Chapter 6

Estimation of Marginal Models

The last chapter introduced several regression approaches longitudinal data. In this chapter we consider in more detail estimation of marginal models based on longitudinal data. The mixed models approach (discussed in the next chapter) is based on the more classical approach of 1) defining an explicit probability model of the data, 2) defining the resulting likelihood of the observed data and 3) maximizing this likelihood with respect to the parameters of interest. This chapter deals with estimators of parameters (coefficients) using a different approach and philosophy than the mixed model approach outlined in the next chapter. Although we will not dive into much technical detail, the approach highlighted here starts with the parameters of interest and derives estimates (estimating equations) of these that are not dependent on knowing the entire probability distribution of the data. It is a much more "targeted" approach, in that one does not fret much over aspects of the data-generating distribution that are not essential to estimating the coefficients of interest (e.g., it can be unnecessary to assert the normality of the data to estimate certain parameters of the data-generating distribution). In addition, the inference that is naturally available from maximums likelihood estimation is not available for these estimators. However, there are methods for deriving standard errors that also are more robust than likelihood estimates - again, they do not rely on correctly specifying the probability distribution of the data. In the end, the approaches discussed in this chapter have gained larger and larger acceptance as users have realized that they are widely applicable, use the structure and models for regression procedures that are familiar (e.g., linear, logistic) and provide robust inference (standard errors) that are not sensitive to underlying models of the data generating mechanism.

6.1 Generalized Estimating Equations

6.1.1 Least Squares and Iteratively Reweighted Least Squares for Linear, Poisson and Logistic Regression Models

Consider the simple marginal regression model (6.1)–repeated here for convenience–that links continuous outcomes Y_{ij} to a single explanatory risk factor X_{ij}:

$$E(Y_{ij}|X_{ij} = x_{ij}) = b_0 + b_1 x_{ij}, \tag{6.1}$$

The sum of squared deviations between the observed outcomes Y_{ij} and the 'predicted' outcomes $b_0 + b_1 x_{ij}$, given by

$$\text{Sum of Squared Deviations} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(y_{ij} - b_0 - b_1 x_{ij}\right)^2, \tag{6.2}$$

yields a familiar measure of the extent to which a particular choice for the regression coefficients, b_0 and b_1, provides a good fit to the data. In matrix form, the sum of squared deviations (6.2) can be written as

$$\text{Sum of Squared Deviations} = (\mathbf{Y} - \mathbf{Xb})^{\mathbf{T}}(\mathbf{Y} - \mathbf{Xb}). \tag{6.3}$$

Estimation of the coefficients b_0 and b_1 then proceeds by choosing those values that optimize the fit of the model to the data, or, equivalently, by minimizing the deviation of the model from the data as quantified by (6.2) or (6.3). Calculus, or direct optimization, shows that such coefficient estimates satisfy

$$\mathbf{X^T}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}) = \mathbf{0}, \tag{6.4}$$

using the matrix notation of Chapter 1.2. Note that both sides of (6.2) are 2×1 matrices reflecting two *estimating equations*, one for b_0 and one for b_1. For example, this least squares estimating equation for b_1 is just

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij} \left(y_{ij} - b_0 - b_1 x_{ij}\right) = 0. \tag{6.5}$$

In this simple linear model, Equation (6.3) can be solved easily to give the explicit form of the coefficient estimators as

$$\hat{\mathbf{b}} = (\mathbf{X^T X})^{-1} \mathbf{X^T Y}. \tag{6.6}$$

With a single explanatory variable as in (6.1), the terms $\hat{\mathbf{b}}, \mathbf{X}$ and $\mathbf{Y}$ have dimension $2 \times 1, N \times 2$ and $N \times 1$, respectively, so that (6.6) yields two equations for the two estimates $\hat{b_0}$ and $\hat{b_1}$. However, equations (6.2) and (6.3) work equally well for the general case where p independent variables occur in the model for $E(Y_{ij})$. This, of course, repeats our description of least squares given in Chapter 3.1.

This approach treats each observation Y_{ij} "equally", with the tacit assumptions, first that the response variables all share exactly the same variance ("homoscedasticity"), and, of more relevance to the discussion here, that the responses are independent (having controlled for the predictors in our model). The second of these assumptions is clearly violated by the longitudinal structure here, where there is almost always some correlation between longitudinal observations on the same individual. Standard regression techniques attack violations of homoscedasticity by weighting observations according to the level of variability; that is, responses that are measured precisely—have low variance–are weighted more heavily than those that are more imprecise—have high variance. As we shall see below, a generalization of this idea of weighting can also be used to exploit the correlation among responses. Before we discuss this further, we first want to expand our discussion of least squares to include more general regression models than (6.1), particularly to include , models for count and binary outcomes, Poisson and logistic regression, respectively. In fact, the primary difference between Poisson and logistic regression methods and the least squares approach described for linear models is that in the former two cases, the variance of the response varies depending on the mean and so cannot be constant across all observations. Thus weighting is needed in these cases, even before we turn to the issue of accommodating longitudinal correlation.

6.1.2 Estimation in Marginal Poisson and Logistic Regression Models

Consider first the simple marginal Poisson regression model akin to (6.1)

$$\log E(Y_{ij}|X_{ij} = x_{ij}) = b_0 + b_1 x_{ij}, \tag{6.7}$$

reflecting the (mean) assumption that the log of the mean number of counts (response) varies linearly with the explanatory variable X. This suggests considering the deviation measure $\sum_{i=1}^{m} \sum_{j=1}^{n_i} \{Y_{ij} - \exp(b_0 - b_1 x_{ij})\}^2$ instead of (6.2). Minimizing this measure of deviation of the data from the model again yields the relevant estimating equations for b_0 and b_1, analogous to (6.3). For example, the estimating equation for b_1 is now given by

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij} \exp(b_0 - b_1 x_{ij}) \left(Y_{ij} - \exp(b_0 - b_1 x_{ij})\right) = 0, \tag{6.8}$$

as compared to (6.5).

