozone. The Ozone data set records ozone over a three-day period during late July 1987 at 20 sites in and around Los Angeles, California, USA. Twelve recordings were taken hourly from 0700 hours to 1800 hours giving us $20 \times 12 \times 3$ ozone readings. Measurement units are in parts per hundred million. Table 2.1 gives the four-letter abbreviation for the sites, the full names of the sites, and the longitude, latitude, and altitude

Site abbr	Site name	Long	Lat	Altitude	Valley
SNBO	San_Bernadino	117.273	34.107	317	SG
RIVR	Riverside	117.417	34	214	SG
FONT	Fontana	117.505	34.099	381	SG
UPLA	Upland	117.628	34.104	369	SG
CLAR	Claremont	117.704	34.102	364	SG
POMA	Pomona	117.751	34.067	270	SG
AZUS	Azusa	117.923	34.136	189	SG
PASA	Pasadena	118.127	34.134	250	SG
BURK	Burbank	118.308	34.183	168	SF
RESE	Reseda	118.533	34.199	226	SF
SIMI	Simi_Valley	118.685	34.278	310	SF
ANAH	Anaheim	117.919	33.821	41	No
LAHB	La_Habra	117.951	33.926	82	No
WHIT	Whittier	118.025	33.924	58	No
PICO	Pico_Rivera	118.058	34.015	69	No
LGBH	N_Long_Beach	118.189	33.824	7	No
LYNN	Lynwood	118.21	33.929	27	No
CELA	Central_LA	118.225	34.067	87	No
HAWT	Hawthorne	118.369	33.923	21	No
WSLA	West_LA	118.455	34.051	91	No

Table 2.1. The Ozone data: general information about sites. The first five columns are the site abbreviation, full name, longitude, latitude, and altitude. Valley is whether the site is in either the San Fernando or Simi Valleys (SF) or San Gabriel (SG) valley. Other sites are adjacent to the ocean without intervening mountain ranges. Abbreviations in names: N North; LA usually means Los Angeles; except La Habra is La Habra.

of each site. Also given is a valley indicator to indicate whether the site is in the Simi or San Fernando Valleys (SF) or San Gabriel Valleys (SG). The remaining sites are adjacent to the ocean or otherwise do not have mountain ranges between them and the ocean. The data was originally collected to compare to output of computer simulations of atmospheric chemistry in Los Angeles. Figure 2.11 shows a map of the site locations. The San Gabriel Valley is on the right or east on the plot. Simi Valley is the left most or western most site and is adjacent to the San Fernando Valley sites of Reseda and Burbank. Each night ozone returns to a baseline value, and we treat the data as having $60 = 20 \times 3$ subjects with 12 longitudinal measures each.

Figure 2.12 plots ozone profiles for the sites separately by day. We see that on day 1 ozone generally increases monotonically up to a peak between 2pm and 4pm before beginning to decrease slightly. There are a range of ozone levels. The ozone peaks appear to be increasing from day 1 to day 2 to day 3. It may be that the peaks are slightly later on day 3 than on day

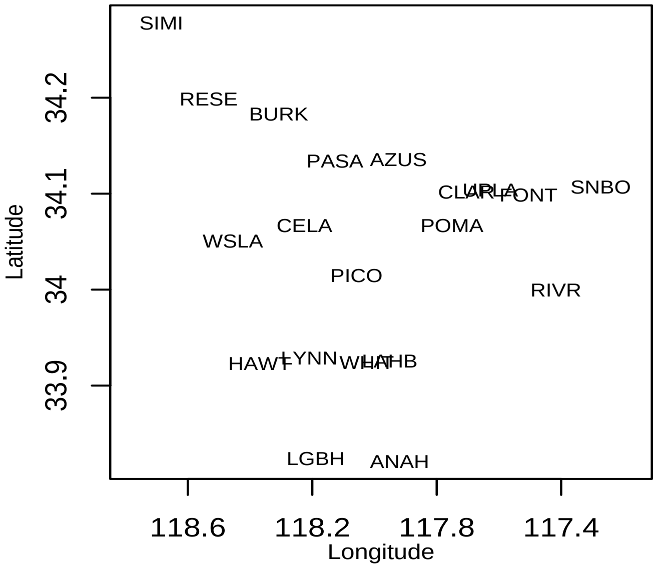

Figure 2.11. Map of Ozone data sites. CLAR overlaps UPLA, which overlaps FONT, and WHIT overlaps LAHB. To place west on the left, longitude increases right to left.

1. The lowest ozone levels appear to be similar across the three days. From this plot we do not know if the same sites are lowest on each of the three days.

Figure 2.13 plots the profiles by site. Sites are ordered by the maximum ozone value over the three days. The individual plots are arranged starting at the bottom left and moving left to right and then from bottom to top. We notice that the sites with the largest ozone concentrations are all in the San Gabriel Valley. The next two sites with high ozone are in the San Fernando Valley. One site not in a valley has higher ozone than Simi Valley, the last valley site. We lumped Simi Valley with the San Fernando Valley to avoid having only one site in that category, but technically it is a different valley from the San Fernando Valley. The sites with the lowest ozone values appear to be rather similar over the three days, while the middle and higher ozone sites have different peak ozone levels over the three days.

We can even plot profiles a single profile per plot and look at as many subjects as we have patience for. In olden times, statisticians would print out a single subject's profile on separate pages, then shuffle the pieces of paper around on a table top like a jigsaw looking for similar patterns among different subjects. In figure 2.14, we print 18 of the 60 cases, with the other 42 on additional pages that are not shown. We now can inspect the patterns

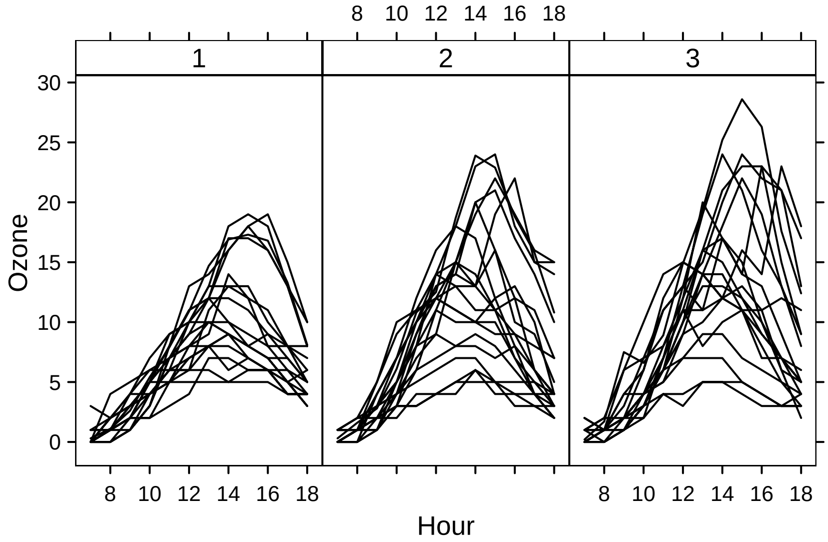

Figure 2.12. Profile plots of the Ozone data separately by days. Days are ordered 1, 2, and 3 from left to right.

of ozone over day as well as over time within a site. In these six sites, we see that the ozone peak is generally increasing from each day to the next, and we again suspect that the peak is moving later within the day from day 1 to day 2 to day 3.

2.4.3 Weight Loss Data and Viewing Slopes

The Weight Loss data consist of weekly weights in pounds from women enrolled in a weight loss trial. Patients were interviewed and weighed the first week and enrolled in the study at the second week. The data from 38 women are plotted in figure 2.15. There are from 4 to 8 measurements per subject. Weights range from roughly 140 pounds to 260 pounds.

Study protocol called for the subjects to visit the clinic at weeks 1, 2, 3, and 6 and weigh themselves on the clinic scale. At weeks 4, 5, 7, and 8, study personnel called subjects at home and asked them to weigh themselves on their home scales and report the measurement. Week 1 was a screening visit; participation in the actual weight-loss regimen did not start until week 2.

Figure 2.15 presents a profile plot of the Weight Loss data. Unfortunately little structure is visible, except for the numerous parallel profiles. This indicates the not surprising result that each woman has her own average weight and her weight varies around that weight over time; this data illustrates a random intercept. We see one slightly heavy subject and one

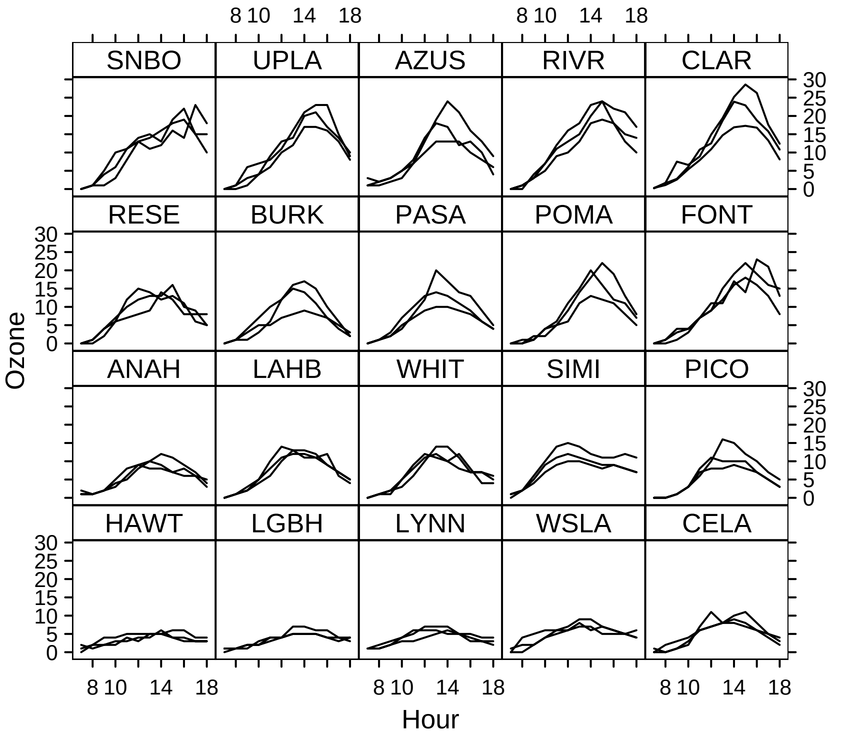

Figure 2.13. Profile plots of the Ozone data separately by site. Sites have been ordered from left to right and bottom to top by their maximum ozone reading over all hours and days.

slightly light subject. It is not obvious whether women are losing weight or not based on this figure.

The problem with figure 2.15 is that we cannot see slopes of individual profiles. Suppose someone lost 5 or even 10 pounds over the 8 weeks; that slope would barely be visible in the plot; a slope of $-1/2$ to -1 pound per week is scientifically quite high yet it would be nearly indistinguishable from a flat profile with slope 0. We want to see slopes of magnitude $-1/2$ to -1, and the question is how to draw the plot so that we can actually see slopes of reasonable magnitude. If we made the figure taller, then a 5 or 10 pound difference would become physically larger on the printed page, and we will be more likely to be able to see it in the plot. The second thing we can do is to make the figure narrower! By narrowing the x axis, we increase the angle of the slopes, so that a slope of -1 pound per week appears steeper on the plot.

