Chapter 5

Longitudinal Data Models

In this chapter, we turn to a broader set of problems where the an outcome for the subjects in a study are repeatedly measured over time, covariates are also measured repeatedly (and some of them might change in time) and the questions of interest involve associations of these outcomes with covariates. We discuss methods that are natural extensions of existing regression procedures (linear, logistic, poisson) that attempt to account for the longitudinal (repeated measures) structure of the data.

It is important at the outset to remind ourselves of the differences between cross–sectional and longitudinal studies. Fundamentally, a cross-sectional study entails only a single observation for each sampled individual. Such observations may all occur at the same time point t or at different time points across individuals. On the other hand, longitudinal studies involve study designs where the value of the outcome and covariates for a typical sampled individual is measured at several different time points. While we do not preclude the possibility that a few individuals may only be observed once, we assume that the bulk of the sample has multiple observations. Since observations on a single individual at different times are very unlikely to be independent we need new methods of analysis that account for such correlation in the data and, in fact, in many situations take advantage of this structure. In this chapter, we introduce three model strategies that are commonly used to analyze longitudinal data: *marginal* models, *mixed effects* models, and *transition* models. The next section gives a brief qualitative introduction to these different approaches. We then focus on application of the models, focusing on marginal models in Chapter 6 and mixed effects models in Chapter 7. Transition models are less commonly used in epidemiology; we discuss them briefly at the end of this chapter. The strategies overlap in their goals to a great extent but the distinction between them and their interpretation are subtle but important.

As we proceed with our discussion of methods we pay close attention to connections between cross-sectional and longitudinal approaches to the same longitudinal data examples.

5.1 Marginal Models

Marginal models are arguably the simplest kind of longitudinal regression model to interpret because they most resemble those used typically with cross-sectional data. In fact, most cross-sectional studies are designed to investigate marginal effects, and therefore can only estimate such effects. We first pose the question: what is a marginal model?

A marginal model attempts to describe variation in population means of subgroups, averaged over all individuals. Thus, such models are sometimes referred to as *population-averaged* models. For example, suppose a study addresses whether and by how much cigarette smoking influences serum cholesterol, based on longitudinal observations on both variables on a random sample of individuals. If smoking is measured by a binary covariate ($X = 1, 0$ represents smokers and non-smokers, respectively), and the outcome cholesterol is denoted by Y, then we can examine the two means $E(Y_{ij}|X_{ij} = 1)$ and $E(Y_{ij}|X_{ij} = 0)$ to determine the difference in mean cholesterol between smokers and non-smokers. Alternatively, $E(Y_{ij}|X_{ij} = 1)$ is the mean of a cholesterol measurement taken on a randomly drawn smoker form the population. Writing this in regression language

$$E(Y_{ij}|X_{ij} = x_{ij}) = b_0 + b_1 x_{ij}, \tag{5.1}$$

where $b_0 = E(Y_{ij}|X_{ij} = 0)$ and $b_1 = E(Y_{ij}|X_{ij} = 1) - E(Y_{ij}|X_{ij} = 0)$. The slope coefficient b_1 associated with the smoking covariate X is thus interpreted as the difference in population mean cholesterol, comparing smokers and non-smokers. Note that a marginal model and its interpretation does not depend in any way on whether the covariate is time-independent or time-dependent. Thus, when no individual ever changes their smoking habit longitudinally (X is time-independent), the model (5.1) involves comparing the cholesterol readings for two distinct groups of individuals, smokers and non-smokers. On the other hand, in a smoking cessation study where every individual is a smoker for the initial reading and a non-smoker for the second measurement (time-dependent), say, the model (5.1) involves comparing the baseline cholesterol measurements for all individuals with the "after" cessation measurements again for all the same individuals. (In practice, the data may reflect a mixture of both with some individuals never altering their smoking and some indicating change.) It may make you uneasy that b_0 means the same in both these extreme cases, arguing that cholesterol differences between lifetime smokers and non-smokers may

be quite different from differences for the same individuals before and after quitting smoking. However, this is not an issue with the use and interpretation of a marginal model, but, in this case, raises the question of whether (5.1) is the appropriate marginal model to use. Referring back to the discussion of Chapter 3, we may prefer the marginal model

$$E(Y_{ij}|X_{ij} = x_{ij}) = b_0 + (b_1)_{CS}x_{ij} + (b_1)_L(x_{ij} - x_{i1}), \tag{5.2}$$

particularly when the covariate X—here, smoking—displays substantial longitudinal variation. The point is, however, that again the slope coefficients, $(b_1)_{CS}$ and $(b_1)_L$, have marginal interpretations. For example, $(b_1)_L$ reflects the comparison of cholesterol observations for baseline non-smokers who smoke at any time (after the first observation, i.e. $j > 1$) so that $x_{ij} - x_{i1} = 1$, with individuals who don't change their smoking habits so that $x_{ij} - x_{i1} = 0$, and also those baseline smokers who subsequently cease smoking so that $x_{ij} - x_{i1} = -1$; in this comparison multiple such observations on the same individual play the same role as observations on different individuals.