However, we now recognize that variability of the response Y_{ij} is likely to vary with the level of the count (which, as discussed in chapter 4 is a property of both the Poisson and negative binomial distribution) and so the individual terms in $\mathbf{X^T}(\mathbf{Y} - \exp \mathbf{X}\hat{\mathbf{b}}) = \mathbf{0}$ need to be weighted to accommodate this. An appropriate choice of weights is to take them to be inversely proportional to the variance of the relevant observation, i.e Y_{ij}. This variance, and hence the weighting, can based on the property of the Poisson distribution for counts, namely that $Var(Y_{ij}) = E(Y_{ij}) = \exp(b_0 - b_1 x_{ij})$. This now leads to adjusted estimating equations, for example with (6.10), for b_1, being altered to yield

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{x_{ij}\exp(b_0 - b_1 x_{ij})\left(Y_{ij} - \exp(b_0 - b_1 x_{ij})\right)}{\exp(b_0 - b_1 x_{ij})} = 0 \tag{6.9}$$

, which immediately simplifies to

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}\left(Y_{ij} - \exp(b_0 - b_1 x_{ij})\right) = 0 \tag{6.10}$$

A similar equation provides the estimator for b_0.

These Poisson regression estimating equations have immediate extensions to cover the case where there are p explanatory variables rather than just one. Unfortunately, the estimating equations, such as (6.10), do not yield an explicit solution for b_1 (and (b_0) as happened with the linear model—see (6.6). Nonetheless it is still easy to solve these equations iteratively to find $\hat{b}_0$ and $\hat{b}_1$. These estimators are therefore often referred to as *iteratively reweighted least squares estimators*. Although we do not pursue this further here, the estimators are, in fact, identical to the maximum likelihood estimators discussed in Chapter 4, based on the assumption that the outcomes Y_{ij} actually follow a Poisson distribution—note that the estimators derived here only used the variance property of a Poisson and thus apply more generally in principle.

A similar development exists for binary responses and logistic regression models. With Y_{ij} a binary outcome, the simple marginal logistic regression model akin to (6.1) and (6.7) is just

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = b_0 + b_1 x_{ij}, \tag{6.11}$$

where $p_{ij} = Pr(Y_{ij} = 1|X_{ij} = x_{ij}) = E(Y_{ij}|X_{ij} = x_{ij})$. An identical strategy that we used for Poisson regression leads to iteratively reweighted least squares estimates for b_0 and b_1, where the weights are this time based on the Bernoulli variance formula $Var(Y_{ij}|X_{ij} = x_{ij}) = p_{ij}(1 - p_{ij})$. Again, the estimators derived in this way are nothing more than the maximum likelihood estimators for logistic regression coefficients (Jewell, 2004).

6.1.3 Properties of Least Squares and Iteratively Re-weighted Least Squares Estimators with Longitudinal Data

As previously discussed in Chapter 3, the least squares estimators of the model coefficients—given by (6.6)–are unbiased even when there is correlation between longitudinal observations, assuming, of course, that the linear model (6.1) is correct. This property extends to the iteratively re-weighted least squares (i.e maximum likelihood) estimators although, technically, the unbiasedness is only true as the sample size grows large.

The important message here is that standard estimators—derived for cross-sectional data— apply equally well for longitudinal data ignoring the longitudinal structure. However, as we already noted in Chapter 3, the analogous standard variance estimates associated with the coefficient estimators can be substantially wrong when the longitudinal correlation is ignored. We thus need a method to calculate, $Var(\hat{b}_1)$, for example, that allows for the possibility of correlation of longitudinal observations for the same individual. Later, in Section 6.x, we also discuss how to exploit such longitudinal correlation to improve the precision of the least squares and iteratively re-weighted least squares estimators.

6.2 Robust Variance Estimates

To develop our intuition we focus first on the linear model (6.1) where we have already seen that the ordinary least squares estimator can be given, in matrix form, by (6.6). Note that this estimator for the regression coefficients is actually just a constant matrix (assuming the independent variables are held fixed) times the random outcome variable $\mathbf{Y}$. We can thus apply one of the simple rules of Chapter 1 to evaluate the variance-covariance matrix of the estimator $\hat{\mathbf{b}}$ in terms of that of $\mathbf{Y}$. That is

$$Var(\hat{\mathbf{b}}) = (\mathbf{X^T X})^{-1}\mathbf{X^T V X}(\mathbf{X^T X})^{-1}, \tag{6.12}$$

where $\mathbf{V}$ is just the variance-covariance matrix of $\mathbf{Y}$. This evaluation did not depend at all on the form of $\mathbf{Y}$, and so is valid in complete generality. Note that naive application of variance estimators for cross-sectional ordinary least squares necessarily assume that $\mathbf{V} = \mathbf{I}$ which simplifies (6.12) to

$$Var(\hat{\mathbf{b}}) = (\mathbf{X^T X})^{-1}\mathbf{X^T I X}(\mathbf{X^T X})^{-1} = (\mathbf{X^T X})^{-1}, \tag{6.13}$$

the standard variance-covariance matrix for coefficient estimators based on the linear model.