Figure 2.15 is square, this is the default shape produced by most statistical software. Instead, if we give instructions to the software to make the

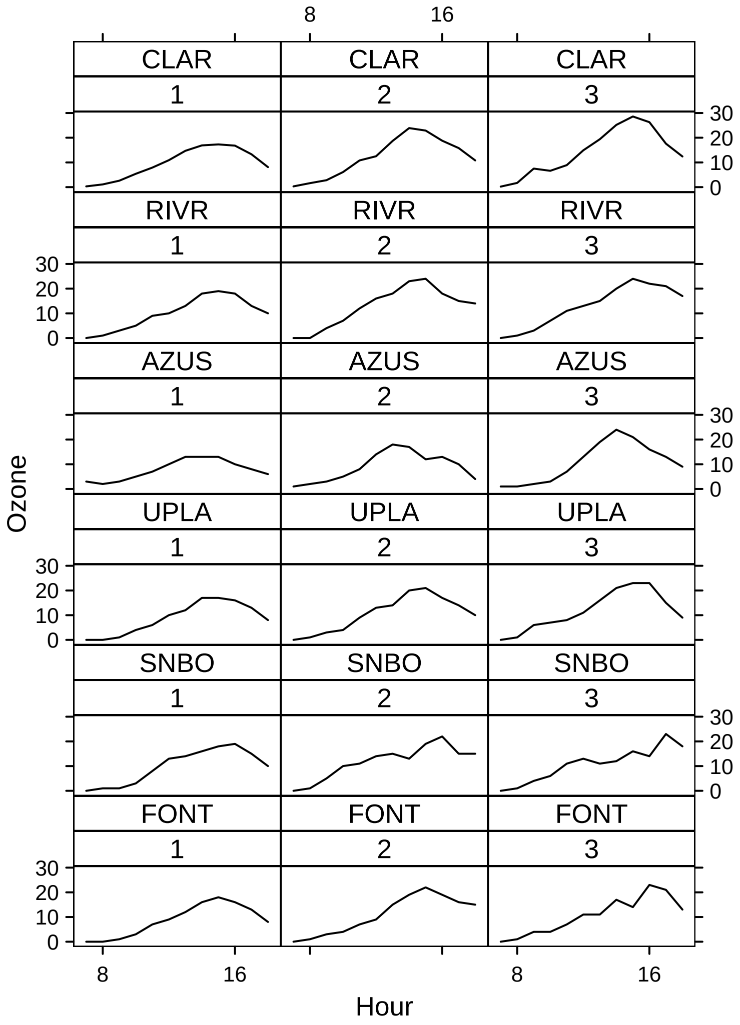

Figure 2.14. Profile plots of Ozone data by day and site. Each row shows profiles from the same site with day increasing left to right. The 6 sites with the highest ozone levels are shown. This is slightly less than 1/3 of a larger display (not presented).

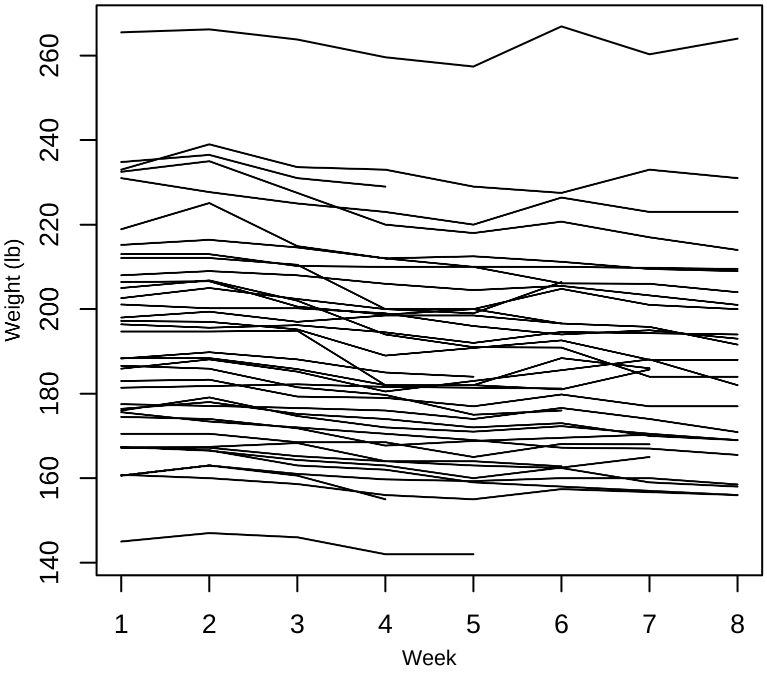

Figure 2.15. Profile plot of the Weight Loss data.

plot taller and narrower, something like either half of figure 2.16 results. In fact, because the taller, narrower figure is larger than the printed page, I broke the *y* axis in half and plotted the lower half of the data from below 140 pounds up to over 200 pounds in one plot on the left and the upper half of the data from approximately 200 pounds to the maximum on the right. Some data has been plotted in both figures, roughly in the range from 198 to 210 pounds. The *shape* of the plot has been modified so that we can see the trends in the weights. Now we can see that people are losing weight; the observation at week 8 generally appears to be lower than that subject's corresponding observation at week 1 or 2. Another feature of the data is also slightly visible. A number of profiles take a steep drop around weeks 4 and 5 then return to a higher level at week 6, and this seems more pronounced for heavier subjects.

Changing the shape of the plot is often necessary when plotting longitudinal data. When the range of the response *across* all of the subjects is large, but the range *within* subjects is small in comparison, then we often have difficulty seeing the time trends of individual subjects on a plot like 2.15. In contrast to the Weight Loss data, the Big Mice data have a large response range within subjects that is almost the same as the response

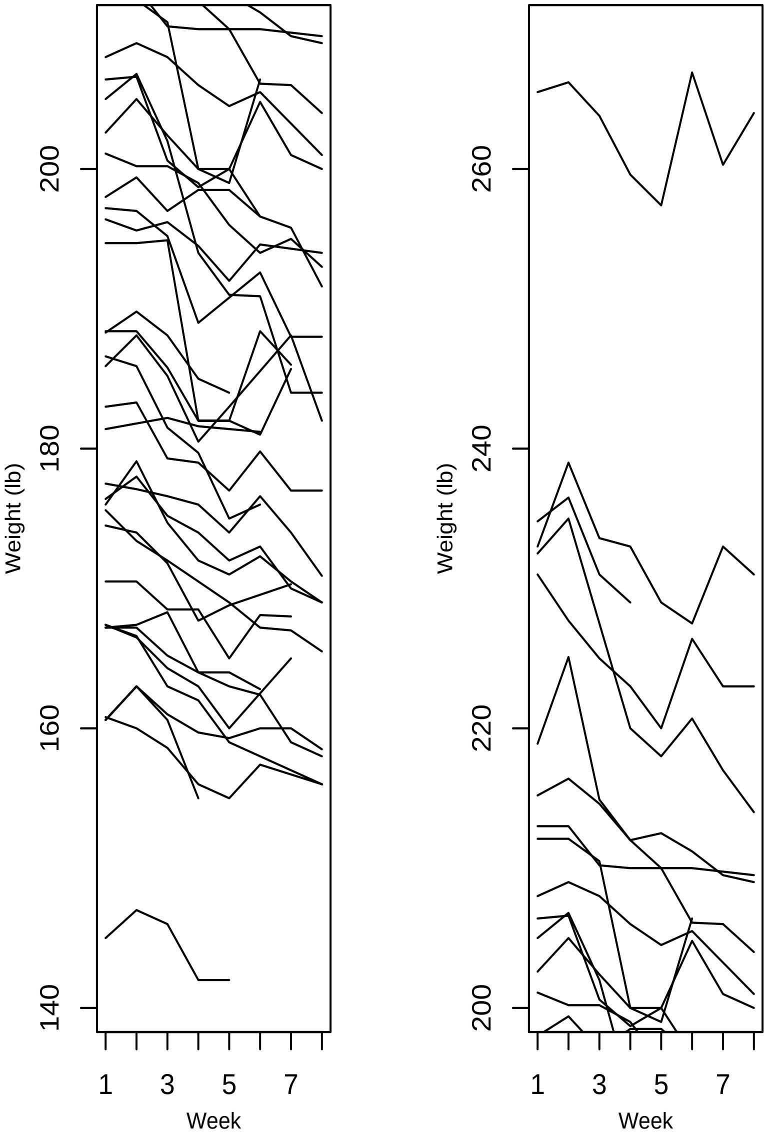

Figure 2.16. Weight Loss data with appropriate shape parameter.

range across subjects. Unfortunately for the Weight Loss data, we would like to make the plot impractically taller. The next subsection provides another solution to the problem of viewing the slopes of individual profiles.

2.4.4 Empirical Within-Subject Residuals

We always begin data analysis with a figure like 2.15. However, this plot was not particularly successful in terms of viewing slopes over time because of the large range in the responses. Reflection suggests that we are not very interested in the absolute weights of the subjects. Rather, we are interested in changes in weight over time. What could we do to focus on the weight changes while not worrying about absolute weight levels?

Ignoring figure 2.16 for now, and just looking at figure 2.15, we see that each subject appears to have her own average weight, and weekly observations vary around these averages. If we estimate the average weight, we can calculate the individual deviations around the average and then consider plotting those in a profile plot. A simple estimator of the intercept for each subject might be the subject's average response. In chapter 9, we will learn about models that produce better estimates of the intercept for each subject. Until then, we consider the subject average $\bar{Y}_i$ as an estimator of the subject's average weight. The difference between the jth observation and the subject mean

$$R_{ij} = Y_{ij} - \bar{Y}_i$$

is an empirical within-subject residual.

Figure 2.17 plots the R_{ij} in a profile plot. At weeks 1 and 2, we see that most residuals R_{ij} are greater than zero and some are as large as 10 pounds, that is, most subject's weights are above their average weight. At weeks 7 and 8, most subject's weights are below their average. Thus we see that yes, subjects do lose weight over the course of the study.

From week 1 to week 2, no weight is lost, if anything a little weight is gained. From week 2 to week 3, the first week of the weight loss treatment, the subjects lose quite a lot of weight. At weeks 4 and 5 they continue to lose weight with a few losing a substantial amount between weeks 3 and 4. At week 6, suddenly, weight is gained, presumably because patients are weighed under supervision with a properly calibrated scale. At weeks 7 and 8, they continue to lose weight. The overall weight loss indicates that a fixed time effect is needed in the model. It is not clear whether the weight loss is strictly linear, an issue we leave for later.

