Chapter 7
Repeated Measures and Longitudinal Data Analysis

Knee radiographs are taken yearly in order to understand the onset of osteoarthritis. Troponin (which is an indicator of heart damage) is measured from blood samples 1, 3, and 6 days following a brain hemorrhage. Groups of patients in a urinary incontinence trial are assembled from different treatment centers. Susceptibility to tuberculosis is measured in family members. All of these are examples of what is called repeated measures data or hierarchical or clustered data. Such data structures are quite common in medical research and a multitude of other fields.

Two features of this type of data are noteworthy and significantly impact the modes of statistical analysis. First, the outcomes are correlated across observations. Yearly radiographs on a person are more similar to one another than to radiographs on other people. Troponin measurements on the same person are more similar to one another than to those on other people. And groups of patients from a single center may yield similar responses because of treatment protocol variations from center-to-center, the persons or machines providing the measurements, or the similarity of individuals that choose to participate in a study at that center.

A second important feature of this type of data is that predictor variables can be associated with different levels of a hierarchy. Consider a study of the choice of type of surgery to treat a brain aneurysm either by clipping the base of the aneurysm or implanting a small coil. The study is conducted by measuring the type of surgery a patient receives from a number of surgeons at a number of different institutions. This is thus a hierarchical dataset with multiple patients clustered within a surgeon and multiple surgeons clustered within a hospital. Predictor variables can be specific to any level of this hierarchy. We might be interested in the volume of operations at the hospital, or whether it is a for-profit or not-for-profit hospital. We might be interested in the years of experience of the surgeon or where she was trained. Or we might be interested in how the choice of surgery type depends on the age and gender of the patient.

Accommodation of these two features of the data, predictors specific to different levels in the data structure, and correlated data, are the topics of the chapter.

E. Vittinghoff et al., *Regression Methods in Biostatistics*, Statistics for Biology and Health, DOI 10.1007/978-1-4614-1353-0_7,

We begin by illustrating the basic ideas in a simple example and then describe hierarchical models through a series of examples. In Sect. 7.4, we introduce the first of the methods of dealing with correlation structures, namely generalized estimating equations. Section 7.4.1 introduces an example that we use throughout the rest of the chapter to illustrate the use of the models. Section 7.5 considers an alternative to generalized estimating equations, called random effects modeling, and the following sections contrast these approaches. We close with a section on power and sample size for some repeated measures and clustered data scenarios (Sect. 7.10).

7.1 A Simple Repeated Measures Example: Fecal Fat

Lack of digestive enzymes in the intestine can cause bowel absorption problems. This will be indicated by excess fat in the feces. Pancreatic enzyme supplements can be given to ameliorate the problem. The data in Table 7.1 come from a study to determine if there are differences due to the form of the supplement: a placebo (none), a tablet, an uncoated capsule (capsule), and a coated capsule (coated).

We can think of this as either a repeated measures dataset, since there are four measurements on each patient or, alternatively, as a hierarchical dataset, where observations are clustered by patient. This simple example has as its only predictor pill type, which is specific to both the person and the period of time during which the measurement was taken. We do not have predictors at the patient level, though it is easy to envision predictors like age or a history of irritable bowel syndrome.

We identify a continuous outcome variable, fecal fat, and a single categorical predictor of interest, pill type. If we were to handle this analysis using the tools of Chap. 3, the appropriate technique would be a one-way ANOVA, with an overall F-test, or, perhaps better, a preplanned set of linear contrasts. Table 7.2 gives the one-way ANOVA for the fecal fat example.

Following the prescription in Chap. 3, the F-test indicates ($p = 0.1687$) that there are not statistically significant differences between the pill types. But this analysis is incorrect. The assumptions of the one-way ANOVA require that all observations be independent, whereas we have repeated measures on the same

Table 7.1 Fecal fat (g/day) for six subjects

Subject number	Pill type				Subject Average
	None	Tablet	Capsule	Coated	
1	44.5	7.3	3.4	12.4	16.9
2	33.0	21.0	23.1	25.4	25.6
3	19.1	5.0	11.8	22.0	14.5
4	9.4	4.6	4.6	5.8	6.1
5	71.3	23.3	25.6	68.2	47.1
6	51.2	38.0	36.0	52.6	44.5
Pill type average	38.1	16.5	17.4	31.1	25.8

Table 7.2 One-way ANOVA for the fecal fat example

```
anova fecfat pilltype

                  Number of obs =      24     R-squared     =  0.2183
                  Root MSE      = 18.9649     Adj R-squared =  0.1010

      Source |  Partial SS    df       MS           F     Prob > F
-------------+----------------------------------------------------
       Model |   2008.6017     3  669.533901       1.86     0.1687
             |
    pilltype |   2008.6017     3  669.533901       1.86     0.1687
             |
    Residual |  7193.36328    20  359.668164
-------------+----------------------------------------------------
       Total |  9201.96498    23  400.085434
```

Table 7.3 Two-way ANOVA for the fecal fat example

```
anova fecfat subject pilltype

                  Number of obs =      24     R-squared     =  0.8256
                  Root MSE      =  10.344     Adj R-squared =  0.7326

      Source |  Partial SS    df       MS           F     Prob > F
-------------+----------------------------------------------------
       Model |  7596.98166     8  949.622708       8.88     0.0002
             |
     subject |  5588.37996     5  1117.67599      10.45     0.0002
    pilltype |   2008.6017     3  669.533901       6.26     0.0057
             |
    Residual |  1604.98332    15  106.998888
-------------+----------------------------------------------------
       Total |  9201.96498    23  400.085434
```

six subjects, which are undoubtedly correlated. The one-way ANOVA would be appropriate if we had collected data on six *different* subjects for each pill type.

Should we have conducted the experiment with different subjects for each pill type? Almost certainly not. We gain precision by comparing the pill types within a subject rather than between subjects. We just need to accommodate this fact when we conduct the analysis. This is analogous to the gain in using a paired *t*-test.

In this situation, the remedy is simple: we conduct a two-way ANOVA, additionally removing the variability between subjects. Table 7.3 gives the two-way ANOVA.

The results are now dramatically different, with pill type being highly statistically significant. In comparing Tables 7.2 and 7.3, we can see that a large portion (about 5,588 out of 7,193 or almost 78%) of what was residual variation in Table 7.2 has been attributed to subject-to-subject variation in Table 7.3, thus sharpening the comparison of the pill types.

This is an illustration of a very common occurrence: failure to take into account the correlated nature of the data can have a huge impact on both the analysis strategy and the results.

7.1.1 Model Equations for the Fecal Fat Example

We next write down model equations appropriate for the fecal fat example to more precisely represent the differences between the two analyses from the previous section. The analysis in Table 7.2 follows the one-way ANOVA model from Chap. 3.

$$\begin{aligned} \text{FECFAT}_{ij} &= \text{fecal fat measurement for person } i \text{ with pill type } j \\ &= \mu + \text{PILLTYPE}_j + \epsilon_{ij}, \end{aligned} \tag{7.1}$$

where, as usual, we would assume $\epsilon_{ij} \sim \text{i.i.d } \mathcal{N}(0, \sigma_\epsilon^2)$.

As noted above, there is no account taken of the effect of each subject. We would expect some subjects to generally have higher values and others to generally have lower values. To accommodate this we include a subject effect in the model, which simultaneously raises or lowers all the measurements on that subject:

$$\begin{aligned} \text{FECFAT}_{ij} &= \text{fecal fat measurement for person } i \text{ with pill type } j \\ &= \mu + \text{SUBJECT}_i + \text{PILLTYPE}_j + \epsilon_{ij}, \end{aligned} \tag{7.2}$$

with

$$\epsilon_{ij} \sim \text{i.i.d } \mathcal{N}(0, \sigma_\epsilon^2).$$

To this we add one more piece. We assume that the subject effects are also selected from a distribution of possible subject effects: $\text{SUBJECT}_i \sim \text{i.i.d } \mathcal{N}(0, \sigma_{subj}^2)$, independently of ϵ_{ij}.

This additional piece serves two purposes. First, it captures the idea that the subjects in our experiment are assumed to be a random sample from a larger population of subjects to which we wish to draw inferences. Otherwise, the conclusions from our experiment would be scientifically uninteresting, as they would apply only to a select group of six subjects. Second, as we will examine in detail in the next section, the inclusion of a subject effect (along with an assigned distribution) models a correlation in the outcomes. Once we added this subject effect to our model, we modified our analysis to accommodate it using a two-way ANOVA.

7.1.2 Correlations Within Subjects

The main reason the results in Tables 7.2 and 7.3 differ so dramatically is the failure of the analysis in Table 7.2 to accommodate the repeated measures or correlated nature of the data. How highly correlated are measurements within the same person? The model given in (7.2) gives us a way to calculate this. The observations on the same subject are modeled as correlated through their shared random subject effect.

The larger the subject effects in relation to the error term, the larger the correlation (relatively large subject effect means the observations on one subject are quite different than those on another subject, but, conversely, that observations *within* a subject tend to be similar). More precisely, there is a covariance between two observations on the same subject:

$$\begin{aligned} \text{cov}(\text{FECFAT}_{ij}, \text{FECFAT}_{ik}) &= \text{cov}(\text{SUBJECT}_i, \text{SUBJECT}_i) \\ &= \text{var}(\text{SUBJECT}_i) \\ &= \sigma^2_{subj}. \end{aligned} \tag{7.3}$$

The first equality in (7.3) is because the μ and pilltype terms are assumed to be fixed constants and do not enter into the covariance calculation. The ϵ_{ij} terms drop out because they are assumed to be independent of the subject effects and of each other. The second equality is true because the covariance of any term with itself is a variance and the last equality is just the notation for the variance of the subject effects.

As we recall from Chap. 3, this is just one ingredient in the calculation of the correlation. We also need to know the standard deviations for the measurements. Model (7.2) also indicates how to calculate the variance and hence the standard deviation:

$$\begin{aligned} \text{var}(\text{FECFAT}_{ij}) &= \text{var}(\text{SUBJECT}_i) + \text{var}(\epsilon_{ij}) \\ &= \sigma^2_{subj} + \sigma^2_{\epsilon} \end{aligned} \tag{7.4}$$

so that

$$\text{SD}(\text{FECFAT}_{ij}) = \sqrt{\sigma^2_{subj} + \sigma^2_{\epsilon}},$$

which is assumed to be the same for all observations. The result, (7.4), is noteworthy by itself, since it indicates that the variability in the observations is being decomposed into two pieces, or components, the variability due to subjects and the residual, or error, variance.

We are now in a position to calculate the correlation as the covariance divided by the standard deviations.

$$\begin{aligned} \text{corr}(\text{FECFAT}_{ij}, \text{FECFAT}_{ik}) &= \frac{\text{cov}(\text{FECFAT}_{ij}, \text{FECFAT}_{ik})}{\text{SD}(\text{FECFAT}_{ij})\text{SD}(\text{FECFAT}_{ik})} \\ &= \frac{\sigma^2_{subj}}{\sqrt{\sigma^2_{subj} + \sigma^2_{\epsilon}}\sqrt{\sigma^2_{subj} + \sigma^2_{\epsilon}}} \\ &= \frac{\sigma^2_{subj}}{\sigma^2_{subj} + \sigma^2_{\epsilon}}. \end{aligned} \tag{7.5}$$

While the methods of the calculations are not so important, the intuition and results are. Namely that subject-to-subject variability simultaneously raises or lowers all the observations on a subject, thus inducing a correlation, and that the variability of an individual measurement can be separated into that due to subjects and residual variance.

Looking at the ANOVA table in Table 7.3, we have an estimate of σ^2_ϵ, which is 106.99888. But what about an estimate for σ^2_{subj}? It would be almost correct to calculate the variance of the subject averages in the last column of Table 7.1, but this would be a bit too large since each subject average also has a small amount of residual variation as well. Taking this into account (see Problem 7.1) gives an estimate of 252.67.

Using this in (7.5) gives a correlation of $0.70 = 252.67/(252.67 + 107.00)$, not a particularly high value. So even a moderate value of the correlation can have a fairly dramatic effect on the analysis, which is why it is so important to recognize repeated measures or clustered-data situations. In this instance, the analysis ignoring the correlation led to nonsignificant results and inflated p-values. Unfortunately, the effect of ignoring the correlation can also make the p-values appear incorrectly small, as will be demonstrated in Sect. 7.4.4. So ignoring the correlation does not always produce a "conservative" result.

In this example, we are mainly interested in comparing the effect of the different pill types and the correlation within subjects must be accommodated in order to perform a proper analysis. The correlation is more of a nuisance. In other studies, the correlation will be the primary focus of the analysis, such as repeatability or validation studies or in analysis of familial aggregation of a disease. In the knee osteoarthritis example, the same radiographs were sent to different reading centers to check consistency of results across the centers. One of the primary parameters of interest was the correlation of readings taken on the same image.

7.1.3 Estimates of the Effects of Pill Type

What about estimating the effects of the various pill types or differences between them? The simple averages across the bottom of Table 7.1 give the estimates of the mean fecal fat values for each pill type. There is nothing better we can do in this balanced-data experiment. The same is true for comparing different pill types. For example, the best estimate of the difference between a coated capsule and an uncoated capsule would be the simple difference in means: $31.07 - 17.42 = 13.65$. That is, we do nothing different than we would with a one-way ANOVA (in which all the observations are assumed independent). This is an important lesson that we extend in the next section: the usual estimates based on the assumption of independent data are often quite good. It is the estimation of the standard errors and the tests (like the F-test) that go awry when failing to accommodate correlated data.

7.2 Hierarchical Data

The data structures we describe in this chapter and the analysis strategies are designed for hierarchical data. This is a somewhat vague term, but we now attempt a more formal definition.

> *Definition*: *Hierarchical data* is data (responses or predictors) collected from or specific to different levels within a study.

