13

Marginal Models: Generalized Estimating Equations (GEE)

13.1 INTRODUCTION

In the previous chapter we considered an approach for extending generalized linear models to longitudinal data that leads to a class of regression models known as *marginal models*. Marginal models have a three-part specification in terms of a regression model for the mean response, supplemented by assumptions concerning the variance of the response at each occasion and the pairwise within-subject association among the responses. In principle, this three-part specification can be extended by making full distributional assumptions about the vector of responses. However, as discussed in Section 12.4, assumptions about the joint distribution of the vector of responses are not necessary for estimation of the marginal model parameters. Indeed, the avoidance of distributional assumptions can be advantageous, since there is no convenient specification of the joint multivariate distribution when the responses are discrete. It also leads to a method of estimation for marginal models known as *generalized estimating equations* (GEE). As we will see, the GEE approach provides a convenient alternative to maximum likelihood (ML) estimation.

The GEE approach for estimating the parameters of marginal models is described in detail in Section 13.2. In Section 13.3, we briefly review some useful residual diagnostics for assessing the fit of marginal models. In Section 13.4, we present three case studies that illustrate the application of marginal models to longitudinal data. Finally, in Section 13.5, we consider estimation and aspects of interpretation of time-varying covariates in marginal models. When a covariate is time-varying, and varies *randomly* over time, subtle issues arise concerning the interpretation and estimation of its effect. Throughout, we adopt the same notation as was used in Chapter 12.

Applied Longitudinal Analysis, Second Edition. By Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware.

13.2 ESTIMATION OF MARGINAL MODELS: GENERALIZED ESTIMATING EQUATIONS

Since there is no convenient specification of the joint multivariate distribution of Y_i for marginal models when the responses are discrete, we require an alternative to maximum likelihood estimation. The generalized estimating equations (GEE) approach provides that alternative. The GEE approach is based on the concept of "estimating equations" and provides a very general and unified approach for analyzing correlated responses that can be discrete or continuous. The essential idea behind the GEE approach is to generalize and extend the usual likelihood equations for a generalized linear model for a univariate response by incorporating the covariance matrix of the vector of responses, Y_i. For the case of linear models (i.e., marginal models with an identity link function), the generalized least squares (GLS) estimator of β discussed in Chapter 4 can be considered a special case of the GEE approach. For marginal models with non-linear link functions, this approach is known as "generalized estimating equations".

Suppose, as in Section 12.2, that the following marginal model has been assumed:

1. The marginal expectation of the response, $E(Y_{ij}|X_{ij}) = \mu_{ij}$, depends on the covariates, X_{ij}, through a known link function

$$g(\mu_{ij}) = \eta_{ij} = X'_{ij}\beta.$$

2. The variance of each Y_{ij}, given the covariates, depends on the mean according to

$$\text{Var}(Y_{ij}|X_{ij}) = \phi\, v(\mu_{ij}),$$

where $v(\mu_{ij})$ is a known "variance function" (i.e., a known function of the mean, μ_{ij}) and ϕ is a scale parameter that may be known or may need to be estimated. For example, when the response is continuous, ϕ is a scale parameter that needs to be estimated. In contrast, with a binary response, ϕ is known and fixed at 1. For count data, ϕ is often estimated from the data at hand to allow for overdispersion relative to Poisson variability.

3. The *pairwise* (or two-way) within-subject association among the vector of repeated responses, given the covariates, is assumed to be a function of the means, μ_{ij}, and an additional set of within-subject association parameters, α. For example, when the vector of parameters α represents the pairwise correlations among the responses, the covariances among the responses depend on $\mu_{ij}(X'_{ij}\beta)$, ϕ, and α. That is, given a model for the pairwise correlations, the corresponding covariance matrix can be constructed as the product of standard deviations and correlations

$$V_i = A_i^{\frac{1}{2}}\text{Corr}(Y_i)A_i^{\frac{1}{2}},$$

where A_i is a diagonal matrix with $\text{Var}(Y_{ij}|X_{ij}) = \phi\, v(\mu_{ij})$ along the diagonal (and $A_i^{\frac{1}{2}}$ is a diagonal matrix with the standard deviations, $\sqrt{\phi\, v(\mu_{ij})}$, along

the diagonal), and $\text{Corr}(Y_i)$ is the correlation matrix (here a function of α). In the parlance of the GEE approach, V_i is known as a "working" covariance matrix to distinguish it from the true underlying covariance among the Y_i. That is, the term "working" acknowledges our uncertainty about the assumed model for the variances and within-subject associations; unless they have been correctly modeled, our model for the covariance matrix may not be correct.