We could consider subtracting off other subject relevant weights such as subject baseline rather than subject average weight from each observation and plot the resulting changes from baseline in a profile plot. Exercise 18 explores what the Weight Loss profile plot looks like if we subtract off the baseline weight measurement instead of the subject average. Exercise 20

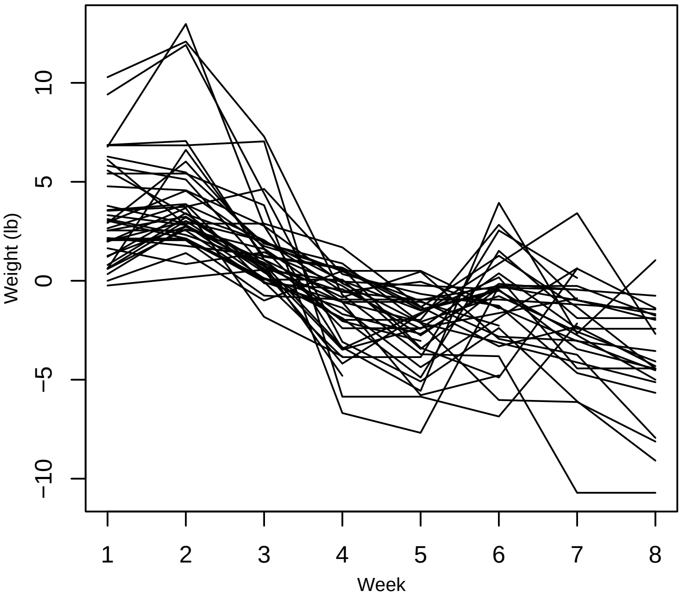

Figure 2.17. Profile plot of empirical within-subject residuals for the Weight Loss data.

explores what happens to the Weight Loss profile plot if we plot differences from one time to the next $Y_{ij} - Y_{i(j-1)}$.

2.4.5 Empirical Population Residuals

Sometimes we wish to see the variation across subjects within a particular time. If there is a large change in the average response over time, then it may be hard to view the individual subject profiles, for example to see if profiles are parallel, or to see if the marginal variance is increasing. In figures 2.1, 2.2, and 2.3, the marginal variance grows quickly initially then appears to stop growing. We can get a better look at the profiles by subtracting off an estimate of the mean at each time point to look at the deviations from the mean. For the Big Mice data, define the sample mean $\bar{Y}_{.j}$ at time j as the mean of all observations at day j and define the empirical population residuals

$$U_{ij} = Y_{ij} - \bar{Y}_{.j}.$$

In figure 2.18, we can now see the bulk of the data better, and we can better see individual profiles. The range of the y axis is approximately 500 mg rather than 1200 mg of figure 2.3. It is easier to tell relative high or low and by how much within a time and to follow the paths of individual

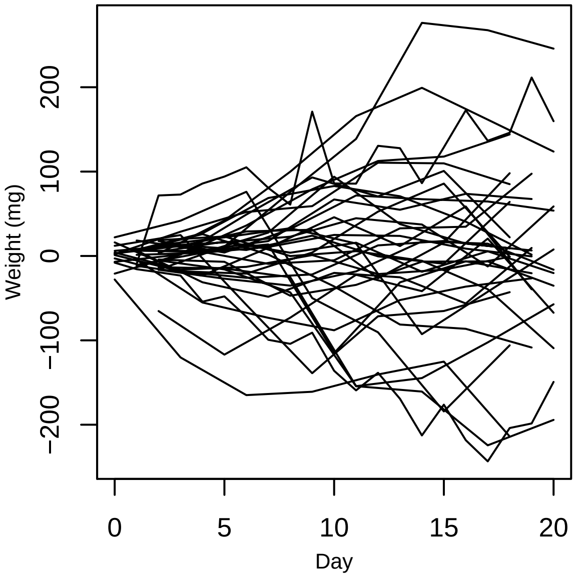

Figure 2.18. Profile plot of empirical population residuals for the Big Mice data.

mice. We lose the population time trend in this plot, the y axis represents differences from the population mean. Positive residuals with a decreasing trend means that the mouse is getting closer to the population mean, not that the weight is decreasing.

2.4.6 Too Many Subjects

Profile plots go by several other names including *spaghetti plots*, *parallel plots*, and *connect-the-dots plots*. The spaghetti plot name describes what profile plots look like when too many profiles are plotted in a single plot. The ink density destroys many features of the data. This can be overcome in several ways. Using a higher quality graphing package and printer can do wonders. Alternatively, a subset of subjects may be plotted. The subset may be a randomly selected subset or, as we did in section 2.4.1, subsets may be specified according to the values of some covariate or, as with the ozone data, we might plot all subjects separately.

2.5 Inspecting Correlations

Longitudinal data are different from linear regression because observations are correlated within subjects. The correlations among observations needs

	Trial			
Trial	1	2	3	4
1	1	.73	.84	.60
2	.73	1	.72	.66
3	.84	.72	1	.76
4	.60	.66	.76	1

Table 2.2. Correlation matrix for the Pediatric Pain data.

to be modeled, and it helps to have summary measures and graphics that help us decide on the covariance model.

The correlations $\hat{\rho}_{jk}$ and variances s_j^2 form many patterns in different data sets; there is not one pattern that describes all data sets. We want to inspect the correlations among our observations to help determine the type of model for the correlations that we will use. A simple and useful summary of the correlations among our longitudinal observations is a table of those correlations.

Table 2.2 gives the correlations among the Pediatric Pain observations. Correlations of observations with themselves are 1, so 1's go down the long diagonal. The correlation between observations at trial 1 and trial 2 is $\hat{\rho}_{12} = \hat{\rho}_{21} = .73$. The pain correlations vary from .60 to .84, with the lowest correlations $\hat{\rho}_{14} = .60$ and $\hat{\rho}_{24} = .66$ being trial 4's correlations with the trial 1 and trial 2 observations. Correlations between trial j and k were calculated by using all subjects that had both observations at times j and k.

Although somewhat different, these 6 correlations are not wildly different, and we might initially consider a model where all correlations among observations are the same. Seeing no other pattern, as an alternative model we might consider a model where all the correlations are different.

An estimate of the uncertainty in a correlation can help with judging the differences in the correlation values. The estimated standard error of a simple correlation is

$$\mathrm{SE}(\hat{\rho}) = \frac{1 - \hat{\rho}^2}{(n-3)^{1/2}}$$

where n is the number of subjects contributing pairs of observation to the computation. For the Pain data, the number of pairs of observations contributing to each correlation ranges from 58 for all correlations involving trial 3 to 62 for $\hat{\rho}_{12}$. The range of standard errors is .04 to .08 for the correlations in table 2.2. We do not have a simple test for the equality of two correlations, but it seems reasonable that differences in correlation less than .1 are not very important. Still, the largest difference in correlations is .24 and that may be significant, suggesting that the six correlations in table 2.2 may be different.

For the Big Mice data, we can calculate sample correlations for any pair of days provided that either group 1, group 2 or group 3 mice were

	2	5	8	11	14	17	20
2	1	.92	.57	.36	.23	.23	.38
5	.92	1	.77	.54	.45	.41	.55
8	.57	.77	1	.86	.80	.76	.81
11	.36	.54	.86	1	.93	.92	.87
14	.23	.45	.80	.93	1	.96	.89
17	.23	.41	.76	.92	.96	1	.92
20	.38	.55	.81	.87	.89	.92	1

Table 2.3. Correlation matrix for the Small Mice data.

measured on both of the days. The group 4 mice do not give us enough observations to calculate correlations for other days. Table 2.3 gives the sample correlations for the Small Mice data based on 14 observations. In inspecting correlation tables like this, we look at the rows beginning at the long diagonal and at the columns also beginning with the long diagonal. We look along diagonals parallel to the long diagonal, looking for simple patterns in the correlations.

Each diagonal away from the long diagonal corresponds to a given *lag* between observations. The first off-the-main diagonal gives the lag 1 correlations; observations being correlated are consecutive observations. The correlation between day 2 and day 5 observations is $\rho_{12} = .92$, indicating a strong relationship between weights at those two early days. The second long diagonal, beginning with correlations .57, then .54 and ending with .89 are the lag 2 correlations; .54 is the correlation between day 5 and day 11 observations. And it continues, until the correlation .38 is the sole lag 6 correlation, the correlation between observations at day 2 and day 20.

To begin more detailed analysis of table 2.3, we inspect the longest and first off diagonal, the lag one diagonal, with correlations of .92, .77, .86, .93, .96, and .92. Although these are not all exactly equal, they are all quite similar, possibly excepting the .77, which is a tad lower than the others. The standard error (se) of .92 is .04 while the SE of .77 is .12. We possibly hypothesize that the lag 1 correlations are all equal, or, approximately equal. Next we go to the diagonal two away from the long diagonal, hoping that this pattern of near equality continues. Here the values start out lower, .57 and .54, then abruptly increase, ranging from .86 to .96, so that either we have increasing correlations along the lag 2 diagonal, or we have two low then three high correlations. The third off diagonal starts low at .37, then steadily increases. The first two values are low, the last two are high, at .76 and .87.

Continued inspection suggests perhaps two groups of observations, the early observations and the late observations. The first four observations at times 2, 5, 8, and 11 have a pattern that has high lag 1 correlations, middling values of lag 2 correlations, and a low lag 3 correlation. The last four observations from day 11 to day 20 have all high correlations,

	7	8	9	10	11	12	1	2	3	4	5	6
7	.63	.27	-.02	-.17	-.24	-.26	-.26	-.27	-.31	-.30	-.26	-.24
8	.13	.73	.53	.33	.03	-.10	-.10	-.19	-.20	-.13	-.07	.03
9	-.02	.56	1.44	.75	.49	.32	.24	.22	.29	.38	.39	.47
10	-.20	.45	2.02	1.87	.79	.59	.39	.25	.33	.37	.37	.47
11	-.37	.05	1.71	3.61	2.43	.86	.60	.45	.51	.49	.51	.54
12	-.54	-.25	1.52	3.61	6.86	3.26	.83	.70	.68	.64	.59	.51
1	-.67	-.30	1.43	3.00	5.95	11.1	4.08	.89	.81	.73	.62	.50
2	-.90	-.75	1.66	2.48	5.86	12.1	19.5	5.35	.94	.86	.77	.63
3	-1.18	-.87	2.55	3.73	7.53	13.5	20.1	30.3	6.07	.93	.84	.75
4	-1.11	-.54	3.18	4.01	7.03	12.3	17.5	26.7	32.9	5.85	.91	.81
5	-.82	-.25	2.81	3.54	6.28	9.75	12.8	20.7	25.7	26.8	5.05	.93
6	-.59	.08	2.68	3.48	5.15	6.61	8.07	13.2	17.9	18.6	18.5	3.94

Table 2.4. Correlation/covariance matrix for the Ozone data. Above the long diagonal are the sample correlations, below the diagonal are sample covariances, and along the diagonal (boxes) are the sample standard deviations.

ranging from .87 to .96. The cross-correlations between the early and late observations generally follow the pattern that the closer in time, the higher the correlation, but that all the later observations have, roughly speaking, similar correlations with any given early time. Day 8 has high correlations with the later observations, but has that mildly lower correlation with day 5, so day 8 may be the dividing day between the early and the late observations.

The lower and upper half of the correlation matrix are the same, and sometimes we omit the lower or the upper half of the matrix. Instead of the format of table 2.2 or 2.3, we can pack information into the table by placing the sample standard deviations down the long diagonal, covariances below (or above) the long diagonal, and correlations in the other half of the table. This is illustrated in table 2.4 for the Ozone data.