Other terminologies for the same or related ideas are repeated measures data, longitudinal data, clustered data, and multilevel data. We next illustrate this definition in the context of two examples.

7.2.1 Example: Treatment of Back Pain

A more complicated example of a hierarchical model was first introduced in Chap. 1. In Korff et al. (1994), 44 primary care physicians in a large HMO were classified according to their practice style in treating back pain management (low, moderate, or high frequency of prescription of pain medication and bed rest). An average of 24 patients per physician were followed for 2 years (1 month, 1 year, and 2 year follow-ups) after the index visit. Outcomes included functional measures (pain intensity, activity limitation days, etc.), patient satisfaction (e.g., "After your visit with the doctor, you fully understood how to take care of your back problem"), and cost. Two possible questions are (1) Do physicians with different practice styles differ in function, satisfaction, or cost? and (2) How much of the variability in the responses is due to physician? In this example, there are three levels to the data structure: physicians, patients, and visits. Predictors could be physician-level variables like practice style and years of experience, patient-level variables like age and reason for the back pain, and visit-level variables like time since index visit. The data set is hierarchical because it has variables that are specific to each of the different levels (physician, patient, or visit) of the data.

7.2.2 Example: Physician Profiling

Common methods for the assessment of individual physicians' performance at diabetes care were evaluated in Hofer et al. (1999). They studied 232 physicians from three sites caring for a total of 3,642 patients, and evaluated them with regard to their ability to control HbA_{1c} levels (a measure of control of blood sugar levels) and with regard to resource utilization. Various methods for obtaining physician level predictions were compared including age- and sex-adjusted averages, the calculation of residuals after adjusting for the case-mix of the patients, and

hierarchical modeling. They found that the first two methods overstate the degree to which physicians differ. This could have adverse consequences in falsely suggesting that some physicians (especially those with small numbers of patients) are over using resources or ineffectively treating patients.

As we will see explicitly later in the chapter, hierarchical analysis is more effective in this situation because it "borrows strength" across physicians in order to improve the predicted values for each physician. Said another way, we can use knowledge of the variation between and within physicians in order to quantify the degree of unreliability of individual physician's averages and, especially for those with small numbers of patients, make significant adjustments.

7.2.3 Analysis Strategies for Hierarchical Data

As has been our philosophy elsewhere in this book, the idea is use simpler statistical methods unless more complicated ones are necessary or much more advantageous. That raises the basic question: Do we need hierarchical models and the attendant more complicated analyses? An important idea is the following. Observations taken within the same subgroup in a hierarchy are often more similar to one another than to observations in different subgroups, other things being equal. Equivalently, data which are clustered together in the same level of the hierarchy (data on the same physician, or on the same patient or in the same hospital) are likely to be correlated. The usual statistical methods (multiple regression, basic ANOVA, logistic regression, and many others) assume observations are independent. And we have seen in Sect. 7.1 the potential pitfalls of completely ignoring the correlation.

Are there simple methods we can use that accommodate the correlated data? Simpler approaches that get around the issue of correlation include separate analyses for each subgroup, analyses at the highest level in the hierarchy, and analyses on "derived" variables. Let us consider examples of each of these approaches using the back pain example.

7.2.3.1 Analyses for Each Subgroup

Analysis for each subgroup would correspond to doing an analysis for each of the 44 doctors separately. If there were sufficient data for each doctor, this might be effective for some questions, for example, the frequency with which patients for that physician understood how to care for their back. For other questions it would be less satisfactory, for example, how much more it cost to treat older patients. To answer this question, we would need to know how to aggregate the data across doctors. For yet other questions it would be useless. For example, comparing practice styles is a between-physician comparison and any within-physician analysis is incapable of addressing it.

7.2.3.2 Analysis at the Highest Level in the Hierarchy

An analysis at the highest level of the hierarchy would proceed by first summarizing the data to that level. As an example, consider the effect of practice style on the cost of treatment. Cost data would be averaged across all times and patients within a physician, giving a single average value. A simple analysis could then be performed, comparing the average costs across the three types of physicians. And by entering into the analysis a single number for each physician, we avoid the complication of having correlated data points through time on the same patient or correlated data within a physician.

There are several obvious drawbacks to this method. First, there is no allowance for differences in patient mix between physicians. For example, if those in the aggressive treatment group also tended to have older, higher cost patients we would want to adjust for that difference. We could consider having additional variables such as average age of the patients for each physician to try to accommodate this. Or a case mix difference of another type might arise: some physicians might have more complete follow-up data and have different proportions of data at the various times after the index visit. Adjusting for differences of these sorts is one of the key reasons for considering multipredictor models.

A second drawback of analysis at the highest level of the hierarchy is that some physicians will have large numbers of patients and others will have small numbers. Both will count equally in the analysis. This last point bears some elaboration. Some data analysts are tempted to deal with this point by performing a weighted analysis where the physician receives a weight proportional to the number of observations that went into their average values or the number of patients that contributed to the average. But this ignores the correlated nature of the data. If the data are highly correlated within a physician then additional patients from each physician contribute little additional information and all physicians' averages should be weighted equally regardless of how many patients they have. At the other extreme, if each patient counts as an independent data point, then the averages *should* be weighted by the numbers of patients.

If the data are correlated but not perfectly correlated, the proper answer is somewhere in between these two extremes: a physician with twice as many patients as another should receive more weight, but not twice as much. To determine precisely how much more requires estimation of the degree of correlation within a physician, i.e., essentially performing a hierarchical analysis.

7.2.3.3 Analysis on "Derived Variables"

A slightly more sophisticated method than simple averaging is what is sometimes called the use of "derived variables." The basic idea is to calculate a simple, focused variable for each cluster or subgroup that can be used in a more straightforward analysis. A simple and often effective example of this method is calculation of a change score. Instead of analyzing jointly the before and after treatment values on a subject (with a predictor variable that distinguishes them), we instead calculate the change score.

Here are two other examples of this methodology. In a pharmacokinetic study, we might sample a number of subjects over time after administration of a drug and be interested in the average value of the drug in the bloodstream and how it changes with different doses of the drug. One strategy would be to analyze the entire data set (all subjects and all times) but then we would need to accommodate the correlated nature of the data across time within a person. A common alternative is to calculate, for each person, the area under the curve (AUC) of the concentration of the drug in the bloodstream versus time. This AUC value would then be subjected to a simpler analysis comparing doses (e.g., a linear regression might be appropriate). In the fecal fat example, the derived variable approach is quite effective. Suppose we were interested in the effect of coating a capsule. We can calculate the six differences in fecal fat measurements between the uncoated and coated capsule (one for each person) and do a one-sample or paired t-test on the six differences. See Problem 7.5. For the back pain example, the derived variable approach is not as successful. The unbalanced nature of the data makes it difficult to calculate an effective derived variable.

In summary, the use of hierarchical analysis strategies is clearly indicated in any of three situations:

(1) When the correlation structure is of primary interest,
(2) When we wish to "borrow strength" across the levels of a hierarchy in order to improve estimates, and
(3) When dealing with highly unbalanced correlated data.

7.3 Longitudinal Data

In *longitudinal* studies, we are interested in the change in the value of a variable within a "subject" and we collect data repeatedly through time. For example, a study of the effects of alcohol might record a measure of sleepiness before and after administration of either alcohol or placebo. Interest is in quantifying the effect of alcohol on the *change* in sleepiness. This is often a good design strategy since each subject acts as their own control, allowing the elimination of variability in sleepiness measurements from person-to-person or even occasion-to-occasion within a person. For this strategy to be effective, the before and after measurements need to be at least moderately strongly positively correlated (otherwise taking differences increases the variability rather than reducing it).

As another example, the Study of Osteoporotic Fractures (SOF) is a longitudinal, prospective study of osteoporosis, breast cancer, stroke, and mortality. In 1986, SOF enrolled 9,704 women and continues to track these women with clinical visits every two years. Data from the first seven visits are now available to the public. The data include measures of BMD, BMI, hormones, tests of strength and function, cognitive exams, use of medication, health habits, and much more.

Some of the questions SOF can be used to answer are:

(1) Is change in BMD related to age at menopause? Considered more generally, this is an analysis relating a time-invariant predictor, age at menopause, with changes in the outcome, BMD.
(2) Is change in BMD related to change in BMI? This is an analysis relating a time-varying predictor, BMI, with changes in the outcome, BMD. BMI varies quite a lot between women, but also varies within a woman over time.
(3) Which participants are likely to maintain cognitive function into their 9th and 10th decades of life? This involves predicting the cognitive trajectory of each of the participants from covariates and previously measured values of cognitive function.

We next consider how longitudinal data can be used to answer questions like (1) and (2) above. We deal with questions of prediction in Sect. 7.7.3.

7.3.1 Analysis Strategies for Longitudinal Data

Including a time variable (such as time since enrollment or visit number if they are approximately equally spaced in time) as a predictor captures the idea of change over time in the outcome. This is because the regression coefficient for a time variable measures the change in the outcome per unit change in time, just as in any regression. For example, if the outcome in a linear regression model was BMD and the time variable was years since enrollment in SOF, the meaning of the regression coefficient for time would be the change in mean BMD per year. If the outcome in a logistic regression model was use of hypertensive medication then the regression coefficient for time would be the change in log odds of using hypertensive medication per year.

But suppose, as above, interest focuses not on the change in BMD *overall*, but instead on whether it is related to age at menopause. Then the regression coefficient for time will vary with age at menopause. In statistical parlance, there will be an *interaction* between time and age at menopause as described in Sect. 4.6. Therefore, to capture the association of change in outcome over time with a time-invariant predictor, we need to include in our model an interaction term with the time variable. For example, to assess whether age at menopause was associated with the change in BMD, the regression model would need to include an interaction between time and age at menopause.

To graphically investigate whether there was an interaction, we divided age at menopause as above or below age 52 and fit a restricted cubic spline in visit with three knots, allowing interactions between age at menopause and visit and derived the predicted values. The commands and results are given in Table 7.4. The fitted model was then plotted versus visit and is given in Fig. 7.1. The relationship between BMD and visit appears curvilinear and those women with age at menopause greater

Table 7.4 Fitting of a restricted cubic spline relating BMD to age at menopause and visit in SOF

```
. mkspline visit_spl=visit, cubic nknots(3)

. regress totbmd i.meno_ov visit_spl* i.meno_ov#c.visit_spl*

      Source |       SS       df       MS              Number of obs =   22372
-------------+------------------------------           F(  5, 22366) =   32.64
       Model |  2.82075406     5  .564150812            Prob > F      =  0.0000
    Residual |   386.56757 22366  .017283715           R-squared     =  0.0072
-------------+------------------------------           Adj R-squared =  0.0070
       Total |  389.388324 22371  .017405942           Root MSE      =  .13147

------------------------------------------------------------------------------
      totbmd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.meno_ov_52 |   .0168294    .008409     2.00   0.045     .0003471    .0333116
  visit_spl1 |  -.0070843   .0011079    -6.39   0.000    -.009256    -.0049127
  visit_spl2 |   .0037694   .0015891     2.37   0.018     .0006546    .0068841
             |
  meno_ov_52#|
c.visit_spl1 |
          1  |  -.0001347   .0025039    -0.05   0.957    -.0050424    .0047731
             |
  meno_ov_52#|
c.visit_spl2 |
          1  |  -.0002443   .0035122    -0.07   0.945     -.0071284   .0066398
             |
       _cons |   .7549819   .0036908   204.56   0.000      .7477477   .762216
------------------------------------------------------------------------------

. predict pred_spl
```

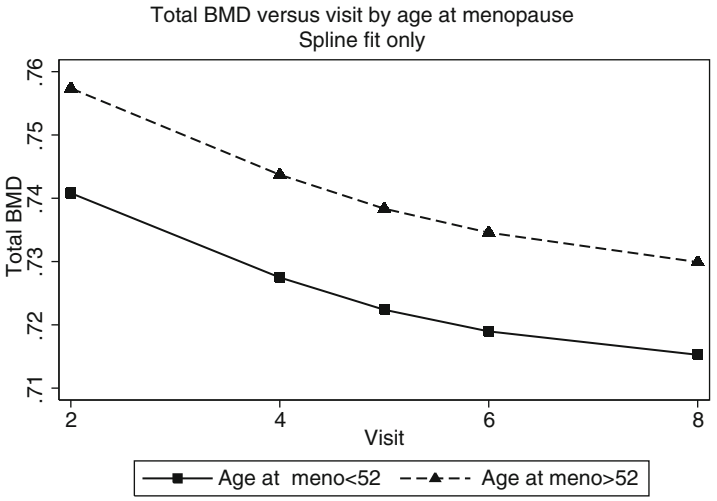


Fig. 7.1 Plot of spline fit to SOF BMD data by age at menopause category

than 52 may have slightly higher BMD values. However, the relationship of the change over time appears remarkably similar between the age at menopause groups, suggesting no time by age at menopause interaction. The analysis in Table 7.4 is

Table 7.5 Summary statistics for first- and last-born babies and the change score

```
summ initwght lastwght delwght

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    initwght |      1000    3016.555    576.2185         815       4508
    lastwght |      1000    3208.195    578.3356        1210       5018
     delwght |      1000      191.64    642.3062       -1551       2700
```

adequate for visualizing the relationship between change in BMD over time and age at menopause, but improper for conducting a formal statistical analysis, since it does not accommodate the repeated measures nature of the data. We return to this example in Sect. 7.7 after we describe appropriate analysis strategies.

7.3.2 Analyzing Change Scores

In simple situations, there is a straightforward approach to analyzing longitudinal data—calculate the change scores (subtract the before measurement from the after measurement) as a derived variable and perform an analysis on the changes. In the alcohol example, we could simply perform a two-sample t-test using the change scores as data to compare the alcohol and placebo subjects.