There are two important features of this specification of a marginal model that are often overlooked. First, recall from Section 12.2 that there is an implicit assumption in the marginal model for the mean response. Marginal models assume that the conditional mean of the j^{th} response, given $X_{i1}, ..., X_{in_i}$, depends only on X_{ij},

$$E(Y_{ij}|X_i) = E(Y_{ij}|X_{i1}, ..., X_{in_i}) = E(Y_{ij}|X_{ij}).$$

The implications of this assumption for time-varying covariates is discussed in detail in Section 13.5. Second, the variance of Y_{ij} at each occasion is specified in terms of a variance function, $v(\mu_{ij})$, and a *single* scale parameter ϕ. In principle, a separate scale parameter, ϕ_j, could be estimated at each occasion for balanced designs; alternatively, the scale parameter could depend on the times of measurement, with $\phi(t_{ij})$ being some parametric function of t_{ij}. In practice, a limitation of many of the implementations of the GEE approach in widely available software is that they assume the scale parameter ϕ is time-invariant. This restriction on the scale parameter makes these implementations of the GEE approach unappealing for analyzing longitudinal data when the response variable is continuous and the variance of the repeated measurements is not constant over the duration of the study.

Next we provide some motivation for the GEE approach. Recall from Chapter 4 that the generalized least squares (GLS) estimator of β for the linear model minimizes the objective function

$$\sum_{i=1}^{N} (y_i - X_i\beta)' \Sigma_i^{-1} (y_i - X_i\beta).$$

Using calculus, it can be shown that if a minimum of this function exists it must solve the following equations:

$$\sum_{i=1}^{N} X_i' \Sigma_i^{-1} (y_i - \mu_i) = 0,$$

where $\mu_i = \mu_i(\beta) = X_i\beta$. (Here $\mu_i(\beta)$ simply denotes that the mean vector, μ_i, depends on β.) For the linear model these equations have the following closed-form solution:

$$\widehat{\beta} = \left\{ \sum_{i=1}^{N} \left(X_i' \Sigma_i^{-1} X_i \right) \right\}^{-1} \sum_{i=1}^{N} \left(X_i' \Sigma_i^{-1} y_i \right),$$

and $\widehat{\beta}$ is known as the GLS estimator of β. The GEE estimator of β for marginal models (or generalized linear models for longitudinal data) can be thought of as arising

from minimizing the following objective function:

$$\sum_{i=1}^{N} \{y_i - \mu_i(\beta)\}' V_i^{-1} \{y_i - \mu_i(\beta)\}, \tag{13.1}$$

with respect to β, where V_i is treated as known (by ignoring its dependence on β through μ_i) and μ_i is the vector of mean responses, with elements

$$\mu_{ij} = \mu_{ij}(\beta) = g^{-1}\left(X_{ij}'\beta\right).$$

Using calculus, it can be shown that if a minimum of the function given by (13.1) exists, then it must solve the following *generalized estimating equations*:

$$\sum_{i=1}^{N} D_i' V_i^{-1} (y_i - \mu_i) = 0, \tag{13.2}$$

where V_i is the so-called "working" covariance matrix and $D_i = \partial\mu_i/\partial\beta$ is the gradient or "derivative" matrix (i.e., the matrix containing the derivative of μ_i with respect to the components of β). By the "working" covariance matrix we mean that V_i approximates the true underlying covariance matrix for Y_i; that is, $V_i \approx \text{Cov}(Y_i)$, recognizing that $V_i \neq \text{Cov}(Y_i)$ unless the models for the variances and the within-subject associations are correct. As before, we let the true covariance matrix for Y_i be denoted by Σ_i. The $n_i \times p$ matrix D_i can easily be derived using calculus and can be thought of as a matrix that transforms from the original units of Y_i (and μ_i) to the units of $g(\mu_{ij})$. Recall that $g(\mu_{ij})$ is the scale on which β has interpretation (e.g., the log odds scale rather than the probability scale when the Y_{ij} are binary and a logit link function has been assumed). The matrix D_i is only a function of β (since the μ_{ij} only depend on β). For example, when a canonical link is used, $D_i' = X_i' A_i$, where A_i is a diagonal matrix with $\text{Var}(Y_{ij}|X_{ij}) = \phi\, v(\mu_{ij})$ along the diagonal. On the other hand, V_i is a function of β, ϕ, and α, since the diagonal elements of V_i are the variances and the off-diagonal terms are the "working" covariances. That is, the variances depend on the means, and hence β, via the variance function, $v(\mu_{ij})$ (they also depend on ϕ); the covariances among the components of Y_i depend on both β and α. As a result the generalized estimating equations are functions of both β and α. For generalized linear models with non-identity link functions, the GEE have no closed-form solution; instead, the solution requires an iterative algorithm.

The generalized estimating equations described above extend in a natural way to marginal models for ordinal responses. Recall from Section 12.3 that when the response is ordinal construction of a marginal model for the cumulative probabilities requires treating each ordinal response as a set of $K-1$ binary variables. With repeated measures of an ordinal response, Y_{ij} is replaced by a $(K-1) \times 1$ vector of binary variables, say $(U_{ij1}, ..., U_{ij,K-1})'$, for the $K-1$ dichotomizations of the ordinal response (where $U_{ijk} = 1$ if $Y_{ij} \leq k$ and $U_{ijk} = 0$ if $Y_{ij} > k$). Therefore the generalized estimating equations for ordinal responses are based on the following $(K-1)\, n_i \times 1$ vector of binary responses: $Y_i = (Y_{i1}', Y_{i2}', ..., Y_{in_i}')'$, where each

$Y_{ij} = (U_{ij1}, ..., U_{ij,K-1})'$ is a $(K-1) \times 1$ vector of binary variables; the dimensions of μ_i, V_i, and D_i in (13.2) are modified accordingly.