We are thus almost at the point where we can obtain an estimator for the true variability of $\hat{\mathbf{b}}$ in the longitudinal setting. We know everything on the right hand side of () except V,

the variance-covariance matrix of $\mathbf{Y}$. If we knew the mean of $\mathbf{Y}$ we could simply subtract it from the observed realization of $\mathbf{Y}$ and calculate empirical variances and covariances in the standard way. But, given $\hat{\mathbf{b}}$, we almost know the mean, that is we have an estimate of it via the assumed model (6.1). This provides us a way to estimate the variance-covariance matrix of $\mathbf{Y}$, assuming the regression model is correct. We look at the details of this before we go further. Given $\hat{\mathbf{b}}$, we calculate the deviation of the observed values of $\mathbf{Y}$ from the estimated mean of $\mathbf{Y}$, the so-called *residuals*

$$r_{ij} = Y_{ij} - \hat{b}_0 - \hat{b}_1 x_{ij}, \tag{6.14}$$

or

$$\mathbf{r} = \mathbf{Y} - \mathbf{xb}, \tag{6.15}$$

in matrix format. We now take advantage of the assumption of independence of subjects which leads to the block-diagonal structure for $\mathbf{V}$, meaning that we can separately estimate each $\mathbf{V_i}$ using the residuals for the i^{th} individuals extracted as follows:

$$\hat{\mathbf{V}}_i = (\mathbf{Y_i} - \mathbf{x_i b})(\mathbf{Y_i} - \mathbf{x_i b})^{\mathbf{T}} \equiv \mathbf{r_i r_i}^{\mathbf{T}}. \tag{6.16}$$

6.3 Using a Working Correlation Structure to Improve Estimation

As we noted above, standard estimating equation estimators, based on the assumption of independent longitudinal observations, may not provide the best use of the data even though they provide consistent estimators of the regression coefficients. This is because they fail to take account and therefore exploit the variance and correlation structure of the longitudinal observations. This suggests the possibility of introducing better estimators that take advantage of an assumed correlation structure that we believe plausibly describes the longitudinal correlation structure.

We first describe the form of weighted least squares estimators of the regression coefficients $\mathbf{b}$. The weight matrix we consider are $n \times n$ block-diagonal matrices where each block matrix $\mathbf{W_i}$ is symmetric. The weighted squared deviations corresponding to (8.2) and (8.3) is then simply

$$(\mathbf{Y} - \mathbf{Xb})^{\mathbf{T}}\mathbf{W}(\mathbf{Y} - \mathbf{Xb}). \tag{6.17}$$

As with unweighted least squares, minimization of (6.17) yields the estimator

$$\hat{\mathbf{b}}_{\mathbf{W}} = (\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{Y}. \tag{6.18}$$

An equivalent derivation to that for (6.12) then yields

$$Var(\mathbf{b_W}) = \left\{(\mathbf{X^T W X})^{-1}\mathbf{X^T W}\right\}\mathbf{V}\left\{\mathbf{W X}(\mathbf{X^T W X})^{-1}\right\}. \tag{6.19}$$

Recall that we assume the following 'block' structure for the covariance matrix for Y_{ij}

$$\mathbf{V} = \text{variance}(\mathbf{Y}) = \begin{pmatrix} \mathbf{V}_1 & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{V}_2 & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{V}_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{V}_m \end{pmatrix},$$

where the only assumption used is that observations on different individuals are independent. The block matrix $\mathbf{V}_i$ is of dimension $n_i \times n_i$ and describes the covariance of repeated observations within the i^{th} individual; allowing each of these matrices to be unspecifed would require a very large number of variance/correlation parameters to be estimated. Our first step at simplification is to therefore assume that each $\mathbf{V}_i$ shares the same basic pattern or structure. The simplest of such assumptions is that, for each i,

$$\mathbf{V}_i = \mathbf{I}, \tag{6.20}$$

the $n_i \times n_i$ identity matrix. This makes the very strong assumption, of course, that longitudinal observations within the same individual are independent. Furthermore, using the identity matrix as a weight matrix in (6.18) makes no difference at all and makes no progress on our goal of more precise estimators.

6.3.1 Exchangeable Working Correlation Structure

A slightly more complex covariance assumption is the so-called *exchangeable*, or *compound symmetry*, assumption that

$$\mathbf{V}_i = \text{variance}(\mathbf{Y}_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix}.$$

Let's take a moment to see what is required to make this structure plausible. First, the common factor σ^2 assumes that each observation Y_{ij}, for the i^{th} individual, has the same

variance as j varies from 1 to n_i. This is reasonable if there is no reason to assume increasing or decreasing variability across the repeated observations, but may be questioned if the scale of Y_{ij} changes substantially over j in a situation where we expect increasing variability with increasing scale. Second, the exchangeable structure (6.3.1) assumes that the correlation between any two repeated observations, j_1 and j_2, is exactly the same value, ρ, however 'close' or 'far apart' j_1 and j_2 might be. This can be a reasonable approximate correlation structure when longitudinal observations are gathered over a relatively short time scale as with the data on teenage sexual activity. However, it is less plausible when repeated observations span very long time scales relative to the interval between 'successive' observations.

The exchangeable correlation structure arises automatically from the simple mixed effects model with random intercepts–see chapter 3. In this case, the correlation, ρ, between any two longitudinal observations has the appealing interpretation of the ratio of between individual variability to total variability.

6.3.2 Time-Decaying Working Correlation Structures

The primary concern with exchangeable correlation is that is does not reflect the possibility that repeated observation closer to each other in the time of measurement are likely to have higher correlation than measurements far apart in time. Several working correlation structures can be used to incorporate the concept that correlation between observations decline the further apart their measurement times are. The first of these we look at is the *autoregressive* correlation structure, yielding the following covariance matrix:

$$\mathbf{V}_i = \text{variance}(\mathbf{Y}_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n_i-1} \\ \rho & 1 & \rho & \cdots & \rho^{n_i-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n_i-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \rho^{n_i-3} & \cdots & 1 \end{pmatrix}.$$

Thus, the correlations between longitudinal observations decline as a power of ρ where the size of the power is just how 'far apart' the observations are in terms of the count of observations for a fixed individual.