We consider three approaches to analysis of before/after data that are commonly used: (1) analysis of change scores, (2) repeated measures analysis, and (3) analysis using the after measurement as the outcome and using the baseline measurement as a covariate (predictor). The justification for this last strategy is to "adjust for" the baseline value before looking for differences between the groups. How do these approaches compare?

7.3.2.1 Example: Birthweight and Birth Order

We consider an analysis of birthweights of first-born and last-born infants from mothers (each of whom had five children) from vital statistics in Georgia. We are interested in whether birthweights of last-born babies are different from first-born and whether this difference depends on the age of the woman when she had her first-born.

For the first question, we begin with the basic descriptive statistics given in Table 7.5, where `lastwght` in the variable containing the last-born birthweights, `initwght` indicates the first-born and `delwght` are the changes between last- and first-born within a woman. These show that last-born tend to be about 191 g heavier than first-born (the same answer is obtained whether you average the differences or take the difference between the averages). To accommodate the correlated data, we either perform a one-sample t-test on the differences or, equivalently, a paired t-test of the first and last births. A paired t-test gives a

Table 7.6 Regression of change in birthweight on centered initial age

```
regress delwght cinitage

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  1,   198) =    0.39
       Model |  163789.382     1  163789.382            Prob > F      =  0.5308
    Residual |  82265156.7   198  415480.589            R-squared     =  0.0020
-------------+------------------------------           Adj R-squared = -0.0031
       Total |  82428946.1   199  414215.809            Root MSE      =  644.58

------------------------------------------------------------------------------
     delwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cinitage |   8.891816   14.16195     0.63   0.531    -19.03579    36.81942
       _cons |     191.64   45.57854     4.20   0.000     101.7583    281.5217
------------------------------------------------------------------------------
```

t-statistic of 4.21, with 199 degrees of freedom (since there are 200 mothers) with a corresponding *p*-value that is approximately 0.

What about the relationship of the change in birthweight to the mother's initial age? For this, we conduct a simple linear regression of the change in birthweight regressed on initial age, where we have centered initial age (`cinitage`) by subtracting the mean initial age. The results are displayed in Table 7.6 with the interpretation that each increase of one year in initial age is associated with an additional 8.9 g difference between the first and last birthweights. This is not statistically significant ($p = 0.53$). When centered age is used, the intercept term (`_cons`) is also the average difference.

To conduct a repeated measures analysis, the data are first reordered to have a single column of data containing the birthweights and an additional column, birth order, to keep track of whether it is a first, second, third, fourth, or fifth birth. The output for the repeated measures analysis using only the first and last births in displayed in Table 7.7, for which we leave the details to the next section. However, many of the elements are similar to the regression analysis in Table 7.6. The term listed under `birthord#c.cinitage` is the interaction of birth order and centered initial age. It thus measures how the *difference* in birthweights between first- and last-born is related to centered initial age, that is, whether the change score is related to initial age, the same question as the regression analysis. As is evident, the estimated coefficient is identical and the standard error is virtually the same. They are not exactly the same because slightly different modeling techniques are being used (regression versus GEE, short for generalized estimating equations). The overall difference between first- and last-born is also displayed in the repeated measures analysis (again with the same coefficient and a very similar standard error and *p*-value) and is associated with the birth order term in the model. Finally, the average for first births is displayed as the intercept (see Problem 7.7). So, at a cost of more complication, the repeated measures analysis answers both questions of interest.

A different sort of analysis is to conduct a multiple regression with two predictor variables, initial age (centered) and first-born birthweight. The idea is to "adjust" the values of last-born weight by the first-born weight and then look for an effect due

Table 7.7 Repeated measures regression of birthweight on birth order and centered initial age

```
 . xtgee bweight i.birthord cinitage i.birthord#c.cinitage
> if birthord==1|birthord==5, i(momid)

GEE population-averaged model                 Number of obs      =       400
Group variable:                       momid   Number of groups   =       200
Link:                                identity   Obs per group: min =         2
Family:                              Gaussian                 avg =       2.0
Correlation:                    exchangeable                  max =         2
                                              Wald chi2(3)       =     26.47
Scale parameter:                     323645.4   Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     bweight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  5.birthord |      191.64   45.35007     4.23   0.000     102.7555    280.5245
    cinitage |   25.13981    12.4992     2.01   0.044     .6418238     49.6378
             |
    birthord#|
  c.cinitage |
           5 |   8.891816   14.09096     0.63   0.528    -18.72596    36.50959
             |
       _cons |   3016.555   40.22719    74.99   0.000     2937.711    3095.399
------------------------------------------------------------------------------
```

Table 7.8 Regression of final birthweight on centered initial age, adjusting for first birthweight

```
regress lastwght cinitage initwght if birthord==5

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  2,   197) =   19.33
       Model |  10961363.1     2  5480681.54           Prob > F      =  0.0000
    Residual |  55866154.3   197   283584.54           R-squared     =  0.1640
-------------+------------------------------           Adj R-squared =  0.1555
       Total |  66827517.4   199   335816.67           Root MSE      =  532.53

------------------------------------------------------------------------------
    lastwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cinitage |   24.90948   11.81727     2.11   0.036     1.604886    48.21408
    initwght |   .3628564   .0660366     5.49   0.000      .232627    .4930858
       _cons |   2113.619   202.7309    10.43   0.000     1713.817     2513.42
------------------------------------------------------------------------------
```

to initial age. Table 7.8 gives the results of that analysis, which are quite different than the previous analyses. Now, initial age has a much larger coefficient and is statistically significant ($p = 0.036$).

The intuitive explanation for why this analysis is so different starts with the observation that the coefficient for birthweight of the first-born is approximately 0.363. So, using BW_k to denote the birthweight of the kth born child, we can think of the fitted model as

$$\mathrm{BW}_5 = 2113.619 + .363\mathrm{BW}_1 + 24.909\,\text{Centered initial age} \tag{7.6}$$

or, taking BW_1 to the left side of the equation,

$$\mathrm{BW}_5 - .363\mathrm{BW}_1 = 2113.619 + 24.909\,\text{Centered initial age.} \tag{7.7}$$

That is, this analysis is not purely looking at differences between last and first birthweight since we are only subtracting off a fraction of the initial birthweight. Since birthweights are more highly correlated with initial age than is the difference, this stronger relationship reflects the fact that the results are close to a regression of BW_5 on initial age.

In observational studies, such as this one, using baseline values of the outcome as a predictor is not a reliable way to check the dependence of the change in outcome on a predictor. In randomized studies, where there should be no dependence between treatment effects and the baseline values of the outcome, this may be a more reasonable strategy.

7.3.2.2 When to Use Repeated Measures Analyses

In the Georgia birthweight example, we see that analysis by change scores or by a repeated measures analysis gives virtually identical and reasonable results. The analysis using the baseline value as a predictor is more problematic to interpret.

If the analysis of change scores is so straightforward, why consider the more complicated repeated measures analysis? For two time points and no (or little) missing data, there is little reason to use the repeated measures analysis. However, in the birthweight example there are three intermediate births we have ignored that should be included in the analysis. In the alcohol example, it would be reasonable to measure the degree of sleepiness at numerous time points post-administration of alcohol (or placebo) to track the speed of onset of sleepiness and when it wears off. When there are more than two repeated measures, when the measurements are recorded at different times and/or when there is missing data, repeated measures analysis can more easily accommodate the data structure than attempting change score analyses. We now consider methods for multiple time points.

7.4 Generalized Estimating Equations

There are two main methods for accommodating correlated data. The first we will consider is a technique called *generalized estimating equations*, often abbreviated GEE. A key feature of this method is the option to estimate the correlation structure from the data without having to assume that it follows a prespecified structure.

Before embarking on an analysis, we will need to consider five aspects of the data:

(1) What is the distributional family (for fixed values of the predictors) that is appropriate to use for the outcome variable? Examples are the normal, binary, and binomial families.
(2) Which predictors are we going to include in the model?

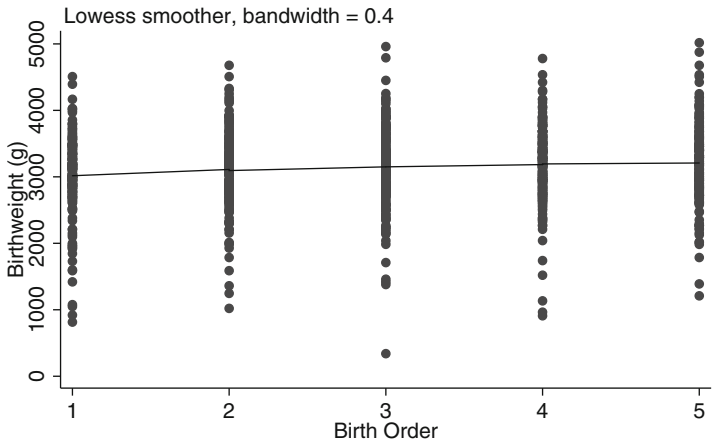


Fig. 7.2 Plot of birthweight (g) versus birth order

(3) In what way are we going to link the predictors to the data? (Through the mean? Through the logit of the risk? Some other way?)
(3) What correlation structure will be used or assumed temporarily in order to form the estimates?
(4) Which variable indicates how the data are clustered?

The first three of these decisions we have been making for virtually every method described in this book. For example, the choice between a logistic and linear regression hinges on the distribution of the outcome variable, namely logistic for binary outcome and linear for continuous, approximately normal outcomes. Chapter 10 discusses the choice of predictors to include in the model (and is a focus of much of this book) and the third has been addressed in specific contexts, e.g., the advantage of modeling the log odds in binary data. The new questions are really the fourth and fifth and have to do with how we will accommodate the correlations in the data. We start by considering an example.

7.4.1 Example: Birthweight and Birth Order Revisited

We return to the Georgia birthweight example and now consider all five births. Recall that we are interested in whether birthweight increases with birth order and mothers' age. Figure 7.2 shows a plot of birthweight versus birth order with both the average birthweights for a given birth order and a LOWESS smooth superimposed. Inspection of the plot suggests we can model the increase as a linear function. A simple linear regression analysis of birthweight versus birth order gives a t-statistic for the slope coefficient of 3.61, which is highly statistically significant. But this analysis would be wrong (why?).

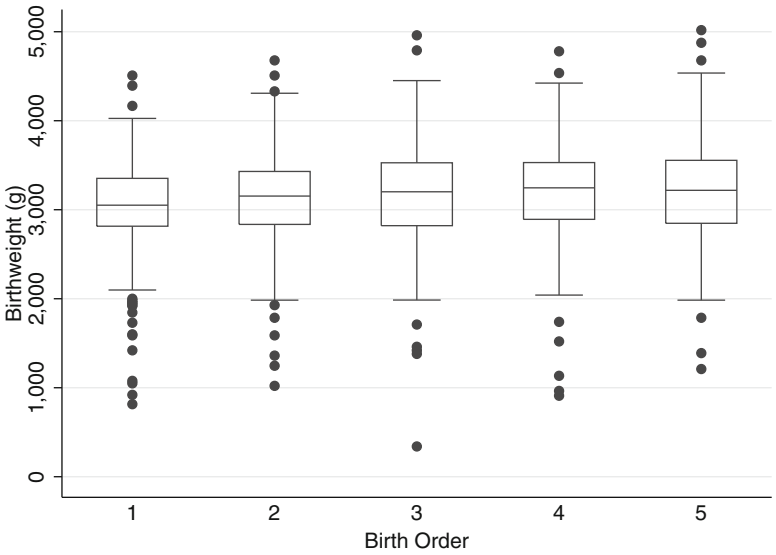


Fig. 7.3 Boxplots of birthweight (g) versus birth order

Recall that the paired t-test using just the first and last births gave a t-statistic of 4.21, even more highly statistically significant. This is perhaps a bit surprising since it discards the data from the three intermediate births.

The explanation for this apparent paradox is that the paired t-test, while using less of the data, does take advantage of the fact that birth order is a within mother comparison. It exploits the correlation of birthweights within a mom in order to make a more precise comparison. Of course, an even better analysis is to use all of the data and accommodate the correlated structure of the data, which we now proceed to do.

7.4.1.1 Analysis

To analyze the Georgia babies dataset, we need to make the decisions outlined above. The outcome variable is continuous, so a logical place to start is to assume it is approximately normally distributed. Figure 7.3 shows boxplots of birthweight by birth order, suggesting that the normality and equal variance assumptions are reasonable. Figure 7.2 has suggested entering birth order as a linear function, which leaves us with the accommodation of the correlation structure.

The data are correlated because five birthweights come from each mother and hence the clustering aspect is clear, leaving us with the decision as to how to model the correlation of measurements taken through time. Figure 7.4 gives a matrix plot of each birthweight against each of the others while Table 7.9 gives the values of the

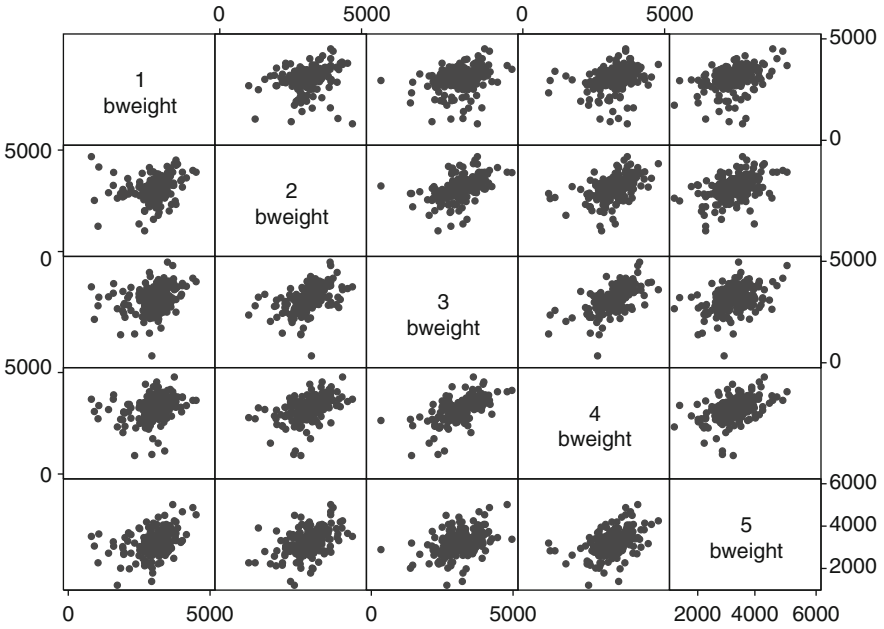


Fig. 7.4 Matrix plot of birthweights for different birth orders

Table 7.9 Correlation of birthweights for different birth orders

```
. corr bweight1 bweight2 bweight3 bweight4 bweight5 (obs=200)

             | bweight1 bweight2 bweight3 bweight4 bweight5
-------------+---------------------------------------------
    bweight1 | 1.0000
    bweight2 | 0.2282 1.0000
    bweight3 | 0.2950 0.4833 1.0000
    bweight4 | 0.2578 0.4676 0.6185 1.0000
    bweight5 | 0.3810 0.4261 0.4233 0.4642 1.0000
```

correlation coefficients. Correlations with the first birthweight might be a bit lower, but the graphs suggest that a tentative assumption of all the correlations being equal would not be far off.