Because the GEE depend on both β and α, the following iterative two-stage estimation procedure is required:

1. Given current estimates of α and ϕ, V_i is estimated and an updated estimate of β is obtained as the solution to the generalized estimating equations given by (13.2).

2. Given the current estimate of β, updated estimates of α and ϕ are obtained based on the standardized residuals

$$e_{ij} = (Y_{ij} - \widehat{\mu}_{ij}) / \sqrt{v(\widehat{\mu}_{ij})}.$$

For example, ϕ can be estimated by

$$\widehat{\phi} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{n_i} e_{ij}^2}{\sum_{i=1}^{N} n_i}.$$

The pairwise association parameters, α, can be estimated in a similar way, depending on the model for the within-subject association in the third component of the marginal model. For example, in a balanced design, when the association is expressed in terms of unstructured correlations, $\alpha_{jk} = \text{Corr}(Y_{ij}, Y_{ik})$ can be estimated by

$$\widehat{\alpha}_{jk} = \left(\frac{1}{\widehat{\phi}\, N} \right) \sum_{i=1}^{N} e_{ij} e_{ik}.$$

Finally, in this two-stage estimation procedure, we usually iterate between steps 1 and 2 until convergence has been achieved; starting or initial estimates of β are usually obtained from fitting a generalized linear model assuming independent observations. This algorithm is computationally quite simple, and the GEE approach has been implemented in many statistical software packages (e.g., PROC GENMOD in SAS, the `gee` and `geepack` packages in R, and the `xtgee` command in Stata).

At convergence, $\widehat{\beta}$, the solution to the generalized estimating equations, has the following properties:

1. $\widehat{\beta}$ is a consistent estimator of β. That is, with very high probability, $\widehat{\beta}$ is close to the population regression parameters β in large samples (i.e., for sufficiently large N). Of note, $\widehat{\beta}$ is a consistent estimator of β whether the within-subject associations have been correctly modeled. That is, for $\widehat{\beta}$ to provide a valid estimate of β we only require that the model for the mean response has been correctly specified. This is an important robustness property of $\widehat{\beta}$ that makes the GEE approach very appealing in many applications.

2. In large samples the sampling distribution of $\widehat{\beta}$ is multivariate normal with mean β and

$$\text{Cov}(\widehat{\beta}) = B^{-1}MB^{-1},$$

where

$$B = \sum_{i=1}^{N} D_i' V_i^{-1} D_i,$$

$$M = \sum_{i=1}^{N} D_i' V_i^{-1} \text{Cov}(Y_i) V_i^{-1} D_i.$$

Note that B and M can be estimated by replacing α, ϕ, and β by their estimates, and by replacing $\text{Cov}(Y_i) = \Sigma_i$ in M by $(Y_i - \widehat{\mu}_i)(Y_i - \widehat{\mu}_i)'$. That is, the expression for $\widehat{\text{Cov}}(\widehat{\beta})$ is given by

$$\left(\sum_{i=1}^{N} \widehat{D}_i' \widehat{V}_i^{-1} \widehat{D}_i\right)^{-1} \left\{\sum_{i=1}^{N} \widehat{D}_i' \widehat{V}_i^{-1} (Y_i - \widehat{\mu}_i)(Y_i - \widehat{\mu}_i)' \widehat{V}_i^{-1} \widehat{D}_i\right\} \left(\sum_{i=1}^{N} \widehat{D}_i' \widehat{V}_i^{-1} \widehat{D}_i\right)^{-1}. \quad (13.3)$$

This is known as the empirical or "sandwich" estimator; the components B and M can be thought of as the "bread" and "meat" of this sandwich estimator of $\text{Cov}(\widehat{\beta})$. Finally, if we model V_i correctly, $V_i = \Sigma_i$, and $\text{Cov}(\widehat{\beta}) = B^{-1}$.