The Ozone data standard deviations, correlations and covariances are more complex than the Pain data. The sample standard deviations begin at a very low level in the morning and steadily increase until 4pm in the afternoon, then decrease slightly by 5pm or 6pm. The lag one correlations start low at .27 then increase rapidly to .53, .75, and so on to over .9, and stay high and roughly constant through the end of the data. The lag two correlations start even lower at $-.02$, then increase to correlations in the .8's, not quite as high as the lag one correlations. The higher lag correlations exhibit a similar pattern, except that the correlation between the first and second observations with the remaining observations remain modest and negative. Along rows, from the row for noon and rows for later times, we see strictly decreasing correlations as the lag increases. For morning rows, at 11am we see a high lag one correlation, then decreasing to a constant correlation. For 9am and 10am, the correlation starts high, decreases to a lower constant correlation and then creeps up slightly at the end, for 7am and 8am, the correlation decreases to negative(!) correlations for most of the day, before starting to creep back up at the end. To summarize, for constant lag, the correlation starts low, then increases to some maximum.

We illustrate two common types of correlation matrices in tables 2.5 and 2.6. In table 2.5, the correlations with constant lag are the same. The lag one correlations are .90, the lag two correlations are all .81, and the lag three

	Trial			
Trial	1	2	3	4
1	1	.9	.81	.73
2	.9	1	.9	.81
3	.81	.9	1	.9
4	.73	.81	.9	1

Table 2.5. Example correlation matrix illustrating banded correlations.

	Trial			
Trial	1	2	3	4
1	1	.85	.84	.86
2	.85	1	.87	.87
3	.84	.87	1	.84
4	.86	.87	.84	1

Table 2.6. Example correlation matrix illustrating approximately equal correlations at all lags.

correlation is .73. Constant correlation for a given lag is called a *banded* correlation matrix; many important correlation structures are banded. As the lag increases, the correlations decrease, and in this particular example, they decrease in a nearly geometric fashion, with $.81 = .9^2$, and approximately $.73 \approx .9^{|4-1|} = .9^3$. We see a decrease in correlation with increasing lags in the mice data and in the Ozone data, but neither example appears to illustrate *banding*.

Table 2.6 illustrates a correlation matrix with approximately equal correlations for all pairs of observations. This would, at least approximately, be called an *equicorrelation* correlation matrix. An equicorrelation correlation matrix says that the lag does not matter in calculating the correlations; no matter how distant in time two observations are, they have the same constant correlation. Equicorrelation is of course a special case of banding, but usually we intend the term banded to mean correlations that are not constant for different lags.

2.5.1 Scatterplot Matrices

A scatterplot of two variables is a graphical illustration of the correlation between the two variables. Additionally it shows whether the relationship between the variables is linear, and whether there are outliers, clusters or other deviations from normality in the data. When we have multiple variables to plot, there are many scatterplots to look at; for J variables, we have $J(J-1)/2$ pairs of variables and in each pair either variable may be on the vertical or horizontal axis. A *scatterplot matrix* organizes all of the pairwise scatterplots into a compact arrangement. For longitudinal data,

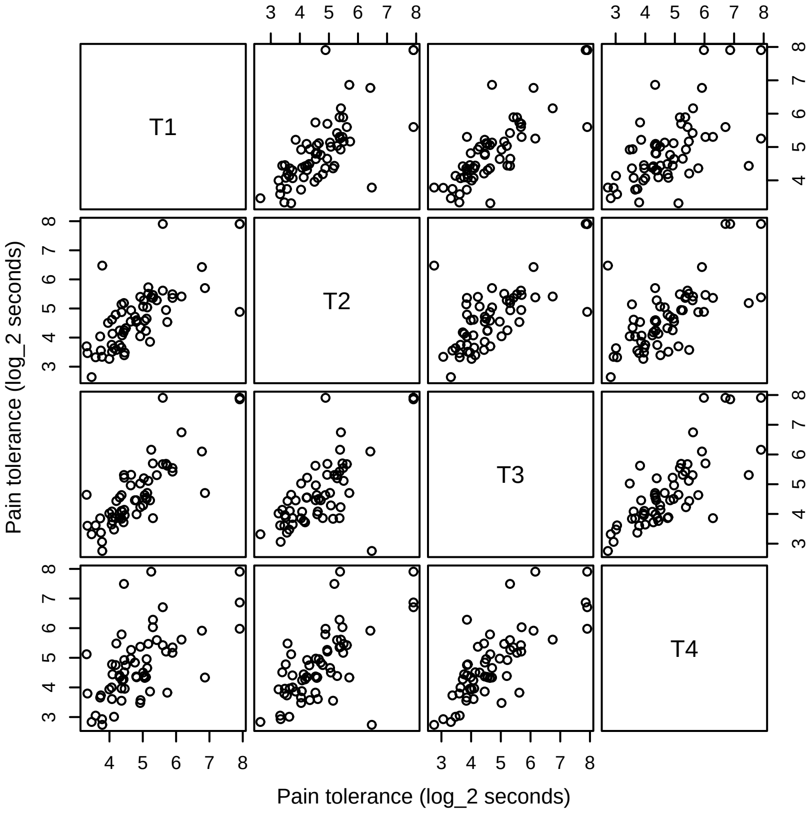

Figure 2.19. Scatterplot matrix for the Pediatric Pain data.

we require balanced or balanced with missing data to conveniently produce a scatterplot matrix.

Figure 2.19 is a scatterplot matrix of the Pain data with responses on the log base two scale. There are 12 scatterplots in the figure. Each plot in row j counting from the top has the response from trial j on the vertical axis. The three plots in a row share the same axis tick marks and tick labels, given either on the left- or right-hand side of the figure. Similarly each plot in a column shares the same variable on the horizontal axis and the same tick marks and labels given either above or below the column. The (j, k)th plot and the (k, j)th plot plot the same data, but with the vertical and horizontal axes reversed. The figures are arranged with the long diagonal going from the upper left to lower right, the same as our correlation matrix in table 2.2. Sometimes the long diagonal goes from lower left to upper right. All i subjects contribute one point to each plot unless they are missing either the jth or kth trial.

The tick mark of 3 $\log_2$ seconds corresponds to $2^3 = 8$ seconds, and the tick marks of 4, 5, 6, 7, and 8 on the log scale correspond to 16, 32, 64, 128, and 256 seconds, respectively. Different rows do not share exactly the same scales or ranges on the axis, although they will be similar because the range of the data at each trial are nearly the same.

A key feature of the scatterplots in figure 2.19 is that the relationship among the pairs of responses is linear. This is important, as it is a major part of the normality assumption that we will use in our analyses. We also see that the observations are generally elliptically distributed, but that there is a halo of points scattered mostly at the higher values; there are some outliers in this data. The correlations in figure 2.19 look approximately equal to us; with experience one can develop the ability to accurately estimate correlations from bivariate normal data to within a few percent. Under closer inspection, it appears perhaps that the lag one correlations are definitely all similar, and that the lag three correlation is lower than the lag 1 correlations. The lag two correlations are difficult to determine, but perhaps, matching table 2.2, the plot of trial 4 against trial 2 is also of lower correlation, whereas trial 3 against trial 1 is of equal or slightly higher correlation to the lag 1 plots.

Figure 2.20 gives the scatterplot matrix for the Small Mice data corresponding to the correlation matrix in table 2.3. The data set is quite small and we often have trouble with identifying both absolute and relative correlations with small sample sizes. Still we see that the lag 1 correlations are all quite high except the plot of days 5 against 8, which seems lower than the other lag 1 plots. Among later days, 8–20, the higher lag plots still have fairly strong positive correlation although the correlation does decrease with increasing lag. Observations from the early days 2 and 5 have low correlations with the observations at later days. We cannot tell the exact values of these correlations, and a correlation less than .4 can be hard to distinguish from independence without actually formally calculating it.

We can also create a scatterplot matrix from randomly spaced data by *binning* the times into convenient intervals and identifying all observations in the same bin as coming from a single *nominal* time. For observations nominally scheduled for every three months, the actual times may vary around the nominal date. We might still use nominal times for plotting in a scatterplot matrix, with the understanding that the variability in times may affect the figure somewhat.

2.5.2 Correlation in Profile Plots

We can identify correlations from profile plots as well. It is easiest to identify the correlation between neighboring observations.

The set of line segments $i = 1, \ldots, n$ between consecutive observations $Y_{i(j-1)}$ and Y_{ij} show the correlation between observations at consecutive times. Figure 2.21(a) illustrates a range of positive correlations between

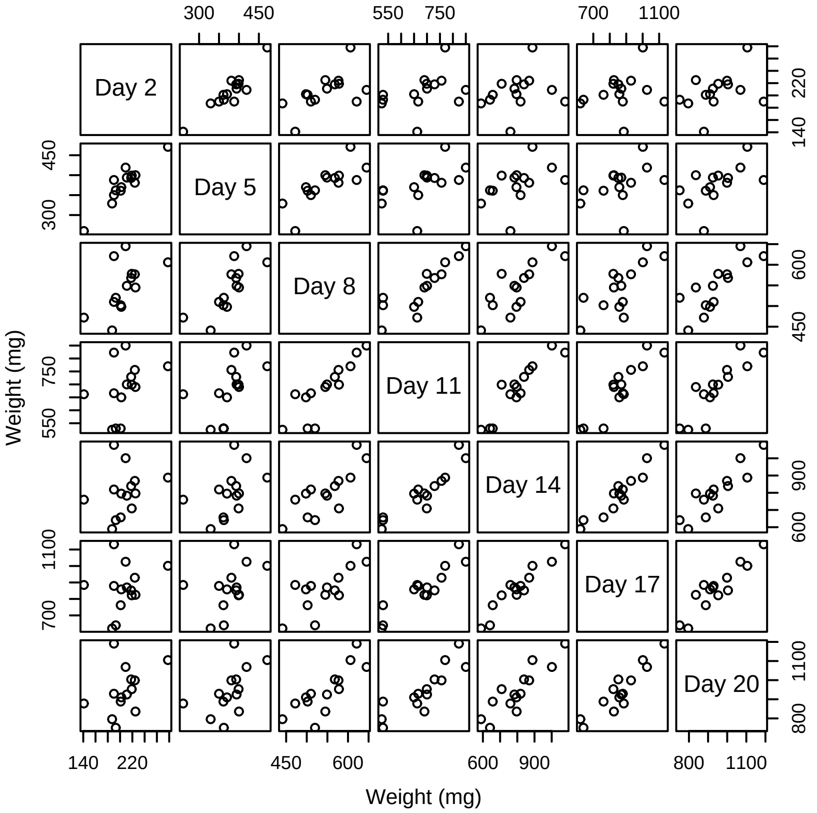

Figure 2.20. Scatterplot matrix for the Small Mice data.

consecutive observations ranging from $\rho = .99$ at the left to $\rho = 0$ at the right, and 2.21(b) shows a range of negative correlations from $\rho = -.99$ at the left to $\rho = 0$ at the right. Both figures 2.21(a) and 2.21(b) show profiles of a sample of 20 subjects observed at times $t = 1, 2, \ldots, 8$. The population mean is zero and the population standard deviation is 1 at all times. The population correlation between time t and $t + 1$ is given at the bottom of the plot. In figure 2.21a, the correlation between observations at $t = 1$ and $t = 2$ is .99. The line segments between $t = 1$ and $t = 2$ rarely cross, and this is indicative of a high positive correlation. The correlation between $t = 2$ and $t = 3$ is lower at $\rho = .95$ than the correlation between the first two times, and there is more crossing of the profiles. As t increases, the correlation between the observations at t and $t + 1$ decreases, and the amount of crossing of the profiles increases from left to right. The last pair of observations are uncorrelated with $\rho = 0$ between observations at $t = 7$ and $t = 8$, and the profiles are at their most haphazard.