7.4.2 Correlation Structures

Dealing with correlated data typically means making some type of assumption about the form of the correlation among observations taken on the same subject, in the same hospital, on the same mouse, etc. For the Georgia babies data set in the previous section, we noted that assuming all the correlations to be equal might be a

reasonable assumption. This form of correlation is termed exchangeable and means that all correlations (except those variables with themselves) are a common value, which is typically estimated from the data. This type of structure is suitable when there is nothing to distinguish one member of a cluster from another (e.g., patients within a physician) and is the genesis for its name (patients within a doctor can be regarded as interchangeable or exchangeable). This sort of assumption is appropriate in the absence of other data structure, such as measurements taken through time or space.

If measurements are taken through time on the same person, it may be that observations taken more closely in time are more highly correlated. Another common correlation structure is the autoregressive structure, which exhibits this feature. In the simplest form of an *a*uto *r*egressive process (first order or AR(1)) the correlation between observations one time unit apart is a given value ρ, that between observations two time units apart ρ^2, three time units apart ρ^3, etc. Simple arithmetic calculation shows this drops off rapidly to zero (e.g., $0.6^5 = 0.08$) so this assumption would only be appropriate if the correlation between observations taken far apart in time was small and would not be appropriate in cases where stable over time characteristics generated the association. For example, SBP would be relatively stable over time for an individual. Even though observations taken more closely together in time would be slightly more highly correlated, an exchangeable correlation structure might come closer to the truth than an autoregressive one.

Other, less structured, assumptions can be made. In Stata, other options are *unstructured*, *non-stationary*, and *stationary*. All are related to the idea of observations within a cluster being ordered, such as by time. As its name suggests, the unstructured form estimates a separate correlation between observations taken on each pair of "times". The non-stationary form is similar, but assumes all correlations for pairs separated far enough in time are zero. The stationary form assumes equal correlation for all observations a fixed time apart and, like non-stationary, assumes correlations far enough apart in time have correlation zero. For example, stationary of order 2 would assume that observations taken at time points 1 and 3 would have the same correlation as time points 2 and 4, but this might be different from the correlation between observations taken at times 2 and 3. Also, correlations for observations 3 or more time periods apart would be assumed to be zero.

If the correlation structure is not the focus of the analysis, it might seem that the unstructured form is best, since it makes no assumptions about the form of the correlation. However, there is a cost: even with a small number of time points, we are forced to estimate quite a large number of correlations. For instance, with measurements on five time points for each subject, there are ten separate correlations to estimate. This can cause a decrease in the precision of the estimated parameters of interest, or, worse yet, a failure in being able to even fit the model.

This is especially true in situations where the data are not collected at rigid times. For example, in the Nutritional Prevention of Cancer trials (Clark et al. 1996), long-term follow-up was attempted every six months. But the intervals varied widely in practice and quickly were out of synchronization. Estimation of the correlations

between all pairs of distinct times would require literally hundreds of estimated correlations. Use of the unstructured, and, to some extent, the stationary and non-stationary correlation assumptions should be restricted to situations where there are large numbers of clusters, e.g., subjects, and not very many distinct pairs of observation times.

Diagnosis and specification of the "correct" correlation structure is very difficult in practice. One method of addressing these problems is via a *working* correlation assumption and the use of "robust" standard errors, which is the next topic.

7.4.3 Working Correlation and Robust Standard Errors

Given the difficulty of specifying the "correct" correlation structure, a compromise is possible using what are called *robust standard errors*. The idea is to make a temporary or working assumption as to the correlation structure in order to form the estimates but to properly adjust the standard errors of those estimates for the correlation in the data. For example, we might temporarily assume the data are independent and conduct a standard logistic regression. The estimates from the logistic regression will be fairly good, even when used with correlated data, but the standard errors will be incorrect, perhaps grossly so. The solution is to use the estimates but empirically estimate their proper standard errors. Another possibility is to make a more realistic assumption, such as an exchangeable working correlation structure; in some circumstances a gain in efficiency may result.

Then, after the model coefficients have been estimated using the working correlation structure, within-subject residuals are used to compute robust standard errors for the coefficient estimates. Because these standard errors are based on the data (the residuals) and not the assumed working correlation structure, they give valid (robust) inferences for large sized samples as long as the other portions of the model (distribution, link and form of predictors) are correctly specified, even if our working correlation assumption is incorrect. Use of robust standard errors is not quite the same as using an unstructured correlation since it bypasses the estimation of the correlation matrix to directly obtain the standard errors. Avoiding estimation of a large number of correlations is sometimes an advantage, though in cases where both approaches can be used they often give similar results.

The key to the use of this methodology is to have sufficient numbers of subjects or clusters so that the empirical estimate of the correlation is adequate. The GEE approach, which goes hand in hand with estimation with robust standard errors, will thus work best with relatively few time points and relatively more subjects. It is hard to give specific guidelines, but this technique could be expected to work well with 100 subjects, each measured at 5 time points but much less well with 20 subjects, each measured at 12 time points, especially if the times were not the same for each subject.

Table 7.10 Generalized estimating equations analysis using robust standard errors

```
. xtgee bweight birthord initage, i(momid) corr(exch) robust

GEE population-averaged model          Number of obs      =  1000
Group variable:           momid        Number of groups   =   200
Link:                     identity     Obs per group: min =     5
Family:                   Gaussian                    avg =   5.0
Correlation:              exchangeable                max =     5
                                       Wald chi2(2)       = 27.95
Scale parameter: 324458.3              Prob > chi2        = 0.000

                          standard errors adjusted for clustering on momid)
------------------------------------------------------------------------------
             |             Semi-robust
     bweight | Coef.       Std. Err.       z    P>|z|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
    birthord | 46.608      10.02134      4.65    0.000     26.96653   66.24947
     initage | 26.73226    10.1111       2.64    0.008     6.914877   46.54965
       _cons | 2526.622    177.2781     14.25    0.000     2179.164   2874.081
------------------------------------------------------------------------------
```

7.4.4 Tests and Confidence Intervals

Hypothesis testing with GEE uses Wald tests, in which the estimates divided by their robust standard errors are treated as approximately normal to form z-statistics. Likewise, approximate 95% confidence intervals are based on normality by calculating the estimate plus or minus 1.96 standard errors. Table 7.10 shows the analysis with an exchangeable working correlation structure and robust standard errors. Some comments are in order about the form of the command. `xtgee` is a regression type command with numerous capabilities. In its basic form, exhibited in Table 7.10, it performs a linear regression (link of identity) of birthweight (`bweight`) on birth order (`birthord`) and mother's age at first birth (`initage`) with an assumed exchangeable correlation structure (`corr(exch)`) within mother (`i(momid)`). The `robust` option requests the use of robust standard errors.

For comparison sake, Table 7.11 gives the analysis without robust standard errors. There is little difference, though this is to be expected since the preliminary look at the data suggested that the exchangeable assumption would be a reasonable one.

Looking at the analysis with the robust standard errors, the interpretation of the coefficient is the same as for a linear regression. With each increase of initial age of one year, there is an associated increase in average birthweight of about 26.7 g. This result is highly statistically significant, with a p-value of 0.008.

Lest the reader think that the analysis is impervious to the correlational assumptions, Table 7.12 shows what happens to the estimates and standard errors under three different correlation structures both with and without the use of robust standard errors. As expected, the estimates are all similar (the independence and exchangeable are equal because of the balanced nature of the data—five observations per mom with the same values of birth order), though there are

Table 7.11 Generalized estimating equations analysis without robust standard errors

```
. xtgee bweight birthord initage, i(momid) corr(exch)

GEE population-averaged model          Number of obs      =  1000
Group variable:           momid        Number of groups   =   200
Link:                     identity     Obs per group: min =     5
Family:                   Gaussian                    avg =   5.0
Correlation:              exchangeable                max =     5
                                       Wald chi2(2)       = 30.87
Scale parameter: 324458.3              Prob > chi2        = 0.000

                          (standard errors adjusted for clustering on momid)
------------------------------------------------------------------------------
             |             Semi-robust
     bweight | Coef.       Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    birthord | 46.608      9.944792     4.69    0.000     27.11657    66.09943
     initage | 26.73226    8.957553     2.98    0.003     9.175783    44.28874
       _cons | 2526.622    162.544     15.54    0.000     2208.042    2845.203
------------------------------------------------------------------------------
```

Table 7.12 Comparison of the estimated coefficients for `initage` and its standard error for various forms of correlation, with and without robust standard errors

Working correlation	Robust SE?	Coefficient estimate	Standard error	*Z*-statistic	*p*-value
Independence	No	26.73	5.60	4.78	0.000
Exchangeable	No	26.73	8.96	2.98	0.003
Autoregressive(1)	No	27.41	7.82	3.51	0.000
Independence	Yes	26.73	10.11	2.64	0.008
Exchangeable	Yes	26.73	10.11	2.64	0.008
Autoregressive(1)	Yes	27.41	9.69	2.83	0.005

slight variations depending on the assumed working correlation. The estimates are unaffected by the use of robust standard errors.

However, the standard errors and hence Wald statistics and *p*-values are quite different. Those using the incorrect assumptions of independence or autoregressive structure (given in the rows without robust standard errors) are too small, yielding Wald statistics and *p*-values that are incorrect. Looking at the rows corresponding to the use of robust standard errors shows how the incorrect working assumptions of independence or autoregressive get adjusted and now have standard errors that are much more alike. As with any different methods of estimation slight differences do, however, remain.

For the `initage` coefficient the *p*-values assuming independence or, to a lesser extent, autoregressive, are falsely small, but standard errors and *p*-values can, in general, be incorrect in either direction. For example, the `birthorder` effect has a standard error of almost 13 assuming independence, but a standard error of about 10 under an exchangeable correlation (Table 7.11) or under a working exchangeable correlation structure using robust standard errors (Table 7.10).

Table 7.13 Generalized estimating equation logistic regression

```
. xtgee lowbrth birthord initage, i(momid) corr(exch) family(binomial) ///
> link(logit) robust ef

GEE population-averaged model                 Number of obs      =      1000
Group variable:                       momid   Number of groups   =       200
Link:                                 logit   Obs per group: min =         5
Family:                            binomial                   avg =       5.0
Correlation:                   exchangeable                   max =         5
                                              Wald chi2(2)       =     10.64
Scale parameter:                          1   Prob > chi2        =    0.0049

                          (standard errors adjusted for clustering on momid)
------------------------------------------------------------------------------
             |             Semi-robust
     lowbrth | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    birthord |   .9204098     .03542   -2.16   0.031     .8535413    .9925168
     initage |   .9148199   .0312663   -2.60   0.009     .8555464    .9781999
------------------------------------------------------------------------------
```

7.4.5 *Use of* `xtgee` *for Clustered Logistic Regression*

As mentioned above, `xtgee` is a very flexible command. Another of its capabilities is to perform logistic regression for clustered data. We again analyze the Georgia birthweight data but instead use as our outcome the binary variable low-birthweight (`lowbrth`) which is one if the birthweight is less than 3,000 g and zero otherwise. Since the data are binary, we adapt `xtgee` for logistic regression by specifying `family(binomial)` and `link(logit)`. As before, we specify `i(momid)` to indicate the clustering, `corr(exch)` for an exchangeable working correlation, and `robust` to calculate robust standard errors; also we add the option `ef` to get odds ratios instead of log odds. Table 7.13 displays the analysis. The estimated odds ratio for birth order is about 0.92, with the interpretation that the odds of a low-birthweight baby decrease by 8% with each increase in birth order. We see that `initage` is still statistically significant, but less so than in the analysis of actual birthweight. This serves as a warning as to the loss of information possible by unnecessarily dichotomizing a variable.

7.5 Random Effects Models

The previous section discussed the use of generalized estimating equations for the accommodation of correlated data. This approach is limited in that

(1) It is restricted to a single level of clustering,
(2) It is not designed for inferences about the correlation structure,
(3) It does not give predicted values for each cluster or level in the hierarchy.

A different approach to this same problem is the use of what are called *random effects* models.