In summary, the GEE approach has a number of appealing properties for estimation of the regression parameters in marginal models. First, in many longitudinal designs the GEE estimator $\widehat{\beta}$ is almost as precise or efficient as the MLE. For example, it can be shown that the GEE has a similar expression to the likelihood equations for β in a linear model for continuous responses that are assumed to have a multivariate normal distribution. That is, the GLS estimator of β can be considered a special case of the GEE approach. The GEE also has an expression similar to the likelihood equations for β in certain models for discrete longitudinal data. As a result, for many longitudinal designs, there is little loss of precision when the GEE approach is adopted as an alternative to maximum likelihood. Second, the GEE estimator $\widehat{\beta}$ is a consistent estimator of β even if the within-subject associations among the repeated measures have been misspecified; this is a very appealing robustness property of the GEE estimator. Although the GEE estimator $\widehat{\beta}$ is a consistent estimator under misspecification of the within-subject associations, the usual standard errors obtained under the misspecified model for the within-subject association are not valid. Fortunately, in many cases valid standard errors for $\widehat{\beta}$ can be obtained using the empirical or "sandwich" estimator of $\text{Cov}(\widehat{\beta})$. In addition to correcting for misspecification of the within-subject association, the "sandwich" estimator also corrects for potential overdispersion (or, indeed, any misspecification of the variance). Another important feature of the GEE approach is that it can readily handle imbalance due to missing data in the response variables. However, in doing so, it does require a strong, and often unrealistic, assumption that data are missing completely at random (MCAR). The GEE estimators can be adapted to provide a valid analysis when data are missing at random (MAR),

but not MCAR, by explicitly modeling the missingness process and weighting the analysis accordingly. This approach to handling missing data is known as the inverse probability weighted (IPW) GEE method and is discussed in detail in Chapters 17 and 18.

Finally, although the main emphasis of this chapter has been on longitudinal analysis of a discrete response, the GEE approach can be applied equally to continuous responses. That is, for the linear regression models described in Part II, the multivariate normal assumption is not crucial. Specifying a linear regression model for the longitudinal responses and a model for the covariance among the responses is sufficient for the purposes of estimating β using the GEE approach. As was mentioned above, the GLS estimator of β can be considered a special case of the GEE approach and the multivariate normal distribution assumption for the responses is not required. The validity of the GLS/GEE estimates of β rests only on having a correct model for the mean response. When the model for the covariance is misspecified, valid standard errors for $\widehat{\beta}$ can be obtained using the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$. Thus, although the GEE approach and the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ are more widely used in marginal models for discrete data, they can also be applied in the linear models for continuous data described in Part II. However, because many of the implementations of the GEE approach in widely available software assume that the scale parameter ϕ is time-invariant, we do not recommend their use for analyzing longitudinal data when the response variable is continuous. Instead, the GEE approach can be implemented using existing software for the general linear model that allows a much wider range of covariance pattern models and/or random effects covariance structures, coupled with the option of calculating standard errors for $\widehat{\beta}$ based on the "sandwich" estimator (e.g., using the `EMPIRICAL` option in PROC MIXED in SAS).

A Note on the "Sandwich" Estimator of $\text{Cov}(\widehat{\beta})$

An appealing property of the GEE estimator $\widehat{\beta}$ is that it is a consistent estimator of β even if the assumed model for the covariances among the repeated measures is not correct. It only requires that the model for the mean response be correct. This robustness property of GEE is important because the usual focus of a longitudinal study is on changes in the mean response. Based either on theoretical grounds (e.g., randomization in an experiment) or subject-matter knowledge of similar data, the data analyst can often specify how changes in the mean response depend on the covariates. On the other hand, much less is usually known about the patterns of two-way and higher-way associations among the responses; moreover these "higher-order moments" are increasingly difficult to estimate from the data.

For inferences about β, valid standard errors can be obtained from the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ given by (13.3). The remarkable property of the "sandwich" estimator is that it is also robust in the sense that it provides valid standard errors when the assumed model for the covariances among the repeated measures is not correct. That is, with large sample sizes, the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ yields correct standard errors. Because of this appealing robustness property, the "sandwich"

estimator is often referred to as the "robust" variance estimator or the "empirical" variance estimator. Maintaining the culinary theme of this section, it would seem that we can have our cake and eat it: we can obtain a valid estimate of β and its sampling variability, even if we have not modeled the within-subject association correctly. Indeed, some readers may see it as a delicious irony that we can disregard the model for the covariances among the repeated measures for the purposes of inference about β.

This raises an important issue. Why bother expending effort to model the within-subject association? For example, naively assuming the responses are independent (i.e., specifying the "working" covariance matrix as diagonal) yields valid estimates of β; valid standard errors can then be obtained using the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$. There are two main reasons for modeling the covariance. First, in general, the closer the "working" covariance matrix (V_i) approximates the true underlying covariance matrix (Σ_i), the greater the efficiency or precision with which β can be estimated. That is, a "working" covariance matrix that approximates the true underlying covariance matrix makes optimal use of the available data for estimation of β. Second, the robustness property of the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ is a large sample (or asymptotic) property. In general, use of the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ is best suited to balanced longitudinal designs where the number of subjects (N) is relatively large and the number of repeated measures (n) is relatively small. Moreover the "sandwich" estimator is less appealing when the design is severely unbalanced and/or when there are few replications to estimate the true underlying covariance matrix. In applications, use of the "sandwich" estimator implicitly relies on there being many replications of the vector of responses associated with each distinct set of covariate values. For example, in a longitudinal clinical trial with two treatment groups there will be many replications of Y_i associated with the two distinct set of covariate values (X_i) for the treatment and control groups. In that case, use of the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ is justified because there is a sufficient number of replications (or number of subjects) to estimate the true underlying covariance matrix within each treatment group. In many observational studies, however, there may be few, if any, replications of Y_i associated with each distinct set of covariate values, especially when X_i includes many covariates and/or quantitative covariates. Similarly, if the longitudinal design is severely unbalanced, with each individual having a unique sequence of measurement occasions, $t_{i1}, ..., t_{in_i}$, there are no replications at each of the measurement occasions. In these cases the use of the "sandwich" estimator is less appealing. In particular, when the number of subjects is relatively small the "sandwich"-based standard errors are biased downward, in the sense that the nominal standard errors are too small and underestimate the variance of $\widehat{\beta}$. Moreover the magnitude of the bias of the "sandwich" variance estimator depends on the covariate design matrix, X_i. In addition, the sampling variability of the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ can be very large, resulting in an unstable estimate of $\text{Cov}(\widehat{\beta})$. This increased variability of the "sandwich" estimator is the price paid for its robustness property. However, the increased variability can lead to problems with the coverage probabilities of confidence intervals constructed from "sandwich" estimates. In par-