In a marginal model such as

$$Y_{ij} = b_0 + b_1 X_{ij} + e_{ij},$$

autoregressive correlations arise when the longitudinal error terms are linked as follows

$$e_{ij} = \rho e_{i(j-1)} + \eta_{ij}$$

where the η_{ij} are independent of each other even with a fixed i. Thus the error term e_{ij} at the j^{th} observation depends, or regresses on, the error time at the immediately preceding observation, that is, $e_{i(j-1)}$.

Be cautious, however, that the level of correlation between given longitudinal observations depends solely on their respective position in the count of repeated measurements for the individual which may not reflect any sense of the actual time between measurements. In fact, only if repeated observations are equally spaced in time will the autoregressive correlation structure truly capture correlations that decline in time according to a simple pattern. As a result, an autoregressive working correlation structure is not applicable to longitudinal observations that are unequally spaced in time. A simple way to accommodate varying time periods between measurements is to use the *exponential* correlation structure where for the correlation between Y_{ij} and Y_{ik}, $\rho^{|j-k|}$ is replaced by $\rho^{|t_j - t_k|}$ where t_j is the time of the j^{th} measurement.

6.4 Example—HAART Therapy on HIV Patients

We now turn to the data, briefly described in Chapter 1.3.1, to investigate the relationship between the log viral load and CD4 cell counts for patients over time. Thus Y_{ij} and X_{ij} represent, respectively the CD4 count and log(viral load) of the i^{th} patient at the j^{th} visit.

Before we examine the full data set we consider some smaller subsets of *paired* data that effectively illustrates the various regression coefficient estimators, with both the relevant non and semi-robust variance estimates, that we described above. By paired data, we mean that we analyze only two longitudinal observations per patient, a situation where it is easy to identify simple estimation strategies for comparison. We further simplify matters by only considering binary covariates so that regression coefficients correspond to differences in two group means of CD4 counts, the two groups defined by the two levels of the relevant binary explanatory factor. We will use these examples to develop our intuition about the role of a working correlation structure, and the associated weighting, in improving the precision of our estimators.

6.4.1 Paired Data—Time-fixed Covariates

For this example, we consider a simple binary measure of age at baseline, dichotomized at 40 years of age—for those patients with complete age information; in this group 306 were younger than 40, and 288 older than 40 at baseline. Consider now the CD4 cell counts

Table 6.1: Two Sample t Test That Compares Mean CD4 Cell Counts Across Two Age Groups 0f HIV Patients Under HAART

Age Group	No. of Observations	Estimate of Mean CD4 Cell Count	Naive SD of Mean Estimate
age $<$ 40 years	306	225.9	9.3475
age $\geq$ 40 years	288	250.1	10.7445

of these individuals at the first two visits with a view to understanding the relationship between the CD4 counts and the simple binary age measure. Specifically, do younger patients tend to have higher or lower CD4 counts than older patients? Since the timing between visits is short the binary measure of age does not change and so is a time-fixed covariate. The advantage of longitudinal over cross-sectional information here is therefore simply the additional precision available in having twice as many observations, although any resultant increase in precision will be modulated by how strongly correlated the observations are on repeated visits.

If there was but a single CD4 measurement, i.e. cross-sectional data, the familiar two-sample t-test provides an immediate assessment of any difference in mean CD4 counts across the two age groups. For example, if we analyze simply the CD4 count from the first visit

Now we extend this analysis, that only used "half" the data, by incorporating both observations per patient. A naive approach simply applies the same two-sample t-test to the 594 observations, ignoring the fact that the repeated observations on individuals are likely to be correlated. This analysis, reported in Table 6.1, reveals a mean CD4 count of 225.9 and 250.1 for the younger and older patients, respectively. There is, therefore, empirical evidence of an *increased* mean CD4 count in the older patients of 24.24. The estimated standard deviation of this mean difference is 14.2 (using the standard pooled estimate of variance), yielding a two-sample t statistic of 1.71. Comparison with a t distribution (with 592 degrees of freedom) provides an associated (two-sided) p-value of 0.09 (equivalent to what is obtained if we simply compare 1.71 to tables of a standard Normal distribution because of the very large number of degrees of freedom).

A similar approach to this question can be based on generalized estimating equations

Table 6.2: GEE Estimates of Marginal Regression Model (6.4.1) Under Independence and Exchangeable Working Correlations 0f HIV Patients Under HAART

Coefficient	Working Correlation	Estimate	Naive SD of Estimate	Robust SD of Estimate
b_0	Independence	225.9	9.87	12.62
b_1	Independence	24.24	14.17	19.26
b_0	Exchangeable	225.9	13.34	12.62
b_1	Exchangeable	24.24	19.16	19.26

derived from the marginal regression model

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

with various working assumptions regarding the within-patient correlation structure and different choices for variance estimation. Summaries of such analyses are shown in Table 6.2.

The GEE estimate of the slope coefficient b_1, given by 24.24, in the marginal model (6.4.1) is equivalent by that derived from the simple difference in sample means between the two age groups as previously seen in Table 6.1. The naive standard deviation is the same. Thus inference from use of a naive t-test or GEE with the single binary age covariate is equivalent when the working correlation is assumed to be independence. The robust variance estimate of $\hat{b}_1$ immediately corrects for the fact that the paired observations show correlation within patients, raising the variance estimate from 14.17 to 19.26. This is a sizeable change and of importance as it changes the p-value to 0.2 from 0.09 when the robust estimate is used instead of the naive.