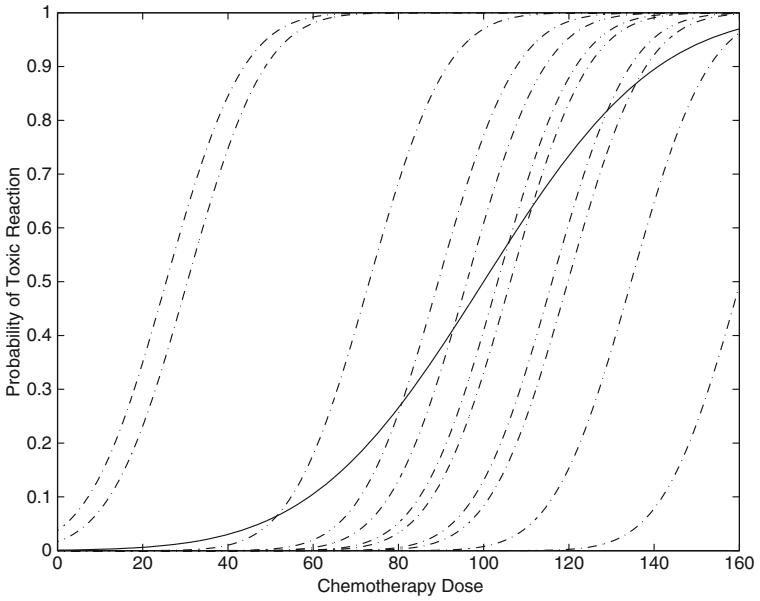


Fig. 7.5 Marginal versus conditional logistic models

First we need to consider two different modeling approaches that go by the names marginal and conditional. These are two common modeling strategies with which to incorporate correlation into a statistical model:

Marginal: Assume a model, e.g., logistic, that holds averaged over all the clusters (sometimes called population averaged). Coefficients have the interpretation as the average change in the response (over the entire population) for a unit change in the predictor. Alternatively, we can think of the coefficient as the difference in the mean values of randomly selected subjects that differ by one unit in the predictor of interest (with all the others being the same).

Conditional: Assume a model specific to each cluster (sometimes called subject-specific). Coefficients have the interpretation as the change in the response for each cluster in the population for a unit change in the predictor. Alternatively, we can think of the coefficient as representing the change within a subject when the predictor of interest is increased by one (holding all the others constant).

In the conditional modeling approach, marginal information can be obtained by averaging the relationship over all the clusters.

On the face of it, these would seem to be the same. But they are not. Here is a hypothetical example. Suppose we are modeling the chance that a patient will be able to withstand a course of chemotherapy without serious adverse reactions. Patients have very different tolerances for chemotherapy, so the curves for individual subjects are quite different. Those patients with high tolerances are shifted to the right of those with low tolerances (see Fig. 7.5). The individual curves are

subject-specific or conditional on each person. The population average or marginal curve is the average of all the individual curves and is given by the solid line in Fig. 7.5 and has quite a different slope than any of the individual curves. This emphasizes that it is important to keep straight which type of model is being used so as to be able to provide proper interpretations and comparisons.

The generalized estimating equations (GEEs) approach most always (always when using `xtgee`) fits a marginal model. Random effects models typically adopt the conditional approach.

Conditional models are usually specified by declaring one or more of the categorical predictors in the model to be *random factors*. (Otherwise they are called *fixed factors*.) Models with both fixed and random factors are called *mixed models*.

> *Definition*: If a distribution is assumed for the levels of a factor, it is a *random factor*. If the values are fixed, unknown constants (to be estimated as model coefficients) it is a *fixed factor*.

The declaration of a factor to be random has several ramifications:

- Scope of inference: Inferences can be made on a statistical basis to the population from which the levels of the random factor have been selected.
- Incorporation of correlation in the model: Observations that share the same level of the random effect are being modeled as correlated.
- Accuracy of estimates: Using random factors involves making extra assumptions but gives more accurate estimates.
- Estimation method: Different estimation methods must be used.

How do we decide in practice as to which factors should be declared random versus fixed? The decision tree in Table 7.14 may be useful in deciding whether the factor is to be considered as fixed or random.

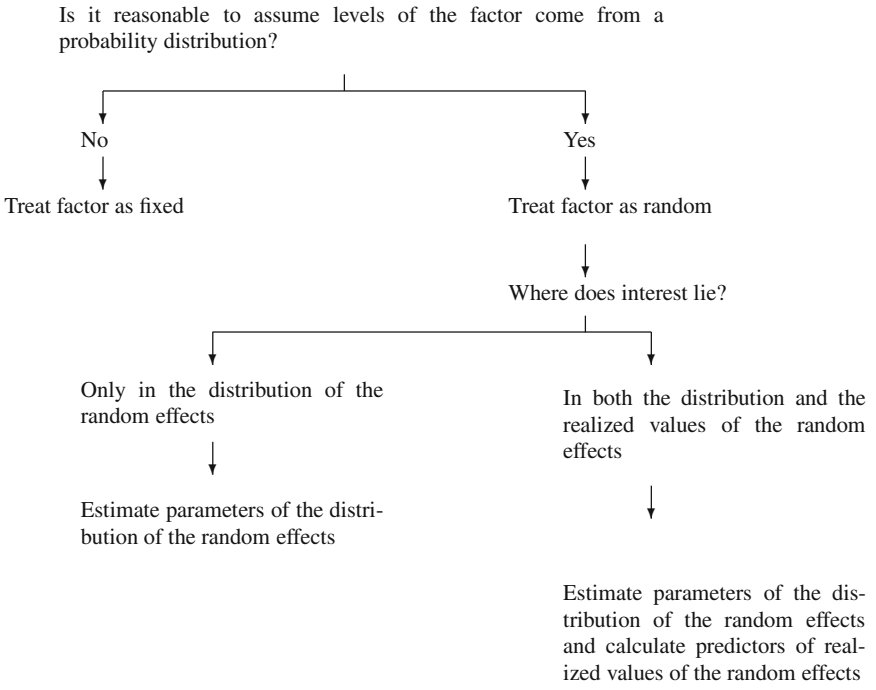
7.6 Re-Analysis of the Georgia Babies Data Set

For the Georgia babies dataset, a random effects assumption for the moms is quite reasonable. We want to regard these particular moms as a sample from a larger sample of moms. Correspondingly the moms' effects on birthweights are easily envisioned as being selected from a distribution of all possible moms.

Stata has a number of commands for conducting random effects analyses; we will focus on two of them: `xtmixed` and `xtmelogit`. The first, `xtmixed`, fits linear mixed models to approximately normally distributed outcomes. The latter, `xtmelogit`, is for mixed models with binary outcomes.

The command syntax is somewhat different from that of `xtgee` because of the need to distinguish the fixed from the random factors. The fixed effect predictors follow the outcome variable in the commands, as is typical of regression commands.

Table 7.14 Decision tree for deciding between fixed and random



However, the random effects are listed after two vertical bars, as displayed in Table 7.15. The colon following the random effect indicates that the model should include random intercepts for each level of that random effect.

The random effects model we fit is similar to that of (7.2):

$$\begin{aligned} \text{BWEIGHT}_{ij} &= \text{birthweight of baby } j \text{ for mom } i \\ &= \beta_0 + \text{MOM}_i + \beta_1 \text{BIRTHORD}_{ij} + \beta_2 \text{INITAGE}_i + \epsilon_{ij}, \end{aligned}$$

with

$$\begin{aligned} \epsilon_{ij} &\sim \text{i.i.d}\, \mathcal{N}(0, \sigma_\epsilon^2) \\ \text{MOM}_i &\sim \text{i.i.d}\, \mathcal{N}(0, \sigma_M^2). \end{aligned} \tag{7.8}$$

Table 7.15 gives the analysis fitting this clustered-data linear regression model. For a linear regression model, the random effects assumption is equivalent to an exchangeable correlation structure as demonstrated in (7.5). Furthermore, for linear models with identity link functions, the marginal and conditional models are equivalent. Hence the random effects analysis reproduces the analysis with an assumed exchangeable correlation structure as given in Table 7.11.

Table 7.15 Linear mixed model analysis of the birthweight data

```
. xtmixed bweight birthord initage|| momid:

Mixed-effects REML regression                   Number of obs      =      1000
Group variable: momid                           Number of groups   =       200
                                                Obs per group: min =         5
                                                               avg =       5.0
                                                               max =         5

                                                Wald chi2(2)       =     30.75
Log restricted-likelihood = -7649.3763          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     bweight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    birthord |     46.608   9.951013     4.68   0.000     27.10437    66.11163
     initage |   26.73226   9.002682     2.97   0.003     9.087332     44.3772
       _cons |   2526.622   163.3388    15.47   0.000     2206.484    2846.761
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.   [95% Conf. Interval]
-----------------------------+------------------------------------------------
momid: Identity              |
                  sd(_cons) |   358.1761   23.71804    314.5799    407.8142
-----------------------------+------------------------------------------------
               sd(Residual) |   445.0228   11.13253    423.7297    467.3859
------------------------------------------------------------------------------
LR test versus linear regression: chibar2(01) = 209.20
Prob >= chibar2 = 0.0000
```

We do, however, have extra output in the random effects analysis. First, the standard deviation of the mom effects, σ_M is equal to 358.1761. This is listed in the output as `sd(_cons)` because it is the standard deviation of the intercepts (or constant terms) associated with each mom. The interpretation of the standard deviation of the mom effects is that it is the standard deviation (across moms) of the true average birthweight per mom. Second is an estimate of the residual standard deviation of 445.0228, from which we can calculate the intramom correlation. Using (7.5), the within mom correlation of any two birthweights is estimated to be $358.1761/(358.1761 + 445.0228) = 0.45$. And third, a test of the null hypothesis of whether the mom-to-mom variation can be considered to be zero, which can be easily rejected using a $\bar{\chi}^2$-test. This is given at the bottom of the Stata output and labeled `chibar2`, short for chi-bar-squared, which has a p-value of approximately 0.

7.7 Analysis of the SOF BMD Data

We return to the Study of Osteoporotic Fractures analysis of the relationship between change in BMD over time and age at menopause (categorized as over or under age 52) that we introduced in Sect. 7.3.1. A primary consideration is how to handle the time variable, visit, which takes on the discrete values, 2, 4, 5, 6, and

8. If we think of the outcome (in this case BMD) evolving smoothly over time, we are naturally led to modeling the trajectory of change using a functional form, for example, a linear trend over time. We would generally like to characterize the trajectory as simply as we can while still using an adequately fitting model. This leads to a natural "ladder" of handling a time predictor like visit, starting from a simple (to model and interpret) linear relationship with time. But this can be quite restrictive and we may need to move up to more flexible models to obtain an adequate fit, for example, also including quadratic functions of time (or even higher degree polynomials) or using a spline (flexible smooth) fit. Failing a simple description with polynomials or splines, and in cases where the times take on a small number of discrete values it may be most expedient to simply handle the time variable as categorical. Moving up the "ladder," we can test statistical significance of the need to utilize the more complicated models.

Recall that the strategy is to include interactions of baseline variables (in this case age at menopause over age 52) with the time variable(s) to check whether there are interactions. Figure 7.1 makes it clear that we need to consider the possibility of a non-linear relationship with visit, so we accommodate visit by using restricted cubic splines. Table 7.16 gives the analysis using GEEs. Neither of the interaction terms with the spline variables is statistically significant so there is no evidence that age at menopause is related to *change* in BMD over time. The spline terms for visit are, themselves, highly statistically significant, indicating that there are changes in BMD over time (unrelated to age at menopause). A comparison with a linear relationship (not shown here) indicates that it is inadequate for describing the changes over time. Consistent with Fig. 7.1, there is a statistically significant difference of about 0.017 between the age at menopause groups across all the visits.

7.7.1 Time Varying Predictors

Age at menopause does not change over time and so is a time-invariant or baseline predictor and we checked for its relationship with changes in BMD by including interactions with the time variables. We next consider the relationship of BMD with BMI, which does change over time within a participant (as well as between participants) and so is a time-varying predictor. How should we include it in the model? Consider a simple model for the measurement on the ith woman at time t with the only predictor being BMI:

$$\text{BMD}_{it} = \beta_0 + \beta_1 \text{BMI}_{it} + \epsilon_{it}. \tag{7.9}$$

Using (7.9) at time $t+1$ and subtracting (7.9) from it gives

$$\begin{aligned} \text{BMD}_{i,t+1} - \text{BMD}_{it} &= (\beta_0 + \beta_1 \text{BMI}_{i,t+1} + \epsilon_{i,t+1}) - (\beta_0 + \beta_1 \text{BMI}_{it} + \epsilon_{it}) \\ &= \beta_1 (\text{BMI}_{i,t+1} - \text{BMI}_{it}) + \epsilon_{i,t+1} - \epsilon_{it}. \end{aligned} \tag{7.10}$$

Table 7.16 Generalized estimating equations analysis of the SOF BMD data

```
. xtgee totbmd i.meno_ov_52 visit_spl* i.meno_ov_52#c.visit_spl*, ///
>    i(id) robust

GEE population-averaged model                 Number of obs      =     22372
Group variable:                         id    Number of groups   =      7004
Link:                             identity    Obs per group: min =         1
Family:                           Gaussian                   avg =       3.2
Correlation:                  exchangeable                   max =         5
                                              Wald chi2(5)       =   2529.68
Scale parameter:                  .0174184    Prob > chi2        =    0.0000

                                  (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Semirobust
      totbmd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.meno_ov_52 |   .0174495   .0040542     4.30   0.000     .0095033    .0253956
  visit_spl1 |  -.0088637   .0003067   -28.90   0.000    -.0094648   -.0082626
  visit_spl2 |   -.000053   .0004967    -0.11   0.915    -.0010265    .0009206
             |
  meno_ov_52#|
c.visit_spl1 |
          1  |   .0000433   .0006456     0.07   0.947    -.0012221    .0013086
             |
  meno_ov_52#|
c.visit_spl2 |
          1  |  -.0001972   .0010286    -0.19   0.848    -.0022132    .0018188
             |
       _cons |    .757436   .0017974   421.40   0.000     .7539131     .760959
------------------------------------------------------------------------------
```

In words, the change in BMD is related to the change in BMI. The import of (7.10) is that, if we fit a model relating the outcome to a time-varying predictor, the regression parameter for the time-varying predictor has the interpretation as the change in the outcome associated with a change in the predictor. That is, it is inherently able to address a longitudinal question.