ticular, use of the "sandwich" estimates can yield confidence intervals with coverage probability well below the desired nominal level.

In summary, the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ is of most practical use when the sample size is relatively large or when the assumed model for the covariances among the repeated measures (the "working" covariance matrix) is questionable. Reliance on the "sandwich" estimator is less appealing when the number of independent subjects is modest (relative to the number of repeated measures), the design is inherently unbalanced, or when subjects cannot be grouped on the basis of having identical covariate design matrices. For any of these cases it can be advantageous to model the covariances among the responses and use a "model-based" estimator of $\text{Cov}(\widehat{\beta})$. The model-based estimator is given by

$$\text{Cov}(\widehat{\beta}) = B^{-1},$$

where

$$B = \sum_{i=1}^{N} D_i' V_i^{-1} D_i,$$

and can be estimated by replacing α, ϕ, and β by their estimates. This estimator of $\text{Cov}(\widehat{\beta})$ is called a "model-based" estimator to remind us that it yields valid standard errors provided that the "working" covariance matrix, V_i, is a close approximation to the true underlying covariance matrix, Σ_i. That is, the "model-based" estimator does require that the model for the covariance, the "working" covariance, be correctly specified.

13.3 RESIDUAL ANALYSES AND DIAGNOSTICS

The adequacy of a fitted marginal regression model can be assessed using various residual analysis techniques similar to those described in Chapter 10 for linear models for longitudinal data. Residuals and other standard regression diagnostics (e.g., Cook's distance and leverage) can also be helpful for identifying outliers and influential observations and/or influential individuals. In this section we briefly review some useful diagnostics for assessing the functional form of the marginal model for the mean response and for detecting observations or individuals that may be having an undue influence on the analysis.

For a marginal regression model it is straightforward to calculate residuals based on the difference between the observed and predicted responses at each occasion,

$$r_{ij} = Y_{ij} - \widehat{\mu}_{ij} = Y_{ij} - g^{-1}(X_{ij}'\widehat{\beta});$$

these can be readily produced by most statistical software packages for fitting marginal models. However, because the variance of the residual, r_{ij}, is a function of the predicted mean response, it is preferable to conduct all model checking using studentized Pearson residuals,

$$e_{ij} = \frac{Y_{ij} - g^{-1}(X_{ij}'\widehat{\beta})}{\sqrt{\phi\, v(\widehat{\mu}_{ij})\,(1 - h_{ij})}},$$

where $v(\mu_{ij})$ is the variance function, ϕ is the scale parameter (either fixed or estimated from the data, depending on the type of response and assumptions about overdispersion), and h_{ij} is the "leverage" of the j^{th} observation on the i^{th} individual. In regression, the leverage, h_{ij}, describes the influence each observation has on its own predicted value. In general, observations that are extreme in terms of the X_{ij}'s have high leverage (albeit in generalized linear models, leverage values also depend on the mean response). These Pearson residuals can be used to check for any systematic departures from the model for the mean response; for example, a scatterplot of e_{ij} against the predicted mean response, $\widehat{\mu}_{ij}$, can be examined for the appearance of systematic trend. The fitting of a smooth curve (e.g., a *lowess* curve) to the scatterplot can often help in judging whether curvature is present. Similarly scatterplots of the residuals against selected covariates from the model for the mean response can be examined for any systematic trends. Such a trend may indicate, for example, the omission of a quadratic term or the need for transformation of the covariate. A scatterplot of the residuals versus time (or age) can be particularly useful for assessing the adequacy of the marginal model assumptions about patterns of change in the mean response over time.