We now examine a weighted GEE estimator based on an exchangeable working correlation model. That is, we use the block-diagonal weight matrix W where each block is of the form

$$\mathbf{W}_i = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}. \tag{6.21}$$

With this definition of $\mathbf{W}$ simple matrix algebra shows that $\mathbf{X^T W X} = (\mathbf{1}+\rho)^{-\mathbf{1}}\mathbf{X^T X}$, and $\mathbf{X^T W Y} = (\mathbf{1}+\rho)^{-\mathbf{1}}\mathbf{X^T Y}$. Thus, from (6.18) the weighted coefficient estimator is:

$$\begin{aligned}\hat{\mathbf{b}}_{\mathbf{W}} &= (\mathbf{X^T W X})^{-\mathbf{1}}\mathbf{X^T W Y} \\ &= ((1+\rho)^{-1}\mathbf{X^T X})^{-\mathbf{1}}(\mathbf{1}+\rho)^{-\mathbf{1}}\mathbf{X^T Y} \\ &= (1+\rho)^{-1}(1+\rho)(\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T Y} \\ &= (\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T Y},\end{aligned}$$

or precisely the same as the unweighted, ordinary least squares estimator (6.6). Thus, using an exchangeable working correlation has no impact at all on estimation in this case, with equal numbers of observations per subject.

Focusing now on the inference, we address the question as to why the naive is so different from the robust estimation in the case where the independence working correlation model is used: the first two rows of Table 6.2, contrasting specifically the standard error of the estimate of b_1 (naive=12.62, robust=19.26). As mentioned above, the ordinary least squares estimate (which is the estimate when a independence correlation model is used) is simple the difference in means in among the older and younger groups, or

$$\hat{b}_1 = \frac{1}{2m_1}\sum_{X_{ij}=1}\sum_{j=1}^{2} Y_{ij} - \frac{1}{2m_0}\sum_{X_{ij}=0}\sum_{j=1}^{2} Y_{ij} \tag{6.22}$$

where the first average is over those with $X_{ij}=1$, or patients over 40, of which there are m_1, and the second average is those among patient $X_{ij}=0$, or patients under 40, of which there are m_0. For now assume that the exchangeable model is true - that is observations on the same subject no matter who the subject have the same underlying correlation, $\rho = cor(Y_{i1}, Y_{i2})$, and same variance, $\sigma^2 = var(Y_{ij})$, and thus the same covariance, $\rho\sigma^2 = cov(Y_{i1}, Y_{i2})$. Using the results from Chapter 3, the naive variance (assuming independence) for this model converges to:

$$\frac{\sigma^2}{2}\left(\frac{1}{m_0}+\frac{1}{m_1}\right) \tag{6.23}$$

whereas the true variance of the difference in means is:

$$\frac{\sigma^2}{2}\left(\frac{1}{m_0}+\frac{1}{m_1}\right)(1+\rho).$$

This implies (given the assumption of the exchangeable correlation model) the naive estimate of the variance of the effect estimate ($\hat{b}_1$) will be biased by a factor $(1+\rho)$. In fact, the estimated residual correlation is $\hat{\rho} = 0.83$. If the exchangeable correlation model were correct, then we would expect the robust estimated standard deviation (also called

the standard error) of $\hat{b}_1$ to be about $\sqrt{1+0.83}$ times the naive standard error of 14.2, and in fact that is precisely what we see $\sqrt{1+0.83} * 14.2 \approx= 19.3$. Further evidence that the exchangeable correlation model works well here is how close the naive and robust standard errors are for b_1 in Table 6.2 (19.16 versus 19.26).

To summarize the lessons from this simple example with a fixed baseline covariate and two measurements per subject are:

- weighting has no impact on estimation with equal numbers of observations per subject and weights based on the exchangeable correlation model,
- the standard error from the naive approach that assumes that the independence working model is true will underestimate the true variability of the effect size,
- the size of this bias increases with increasing correlation of measurements made on the same subject.

.

6.4.2 Paired Data—Time-varying Covariates

Now we consider a similar data structure (with the same outcome, CD4 count) that delivers a very different message about the pitfalls of ignoring correlation and the virtues of a GEE approach. As above, in this case we also have only two measurements per subject. The difference is that as opposed to a fixed covariate (like age was treated above), there is a time varying covariate: viral load (VL) high ($X_{ij} = 1$ if $VL > 2000$) versus low ($X_{ij} = 0$ if $VL < 2000$). In this case, a data set has been constructed that ensures every subject has CD4 count recorded at exactly one high and one low time of viral load. In this case, the model for the effect of viral load looks identical to (6.4.1) and the ordinary least squares estimate for b_1 is again a simple mean difference. Assuming that we have ordered the data so $j = 1$ and $j = 2$ is the low and high viral load measurement for all subjects, then we can write the OLS estimate as:

$$\hat{b}_1 = \frac{1}{m}\sum_{i=1}^{m}(Y_{i2} - Y_{i1}). \tag{6.24}$$

We present the results of fitting model (6.4.1) providing both the naive and robust standard errors and using both independence and exchangeable working correlation. One can think

Table 6.3: GEE estimates of marginal Regression model (6.4.1) for binary viral load (high versus low) under independence and exchangeable working correlations of HIV patients under HAART