In figure 2.21(b), the correlation decreases in absolute value from left to right, but this time the correlations are negative. The crossing of the line segments between observations increase from the right-hand side to the left-hand side. The difference from right to left is that the crossing is less and less haphazard and becomes more and more focused in a smaller and smaller region as the negative correlation increases. Between $t = 1$ and $t = 2$, the correlation is highest, and the line segments all intersect in a very narrow region near a point $(1.5, 0)$. Exactly where this point is in general depends on the mean and variances of the two observations but the intersection point is between the two times when the observations are negatively correlated and is outside the two times when the correlation is positive. If the points were perfectly negatively correlated, then all line segments would intersect exactly at a single point.

When the correlation between consecutive points is positive, the line segments between observations tend not to intersect. The stronger the correlation the fewer the intersections. The line segments will not be parallel unless the variances at the two times are equal. Generally for positively correlated data, the line segments would intersect if we were to extend the lines out toward the direction of the time with the smaller variance. The closer the variances, the farther we must extend the lines to see the intersections. And if the variances are equal, the lines are parallel and the intersection points go out to infinity.

2.5.3 The Correlogram

For equally spaced data, an empirical correlogram is a plot of the empirical correlations ρ_{jk} on the vertical axis against the lag $|j-k|$ on the horizontal axis. If our observations are balanced but not equally spaced, we might instead plot ρ_{jk} against $|t_j - t_k|$. As when we wish to draw a scatterplot matrix, we must bin the data for randomly spaced data to create a correlogram. Various enhancements to the basic correlogram plot are possible. Figure 2.22 shows a correlogram for the Ozone data. Correlations ρ_{jk} whose j are equal are connected by line segments. This correlogram tells us the same information as the correlation matrix in table 2.4, in a different form. It is easy to see that correlations decrease with increasing lag, and that they tend to level out once the lag reaches 3 or 4, and that the correlations may even begin to increase for still greater lags.

2.6 Empirical Summary Plots

An empirical summary plot presents information about the average response over time. One simple way to estimate the mean at a given time is to take the average $\bar{Y}_{\cdot j}$ of all observations at a given time t_j. We can plot

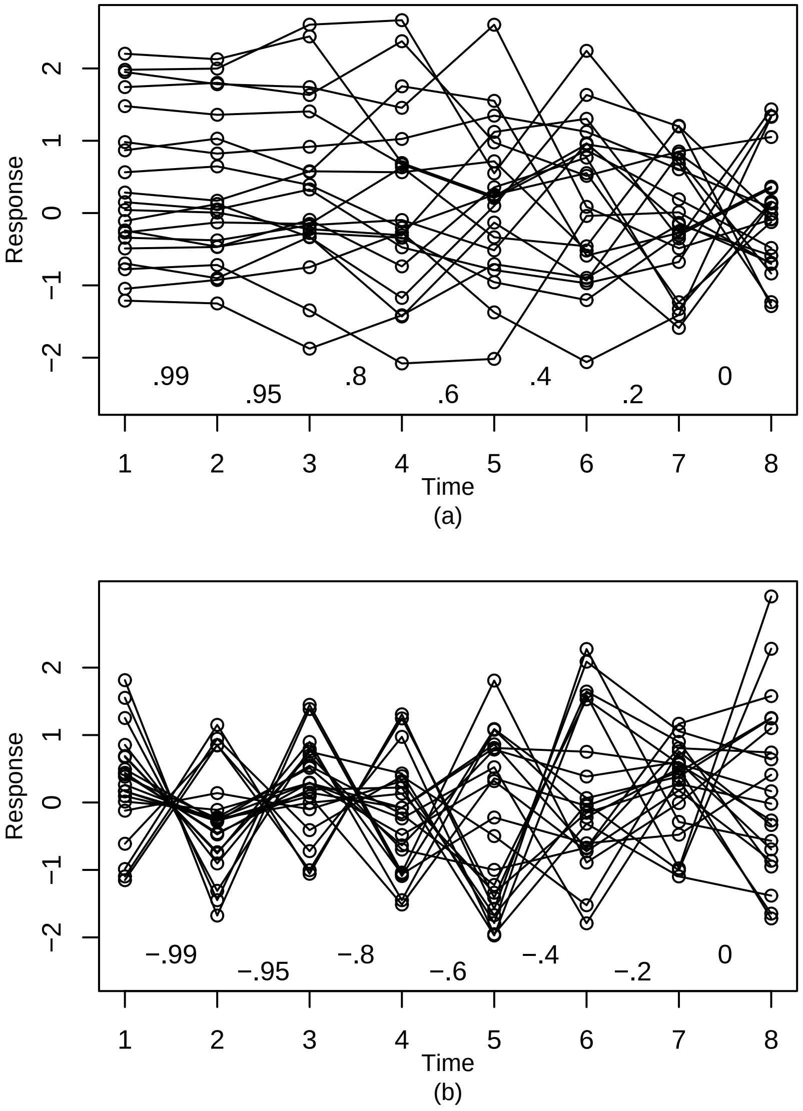

Figure 2.21. Example profile plots for 20 subjects illustrating decreasing correlations between consecutive observations. Part (a) shows positive correlations .99, .95, .8, .6, .4, .2, 0 between consecutive times. Part (b) shows negative correlations −.99, −.95, −.8, −.6, −.4, −.2, and 0.

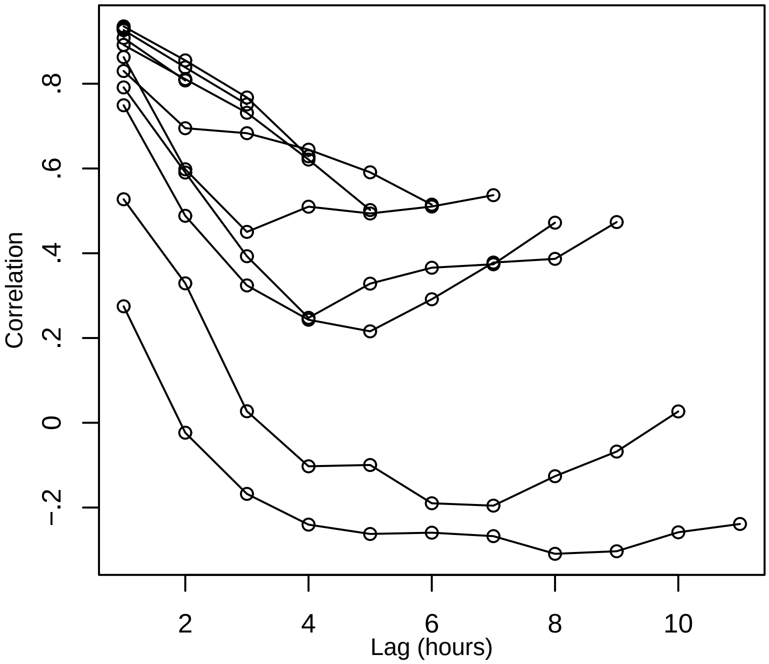

Figure 2.22. Correlogram for the Ozone data.

the $\bar{Y}_{.j}$ against t_j; usually we connect the dots between consecutive time points, as with the profile plots. When we plot the means $\bar{Y}_{.j}$, we often want to show a measure of uncertainty. When we average n independent observations with sample standard variance $s_{jj}^2 = (n-1)^{-1}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_{.j})^2$, then the standard error of the mean is $\mathrm{SE}(\bar{Y}_{.j}) = n^{-1/2}s_{jj}$. We may plot the means along with error bars that illustrate plus and minus 1 standard error to show the size of the standard error. More often we plot plus and minus 2 standard errors to show approximate 95% confidence intervals around $\bar{Y}_{.j}$, which we call an *empirical summary plot*. A third possibility is to show an interval that covers most of the data. We might show error bars that are plus and minus 2 sample standard deviations, $\pm 2s_{jj}$. This is an approximate 95% prediction interval and we call this plot an *empirical prediction plot*.

Figure 2.23(a) gives an empirical summary plot, and 2.23(b) gives an empirical prediction plot for the Big Mice data. The prediction intervals are much wider than the inference intervals; predictions address the prediction of a new observation with all the variability an individual observation has.

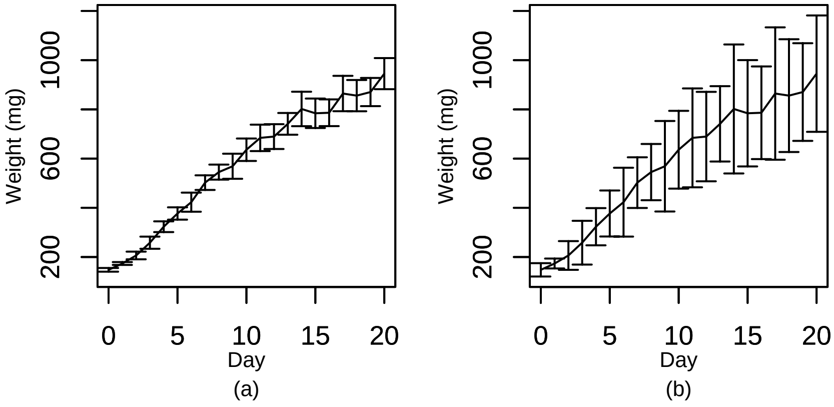

Figure 2.23. (a) Empirical summary plot and (b) empirical prediction plot for the Big Mice data.

The empirical summary interval is making an inference about the *average* response at a given time.

We often wish to distinguish between subjects with different covariate values when we plot empirical summary plots. We may plot the mean profiles from different covariate values either in different plots, or on the same plot, slightly offset from each other so that neither set of intervals obscures the other. Figure 2.24(a) illustrates inference profiles for the attenders and distracters on a single plot. The error bars for attenders and distracters are slightly offset from each other to avoid overplotting. The units are log base two seconds. The means and standard deviations at each time are calculated using either attender or distracter observations at the given time.

Figure 2.24(b) is a back-transformed version of the Pain data empirical summary plot. The means and the interval endpoints have been transformed back to the original seconds scale prior to plotting. The log and back-transformed plots look similar visually. The advantage of the original seconds scale is that we can read off numbers in convenient units for the centers and endpoints of intervals.

For a continuous covariate, we might do a median split before creating our empirical summary plots. Then we create two of the desired plots, one for subjects above the median and one for subjects below the median.