An acknowledged difficulty with the interpretations of these conventional residual diagnostic plots is that they are inherently subjective. In Chapter 10 (Section 10.4) we discussed how this problem can be overcome by basing model assessment on "cumulative sums" and "moving sums" of residuals. The exact same approach can be extended to residuals from marginal regression models. That is, we can compare the *observed* sum of the residuals, both graphically and numerically, to a reference distribution under the assumption of a correctly specified marginal model for the mean response. This allows us to determine whether any apparent pattern in the observed sum of the residuals is evidence of a systematic trend or simply due to natural variation. For example, to check the functional form of the k^{th} covariate, we can define the cumulative sum of the residuals over values of X_{ijk},

$$W_k(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{j=1}^{n_i} I(X_{ijk} \leq x) r_{ij},$$

where $I(\cdot)$ is the indicator function. In addition we can construct the cumulative sum of residuals over the fitted values, denoted by $W_f(x)$,

$$W_f(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{j=1}^{n_i} I(X'_{ij}\widehat{\beta} \leq x) r_{ij}.$$

Recall from Section 10.4 that if the assumed model for the marginal mean has been correctly specified, the cumulative residual process should be centered at zero and behave like a zero-mean Gaussian (or normal) process. This zero-mean Gaussian process provides a reference for deciding whether any pattern in the observed cumulative residual process is systematic or simply due to the natural variation of the process. An assessment of model adequacy is based on comparing the pattern of the observed cumulative residual process with computer simulated realizations from

the zero-mean Gaussian process (the null distribution). In particular, the cumulative sum, $W_k(x)$, can be used to provide both a graphical and numerical assessment of the functional form of the covariate; the cumulative sum, $W_f(x)$, can be used to provide an assessment of the adequacy of the link function. An omnibus test ("supremum" test) of the adequacy of the marginal regression model with respect to the relevant coordinate (e.g., a particular covariate or the fitted values) can be obtained by comparing the maximum absolute value of the observed cumulative sum to a large number of realizations (e.g., 10,000) from the null distribution (see Section 10.4 for additional details concerning the supremum test).

An appealing property of the graphical and numerical methods based on cumulative (and moving) sums of residuals is that they are valid regardless of whether the covariance among the responses has been correctly specified. As such, these graphical and numerical techniques for assessing the marginal model for the mean response are relatively robust to the working correlation assumption.

Residual analyses are also useful for detecting outlying *observations*. The detection of outlying observations is important because they can potentially have an inordinate influence on the analysis. In addition residuals can be used to identify outlying *individuals* who have unusual patterns of responses. Specifically, for each individual we can calculate a summary measure of multivariate distance between their vector of observed and fitted responses,

$$d_i = r_i' \widehat{V}_i^{-1} r_i,$$

where r_i denotes the $n_i \times 1$ vector of residuals for the i^{th} subject,

$$V_i = A_i^{\frac{1}{2}} \text{Corr}(Y_i) A_i^{\frac{1}{2}},$$

and $A_i^{\frac{1}{2}}$ is a diagonal matrix with the standard deviations, $\sqrt{\phi v(\mu_{ij})}$, along the diagonal. If the model for the mean is correctly specified, and $V_i \approx \text{Cov}(Y_i)$, d_i has an approximate chi-squared distribution with degrees of freedom equal to n_i. Outlying individuals will have distances, d_i, that have relatively small associated p-values. These p-values provide a common metric for comparing d_i when the number of repeated measurements varies across subjects; however, it must be recognized that relatively small p-values (e.g., p-values less than 0.05) are expected to occur with predictable regularity.

Finally, both influential observations and influential individuals can be detected using "deletion diagnostics" for marginal regression models fitted by generalized estimating equations. Various measures of influence, initially developed for standard linear regression (e.g., Cook's distance and leverage), have been extended to marginal regression models. For example, Cook's distance is a widely used statistic in standard linear regression (Cook, 1977) that measures the change in the regression parameter estimates caused by deleting each observation in turn. Cook's distance can also be applied to marginal regression models to detect either influential individuals or influential observations. For detecting influential individuals, Cook's distance is based on $(\widehat{\beta} - \widehat{\beta}_{[i]})$, where $\widehat{\beta}$ is the vector of parameter estimates obtained from the

analysis of data on all individuals and $\widehat{\beta}_{[i]}$ denotes the vector of parameter estimates obtained after deleting all of the observations on the i^{th} individual. For detecting influential observations, Cook's distance is based on $(\widehat{\beta} - \widehat{\beta}_{[ij]})$, where $\widehat{\beta}_{[ij]}$ denotes the vector of parameter estimates obtained after deleting the j^{th} observation on the i^{th} individual. Exact values for these diagnostics can be obtained by deleting the observation (or the set of correlated observations on an individual) and re-fitting the marginal model to the remaining observations. However, this requires iterating the GEE fitting algorithm until convergence has been achieved. Therefore, to greatly reduce the computational burden, one-step approximations to these diagnostics are commonly used. These one-step approximations are sufficiently accurate for most practical purposes. Other measures of influence, such as leverage, have also been extended to marginal regression models, and these statistics can be defined at the levels of individuals and observations. For example, leverage defined at the level of an individual is simply the sum of the observation leverages, $\sum_{j=1}^{n_i} h_{ij}$.