Marginal models can be similarly applied to longitudinal binary outcome data where the logistic regression model is often used to study patterns in the odds of an outcome under various levels of a set of risk factors. The data of Chapter 1.3.3 concerns the influence of drug use on teenage sexual activity. The latter variable is binary, reflecting simply whether a respondent reported sexual activity in the past 24 hours ($Y_{ij} = 1$) or not ($Y_{ij} = 0$). With cross-sectional data—one observation per respondent—we might use a logistic regression model to link p_{ij} to explanatory factors including drug use in the past 24 hours (X_{ij}). Extending this approach longitudinally—multiple observations per respondent—produces a simple *marginal* logistic regression model:

$$log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = b_0 + b_1x_{ij}, \tag{5.3}$$

where $p_{ij} = P(Y_{ij} = 1 \mid X_{ij} = x_{ij})$.The term *marginal* again refers to the appropriate interpretation of the regression coefficients in (5.3). First, the intercept, b_0, corresponds to the log odds of sexual activity in a given day, amongst all individual days for whom $x_{ij} = 0$; that is, b_0 is the log odds of sexual activity across individual days with no reported drug use, whether such days are on the same or different individuals. The slope coefficient, b_1, is interpreted as the log odds ratio associated with a unit increase in x_{ij}—that is, comparing the probability of sexual activity on days with drug use to days without. As with the continuous model, the marginal model and interpretation applies if the data includes only individuals who always exhibit the same drug/alcohol use for every observation day, so that b_1 refers to comparing incidence of sexual activity across two distinct groups of teenagers, the users and non-users; and, in cases where individuals always change their use pattern so that we are comparing incidence of sexual activity in the same teenagers but under two

different scenarios–the days they use drugs/alcohol and the days they don't. Of course, in many cases, the data reflects a mixture of these two types; the point being that coefficients of a marginal model like (5.3) have the same interpretation under very different longitudinal situations for the risk factors.

In either case–continuous or binary outcomes–or for longitudinal count data for that matter (where a marginal Poisson model may apply), the parameters of a marginal model carry the same meaning as they would in analogous models for cross-sectional—non-longitudinal—data. Note, however, that marginal models can still exploit longitudinal data to study longitudinal effects, like $(b_1)_L$ in (5.2), that cannot be tackled with cross-sectional information. Further, at this stage, the marginal models do not describe the correlation or covariance structure of longitudinal observations, only properties of the (marginal) means of outcome variables. We will certainly have to address this structure in estimation of regression coefficients—specifically their sampling variance as already introduced in Chapter 3–but correlation patterns are not explicitly invoked by any marginal model. In this sense, marginal models do not attempt to explain or model correlation among the repeated observations for an individual, at least not directly. As we shall see, mixed effects and transition models pursue a somewhat more aggressive strategy by trying to directly model correlations directly, at the same time as the means of our outcome of interest.

5.2 Mixed Effects Models

Marginal models describe the relationship of the mean of the outcome Y_{ij} and risk factors X_{ij} across all observations simultaneously, that is, without immediate regard as to whether the observations come from a common individual or note. *Mixed effects models* focus initially on the regression relationship *restricted to observations on a single individual.* The model is then extended to multiple individuals by allowing some pieces of the model to vary from individual to individual in a proscribed manner, while other components remain the same. To use the common jargon, then , constructing a mixed effects model focuses on the introduction of *random effects*—these are the pieces of the model that vary across individuals—in addition to *fixed* effects of cofactors of interest, the relationships that are assumed identical for every subject. As we shall see, this approach indirectly describes and interprets the covariance structure for longitudinal observations. In fact, this is the *fundamental* difference in approach in using a mixed effects as compared to a marginal model.

As noted, a mixed model usually includes two types of effects, *fixed* and *random.* Fixed effects describe the impact of known measured covariates, such as time, sex, weight, etc,

where such effects are assumed to hold over a broad population of individuals. On the other hand, a random effect measures the impact of known variables where effects are assumed to vary across individuals in the population. Understanding these two types of effects is best illustrated by considering some simple examples.