When we have substantial missing data, we must be careful in drawing conclusions from empirical summary plots; we need to try to confirm that observations are not missing differentially in one group or another, or that high or low observations are not differentially missing. Chapter 12 discusses this at length. Fitting a statistical model to the data and then

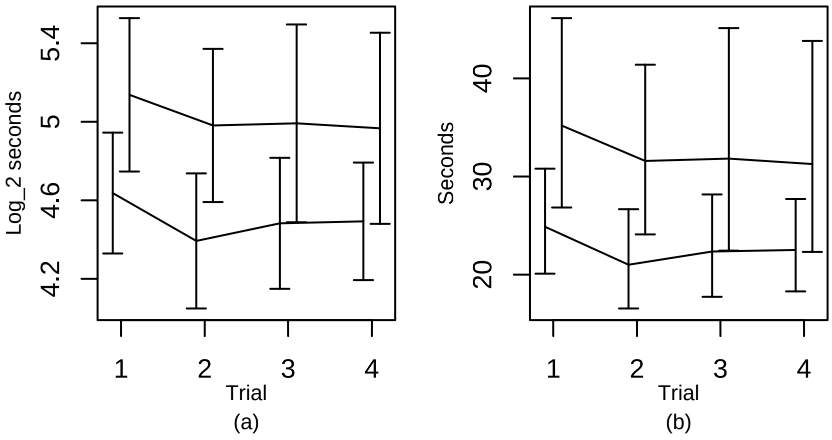

Figure 2.24. Empirical summary plots on the (a) log base two seconds scale and (b) back-transformed seconds scale for the Pain data, separately by coping styles. The upper intervals are for distracters.

plotting an inference plot based on the fitted model can sometimes, but not always, overcome the problems caused by missing data. Thus the empirical summary plot is something we draw early in a data analysis to help us understand the general population time trend and what the effects of covariates might be on the mean. It is not necessarily a good source of conclusions.

2.7 How Much Data?

With longitudinal data, we need to understand how much data we have. How many observations do typical subjects have? For otherwise balanced data except for some missing observations, we tabulate the number of missing (or observed) observations at each time point. An observed observation sounds redundant. Here, observation refers to data we intended to collect by design, and observed means that we actually collected that data from the subject. Similarly, a missing observation is an observation we intended but failed to collect. For randomly observed data there are no missing or observed observations, rather, there are just the observations that we managed to collect. For data with actual times that can be different from the nominal times, we compare the two sets of times to see how they differ. A histogram of actual times can be helpful to give an idea of the times of observations.

Round	# obs
1	543
2	510
3	509
4	497
5	474

Table 2.7. At each round of the Cognitive data, the number of subjects with a Raven's score.

n_i	# obs
0	7
1	8
2	21
3	16
4	40
5	455

Table 2.8. Number of Kenya subjects with from 0 to 5 Raven's observations.

2.7.1 Cognitive Data: Raven's

The Cognitive data is from a school lunch intervention in rural Kenya. The school lunch intervention began at time $t = 0$ in 9 out of 12 schools in the study. Students at the other three schools formed a control group. A number of different measurements were taken. Here we study a particular Cognitive measure called Raven's colored matrices®, a measure of cognitive ability. Up to 5 rounds of data were collected on children in the first form (first grade) in the schools. Round 1 data is baseline data collected in the term before the onset of intervention. Round 2 was taken during the term after the intervention started, rounds 3, 4, and 5 were during the second, fourth, and sixth terms after intervention started. We explore here how much data was collected and when it was collected.

Table 2.7 gives the number of observations taken at each round. We see a steadily decreasing number of observations. This is a common pattern in longitudinal data as subjects drop out of the study or get tired and decline to answer questions or supply information. There are 547 subject identification numbers (ids) in the data set, but only 540 have data in the Raven's data set. How did we get 543 observations at baseline? Further inspection of the data shows that only 530 subjects had baseline data. There are 13 subjects with a second observation before $t = 0$. Other than those 13 subjects, no subject had two Raven's observations during a single round.

Table 2.8 gives the numbers of subjects with from 0 up to 5 observations. Most subjects (83%) have a full 5 observations. The average number of observations per subject is $4.6 = (7 \times 0 + 8 \times 1 + 21 \times 2 + 16 \times 3 + 40 \times 4 + 455 \times 5)/547$.

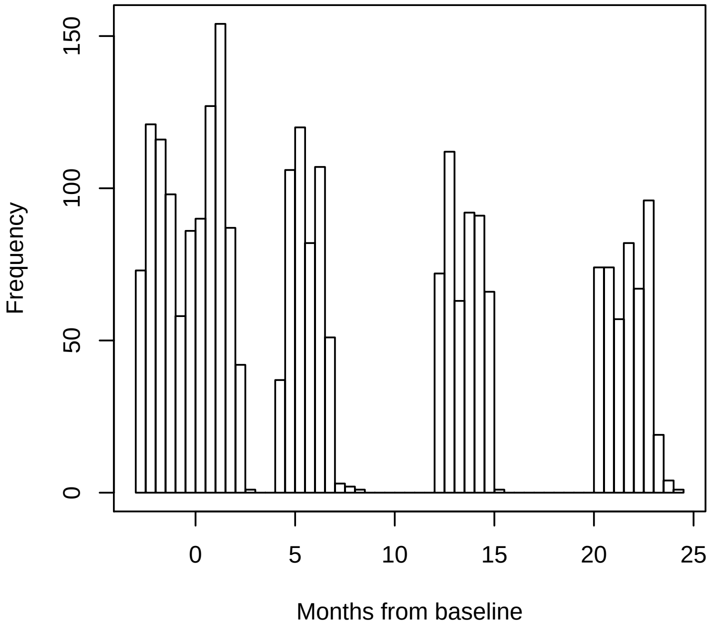

Figure 2.25. Histogram of the actual times of observations for the Kenya Cognitive data.

Inspection of the times of the actual observations can be useful. Figure 2.25 is a histogram of all actual times of observations. We see round 1 observations with times less than zero continuing smoothly into the second round of observations from zero to three months. The third round is clearly separated in time from the second round with observations taken between $t = 4$ up to $t = 8$ months. The fourth and fifth rounds are also clearly separated from each other and from round three.

Figure 2.26 plots the actual times of observations for a selection of subjects. Plotting all subjects requires several figures to create an adequate display, and so figure 2.26 shows 80 subjects. Vertical lines are drawn at months 0 and 3 to show breaks between different rounds of data collection. In inspecting these plots, I redrew them several times adding various vertical lines to aid in drawing conclusions about the times. We see that observations appear to have been taken in clusters. Most of the round 1 observations were taken between -3 and -2 months, with fewer observations taken between -2 and -1 months, and about 11 observations taken between -1 and 0 months. Most of the round 2 observations in this set of subjects were taken right at 1 month. The remainder were taken a bit after month 2. Round 3 has two observations taken right at month 4 but most observations were taken after month 5, with the remainder taken after

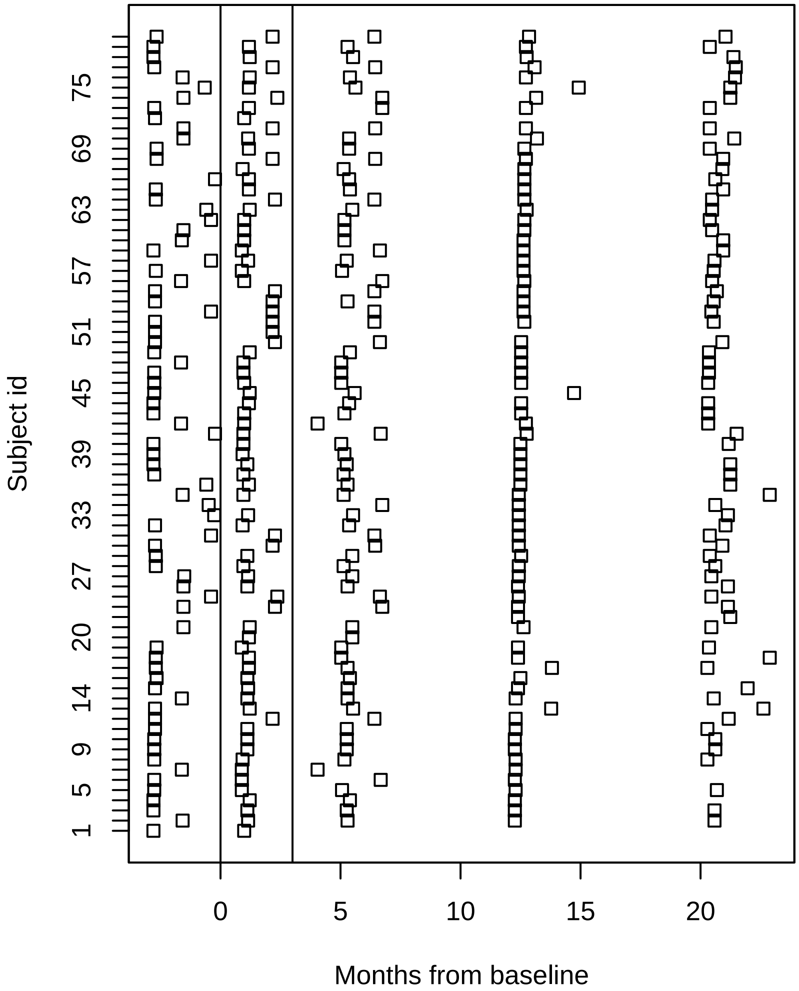

Figure 2.26. Plot of times of observations against ID number for Cognitive data.

month 6. Round 4 has observations taken between month 12 and 13, but there are several straggler observations taken almost to month 15. These comments apply only to the 80 subjects we see here. Figure 2.26 is an *event chart*.

2.8 Discussion

Our plots so far are exploratory; they are designed for investigating the basic distribution and characteristics of the data. Profile plots and scatterplots plot the raw data, empirical correlograms and empirical summary plots plot simple summaries of the data.

Our primary inferences from longitudinal data are about the average response over time and how it varies as a function of covariates. Our models for longitudinal data require us to specify how profiles vary over time and as functions of covariates; profile plots and empirical summary plots can help us with the initial model specification. We also need to specify models to describe the variances of the observations and the correlations among observations. Based on the fitted values from our model, we may plot inferred correlations in a correlogram of the model fit, and we may plot fitted means as functions of time and covariates in inference and prediction plots. These plots may be compared informally to the empirical correlogram and empirical summary plots that we drew in this chapter for model checking. Our model based inferences will usually be more accurate than the empirical summary inferences, but if our model assumptions are incorrect, then the empirical plots can show us how to fix them.

2.9 Problems

1. Consider finding the number of times where a mouse's weight decreases between consecutive measurements.
    (a) In figures 2.1 and 2.2, for how many observations can one be absolutely certain that there was a decrease in weights between one observation and the next for any mouse?
    (b) How about in figure 2.3?
    (c) Inspect the raw data and identify the mouse and the times at which a mouse weight decreases from one observation to the next. How many of these observations can be found in the first three plots of this chapter?
    (d) How difficult is it to identify all circumstances of consecutive measurements with decreasing mouse weight by hand? How long does it take? Is it easier if the Big Mice data were in long or wide format?