In summary, the GEE estimator $\widehat{\beta}$ provides a valid estimate of β when the model for the mean response has been correctly specified. To assess the adequacy of the marginal regression model residual analysis techniques, similar to those described in Chapter 10, can be used. In addition many of the standard regression diagnostics for identifying outliers and influential observations have been extended to marginal models fit by generalized estimating equations.

13.4 CASE STUDIES

Next we illustrate the main ideas presented in this and the previous chapter by considering marginal models for analyzing longitudinal data from three different studies. The first illustration employs marginal models to analyze obesity data in a sample of school-age children from the Muscatine Coronary Risk Factor (MCRF) study. The second illustration considers marginal models for analyzing count data from a study comparing two antibiotics to a placebo for the treatment of leprosy. The third illustration considers marginal models for analyzing treatment related changes in an ordinal response measuring patients' self-assessment of their arthritis.

Muscatine Coronary Risk Factor Study

The Muscatine Coronary Risk Factor (MCRF) study was a longitudinal survey of school-age children in Muscatine, Iowa (Woolson and Clarke, 1984; Lauer et al., 1997). The goal of the study was to examine the development and persistence of risk factors for coronary disease in children. In the MCRF study, weight and height measurements of five cohorts of children, initially aged 5–7, 7–9, 9–11, 11–13, and 13–15 years, were obtained biennially from 1977 to 1981. In total, data were collected on 4856 boys and girls. Although each child was eligible to participate in all three surveys, the data are incomplete for many children.

Table 13.1 Percentage of children from the Muscatine Coronary Risk Factor study classified as obese in 1977, 1979, and 1981.

Gender	Age Cohort	Percentage Obese		
		1977	1979	1981
Males				
	5–7	7.9	15.4	21.2
	7–9	18.8	20.5	23.7
	9–11	21.2	22.7	22.5
	11–13	24.3	21.8	19.4
	13–15	19.2	21.1	18.2
Females				
	5–7	14.0	17.2	25.1
	7–9	16.5	24.0	24.9
	9–11	25.4	26.2	22.2
	11–13	23.8	22.1	19.9
	13–15	22.9	25.8	20.9

In this section we present longitudinal analyses of a binary response, indicating whether the child is obese. At each occasion, on the basis of a comparison of their weight to age–gender-specific norms, children were classified as obese or not obese. The goal of the analyses is to determine whether the risk of obesity increases with age and whether patterns of change in obesity are the same for boys and girls. The percentages of the children classified as obese at each of the three measurement occasions are displayed in Table 13.1. These percentages were calculated based on the available data in each age-gender cohort at each occasion. These descriptive statistics suggest that the rates of obesity increase from ages 6 to 12, but decline thereafter. They also suggest that the rates of obesity are higher for girls at all ages.

Initially our analysis of these data assumes that there are no cohort effects. The marginal probability of obesity is modeled as a logistic function of the covariates: linear and quadratic age, gender, and the gender-age interactions. Here age is the midpoint of the age cohort that a child belongs to (e.g., 6, 8, and 10 years at the first, second, and third occasions for the cohort of children initially aged 5–7 years). Letting $Y_{ij} = 1$ if the i^{th} child is classified as obese at the j^{th} occasion, and $Y_{ij} = 0$ otherwise, we assume that the marginal probability of obesity at each occasion follows the logistic model,

$$\log\left\{\frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)}\right\} = \beta_1 + \beta_2\,\text{Gender}_i + \beta_3\,\text{Age}_{ij} + \beta_4\,\text{Age}_{ij}^2 + \beta_5\,\text{Gender}_i \times \text{Age}_{ij} + \beta_6\,\text{Gender}_i \times \text{Age}_{ij}^2,$$

where Age_{ij} = midpoint of age cohort at the j^{th} occasion $-$ 12 years; $\text{Gender}_i = 1$ if the i^{th} child is female, and $\text{Gender}_i = 0$ otherwise. This specifies the first component of a marginal model, the model for the mean response. It is assumed that the log odds of obesity changes curvilinearly with age (i.e., quadratic age trend), but the trend over time is allowed to be different for girls and boys. Next we assume that

$$\text{Var}(Y_{ij}) = \mu_{ij}(1-\mu_{ij}).$$

This specifies the second component, the variance function and known scale parameter ($\phi_j = 1, j = 1, \ldots, 3$). Finally, we need to make assumptions about the pairwise within-subject associations among the binary responses. Because the response is binary, correlation is not the most appealing metric for association. As was mentioned in Section 12.2, with binary responses the correlations are constrained and must satisfy certain linear inequalities determined by the marginal probabilities. Instead, we specify the association in terms of pairwise log odds ratios, a more natural measure of association between pairs of binary responses.[1] Specifically, the within-subject association among the three repeated binary responses is assumed to have the following unstructured pairwise log odds ratio pattern,

$$\log \text{OR}\,(Y_{ij}, Y_{ik}) = \alpha_{jk},$$

where

$$\text{OR}(Y_j, Y_k) = \frac{\Pr(Y_j=1, Y_k=1)\Pr(Y_j=0, Y_k=0)}{\Pr(Y_j=1, Y_k=0)\Pr(Y_j=0, Y_k=1)}.$$