Coefficient	Working Correlation	Estimate	Naive SD of Estimate	Robust SD of Estimate
b_0	Independence	377.4	21.4	22.9
b_1	Independence	-98.3	30.3	16.5
b_0	Exchangeable	377.4	21.4	22.9
b_1	Exchangeable	-98.3	16.4	16.5

of this as a "paired" design as within each person there is a natural pairing: a low viral load measurement is paired with a high viral load. It is well known that matching like this can reduce the variability of estimation by reducing the impact of between subject variance. In this case, using the naive approach will over-estimate the variability and so standard errors will be too big if one uses a complete naive approach that assumes observations on the same subject are independent both in estimation and inference. Using the same assumptions as the previous example (exchangeable correlation model), the variance of $\hat{b}_1$ is

$$var(\hat{b}_1) = \frac{2\sigma^2(1-\rho)}{m}$$

with the naive variance estimate providing an estimate of (6.23). Thus, in this case, the naive estimate of the variance is biased by the factor $(1-\rho)$, or is biased too high if, as is typically the case, measurements on the same subject are positively correlated ($\rho > 0$). The results verify 1) that weighting again has no impact (for reasons discussed in the previous example) and 2) the naive standard error for the independence working model is much larger (nearly double) that of the robust standard error. As above, the naive and robust standard errors for the exchangeable working correlation model are nearly the same, again indicating a good fit of that model to the data in this example. The bottomline: the robust standard error in GEE automatically accounts for the correlation and the results vary by the data structure: sometimes naive standard errors will under-estimate the standard error (fixed covariates) and sometimes they can over-estimate (time-varying covariates). Adding more covariates and combinations of fixed covariates and time-varying makes the precise relationship between naive and robust more complicated.

6.5 Example—The Effect of Drug and Alcohol Use on Teenage Sexual Activity

Recall that, for this data set—see Chapter 1.3.2—we use the indices i for a teenager, and j for a particular day of report. Thus, Y_{ij} refers to reported sexual activity for the i^{th} teenager on the j^{th} day of reporting with $Y_{ij} = 1$ indicating sexual activity that day (that is, in the past 24 hours) and $Y_{ij} = 0$ meaning no sexual activity in the previous 24 hours. Similarly, X_{ij} describes reported drug and/or alcohol use in the past 24 hours for the i^{th} teenager on the j^{th} day of reporting, with again $X_{ij} = 1$ indicating drug and/or alcohol use in the past 24 hours) and $X_{ij} = 0$ meaning no such use in the previous 24 hours.

In constructing a marginal regression model for this data, recall that the number of reporting days per teenager ranges from 1 to 33, with the average being 15.7. Of more relevance, the period of time covered by the multiple reporting days is short, almost always less than a month. The significance of this is that we do not anticipate any major longitudinal shifts in either the covariate X or the outcome Y here. It thus seems reasonable to focus entirely on the cross-sectional effects of drug/alcohol use on sexual activity. This could have been measured buy a simple cross-sectional study with one observation per teenager; however, the longitudinal data should provide additional precision in estimation of regression effects depending on how much correlation there is between observations on different reporting days for the same teenager. Only in the unlikely scenario that every teenager reports exactly the same behavior on each day would no additional precision be gained. Of importance also is that, typical of many observational data longitudinal studies, data is missing for some times, for some subjects. In this case, a teenager who called in for 30 days is compliant - any less means that the teenager decided for whatever reason not to call in and report their activities. For these subjects, their data for those days is missing. This will have consequences on interpreting the different estimates depending on the working correlation model.

We turn to the simplest marginal model that relates drug and alcohol use to sexual activity on the same day. We let $p_{ij} = Pr(Y_{ij} = 1|X_{ij} = x_{ij})$, the probability of observing reported sexual activity on the i^{th} teenager on the j^{th} day of reporting, given the observed information on drug and alcohol use as measured by x_{ij}. Then, a simple marginal logistic regression model is given by

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = a + bx_{ij}. \tag{6.25}$$

We can fit this marginal model using independence of longitudinal observations as our working assumption—this is, of course, equivalent to simply fitting a logistic regression

Table 6.4: Estimates Based on a Simple Marginal Logistic Regression Model Relating Drug and Alcohol Use to Teenage Sexual Activity: Working Correlation Structure is Independence

Model	Parameter	Estimate	Naive SD	Robust SD
$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = a + bx_{ij}$	a	-1.0679	0.0648	0.1105
	b	0.5536	0.1164	0.1811

model to the entire data set, ignoring the fact that observations are grouped by teenager. The results of this analysis are shown in Table 6.4. Thus, the estimated marginal probability of teenage sexual activity on a day with no reported drug/alcohol use is simply calculated from the estimate of the intercept, a as $\frac{e^{-1.0679}}{1+e^{-1.0679}} = 0.26$, so that this data represents a very sexually active teenage population. This is therefore the "baseline" rate of sexual activity per day.

Variability in this estimate can be assessed through the estimated standard deviation (SD) of the estimate. Two estimated standard deviations are given in Table 6.4, the first naively assumes that the working assumption of independence is actually correct and so ignores the longitudinal structure in variability estimation. Although not directly shown in the table, this naive standard deviation leads to a 95% confidence interval for the probability of sexual activity in a day as $\left(\frac{e^{-1.1949}}{1+e^{-1.1949}}, \frac{e^{-0.9409}}{1+e^{-0.9409}}\right) = (0.23, 0.28)$, arising from the equivalent interval for a. Undoubtedly this inference will be somewhat biased since it assumes that all longitudinal observations are independent, an unlikely possibility. To address this issue, we must account for the within teenager correlation in responses even when using the simple logistic regression model based on the working independence assumption. We do this using the robust standard deviation estimator described above. For the intercept, the robust estimate of the standard deviation rises from 0.0648 to 0.1105, an increase of about 71%, reflecting presumably that there is substantial correlation across the longitudinal responses.