Table 7.17 gives a mixed model analysis of the relationship between BMD and BMI. Several comments are in order. The model being fit allows for flexible trends over visit, by using a restricted cubic spline in visit. It also allows each participant to have their own intercept and linear trend over visits, through the `id:visit` option. The `cov(uns)` option allows those random intercepts and trends to have arbitrary (unstructured) standard deviations and correlations. This is generally appropriate: the intercepts and trends are measured on completely different scales and are unlikely to have the same standard deviation and, in practice, they are often correlated.

The analysis indicates that there is a highly statistically significant relationship between BMD and BMI. BMD (and perhaps BMI) are measured in units that are unfamiliar to many and so it is difficult to interpret the value of the coefficient for BMI in Table 7.17. It is sometimes easier to consider the changes measured in standard deviation units, namely, the change in BMD (measured in standard deviations of BMD) associated with a single standard deviation change in BMI. This can be simply derived by multiplying the regression coefficient by the standard

Table 7.17 Mixed model analysis of the SOF BMD and BMI data

```
. xtmixed totbmd bmi visit_spl* || id: visit, cov(uns)

Computing standard errors:

Mixed-effects REML regression                   Number of obs      =     26829
Group variable: id                              Number of groups   =      8468
                                                Obs per group: min =         1
                                                               avg =       3.2
                                                               max =         5

                                                Wald chi2(3)       =   7837.49
Log restricted-likelihood =  41482.617          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      totbmd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .0080668   .0001296    62.26   0.000     .0078128    .0083207
  visit_spl1 |  -.0091991   .0002191   -41.98   0.000    -.0096286   -.0087696
  visit_spl2 |  -.0008798   .0002741    -3.21   0.001    -.001417    -.0003425
       _cons |   .5538826   .0036393   152.19   0.000     .5467496    .5610156
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.    [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
                   sd(visit) |   .0096978   .0001461     .0094156    .0099883
                   sd(_cons) |   .1146832   .0009979      .112744    .1166559
              corr(visit,_cons) |  -.0491474   .0169775    -.0823559   -.0158298
-----------------------------+------------------------------------------------
                sd(Residual) |    .023423   .0001517     .0231275    .0237224
------------------------------------------------------------------------------
LR test versus linear regression:  chi2(3) = 45049.07   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

deviation of BMI (which is about 4.70 for this data set) and dividing it by the standard deviation of BMD (which is about 0.13), giving a result of 0.28. So a change in BMD of a single standard deviation is associated with a change of 0.28 standard deviations in BMD, a practically important effect.

7.7.2 Separating Between- and Within-Cluster Information

One could just as well fit a model like (7.9) to a time-invariant predictor, in which case it would not address a longitudinal question. A variable like BMI does vary within an individual over time, but varies even more between individuals. This raises the concern that the coefficient in Table 7.17 might mostly reflect differences between individuals rather than the association of the change in BMD with the associated change of BMI *within an individual*. Between individual associations are often more susceptible to confounding.

Table 7.18 Mixed model separating within and between person BMI

```
*Separate between and within person changes in BMI
bysort id: egen meanbmi=mean(bmi)
gen bmi_dev=bmi-meanbmi

xtmixed totbmd meanbmi bmi_dev visit_spl* || id: visit, cov(uns)

Mixed-effects REML regression                   Number of obs      =     26829
Group variable: id                              Number of groups   =      8468

                                                Obs per group: min =         1
                                                               avg =       3.2
                                                               max =         5

                                                Wald chi2(4)       =   8282.99
Log restricted-likelihood =  41683.621          Prob > chi2        =    0.0000
------------------------------------------------------------------------------
      totbmd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     meanbmi |   .0130329   .0002722    47.88   0.000     .0124994    .0135664
     bmi_dev |    .006695   .0001454    46.04   0.000       .00641      .00698
  visit_spl1 |  -.0090266   .0002192   -41.17   0.000    -.0094562   -.0085969
  visit_spl2 |   -.001178   .0002732    -4.31   0.000    -.0017134   -.0006425
       _cons |   .4226782   .0072925    57.96   0.000     .4083853    .4369712
------------------------------------------------------------------------------
```

Fortunately there are simple ways to isolate the within individual (or more generally within cluster) changes. The first step is to decompose the predictor into two pieces:

$$\begin{aligned} \text{BMI}_{it} &= \left(\text{BMI}_{it} - \overline{\text{BMI}_i}\right) + \overline{\text{BMI}_i} \\ &= \text{BMI\_dev}_{it} + \overline{\text{BMI}_i}, \end{aligned} \tag{7.11}$$

where $\overline{\text{BMI}_i}$ represents the average BMI for person i and BMI_dev_{it} is the deviation of the BMI measurement at time t from their mean BMI. In Stata, the mean and deviation forms of the predictor can easily be calculated using the `bysort` and `egen` commands. Next, both of them are entered in the model as predictors. The deviation form of the predictor represents the within-cluster association and the mean form of the predictor represents the between-cluster association. Another approach that works equally well is to describe the between-cluster portion of the predictor using its baseline value and the within-cluster portion using the difference between each value of the predictor and the baseline value. That is, use BMI_{i1} for between and $\text{BMI}_{it} - \text{BMI}_{i1}$ for within, again entering both predictors in the model.

Table 7.18 shows how to calculate the between and within forms of the predictor and displays a portion of the output from the analysis.

Both the within and between coefficients are highly statistically significant, though this is not surprising given the large sample size. But the between coefficient is about 0.013 and almost twice the size of the within person coefficient, which is about 0.007. This could easily be due to confounding at the person level. The previous analysis reported a weighted average of these two coefficients.

Even in situations in which confounding is not an issue, it may be of substantive interest to conduct such a decomposition of a predictor. For example, Haas et al. (2004) studied the influence of county level race and ethnic composition on access to health care over and above the influence of an individual's race or ethnicity. In this case, interest focused on separating the county level effect (cluster-level effect) of race and ethnicity from the individual level effects.

7.7.3 Prediction

One of the advantages of the random effects approach is the ability to generate predicted values for each of the random effects, which we do not get to observe directly. Returning to the Georgia babies data set, we consider obtaining predicted values for each of the mom effects, MOM_i.

First, let us consider how we might go about estimating the mom effect from first principles. The first mom in the data set had an initial age of 15 and hence, using the estimated coefficients from Table 7.15, has predicted values for the five births (in grams) of 2974.2, 3020.8, 3067.4, 3114.0, and 3160.6 (for example, the first of these is $2974.214 = 2526.622 + 46.608(1) + 26.732(15)$) and actual values of 3720, 3260, 3910, 3320, and 2480, respectively. Her residuals, defined as actual minus predicted, were 745.8, 239.2, 842.6, 206.0, and −680.6 with an average of 270.6. So we might guess that this mom has babies that are, on average, about 271 g heavier than the "average" mom.

Using software to get the predicted effect (deviation from average) for the first mom gives 206.7, only about 76% of the raw data value. Calculation for the other moms shows that all the predicted values are closer to zero than the raw data predicts. Why?

Predicted values from random effects models are so-called *shrinkage estimators* because they are typically less extreme than estimates based on raw data. The shrinkage factor depends on the degree of similarity between moms and, for simple situations, is given by

$$\text{shrinkage factor} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2/n_i}, \tag{7.12}$$

where n_i is the sample size for the ith cluster, σ_u^2 is the between cluster variance, and σ_ϵ^2 is the error variance. In our case, this factor is equal to (taking the estimates from Table 7.15)

$$\begin{aligned}\text{shrinkage factor} &= \frac{358.1761^2}{358.1761^2 + 445.0228^2/5} \\ &= \frac{128,290.1}{128,290.1 + 39,609.1} = 0.76.\end{aligned} \tag{7.13}$$

It is instructive to consider the form of (7.12). Since all the terms in the equation are positive, the shrinkage factor is greater than zero. Further, since the denominator is bigger than the numerator by the factor σ_ϵ^2/n_i, the shrinkage factor is less than 1. So it always operates to shrink the estimate from the raw data to some degree.

What is the magnitude of the shrinkage? If σ_u^2 is much larger than σ_ϵ^2/n_i then the shrinkage factor is close to 1, i.e., almost no shrinkage. This will occur when (a) subjects are quite different (i.e., σ_u^2 is large), and/or (b) results are very accurate and σ_ϵ^2 is small, and/or (c) when the sample size per subject, n_i, is large. So little shrinkage takes place when subjects are different or when answers are accurate or when there is much data.

On the other hand, in cases where subjects are similar (and hence σ_u^2 is small) there is little reason to believe that any individual person deviates from the overall. Or in cases of noisy data (σ_ϵ^2 large) or small sample sizes, random fluctuations can make up the majority of the raw data estimate of the effect and are naturally de-emphasized with this shrinkage approach.

The advantage of the shrinkage predictions are twofold. First, they can be shown theoretically to give more accurate predictions than those derived from the raw data. Second (which is related), they use the data to balance the subject-to-subject variability, the residual variance and the sample size to come up with the best combination of the subject-specific information and the overall data.

Examples of uses of this prediction technology include prediction for prostate cancer screening (Brant et al. 2003) and the use of shrinkage estimators in the rating of individual physicians (Hofer et al. 1999) in treatment of diabetes.

7.7.4 A Logistic Analysis

Turning to the binary outcome variable `lowbrth`, we use the Stata command `xtmelogit`. This model is similar to (7.8) with the needed changes for a logistic model for binary data. This model is:

$$\begin{aligned} \text{LOWBRTH}_{ij} &= 1 \text{ if baby } j \text{ for mom } i \text{ is } < 3{,}000\,\text{g and 0 otherwise} \\ &\sim \text{Bernoulli}(p_{ij}) \end{aligned}$$

with

$$\text{logit}(p_{ij}) = \beta_0 + \text{MOM}_i + \beta_1 \text{BIRTHORD}_{ij} + \beta_2 \text{INITAGE}_i, \quad (7.14)$$

and

$$\text{MOM}_i \sim \text{i.i.d}\, \mathcal{N}(0, \sigma_u^2).$$

This analysis is given in Table 7.19 with a syntax similar to that of `xtmixed`. The fixed effects are listed after the outcome and the vertical bar notation separates the fixed effects from the random effects, again with the `momid:` indicating the

Table 7.19 Random effects logistic regression analysis for the birthweight data

```
. xtmelogit lowbirth  birthord initage|| momid:, or

Mixed-effects logistic regression               Number of obs      =      1000
Group variable: momid                           Number of groups   =       200
                                                Obs per group: min =         5
                                                               avg =       5.0
                                                               max =         5

Integration points =   7                        Wald chi2(2)       =     11.85
Log likelihood = -588.07113                     Prob > chi2        =    0.0027

------------------------------------------------------------------------------
    lowbirth | Odds Ratio   Std. Err.      z    P>|z|   [95% Conf. Interval]
-------------+----------------------------------------------------------------
    birthord |   .8872745   .0500702    -2.12   0.034   .7943711    .9910432
     initage |   .8808974   .0406081    -2.75   0.006   .8047967    .9641941
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.   [95% Conf. Interval]
-----------------------------+------------------------------------------------
momid: Identity              |
                  sd(_cons) |    1.60859   .1676556     1.31138    1.973158
------------------------------------------------------------------------------
LR test versus logistic regression: chibar2(01) = 123.21 Prob>=chibar2 = 0.0000
```

inclusion of random intercepts for each mother. The option `or` requests odds ratios in the output table as opposed to log odds. This gives somewhat different results than the GEE analysis, as expected, since it is fitting a conditional model. More specifically (as predicted from Fig. 7.5), the coefficients in the conditional analysis are slightly farther from 1 than the marginal coefficients, for example the odds ratio for birth order is now 0.89 as compared to 0.92 in the marginal model. The tests are, however, virtually the same, which is not unusual.

The interpretation of the `birthord` coefficient in the conditional model is that the odds of a low-birthweight baby decreases by about 11% for each increase of birth order of one for each woman.

This is opposed to the interpretation of the odds-ratio estimate from the marginal fit given in Table 7.13 of 0.92. The interpretation in the marginal model is the decrease in the odds (averaged across all women) is about 8% with an increase in birth order of one.

7.8 Marginal Versus Conditional Models

The previous section has demonstrated that, for non-linear models like the logistic model, it is important to distinguish between marginal and conditional models since the model estimates are not expected to be equal. Conditional models have a more mechanistic interpretation, which can sometimes be useful (being careful, of course, to remember that many experiments do not strongly support

mechanistic interpretations, no matter what model is fit). Marginal models have what is sometimes called a "public health" interpretation since the conclusions only hold averaged over the entire population of subjects.

7.9 Example: Cardiac Injury Following Brain Hemorrhage

Heart damage in patients experiencing brain hemorrhage has historically been attributed to preexisting conditions. However, more recent evidence suggests that the hemorrhage itself can cause heart damage through the release of norepinephrine following the hemorrhage. To study this, Tung et al. (2004) measured cardiac troponin, an enzyme released following heart damage, at up to three occasions after patients were admitted to the hospital for a specific type of brain hemorrhage (subarachnoid hemorrhage or SAH).

The primary question was whether severity of injury from the hemorrhage was a predictor of troponin levels, as this would support the hypothesis that the SAH caused the cardiac injury. To make a more convincing argument in this observational study, we would like to show that severity of injury is an independent predictor, over and above other circulatory and clinical factors that would predispose the patient to higher troponin levels. Possible clinical predictors included age, gender, body surface area, history of coronary artery disease (CAD), and risk factors for CAD. Circulatory status was described using systolic blood pressure, history of hypertension (yes/no) and left ventricular ejection fraction (LVEF), a measure of heart function. The severity of neurological injury was graded using a subject's Hunt–Hess score on admission. This score is an ordered categorical variable ranging from 1 (little or no symptoms) to 5 (severe symptoms such as deep coma).

The study involved 175 subjects with at least one troponin measurement and between 1 and 3 visits per subject. Figure 7.6 shows the histogram of troponin levels. They are *severely* skewed right with over 75% of the values equal to 0.3, the smallest detectable value and many outlying values. For these reasons, the variable was dichotomized as being above or below 1.0, as is labeled in the output as `CTover1`. Table 7.20 lists the proportion of values above 1.0 for each of the Hunt–Hess categories and Table 7.21 gives a more formal analysis using GEE methods, but including only the predictor Hunt–Hess score and not using data from visits four or greater (there were too few observations to use those later visits).

The reference group for the Hunt–Hess variable in this analysis is a score of 1, corresponding to the least injury. So the odds of heart damage, as evidenced by troponin values over 1, is over two times higher for a Hunt–Hess score of 2 as compared to 1 and the odds go up monotonically with the estimated odds of heart damage for a Hunt–Hess score of 5 being over 70 times those of a score of 1. Even though the odds ratio of a score of 5 is poorly determined, the lower limit of the 95% CI is still over 16.

The primary goal is to assess the influence of a single predictor variable, Hunt–Hess score, which is measured only once per subject. Since it is only measured once,

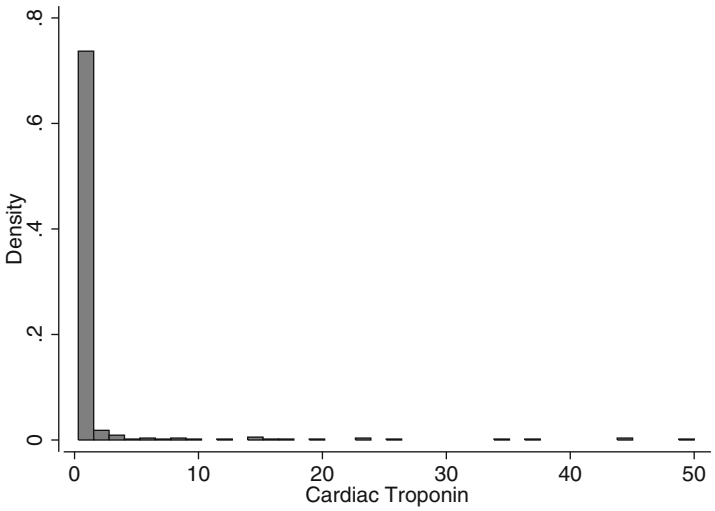


Fig. 7.6 Histogram of cardiac troponin levels

Table 7.20 Proportion of troponin levels over 1.0 and sample size versus Hunt–Hess score