Consider first the relationship between CD4 cell count (Y) and viral load (X) where interest focuses on how changes in viral load affect the immunological system as measured by CD4 cell count. Suppose, for the i^{th} individual, we describe this association through the following linear model

$$E(Y_{ij}|X_{ij} = x_{ij}) = b_{0i} + b_1 x_{ij},$$

or, alternatively,

$$Y_{ij} = b_{0i} + b_1 X_{ij} + e_{ij}, \tag{5.4}$$

where the error terms, e_{ij}, are assumed independent and identically distributed in particular independent of the explanatory variable, X_{ij}. This mixed effects model includes (i) the random effect b_{0i}, in this case a random intercept, that is assumed to vary from individual to individual, and (ii) the fixed effect, b_1, which quantifies the association between viral load and CD4 cell count, and is assumed to be the same for all individuals. Note that (5.4) is a model for the i^{th} person only. To fully specify the model for the population, we must now introduce assumptions which link together the random effects across individuals. For the model given in (5.4) we add the assumption that the random effect terms, b_{0i}, are themselves independently and identically distributed across the population, according to a *known* distribution, often taken to be a Normal distribution with mean b_0 and variance τ^2. Further we assume that the random effect b_{0i} is independent of the error terms e_{ij} for all i and j. That is, the error around the regression line does not depend on the inherent level of response for the i^{th} individual. Note that the interpretation of the fixed effects term b_1 is now specific to the i^{th} person; that is, b_1 is the change in the mean of CD4 count associated with a unit increase in viral load for the i^{th} individual.

It is important to recognize that the model (5.4), together with the assumptions about e_{ij} and b_{0i}, automatically determines a correlation structure for the longitudinal responses Y_{ij}. In fact, with this simple mixed effects model, we have already examined this structure in detail in Chapter 3 where we showed that the correlation between any two observations on the same individual is just $\rho = \frac{{\sigma_b}^2}{{\sigma_b}^2 + {\sigma_e}^2}$.

As we have noted, the model (5.4) assumes that the effect of viral load on CD4 account is the same for everyone in the population. It is also straightforward to expand (5.4) to allow for a random rather than fixed effect for viral load. In particular, consider the model

$$E(Y_{ij}|X_{ij} = x_{ij}, b_{0i}, b_{1i}) = b_{0i} + b_{1i} x_{ij}.$$

Now we assume that the regression association between viral load and CD4 count—as measured by the coefficient b_{1i}—differs from person to person.

For our second example, we consider a binary outcome, using the example on drug/alcohol use and sexual activity for illustration. The analogue of the simple mixed effects model (5.4) for binary outcome data is the following logistic regression model linking drug/alcohol use to sexual activity on a given day:

$$\text{logit}\{\Pr(Y_{ij}=1|b_{0i},X_{ij}=x_{ij})\}=\log\left\{\frac{\Pr(Y_{ij}=1|b_{0i},X_{ij}=x_{ij})}{\Pr(Y_{ij}=0|b_{0i},X_{ij}=x_{ij})}\right\}=b_0+b_{0i}+b_1x_{ij}. \quad (5.5)$$

Similar assumptions on the longitudinal observations are usually invoked, namely that repeated observations on the i^{th} teenager are independent, given b_{0i} and X_{ij}. Furthermore, we once more link the logistic regressions (5.5) for each individual via an assumption regarding the random intercept terms b_{0i} by demanding that these terms are independently and identically distributed across the teenagers according to a known distribution, for example, a Normal distribution with mean 0 and variance ${\sigma_b}^2$.

As with the continuous model, the interpretations of the coefficients in (5.5) are subject-specific to a specific teenager. In particular, b_0+b_{0i} is the log odds of sexual activity on any given drug/alcohol free day for the i^{th} teenager. The model allows this baseline measure of sexual activity to vary across teenagers, thereby allowing a different inherent rate of sexual activity for each person. This is the random effect of the model (5.5). On the other hand, the fixed effect term b_1 describes the log odds ratio comparing the frequency of sexual activity between days with and without drug/alcohol use *for the* i^{th} *teenager*. While this odds ratio is assumed the same for every teenager, this does not mean that it can be directly interpreted marginally, that is comparing days with no drug/alcohol use with days with drug/alcohol use for all teenagers. We discuss this important distinction further below.