Note that this estimate of the population level of sexual activity on days with no drug/alcohol use arises solely from the data measured on such days–this follows because the only covariate in the regression model (6.25) is binary. Thus, an alternative view of the

variability increase in the intercept estimate is that the naive estimate of standard deviation assumes a total sample of 1,251 independent observations (ignoring that these arise from only 104 independent teenagers), the days with no observed drug/alcohol use; the robust standard deviation estimate indicates that the information on the "baseline" probability of sexual activity per day corresponds to what one might obtain from approximately 430 independent observations (still substantially more than 104, reflecting that there is additional information in the longitudinal repeated responses for a teenager, but far less than if such observations were all truly independent). If you wonder where the number 430 came from, this reflects that standard deviations of estimated means, or probabilities, based on independent observations, are inversely proportional to the square root of the sample size: thus the ratio of the estimated standard deviations reflects approximately the ratio in the square roots of the "effective" sample size, yielding $\sqrt{1251} \times \frac{0.0648}{0.1105} \approx \sqrt{430}$. Note that the more plausible 95% confidence interval for the probability of sexual activity in a day, assuming no drug/alcohol use, is $\left(\frac{e^{-1.2845}}{1+e^{-1.2845}}, \frac{e^{-.8513}}{1+e^{-.8513}}\right) = (0.22, 0.30)$, using the robustly estimated standard deviation.

The slope coefficient, b, provides quantitative assessment of how much the use of drugs and/or alcohol changes this rate of sexual activity across this population. The coefficient corresponds to the log odds ratio, so that Table 6.4 gives the estimated odds ratio for sexual activity, associated with drug/alcohol use to be $e^{0.5536} = 1.74$. That is, comparing "average" teenage days with no drug/alcohol use to those with drug/alcohol use, shows an increase of 74% in the odds of sexual activity. Note this odds ratio cannot be interpreted as a relative risk since the underlying baseline probability is quite high. As for the intercept, we can translate confidence intervals for b into equivalent intervals for this odds ratio; in particular, the naive 95% confidence interval for the odds ratio is $(e^{0.3255}, e^{0.7817}) = (1.38, 2.19)$, whereas the robust version is $(e^{.1987056}, e^{.9085145}) = (1.22, 2.48)$. The robust interval reflects substantially more variability than the naive version, again reflecting the presence of considerable within teenager correlation in their responses on different days. In fact, the ratio of the standard deviations reflects an increase of about 56% from the naive to the robust estimates.

This preliminary analysis already illustrates the principal advantage of longitudinal observations in this study, namely increased precision due to the additional "sample size", albeit not quite as large as the total number of observations as just discussed. For example, suppose only the first observation had been available for each teenager, corresponding to a simple cross-sectional investigation of the relationship between drug/alcohol use and sexual activity. Standard logistic regression techniques (Jewell, 2003) then yield estimates of a and b in (6.25) of -1.2098 and 0.1480, respectively. The standard deviation of the slope estimate, no longer complicated by within teenager correlation as there is only one observation per person, is easily estimated to be 0.4753, considerably larger than the robust version from

the full longitudinal data in Table 6.4. This translates into a cross-sectional estimate of the odds ratio for sexual activity, associated with drug/alcohol use, of 1.16 with an associated 95% confidence interval of $(0.46, 2.95)$. Note that this confidence interval is much larger than—in fact completely contains—the robust longitudinal version of $(1.22, 2.48)$ from the last paragraph. In this case, a cross-sectional study with the same number of individuals, here 109, is completely inadequate in estimating this odds ratio, whereas the longitudinal version provides much more accuracy. A subtler advantage of the longitudinal study is evident from the point estimate of the odds ratio which is itself smaller than that obtained when all observations are used, 1.16 as compared to 1.74. While this discrepancy can be explained by the high variability in the cross-sectional estimate, it raises the possibility that the first observation on a teenager might be somewhat different than later ones. This can be examined by adding an indicator variable for being a first observation or not to (6.25) and, more to the point, an interaction term of this indicator with the drug/alcohol use variable. Still using a working independence assumption and robust standard deviations with the longitudinal observations yields estimates of the odds ratio, associating drug/alcohol use to sexual activity, of 1.16 amongst the first and 1.79 for later observations. Although the data is insufficient to definitively claim that the relationship is different for the first observations (the p-value for the interaction term is 0.36), the point is that this issue cannot even be considered with only cross-sectional observations.

Now we return to further longitudinal analysis of the data, attempting to exploit the correlation structure to improve the precision of our estimates. It is apparent from the comparison of naive and robust variability estimates in Table 6.4 that the independence working assumption is not reflected by the data. This suggests that it may be possible to improve precision by exploiting the true correlation structure of the longitudinal observations. A simple first attempt at this approach is to use a working exchangeable correlation structure that assumes that all repeated observations, on the same teenager, have the same level of correlation. This working assumption may not be entirely plausible since it is likely that observations further apart in time have lower correlation than those closer together. However, it is a reasonable first step beyond independence and may be appropriate in this case since all longitudinal observations are quite close in time since the total span of repeated observations are all usually within one month of each other. Table 6.5 reports the results of using a weighted GEE algorithm with a working exchangeable correlation matrix for weighting. The estimated exchangeable correlation weight is 0.16.

The results from Table 6.5 can easily be translated into a point estimate and confidence interval for the odds ratio linking drug/alcohol use to sexual activity. Specifically, the odds ratio estimate is given by $e^{0.3320} = 1.39$, with an associated 95% confidence interval $(1.06, 1.83)$. It is important to stress that, these weighted estimates are somewhat lower than the unweighted versions discussed above and we discuss why this might be the case

below. For now, we focus on the differences in standard deviation, given we hoped that a working correlation structure might squeeze out additional precision. Comparing the valid robust estimates of standard deviation in Tables 6.4 and 6.5 we see that the standard deviation of the estimate of b has declined from 0.1811 to 0.1377, a significant drop of 32%. This is reflected in the narrower range of the robust 95% confidence interval based on the weighted as compared to the unweighted estimator: $(1.06, 1.83)$ instead of $(1.22, 2.48)$. So the decision to use a working exchangeable correlation is a good one.