(e) Write a program to find the observations where weight decreases from one measurement to the next and present the results.

(f) Compare the difficulty of doing the task by hand as compared to finding all decreases by looking at a figure like 2.3.

2. For the Ozone data,

(a) How often does ozone decrease from one hour to the next during the morning? What locations and what times?

(b) Construct a scatterplot of ozone versus time. Compare it to the profile plot. Which shows more information? What can we figure out about the data using the profile plot that we can't from the scatterplot?

3. Assume balanced data and exactly two time points. Show that the average of the individual slopes is the slope of the line segment connecting the sample means.

4. Does the within-subject average of γ_{ij} from equation 2.1 estimate anything of interest? What is it? To answer these questions:

(a) First assume equally spaced observations and answer the questions.

(b) Second consider unequally spaced and answer the questions.

(c) For unequally spaced observations, is there a weighted average that estimates something similar to what we get for equally spaced observations? Are the weights interpretable?

5. Is the mean over all observations Y_{ij} equal to the mean of the individual averages $\bar{Y}_i$? Explain under what conditions it is and when it is not.

6. Two profiles are said to have similar patterns over time when $Y_{ij} - \bar{Y}_i = Y_{lj} - \bar{Y}_l$, assuming the two subjects' times are the same $t_{ij} = t_{lj}$. Assuming same times, show that this is equivalent to $\gamma_{ij} = \gamma_{lj}$, where γ_{ij} is the slope between times t_{ij} and $t_{i(j-1)}$ for subject i, defined in equation 2.1.

7. For each of the following statements, state whether it is true or false, and come up with a rationalization (or proof) if true or a counterexample if false.

(a) Suppose that all the individual profiles are flat. Then the sample average will be flat over time.

(b) Suppose that all the individual profiles have the same non-zero slope. Then the empirical summary plot slope will be equal to the slope of the individual profiles.

(c) Suppose the data are balanced and the individual profiles are flat. Then the sample average is flat.

(d) Suppose the data are balanced and the individual profiles have the same pattern over time. Then the empirical summary plot pattern over time will be equal to the individual patterns over time.

8. For the data sets in figures 2.5(a) and (c), calculate the mean, sd, min, and max of the observations in each window of width 1 beginning at $t_L = 1$ and increasing t_L in steps of .5 up to $t = 5$. Plot each summary statistic against the midpoints of the window, connecting the dots. For (c), do the calculations both with and without the outlier. In 2.4(a) the means vary quite a lot, even though from looking at the plot we think that they should be approximately constant across time. Explain why the means vary so much in this plot. Explain why dropping the outlier makes the plots for (c) smoother.

9. Create data sets that illustrate the following points. Use 8 subjects with 5 observations per subject unless otherwise specified. Data sets may be sketched with paper and pencil or generated using a statistics package with a random number generator.

   (a) Illustrate a data set where each subject has its own intercept and slope.
   (b) Continuing from the previous example, have the subjects with higher intercepts have the higher slopes also.
   (c) Continuing, have the overall population slope be negative.
   (d) Illustrate subjects who all start low, grow high with separate growth rates, then level off at different heights. You may wish to have more than 5 observations per subject.
   (e) Illustrate subjects who start low, go up to a subject-specific high, then come down low again.
   (f) Invent two more patterns. Plot sample data, and describe the patterns in a sentence.

10. Dropout. Consider three hypothetical weight-loss studies. The first study has only one group of subjects, all treated the same with some intervention that begins immediately after the baseline measurement. The second study is a double-blinded randomized chemical weight loss intervention study. The treatment group gets a diet pill, whereas the control group gets a placebo pill. The third study is also randomized. It is a study of a behavioral weight-loss intervention. The treatment group gets regular meetings with an attitude control specialist, group therapy, and weekly phone calls from a nurse practitioner. The control group gets a pamphlet on weight loss.

    (a) For the three studies, if the treatment is successful, what results would we expect to see? Sketch empirical summary plots for the groups in the study. Describe the plots in one sentence. Do not include error bars, just the mean is fine for these problems.

(b) In general, in a weight loss study, who would be more likely to drop out, those who lose a lot of weight or those who do not lose weight? What effect will this have on a profile plot of weights we drew?

(c) Study 2. Suppose that the pill does not work. What will the profile plots for subjects in the two groups look like?

(d) Study 2, cont. Suppose the pill works. Consider drawing conclusions from the empirical summary plot. Would the apparent conclusions be stronger if there was dropout as compared to if there was no dropout?

(e) Study 3. Which group is likely to stay with the trial longer, which is likely to drop out sooner?

(f) Study 3, cont. Assume both groups lose weight equally, and assume that subjects who don't lose weight drop out differentially in the two groups. Which group, will appear from the empirical summary plot to have better results?

(g) Study 3, cont. Suppose that the treatment group loses weight, whereas the control group does not. Suppose dropout is solely related to the treatment group but not to the amount of weight loss. Will the empirical summary plots make it look as if one group is doing better than it really is?

11. Plot histograms of the Pain data for each trial.

    (a) Does the original scale appear to be skewed? Try various transformations to improve the normality of the data. What is your preferred transformation?

    (b) Does the time point you choose to plot affect the choice of transformation?

    (c) Without plotting histograms of the mice data (or you may if you want to!), describe the differences between two plots, one of which has the data from a specific time as opposed to another that includes data from all trials.

12. Draw profile plots of the Pain data using different line types for the attenders and distracters. Can you tell that the distracters have greater pain tolerance on average? Try using different colors for the lines instead and answer the question.

13. Plot the Pain data with one subject per plot. Make sure that the y axis has the same log base 2 scale for all subjects. Do you observe that low average subjects seem to have less within-subject variability than high average subjects?

14. Take averages of all the observed log base 2 Pain data responses for (a) the attenders, (b) distracters, and (c) attenders omitting the high outlier subject. Transform the averages back to the mean scale.

How well do our eyeball judgments compare with the estimates from subsection 2.4.1? Explain any discrepancies.

15. On a logarithmic scale, suppose that two tick marks are labeled c and d with $c < d$, and you estimate that an observation is $x(100)\%$ of the way from c to d. Show that the observation value is at $c(d/c)^x$. If the proof isn't easy, try plugging in some values for c and d and x. For the Pain data, we already have $d/c = 2$, which is why we used 2 as the power in those calculations.

16. Draw pictures of the Weight Loss data, one profile to a plot.
    (a) There are several choices to be made.
        i. One may construct the y axis of each plot so that they all cover the entire range of the data.
        ii. One may draw each profile's plot so the y axis covers just the range of the given profile.
        iii. One may draw each profile's plot so that the range of the y axis is the same for each profile but is as small as possible.

        For each choice of y axis, is the plot most akin to (i) figure 2.15 or (ii) figure 2.16 or (iii) neither? Use each of the three answers exactly once!
    (b) One can also plot our empirical within-subject residuals $Y_{ij} - \bar{Y}_i$ instead of Y_{ij}. Is there any advantage to the residual profiles one to a plot instead of the original Y_i?

17. In subsection 2.4.4, we plotted profiles of the empirical within-subject residuals $Y_{ij} - \bar{Y}_i$. Would this be of much value for the (i) Big Mice data? (ii) How about the Ozone data? (iii) The Pediatric Pain data? Calculate the empirical within-subject residuals and draw the plots. What do you learn, if anything?

18. In subsection 2.4.4, instead of plotting the empirical within-subject residuals $Y_{ij} - \bar{Y}_i$, suppose that we instead subtract off the baseline measurement and define $W_{ij} = Y_{ij} - Y_{i1}$.
    (a) Plot all of the W_{ij} in a profile plot.
    (b) From this plot, what characteristics of the plot tell you
        i. that subjects are losing weight over the duration of the trial?
        ii. that subjects lose weight at different rates?
        iii. that there is something odd going on from trial 5 to 6?
    (c) Is it easier or harder to detect these three items in this plot as compared to the plot of empirical within-subject residuals? Which plot is better?

19. How could you produce an estimate of a single subject's intercept that is better than the mean $\bar{Y}_i$??? By better, I mean closer to the true value on average.

20. For the Weight Loss data, suppose that we took each observation Y_{ij} for $t_j > 1$ and subtracted off the previous observation $Y_{i(j-1)}$ giving consecutive differences.

$$W_{ij} = Y_{ij} - Y_{i(j-1)}$$

    (a) For a fully observed Y_i, what is the length of $W_i = (W_{ij})$ the vector of all W_{ij} for subject i?
    (b) Plot the W_{ij} in a profile plot.
    (c) What features do you see in the profile plot?
    (d) What do these features imply about the original data Y_{ij}?
    (e) If the Y_{ij} have a random intercept, i.e., $Y_{ij} = \mu_i + \epsilon_{ij}$, what will the W_{ij} profiles look like?
    (f) If the Y_{ij} fall on a subject-specific line $Y_{ij} = a_i + b_i j + \epsilon_{ij}$, then what will the W_{ij} profiles look like?
    (g) If the Y_{ij} follow a subject-specific quadratic $Y_{ij} = a_i + b_i j + c_i j^2 + \epsilon_{ij}$, what will the differences look like?
    (h) What is the problem with this differences plot if (a) There is missing data in the middle of the times? (b) The data are observed at random times?

21. The empirical population residuals were of some use for the mice data, allowing us to look at the individual observation to observation variation. In contrast, it seems implausible that the empirical within-subject residuals would be useful for the mice data.

    (a) For the Pain data, without calculating the two types of residuals and without drawing the plots, one of the residual plots is very unlikely to show us interesting structure, and one might or might not show us interesting structure. Which is which, and briefly, why?
    (b) Answer the same question for the Ozone data.
    (c) Draw both residual plots (empirical within-subject, and empirical population) for the Pain data and illustrate your conclusion from problem part 21(a).
    (d) Draw both plots for the Ozone data and illustrate your conclusion from problem 21b(b).

22. Occasionally, the profile plot plan of connecting the dots may obscure the actual trends in the data. This tends to happen when there is a combination of rapid changes in responses over time and missing data. Table 2.9 presents the Vagal Tone data. Vagal tone is supposed to be high and in response to stress it gets lower. The subjects in this study were a group of 21 very ill babies who were undergoing cardiac catheterization, an invasive, painful procedure. The columns in the table give the subject id number, gender, age in months, the duration of the catheterization in minutes, up to 5 vagal tone measures, and a measure of illness severity (higher is worse). The first

Id	Gender	Age	Dur	Pre1	Pre2	Post1	Post2	Post3	Med
		(m)	(min)						sev
1	F	24	150	1.96	3.04		3.18	3.27	4
2	M	8	180						13
3	M	3	245				0.97		20
4	M	14	300						18
5	F	10	240	3.93	3.86		3.59	3.27	12
6	M	23	240		2.24		2.55	2.27	20
7	M	5	240	2.57	2.52	3.92	3.22	1.28	12
8	M	4	210	4.44	3.07	1.7	3.43	3.43	3
9	M	15	180				1.15	1.03	19
10	F	6.5	330	1.74		1.23	1.49	0.92	5
11	M	15.5	180			2.14		4.92	22
13	F	11	300		3.06	1.94	2.6	1.18	7
14	F	5.5	210						11
15	M	10	300			2.97	5.23		4
16	M	19	330	4.4	3.92	0			29
17	M	6.5	540	1.96	1.95		2.51	1.25	23
18	F	9	210		3.51	1.88	2.14	2.42	13
19	M	15	120	3.63	3.11	1.87	4.46	3.7	15
20	F	3	80	2.91	2.91	0.61			5
21	F	23	65		5.03				5

Table 2.9. Vagal Tone data. Columns are subject number, gender, age in months, length of time in minutes of cardiac catheterization procedure, five vagal tone measures, and a medical severity measure. Blank indicates missing measurement. Subject number 12 has no data.

two measures are before the catheterization, the last three are after. The first measure was taken the night before, the second measure the morning before, then the catheterization; the third measure was taken right after the catheterization, the fourth was taken the evening after, and the last measure was taken the next day. There is a substantial amount of missing data; blanks in the table indicate missing data; subject 12 is missing all variables.