5.2.1 Connection between Marginal and Mixed Effects Models

A natural question now arises regarding the connection between marginal and mixed effects models for the same data for example, consider the simple linear mixed effects model (5.4) for continuous outcome Y_{ij}. This model is subject-specific as we previously indicated; now, we average over all subjects to obtain a marginal model. Specifically, this is achieved by taking the expectation, or average, of (5.4) over the distribution of b_{0i}. Since the expectation of sums of random variables is just the sum of the expectations, and $E(b_{0i})=0$, it follows that (5.4) yields

$$E(Y_{ij}|X_{ij}=x_{ij})=b_0+b_1x_{ij}. \quad (5.6)$$

The key point here is that this marginal model has exactly the same slope coefficient b_1 as the subject-specific model (5.4). Thus, when using a mixed effects model like (5.4) we can interpret the regression coefficient directly specific to a given subject *or* in terms of a population averaged effect in the marginal model (5.1). In other words, in linear models, the introduction of random effects does not alter the interpretation of fixed affects. Give example.

In summary, with linear models for continuous outcomes, the primary difference between the mixed effects and marginal approaches is that the former specifically models the correlation structure between longitudinal Y_{ij} whereas the latter model leaves this unspecified.

In contrast to the linear model, the use of mixed effects in *some* generalized linear models specifically modifies the interpretation associated with fixed effects. For example, consider the following two logistic regression models, modeling the association of disease ($Y_{ij} = 1$ means the patient has the disease, 0 otherwise) with ($b_0 = 0, b_1 = 4$) and X_{ij} a binary exposure ($= 1$ then Exposed, $= 0$ then Not $\bar{E}$xposed)

$$\begin{aligned}\text{logit}\{\Pr(Y_{ij} = 1|b_{0i} = -1.386, X_{ij} = x_{ij})\} &= b_0 - 1.386 + b_1 x_{ij} \\ \text{logit}\{\Pr(Y_{ij} = 1|b_{0i} = 0, X_{ij} = x_{ij})\} &= b_0 + 0 + b_1 x_{ij},\end{aligned}$$

where this represents logistic models relating the effect of exposure for two different kinds of people: one with the underlying rate of the disease when not exposed of $1/(1+exp(-(-1.386))) = 0.20$ and one with $1/(1 + exp(-(0))) = 0.50$. Note, both have an odds ratio of 4, which is the ratio of odds of the disease comparing the person when they are exposed to when they are not exposed, that is, b_1 representing the individual impact of the exposure and disease. In this simple case, we can straightforwardly compare the marginal effect (average effect) with the individual effect. Let's assume these are the only two kinds of people in the world and we want the marginal odds ratio. Then, we can simply take a random sample of people/times and measure the corresponding outcomes and exposures - the marginal OR is simply the one from a pooled 2x2 table. Table 5.2.1 has an example of just such a scenario; the first two tables represent the data collected from repeated observations of individual 1 (with intercept $b_{0i} = -1.386$) and the second table from similar data on individual 2 ((with intercept $b_{0i} = 0$). As one can see, the random effects OR (those for each individual) are 4.0, whereas the marginal OR (from pooled table) is 3.45. Neither of these OR's is either right or wrong, they are simply different estimates of the association of Y and X; the random effects model gives the individual effect, the marginal OR compares different groups (the subjects with and without exposure).

Table 5.1: Example Showing The Difference Between Subject-Specific and Marginal Odds Ratios

$b_{0i} = -1.386$				$b_{0i} = 0.000$				Pooled			
	D	$\bar{D}$			D	$\bar{D}$			D	$\bar{D}$	
E	50	50	100	E	80	20	100	E	130	70	200
$\bar{E}$	20	80	100	$\bar{E}$	50	50	100	$\bar{E}$	70	130	200
$OR = 4.00$				$OR = 4.00$				$OR = 3.45$			

Figure 5.1: Difference Between Subject-Specific Logistic Regressions and Marginal Version: the individual thin lines represent curves for different individuals - the thick black line represents the marginal probability of disease, averaged over the individual curves

figure51.pdf

As additional illustration consider the case where exposure X_{ij} is continuous, and we now have a whole sample of different kinds of people, each with their own underlying (baseline) rate of disease. Figure 5.1 plots the probability of disease for a random subset of people, each with their own individual intercept, b_{0i}, versus the exposure, X_{ij}: $P(Y_{ij} = 1 \mid X_{ij} = x_{ij}, b_{0i})$. The thick line is the average probability of disease at each level of exposure or $P(Y_{ij} = 1 \mid X_{ij} = x_{ij}) = E_{b_{0i}}[P(Y_{ij} = 1 \mid X_{ij} = x_{ij}, b_{0i})]$. This last equation simply means one derives the average probability at each x_{ij} by averaging the probabilities among all individuals in the population. The result is an essential characteristic of random effects logistic regression models versus marginal models: the effect size is "smaller" (closer to the null) in marginal versus random effects models. One can see this in the figure as a shallower slope of the thick line (marginal) versus the thinner (individual) lines. This is in contrast to a linear or log-linear model where both the marginal and random effects estimates of the association are equivalent. For instance, consider a log-linear models for two individuals:

$$\log\{\Pr(Y_{ij} = 1 | b_{0i} = -2.996, X_{ij} = x_{ij})\} = b_0 - 2.996 + b_1 x_{ij}$$
$$\log\{\Pr(Y_{ij} = 1 | b_{0i} = -1.609, X_{ij} = x_{ij})\} = b_0 - 1.609 + b_1 x_{ij}.$$

The corresponding theoretical 2x2 tables from data on repeated measures on both these individuals (table 5.2.1) demonstrates that, opposed to the logistic regression example, the random effects and marginal relative risks (RR) are the same.

Table 5.2: Example Showing No Difference Between Subject-Specific and Marginal Relative Risks for relative risk – or log-linear model

$b_{0i} = -2.996$

	D	$\bar{D}$	
E	20	80	100
$\bar{E}$	20	80	100

$RR = 4.00$

$b_{0i} = -1.609$

	D	$\bar{D}$	
E	20	80	100
$\bar{E}$	5	95	100

$RR = 4.00$

Pooled

	D	$\bar{D}$	
E	100	100	200
$\bar{E}$	25	175	200

$RR = 4.00$

5.2.2 Motivation for Mixed Effects Models

We summarize the discussion of mixed effects models by focusing on several appealing features of this approach. First, mixed effects models directly incorporate natural variability in (random) regression coefficients that are allowed to vary from individual to individual. Consequently, at least with continuous outcomes, this leads to simple ways of partitioning the variability in response information, thereby providing useful insight into the various sources of variation. In particular, model estimates can be used to describe the assumed probability distribution of random coefficients. As a side benefit of such model description of variability, mixed effects models automatically proscribe natural covariance or correlation structures for longitudinal observations, and as such provide putative explanations for such correlation properties that are exhibited by the data. Finally, mixed effects models are most useful when we want to make inference about individuals, for example, the link between a risk factor and outcome, rather than comparing population averages. One drawback is that some methods for fitting mixed effects models require detailed assumptions about the distributions of random effects which may be hard to verify with existing or external data.

5.3 Transition Models

For time-sequenced repeated measures, transition are defined by using previous outcomes as predictors of future outcomes, putting into practice the adage that the best predictor of the future is the past. Technically, transition (sometimes called Markov) models build the joint distribution of the data by specifying a sequence of distributions that are conditioned

on previous measurements on the individual. Consider the study of teenage sex (Chapter 1.3.3), where Y_{ij} is an indicator of whether (=1) or not (=0) a teenager, i has sex at day j and X_{ij} is a similar indicator of whether or not the same teenager on the same day either took illegal drugs or consumed alcohol. A possible transition model:

$$E(Y_{ij}|X_{ij} = x_{ij}, Y_{i(j-1)}, Y_{i(j-2)}, \ldots, Y_{i1}) = b_0 + b_1 x_{ij} + \eta Y_{i(j-1)}, \tag{5.7}$$

where Y_{i1} is outcome at time t_{i1}, Y_{i2} at t_{i2}, ..., and $t_{i1} < t_{i2} < \cdots < t_{in_i}$. Thus, this model accounts for the correlation of subsequent measurements on the same subject using a relatively simple model that the observations are conditionally independent of one another given only the previous time. For instance, consider the case of no covariate, X, then a simple (equivalent) transition model will imply that one can write down the joint distribution of the data on one individual as:

$$Pr(Y_{i1}, Y_{i2}, \ldots, Y_{in_i}) = Pr(Y_{i1})Pr(Y_{i2} \mid Y_{i1}) \cdots Pr(Y_{in_i} \mid Y_{i(n_i-1)}).$$

The most important feature of models like (5.7) is the interpretation of the coefficients; in this case b_1 will be the log(OR) of having sex, comparing subjects who do versus do not engage in drug/alcohol consumption, keeping the sexual activity the previous day fixed. η is the odds ratio of sex comparing subjects who did versus did not have sex the preceding day, keeping drug/alcohol status on the same day fixed. It only takes a little imagination to extend (5.7) to include variables that describe a longer history of either the outcome (e.g., the number of days having sex in the past week) or drug/alcohol use (e.g., the number of days in the last week engaging in activities to which Nancy Reagan would not approve).