```
. table hunt, c(mean CTover1 n CTover1)
----------------------------------------
Initial     |
Hunt--Hess  | mean(CTover1)    N(CTover1)
------------+---------------------------
          1 |      .0318471           157
          2 |      .0615385            65
          3 |      .1269841           126
          4 |      .1692308            65
          5 |      .6818182            22
----------------------------------------
```

rather than repeatedly, a marginal model and the use of GEE methods is attractive. Since we are interested in a single predictor, we will be more liberal in including predictors for adjustment. We certainly would like to adjust for the amount of time after the SAH occurred, as captured by the visit number, `stday`, since troponin levels drop over time. We also want to adjust for fundamental differences that might be due to age, sex, and body surface area (`bsa`), which may be related to troponin levels.

In addition, we choose to adjust for preexisting conditions that might influence the troponin levels, including left ventricular ejection fraction, standardized (`lvef_std`), SBP (`sbp`), heart rate (`hr`), and history of hypertension (`hxhtn`). Quadratic functions of left ventricular ejection fraction (`lvef_std2`) and SBP (`sbp2`) are included to model non-linear (on the logit scale) relationships.

Table 7.22 gives the output after dropping some nonstatistically significant predictors from the model and using the `xtgee` command. It also gives an overall test of whether troponin levels vary with Hunt–Hess score.

Table 7.21 Effect of Hunt–Hess score on elevated cardiac troponin levels

```
. xtgee CTover1 i.hunt if stday<4, i(stnum) family(binomial) ef

GEE population-averaged model                 Number of obs      =       434
Group variable:                       stnum   Number of groups   =       168
Link:                                 logit   Obs per group: min =         1
Family:                           binomial                   avg =       2.6
Correlation:                  exchangeable                   max =         3
                                              Wald chi2(4)       =     39.03
Scale parameter:                         1    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     CTover1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        hunt |
          2  |   2.036724   1.669731    0.87   0.386     .4084194    10.15682
          3  |   4.493385   2.820396    2.39   0.017     1.313088    15.37636
          4  |   6.542645   4.347658    2.83   0.005     1.778774      24.065
          5  |   70.66887   52.16361    5.77   0.000     16.63111     300.286
------------------------------------------------------------------------------
```

Even after adjustment for a multitude of characteristics, the probability of an elevated troponin level is associated with Hunt–Hess score. However, the picture is a bit different as compared to the unadjusted analysis. Each of the categories above 1 has an estimated elevated risk of troponin release, but it is not a monotonic relationship. Also, only category 5, the most severely damaged group, is statistically significantly different from category 1.

What is the effect of adjusting for the large number of predictors is this model? We might be worried that CIs for some of the coefficients have gotten quite wide due to correlations among the predictors and the Hunt–Hess score. Table 7.23 gives the analysis after minimally adjusting for just `stday`.

While it is not clearly evident from the output on the odds-ratio scale, standard errors for the log odds values are not appreciably larger in the adjusted analysis (see Problem 7.10). The minimally adjusted and unadjusted analyses have similar pattern of estimated odds ratios. However, both of them may have overestimated the association with Hunt–Hess score slightly and so the adjusted analysis reported in Table 7.22 would be preferred.

7.9.1 Bootstrap Analysis

We might also be concerned about the stability of the results reported in Table 7.22 given the modest sized dataset with a binary outcome and the large number of predictors. This is exactly a situation in which bootstrapping can help understand the reliability of standard errors and CIs.

Correspondingly, we conducted a bootstrap analysis and we focus on the stability of the result for the comparison of Hunt–Hess score of 5 compared to a value of 1. Bootstrapping is conducted for the log odds (which can be transformed easily back to the odds scale) since that is the basis of the calculation of CIs.

Table 7.22 Adjusted effect of Hunt–Hess score on elevated troponin levels

```
. xtgee CTover1 i.hunt i.stday sex lvef_std lvef_std2 hxhtn sbp sbp2 if
  stday<4, i(stnum)
> family(binomial) ef

GEE population-averaged model                 Number of obs      =       408
Group variable:                       stnum   Number of groups   =       165
Link:                                 logit   Obs per group: min =         1
Family:                            binomial                   avg =       2.5
Correlation:                   exchangeable                   max =         3
                                              Wald chi2(12)      =     44.06
Scale parameter:                          1   Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     CTover1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        hunt |
          2  |   1.663476   1.334533     0.63   0.526     .3452513    8.014895
          3  |   1.830886   1.211796     0.91   0.361     .5003595     6.69947
          4  |   1.560879   1.241708     0.56   0.576     .3282637    7.421908
          5  |   74.99009   69.48431     4.66   0.000     12.19825    461.0097
             |
       stday |
          2  |   .5258933   .2163491    -1.56   0.118     .2348112    1.177813
          3  |    .374303   .1753685    -2.10   0.036     .1494233    .9376232
             |
         sex |   8.242847   6.418324     2.71   0.007     1.791785    37.92002
    lvef_std |   .5438802   .1290215    -2.57   0.010     .3416472    .8658223
   lvef_std2 |   1.388986   .1863399     2.45   0.014     1.067836    1.806721
       hxhtn |    3.11661   1.572135     2.25   0.024      1.15959    8.376457
         sbp |   1.143139   .0771871     1.98   0.048     1.001438     1.30489
        sbp2 |   .9995246   .0002293    -2.07   0.038     .9990753    .9999742
------------------------------------------------------------------------------

. testparm i.hunt

 ( 1)  2.hunt = 0
 ( 2)  3.hunt = 0
 ( 3)  4.hunt = 0
 ( 4)  5.hunt = 0
           chi2(  4) =   23.87
         Prob > chi2 =    0.0001
```

A complication with clustered data is what to resample. By default, bootstrapping will resample the individual observations. However, the basis of sampling in this example (which is common to clustered-data situations) is subjects. We thus need to resample *subjects* not observations. Fortunately, this can be controlled within Stata by using a `cluster` option on the bootstrap command. The analysis was run using a robust variance estimate and independence working correlation, which improved the stability of the estimates. Table 7.24 gives the portion of the output associated with the Hunt–Hess scores. The bias-corrected bootstrap (using the `ef` option to generate odds ratios) gives a CI for the odds ratio for a Hunt–Hess of 2 compared to 1 of 0.23–7.91. This compares with the interval from 0.35 to 8.01 from Table 7.22 in the original analysis. For comparing a Hunt–Hess score of 5 to that of 1, the bootstrap analysis gives a CI of 14.66–472.09 compared to 12.19–461.00. The results are quite similar and give qualitatively the same results, giving us confidence in our original analysis.

Table 7.23 Effect of Hunt–Hess score on elevated troponin levels adjusting only for `stday`

```
. xtgee CTover1 i.hunt i.stday if stday<4, i(stnum) family(binomial) ef

GEE population-averaged model                 Number of obs      =       434
Group variable:                      stnum    Number of groups   =       168
Link:                                logit    Obs per group: min =         1
Family:                          binomial                   avg =       2.6
Correlation:                 exchangeable                   max =         3
                                              Wald chi2(6)       =     40.75
Scale parameter:                         1    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     CTover1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        hunt |
          2  |   2.136339   1.711752    0.95   0.343     .4442634    10.27306
          3  |   4.312505    2.68268    2.35   0.019     1.274157    14.59609
          4  |    6.41448   4.228072    2.82   0.005     1.762367    23.34676
          5  |   60.09793   44.25148    5.56   0.000     14.19385    254.4595
             |
       stday |
          2  |   .5564922   .1968294   -1.66   0.098     .2782224    1.113079
          3  |   .5170812   .2016593   -1.69   0.091     .2407654    1.110512
------------------------------------------------------------------------------
```

Table 7.24 Bootstrap analysis of adjusted Hunt–Hess model

```
. bootstrap _b, reps(1000) cluster(stnum) seed(2718):xtgee CTo i.hunt i.stday sex
> lvef_std lvef_std2 hxhtn sbp sbp2 if stday <4, i(stnum) family(bin) robust corr(inde)
. estat boot, ef

Bootstrap results                             Number of obs      =       408
                                              Replications       =       921

                              (Replications based on 165 clusters in stnum)
 ( 1)  1b.hunt = 0
 ( 2)  1b.stday = 0
------------------------------------------------------------------------------
             |   Observed               Bootstrap
     CTover1 |     exp(b)       Bias    Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1b.hunt |          1          0           0          .           .  (BC)
      2.hunt |  1.5943809   .2109551   1.4253662    .233916   7.909329  (BC)
      3.hunt |  2.2787955   .1224353   1.8531043    .461805   11.75678  (BC)
      4.hunt |  2.3847466   .0586308   2.4034528    .219347    14.8611  (BC)
      5.hunt |  78.628816   66.77691   99.854523   14.66172   471.0894  (BC)
    1b.stday |          1          0           0          .           .  (BC)
     2.stday |  .57482247  -.0672446   .26831764   .2518986   1.369652  (BC)
     3.stday |  .41069747  -.0548035   .23561377   .1419608   1.170843  (BC)
------------------------------------------------------------------------------
(BC)   bias-corrected confidence interval
Note: one or more parameters could not be estimated in 79 bootstrap
      replicates; standard-error estimates include only complete
      replications.
```

7.10 Power and Sample Size for Repeated Measures Designs

Planning the sample size or calculating power for a repeated measures analysis can be challenging, due to the need to specify the correlation structure (which can be difficult) and because the calculations are different for different types of predictors. We present some results here for the simple situation in which there is a single level of clustering, observations within a cluster are equally correlated and all have the same variability, and the sample size per cluster is the same. This serves as the starting point for many calculations and illustrates some features of power and sample size for repeated measures designs.

An important distinction is whether the sample size calculation is for a *between-* or *within-cluster predictor*. A purely between-cluster predictor is one that may vary between clusters but is constant within a cluster. A within-cluster predictor is one that may vary within a cluster, but whose average is constant across clusters. For example, in a longitudinal study in which the clusters are participants, the participant's race, age at entry to the study, and genetic information are all between-cluster predictors. If every participant was measured at every visit, then visit would be a purely within-cluster predictor. In practice, most predictors that vary within a cluster are not purely within-cluster predictors; their average varies at least somewhat across clusters. Section 7.7.2 shows how to separate a predictor into its purely between and purely within components.

7.10.1 Between-Cluster Predictor

In the situation in which the cluster sample sizes are equal, the analysis of between-cluster predictors are, in essence, based on the cluster level means. This realization also serves to temper the number of between-cluster predictors that can be included in an analysis, because the effective sample size is the number of clusters.

When the data are equally correlated, the variance of a cluster-level mean is given by $\sigma^2[1 + (n - 1)\rho]/n$, where σ^2 is the variability of the outcome, ρ is the within-cluster (intraclass) correlation, and n is the sample size per cluster. In contrast, when the data are independent, the variance would be σ^2/n. That is, the cluster-level mean has a variance that is larger by a factor of $[1 + (n - 1)\rho]$. Since required sample sizes are proportional to the variability of the measurements, the consequence is that sample sizes must be larger by this factor, compared to an experiment using independent data. Because of the central role this factor plays, it has been named the *design effect* and is often abbreviated as DEFF, i.e., $\text{DEFF} = 1 + (n - 1)\rho$. This also gives a convenient way to do sample size calculations. Namely, a calculation is conducted assuming independent data, then it is multiplied by the DEFF to find the required sample size for the repeated measures design.

Here is an illustration of planning a new study, but patterned after Whelan et al. (2004), which was a randomized controlled trial of a decision-making aid (versus

Table 7.25 Sample size and power calculation examples