Figure 2.27 shows the Vagal Tone profile plot drawn in two ways. In 2.27(a) we draw the usual plot and connect the dots between all observations within a subject, even if they are not consecutive observations; in 2.27(b), points are connected only if they are consecutive observations from the same subject.

(a) Describe the impressions one gets from the two plots. How are the impressions different?

(b) Which plot do you prefer?

23. Plot the Weight Loss data one profile to a plot. What fraction of subjects appear to be losing weight?

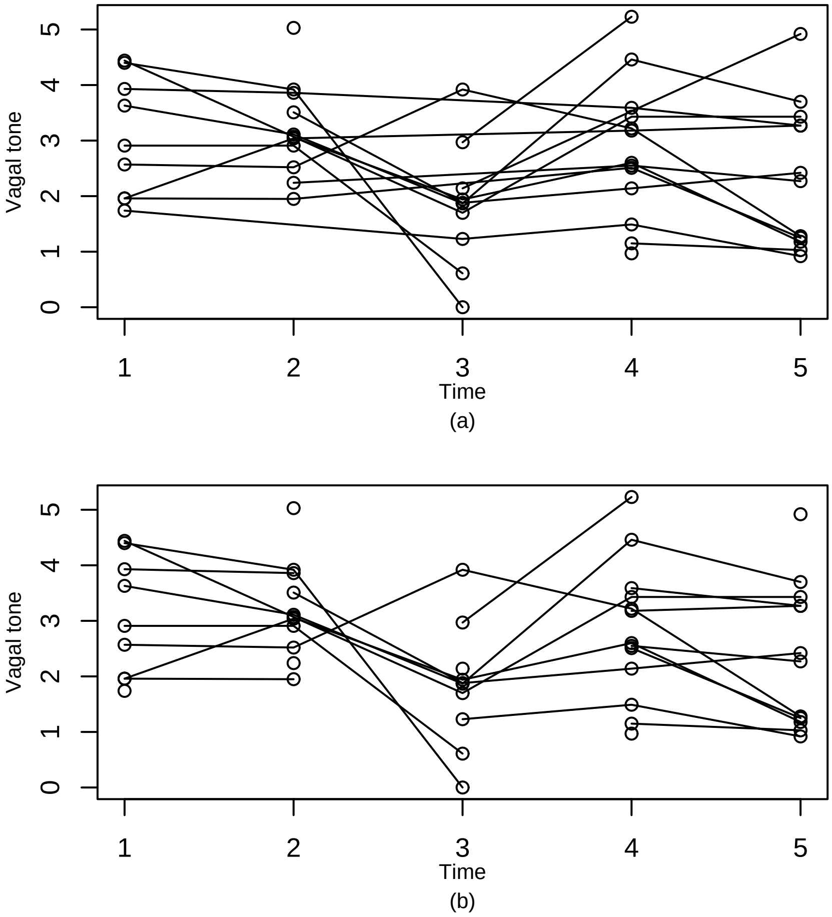

Figure 2.27. Vagal Tone data, plots of vagal tone by measurement. (a) Profile plot. (b) Profile plot, but non-consecutive observations within subject are not connected.

24. Inspect the profile plot of the Weight Loss data.
    (a) Make a rough guess of the correlation between any two observations. Does your guess depend on the specific times that you choose?
    (b) Calculate the correlation matrix for the Weight Loss data.
    (c) Plot the scatterplot matrix of the data.
    (d) Describe your conclusions about the correlations.
    (e) Will the Weight Loss residuals have a greater, lesser, or equal variety of the correlations as compared to the raw Weight Loss data?

(f) Look at figure 2.17. Between what sets of observations do you expect to find

i. a strong positive correlation,

ii. a strong negative correlation, and

iii. a low or zero correlation?

Briefly explain your reasoning.

(g) Calculate the correlation matrix and draw the scatterplot matrix for the Weight Loss residuals. Describe your findings.

25. In the Pain data, draw separate scatterplot matrices for the attenders and distracters. How do the scatterplots differ? What conclusion do you draw?

26. Dental data. The Dental data set is a classic data set for longitudinal data analysis. The response is the length in millimeters from the center of the pituitary gland to the pteryomaxillary fissure for 11 girls and 16 boys. The measurements were taken every two years at ages 8, 10, 12, and 14. There is a single covariate, gender. The purpose of this analysis is to correctly describe the important characteristics of the data.

(a) Create a profile plot of the data. Use separate line types (or colors or plots) for the boys and the girls.

(b) Briefly report your findings. What is the overall pattern? Are boys and girls different? In what ways?

(c) Calculate the correlations among the observations, and draw a scatterplot matrix. Use separate plotting characters for boys and girls.

(d) Report any additional findings.

(e) Draw an empirical summary plot, and repeat separately for boys and girls. Is there a difference in level between boys and girls?

(f) Is there a difference in average slope between boys and girls?

(g) Inspect the profile plot of empirical within-subject residuals. What do you learn about the data? There are four important items to identify about this data set. What are they? You may or may not have seen all of them in the original profile plot.

27. Draw a correlogram for the Small Mice data. Interpret the results.

28. Calculate the correlation matrix and draw a correlogram for the Dental data. What are your conclusions about the correlations?

29. Draw a correlogram for the Pain data. Draw it separately for the attenders and distracters. Describe your conclusions.

30. The standard deviation of an estimated correlation when the true correlation is zero is SE $= (n-3)^{-1/2}$. Often we add horizontal lines

Subject		Age at measurement			
no.	Gender	8	10	12	14
1	Girl	21	20	21.5	23
2	Girl	21	21.5	24	25.5
3	Girl	20.5	24	24.5	26
4	Girl	23.5	24.5	25	26.5
5	Girl	21.5	23	22.5	23.5
6	Girl	20	21	21	22.5
7	Girl	21.5	22.5	23	25
8	Girl	23	23	23.5	24
9	Girl	20	21	22	21.5
10	Girl	16.5	19	19	19.5
11	Girl	24.5	25	28	28
12	Boy	26	25	29	31
13	Boy	21.5	22.5	23	26.5
14	Boy	23	22.5	24	27.5
15	Boy	25.5	27.5	26.5	27
16	Boy	20	23.5	22.5	26
17	Boy	24.5	25.5	27	28.5
18	Boy	22	22	24.5	26.5
19	Boy	24	21.5	24.5	25.5
20	Boy	23	20.5	31	26
21	Boy	27.5	28	31	31.5
22	Boy	23	23	23.5	25
23	Boy	21.5	23.5	24	28
24	Boy	17	24.5	26	29.5
25	Boy	22.5	25.5	25.5	26
26	Boy	23	24.5	26	30
27	Boy	22	21.5	23.5	25

Table 2.10. The Dental data. Columns are subject number, gender, and then the four repeated measurements. Responses are the length in millimeters from the center of the pituitary gland to the pteryomaxillary fissure on each subject. Measurements were taken at ages 8, 10, 12, 14.

at $\pm 2(n-3)^{-1/2}$ to a correlogram to identify correlations that are not significantly different from zero.

(a) Suppose you were to add these lines to a correlogram of the Pain data. Would it change any conclusions? Explain why you can answer this question without actually drawing the correlogram.

(b) Add these lines to the Ozone data correlogram. What does it suggest on an individual correlation basis? Still there are many correlations all of similar size, so perhaps all of those correlations are not equal to zero.

31. Draw an empirical summary plot for the Ozone data. Then draw one separately for valley and non-valley sites. Finally, draw a third for each of the three days and briefly summarize your conclusions.

32. Suppose that some subjects drop out of your study early. Could this cause the empirical summary plot to be misleading? Consider the following examples. For each, (a) sketch and describe how the empirical summary plot will look as compared to how it would look if you had full data, (b) whether the empirical summary plot is misleading, and (c) if it is misleading, how it would be misleading.
    (a) Subject profiles follow a random intercept pattern. All subjects have a 25% chance of not appearing for any given observation.
    (b) Subject profiles follow a random intercept pattern. Subjects who are below average are much more likely to drop out than those above average.
    (c) You have two groups. Subject profiles all start from similar starting points and have different, random, slopes. Subjects who score too high are cured and then tend to drop out permanently. Assume the average slope is the same in both groups. You are interested in either the trend over time or the differences in trend between the two groups; how are these inferences affected by the dropout?
    (d) The same situation as the previous part, but now, subjects in group 1 have a higher slope than subjects in group 2.

33. In the construction of the empirical summary plot, we plotted plus and minus two standard errors of the mean, or plus and minus two sample standard deviations to make an empirical prediction plot. Assuming normally distributed data, how might you improve on these two plots to show (a) an exact 95% confidence interval for the mean, and (b) an exact 95% prediction interval for future data? (Hint: we used the number 2 in constructing our plots. What number should you use instead?) Draw your improved plots for the mice data, can you tell the difference between your plots and figure 2.23?

34. Sketch by hand how your empirical summary plots would look like in figures 12.1(a)–(d). From data that looked like those in (b) and (d), how might you figure out that subjects with low responses were dropping out more than subjects with high responses?

35. The data in figures 12.1(b) and (d) seem troubling. However, approximately, what can happen when we fit a model is that the model first estimates intercepts and slopes for each subject, then averages subjects' intercepts or slopes to get a estimate of the population intercept and slope. Explain why this might be sufficient to get your inference plot from the fitted model in these two figures to look more like the desired empirical summary plot from 12.1(a) and (c).

36. For each of the following data sets, produce a table of the number of observations at each nominal time of observation.

    (a) Small Mice
    (b) Big Mice
    (c) Pediatric Pain
    (d) Weight Loss
    (e) BSI total
    (f) Cognitive
    (g) Anthropometry weights

    Describe each data set observed pattern in a few words.

37. For each of the following data sets, produce a table of the number of subjects with each possible number (i.e., 0 up to J) of observations.

    (a) Small Mice
    (b) Big Mice
    (c) Pediatric Pain
    (d) Weight Loss
    (e) BSI total
    (f) Cognitive
    (g) Anthropometry weights

38. For each of the following data sets, produce a histogram of the actual times t_{ij} that observations were taken. On your plot, mark the nominal times that observations were taken.

    (a) Weight Loss
    (b) BSI total
    (c) Cognitive
    (d) Anthropometry weights