Note, however, the difference between the naive and robust standard deviations in Table 6.5 for the weighted estimator, an increase of about 13%. While this is a smaller discrepancy than occurred with the unweighted estimator—based on a working independence assumption—it indicates that exchangeable correlation does not fully account for the correlation reflected in the data. This, of course, opens the possibility of trying to extract even more precision in our estimators by using an even more accurate description of the actual correlation, we do not pursue this further here since any gains are unlikely to be substantial. Nevertheless this observation reflects the importance of using robust variance calculations even when more appropriate working correlation structures are invoked.

Unfortunately, comparing the estimates of Tables 6.4 and 6.5 raises even more complicatedness, given they change from $\hat{b} = 0.55$ (independence working correlation) to $\hat{b} = 0.33$ (exchangeable), a difference that is not dwarfed by the robust estimates of the variability. This implies that these estimates are simply not competing estimates of the same parameters, but appear to estimate "different things". What are these different things? Consider the difference in the two estimates - in the independence working model, every observation is weighted identically, whether it is the only observation on a person or it is one of 33 observations. However, in the exchangeable working model, observations on a person with fewer total observations will be weighted more than observations on a person with more data (more compliant teenagers). Thus, in the independence working model, compliant teenager have a bigger impact on the estimate than the exchangeable correlation model. It turns out in this case, compliant teenagers also tend to be those that report little drug/alcohol use and sexual activity (surprise, surprise). Thus, if these observations are not down-weighted (as they are in the exchangeable working correlation model), then it will re-enforce a larger association of drug/alcohol use and sexual activity (in this case, given they are positively associated). This is an example of data not missing completely at random and as is well known, this can have a large impact on the estimate if not properly treated (and sometimes, it's not possible to "adjust" for missing data). Thus, the missing data now results in these two estimates estimating different associations: the independence working correlation model is the association in a data-generating model where a randomly drawn teenager then complies according to certain inherent characteristics. It is not as easy to describe the data-generating mechanism for which the exchangeable working correlation

Table 6.5: Estimates Based on a Simple Marginal Logistic Regression Model Relating Drug and Alcohol Use to Teenage Sexual Activity: Working Correlation Structure is Exchangeable

Model	Parameter	Estimate	Naive SD	Robust SD
$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = a + bx_{ij}$	a	-0.8751	0.1101	0.1168
	b	0.3320	0.1221	0.1377

model is an estimate. In this case, suffice to say it is closer to that model where teenagers are randomly drawn and then they all comply (all in the same amount). In this case, this reflects more closely the question of interest - how drug-use affects the risk of sexual activity among a randomly drawn teenager. The lesson is that once data are missing, then the competing working correlation models can no longer be expected to estimate the same quantity (parameter). Ironically, it is also in this case that different correlation structures often have the biggest impact on the estimates.

We end this section by briefly considering a more complex regression model than (6.25) which examines the possibility that the observed association between drug/alcohol use and teenage sexual activity may be due to the role of confounding factors, in particularly, the day of the week of the measurement. That is, it is plausible that the apparent relationship is not causal but due to the fact that teenagers may tend to engage in both activities on particular days of the week, especially Friday and Saturday. While we defer a fuller discussion of confounding and causal inference to Chapter X, we briefly consider the role of the day of report in the current analysis using standard adjustment techniques, originally developed for cross-sectional data (see Jewell, 2003, Chapters 8 and 14). Specifically, we now consider the following marginal logistic regression model

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = a + bx_{ij} + c_1(Tues) + c_2(Wed) + \cdots + c_6(Sun). \tag{6.26}$$

In this model, the day of report is captured using six indicator variables with "Monday" being the reference group. Table 5.x provides estimates based on (6.26), along with naive and robust standard deviations, using a working exchangeable correlation structure for weighting.

Table 6.6: Estimates Based on a Simple Marginal Logistic Regression Model Relating Drug and Alcohol Use to Teenage Sexual Activity and Weekday of Report: Working Correlation Structure is Exchangeable

Model	Parameter	Estimate	Naive SD	Robust SD
$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = a + bx_{ij} + c_1(Tues) + \cdots + c_6(Sun)$	a	-1.0524	0.1747	0.1913
	b	0.3170	0.1240	0.1344
	c_1	0.2145	0.1915	0.1873
	c_2	0.2106	0.1885	0.2044
	c_3	0.0951	0.1875	0.2120
	c_4	-0.0168	0.1941	0.1967
	c_5	0.2448	0.1897	0.1950
	c_6	0.4561	0.1858	0.1942

The estimate of b is very similar in Table 6.6 to that obtained from model (6.25) and using the exchangeable working correlation model(Table 6.5) reflecting that the reporting day induces very little confounding in assessment of the impact of drug/alcohol use on sexual activity. The adjusted robust 95% confidence interval for the odds ratio for this relationship is $e^{0.3170\pm(1.96\times0.1344)} = (1.06, 1.79)$, essentially equivalent to what we obtained previously without adjustment. The lack of confounding results from the fact that there is little dependence between report day and drug/alcohol, as the association of report day and sexual activity is quite strong, at least for certain days, as shown by the coefficients $c_1, \ldots, c_6$ in Table 6.6. By far the strongest such association—and the only statistically significant one—is for Sunday as the reporting day: note that the odds ratio for sexual activity comparing Sunday with the baseline report day, Monday, is $e^{0.4561} = 1.58$, reflecting a substantial increase in the likelihood of sexual activity reported on Sunday. Recall that the report day refers to the *prior* 24 hour period so that this effect describes increased levels of activity on Saturdays as might be expected.