```
. sampsi 1.7 1.4, sd1(0.5) power(0.8)

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                    and m2 is the mean in population 2
Assumptions:
         alpha =   0.0500  (two-sided)
         power =   0.8000
            m1 =      1.7
            m2 =      1.4
           sd1 =       .5
           sd2 =       .5
         n2/n1 =     1.00

Estimated required sample sizes:
            n1 =       44
            n2 =       44

. sampsi 1.546 1.4, sd1(0.5) n1(130)

Estimated power for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                    and m2 is the mean in population 2
Assumptions:
         alpha =   0.0500  (two-sided)
            m1 =    1.546
            m2 =      1.4
           sd1 =       .5
           sd2 =       .5
sample size n1 =      130
            n2 =      130
         n2/n1 =     1.00

Estimated power:
         power =   0.6533
```

not) for physicians to help them counsel breast cancer patients on surgical options. The outcome is *decisional conflict* and will be assessed using a numerical scale and measures the degree to which patients are well-informed about their choices concerning treatment for breast cancer. This is a repeated measures design because the outcome will be measured at the patient level and there will be multiple patients per physician. The predictor (decision aid or not) is a between-cluster (physician-level) predictor. We use input values from Whelan et al. (2004): an average of about 7.5 patients per physician, standard deviation of the outcome of 0.5 (measured across patients and physicians), and an intraclass correlation, ρ, of 0.3.

We use the corresponding independent samples comparison (a two sample t-test), with a detectable effect size of 0.3 and a desired power of 0.8. Using the `sampsi` command as illustrated in Table 7.25 shows that 44 observations per group would be needed. With 7.5 patients per physician, the design effect is DEFF $= 1+(n-1)\rho =$ $1 + (7.5 - 1)0.3 = 2.95$ and about 2.95(44) or about 130 patients would be needed per treatment group, working out to about 130/7.5 or 18 physicians for each of the two treatment groups.

While the calculation of the required sample size for a between-cluster predictor is not numerically difficult, in the absence of preliminary data, specifying the intraclass correlation coefficient can be an issue. It is sometimes slightly easier to consider the within-cluster variability in the outcome across observations and the variability in the true cluster-level means across clusters. In the decision aid example, this would mean considering the variation in the decisional conflict scale across patients within a physician and the variation in the true physician level means (i.e., the average value if an unlimited number of patients were measured for each physician). The intraclass correlation coefficient can then be calculated (see Sect. 7.1.2) as the ratio of the between-cluster variability and the sum of the between- and within-cluster variances. For the decision aid example, the within-cluster variance in the outcome is $\sigma_e^2 = 0.175$ and the between-cluster variation is $\sigma_u^2 = 0.075$, giving an intraclass correlation coefficient of $0.3 = 0.075/(0.075 + 0.175)$ and an overall variance of $0.25 = 0.075 + 0.175$, corresponding to the standard deviation of $0.5 = \sqrt{0.25}$.

The design effect can be used for power calculations, as opposed to sample size calculations, by reducing the detectable effect size by the square root of the design effect and using power calculations assuming independent data. Continuing the decision aid example, suppose the effect size was 0.25 instead of 0.3 and we have 130 patients per treatment group. How much would the power decrease? The reduced detectable effect size would be $0.25/\sqrt{2.95} = 0.146$. The power calculation for the *t*-test shown in Table 7.25 gives a power of 0.65.

The calculations above have been illustrated for a single predictor and for a numerical outcome, but the general principle extends to the other scenarios described in the book, including different outcome types and the use of multiple predictors. That is, for sample size, preliminary calculations are performed assuming independent data, the result of which is multiplied by the design effect to find the required sample size for the repeated measures design. For power, detectable effect sizes are reduced by dividing by the square root of the design effect and that is used in an independent sample size power calculation to find the power for the repeated measures design. In the binary outcomes outcome case, the design effect can be applied to (5.16) which also would accommodate multiple, correlated, between-cluster predictors through the factor $1/\left(1 - \rho_j^2\right)$.

7.10.2 Within-Cluster Predictor

In the typical case where the correlation within a cluster is positive and for the same sample size and detectable effect size, power for within-cluster predictors is higher than for between-cluster predictors; for the same power and detectable effect size, required sample sizes are smaller. In fact, this is a common rationale in using cluster designs such as longitudinal studies, which are often described as

"using each person as their own control" in order to increase precision. In contrast to between-cluster predictors, the effective sample size for purely within-cluster predictors is the total sample size, not the number of clusters.

As with between-cluster predictors, sample size or power calculations can be obtained by modifying the results from an independent sample size calculation. Again, with ρ being the intraclass correlation coefficient, the sample size can first be calculated assuming independent data and then reduced by the factor $1 - \rho$. Alternatively, if the within-cluster standard deviation is known, this can be used directly to perform a sample size calculation, ignoring the clustered design. Going back to the decision aid example, suppose we are interested in a within-physician predictor, such as the age of the patient, and we divide patients according to whether they are above or below the median value for that physician. A power of 0.8 is desired and the detectable effect size is 0.2. Using the `sampsi` command indicates that 99 observations are needed per group. But this can be reduced by $1 - \rho$ to arrive at a final sample size of 70 per group. That would mean that we would need an overall sample size of 140. Equivalently, we can use the within-cluster standard deviation of $0.418 = \sqrt{0.175}$ to directly perform a sample size calculation (see Exercise 7.12).

For calculating power for a within-cluster predictor, the detectable effect size is *increased* by multiplying it by $1/\sqrt{1-\rho}$ and then an independent sample size calculation is conducted. Or, if the within-cluster standard deviation is available, this is used to directly perform the power calculation, ignoring the clustered nature of the design.

As with the between-cluster calculations, this approach extends to other scenarios covered in this book. That is, for sample size, preliminary calculations are performed assuming independent data, the result of which is reduced by $1 - \rho$ to find the required sample size for the repeated measures design. Again, in the binary outcomes outcome case, the multiplier can be applied to (5.16) which handles multiple, correlated, within-cluster predictors through the factor $1/(1 - \rho_j^2)$. Or the calculations are conducted using the within-cluster standard deviation, which is smaller than the overall standard deviation. For power, detectable effect sizes are increased by multiplying by $1/\sqrt{1-\rho}$ and that is used in an independent sample size power calculation to find the power for the repeated measures design. Or, the within-cluster standard deviation is used directly to calculate the power assuming independent samples.

7.11 Summary

The main message of this chapter has been the importance of incorporating correlation structures into the analysis of clustered, hierarchical, longitudinal, and repeated measures data. Failure to do so can have serious consequences. Two main methods have been presented, GEEs and random effects models.

A primary advantage of the GEEs approach is the availability of the robust variance estimate, which provides valid standard errors without having to explicitly model the nature of the correlations within a cluster. GEEs approaches typically fit models for estimating effects averaged across a population, called marginal models.

In contrast, mixed models incorporate correlation by introducing random effects. This may require more careful modeling and assessment of assumptions, but yield extra capabilities in the form of partitioning the variability, enabling calculation of intraclass correlation coefficients, testing for the presence of clustering, and generating predicted values of random effects. Mixed model approaches typically fit models for estimating effects specific to a cluster (e.g., an individual) and are conditional models.

For simple clustered-data situations, power and sample size calculations can be based on straightforward modifications of the calculations for independent data. These modifications depend on whether the predictor of interest is a between- or within-cluster predictor and require knowledge of the within-cluster correlation (or equivalent quantities).

7.12 Further Notes and References

For those readers desiring more detailed information on longitudinal and repeated measures analyses, there are a number of book length treatments, especially for continuous, approximately normally distributed data. Notable entries include Raudenbush and Bryk (2001), Goldstein (2003), Verbeke and Molenberghs (2000), Diggle et al. (2002), Fitzmaurice et al. (2004), and McCulloch et al. (2008). Unfortunately, many are more technical than this book.

7.12.1 Missing Data

The techniques in this chapter handle unequal sample sizes and unequal spacing of observations in time with aplomb. However, sample sizes are often unequal and observation times unequal because of missing outcome data. And data are often missing for a reason related to the outcome under study. As examples, sicker patients may not show up for follow-up visits, leading to overly optimistic estimates based on the data present. Or those patients staying in the hospital longer may be the sicker ones (with the better-off patients having been discharged). This might lead us to the erroneous conclusion that longer stays in the hospital produce poorer outcomes, so why check-in in the first place?

To a limited extent, the methods in this chapter cover the situation in which the missing data are systematically different from the data available. If the fact that data are missing is related to a factor in the model (i.e., more missing data for males, which is also a factor in the model) then there is little to worry about. However, the

methods described here do *not* cover the situation where the missing data are related to predictors not in the model and can give especially misleading results if the fact that the data are missing is related to the value of the outcome that would have been measured.

See Chap. 11 for much more detail.

7.12.2 Computing

Stata has a wide array of clustered-data techniques. The commands `xtmixed`, `xtmelogit`, and `xtmepoisson` can fit mixed models for multilevel hierarchical data structures. The generalized estimating equations methods are limited to one level of clustering. So, for example, they can explicitly model repeated measures data on patients, but not repeated measures data on patients clustered within doctors. Of course, with sufficient numbers of doctors, even the clustering of patients within doctors could be accommodated with robust standard errors.

Other software packages can also conduct these analyses. For continuous, approximately normally distributed data, SAS Proc MIXED can handle a multitude of models (Littell et al. 1996) and SAS Proc GENMOD can fit models using GEEs and, for binary data, can fit two-level clustered binary data with a technique called alternating logistic regression (Carey et al. 1993). MLWin and HLM are two other clustered data packages with additional capabilities.

7.13 Problems

Problem 7.1. Using the fecal fat data in Table 7.1, calculate the sample variance of the subject averages. Subtract from this the residual variance estimate from Table 7.3 divided by four (why four?) to verify the estimate of σ^2_{subj} given in the text.

Problem 7.2. Using the fecal fat data in Table 7.1, verify the F-tests displayed in Tables 7.2 and 7.3.

Problem 7.3. From your own area of interest, describe a hierarchical dataset including the outcome variable, predictors of interest, the hierarchical levels in the dataset and the level at which each of the predictors is measured. Choose a dataset for which not all of the predictors are measured at the same level of the hierarchy.

Problem 7.4. Could you successfully analyze the data from the fecal fat example using the idea of "analysis at the highest level of the hierarchy?" Briefly say why or why not.

Problem 7.5. For the fecal fat example of Table 7.1, analyze the difference between capsule and coated capsules in two ways. First, use the "derived variable" approach

to perform a paired t-test. Second, in the context of the two-way ANOVA of Table 7.3, test the contrast of coated capsule versus capsule. How do the two analyses compare? What differences do you note? Why do they come about? What are the advantages and disadvantages of each?

Problem 7.6. Consider an example (like the Georgia birthweight example) with before and after measurements on a subject. If the variability of the before and after measurements each have variance σ^2 and correlation ρ then it is a fact that the standard deviation of the difference is $\sigma\sqrt{2(1-\rho)}$.

(1) The correlation of the first and last birthweights is about 0.381. Using Table 7.5, verify the above formula (approximately).
(2) If we were to compare two groups, based on the difference scores or just the last birthweights (say, those with initial age greater than 17 versus those not), which analysis would have a larger variance and hence be less powerful? By how much?

Problem 7.7. The model corresponding to the analysis for Table 7.7 has an intercept, a dummy variable for the fifth birth, a continuous predictor of centered age (age minus the average age), and the product of the dummy variable and centered age.

(1) Write down a model equation.
(2) Verify that the intercept is the average for the first-born, and that the coefficient for the dummy variable is the difference between the two groups, both of these when age is equal to its average.
(3) Verify that the coefficient for the product measures how the change in birthweight from first to last birth depends on age.

Problem 7.8. Reproduce the standard error calculations in Table 7.12, but for the coefficient of `birthorder`. How different are the standard errors when not using the robust option? When using the robust option? Are any of the analyses likely to give misleading results? If so, which ones?

Problem 7.9. Verify the calculation of the predicted values and residuals in Sect. 7.7.3.

Problem 7.10. Using the CIs for the odds ratios for the Hunt–Hess scores in Tables 7.22 and 7.21, calculate the confidence intervals for the log-odds ratios. Show that the width of the CIs in the adjusted analysis (Table 7.22) are not appreciably larger than those in the unadjusted analysis (Table 7.21).

Problem 7.11. Compare the bootstrap-based CI for the comparison of study day 1 and study day 2 from Table 7.24 to the CI from the original analysis reported in Table 7.22. Do they agree substantively? Do they lead to different conclusions?

Problem 7.12. Verify that a two independent sample t-test sample size calculation with a standard deviation of 0.5 when reduced by the factor $1-\rho = 1-0.3 = 0.7$

gives virtually the same answer as a direct calculation using the standard deviation of 0.418.

7.14 Learning Objectives

(1) Recognize a hierarchical data situation and explain the consequences of ignoring it.
(2) Decide when hierarchical models are necessary versus when simpler analyses will suffice.
(3) Define the terms hierarchical, repeated measures, clustered, longitudinal, robust variance estimator, working correlation structure, generalized estimating equations, fixed factor, and random factor.
(4) Interpret Stata output for GEE and random effects analyses in hierarchical analyses for linear regression or logistic regression problems.
(5) Explain the difference between marginal and conditional models.
(6) Decide if factors should be treated as fixed or random.
(7) Explain the use of shrinkage estimators and best prediction for random factors.
(8) Perform power or sample size calculations for simple clustered-data situations.