The estimated regression coefficients and pairwise log odds ratios for the within-subject association obtained using the GEE approach are presented in Table 13.2. A test of the hypothesis that changes in the log odds of obesity are the same for boys and girls, H_0: $\beta_5 = \beta_6 = 0$, can be constructed using a multivariate Wald statistic. This test produces a Wald statistic, $W^2 = 0.91$, with 2 df ($p > 0.60$), and the null hypothesis cannot be rejected at the 0.05 significance level. Thus a marginal logistic regression model without the gender $\times$ age interactions is defensible. Note that the $\widehat{\alpha}_{jk}$ have interpretation in terms of the pairwise log odds ratio for the responses at the j^{th} and k^{th} occasions. The pairwise log odds ratios between adjacent occasions are

[1] Although the (log) odds ratio is a preferable metric for within-subject association among pairs of binary responses, implementations of GEE with odds ratios for within-subject association are currently incorporated in only a few statistical software packages (e.g., the `LOGOR` option in PROC GENMOD in SAS, the `ordgee` function in the `geepack` package in R, and the `alr` package in R and S-Plus). However, because statistical software is contantly evolving, we anticipate that implementations of GEE with odds ratios for within-subject association will soon become available within most of the major statistical packages.

Table 13.2 Parameter estimates and standard errors from marginal logistic regression model for the obesity data from the Muscatine Coronary Risk Factor study.

Variable	Estimate	SE[a]	Z
Intercept	−1.2135	0.0506	−24.00
Gender	0.1159	0.0711	1.63
Age	0.0378	0.0133	2.85
Age^2	−0.0175	0.0034	−5.19
Gender × Age	0.0075	0.0182	0.41
Gender × Age^2	0.0039	0.0046	0.85
α_{12}	3.1528	0.1280	24.63
α_{13}	2.5975	0.1353	19.20
α_{23}	2.9868	0.1236	24.17

[a]SE based on "sandwich" variance estimator.

very similar and approximately equal to 3, indicating that the odds ratio for within-subject association is approximately 20 (or e^3). As expected, there is strong positive association among the indicators of obesity status at the three measurement occasions.

Recall that these data on obesity are from five cohorts of children, initially aged 5–7, 7–9, 9–11, 11–13, and 13–15 years. Our analysis of the trends in the risk of obesity with age implicitly assumes that there are no cohort effects. That is, the logistic model for the probability of obesity assumes that the cross-sectional and longitudinal effects of aging are identical (see Chapter 9, Section 9.5). Following the approach used in the analysis of the FEV_1 data in Section 9.6, we can conduct a formal test of equality of the cross-sectional and longitudinal effects of aging by including linear and quadratic effects of both mean age (where averaging is over time) and current age minus mean age (and also their interactions with gender) in the logistic model,

$$\log\left\{\frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)}\right\} = \beta_1 + \beta_2 \text{Gender}_i + \beta_3 \overline{\text{Age}}_i + \beta_4 \overline{\text{Age}_i^2} + \beta_5 \text{Gender}_i \times \overline{\text{Age}}_i$$
$$+ \beta_6\, \text{Gender}_i \times \overline{\text{Age}_i^2} + \beta_7\,(\text{Age}_{ij} - \overline{\text{Age}}_i) + \beta_8\,(\text{Age}_{ij}^2 - \overline{\text{Age}_i^2})$$
$$+ \beta_9\, \text{Gender}_i \times (\text{Age}_{ij} - \overline{\text{Age}}_i) + \beta_{10}\, \text{Gender}_i \times (\text{Age}_{ij}^2 - \overline{\text{Age}_i^2}),$$

where $\overline{\text{Age}}_i = \frac{1}{3}\sum_{j=1}^{3} \text{Age}_{ij}$ and $\overline{\text{Age}_i^2} = \frac{1}{3}\sum_{j=1}^{3} \text{Age}_{ij}^2$. This model distinguishes between the cross-sectional effects of aging ($\beta_3, \beta_4, \beta_5$, and β_6) and the longitudinal

Table 13.3 Parameter estimates and standard errors from marginal logistic regression model for the obesity data from the Muscatine Coronary Risk Factor study, omitting gender × age and gender × age^2 interactions.

Variable	Estimate	SE[a]	Z
Intercept	−1.2283	0.0477	−25.75
Gender	0.1449	0.0627	2.31
Age	0.0418	0.0091	4.60
Age^2	−0.0155	0.0023	−6.73
α_{12}	3.1496	0.1280	24.61
α_{13}	2.5931	0.1352	19.17
α_{23}	2.9878	0.1236	24.18

[a] SE based on "sandwich" variance estimator.

effects of aging ($\beta_7, \beta_8, \beta_9$, and β_{10}). Note that, when $\beta_3 = \beta_7$, $\beta_4 = \beta_8$, $\beta_5 = \beta_9$, and $\beta_6 = \beta_{10}$, we obtain the logistic model considered previously. A test of equality of the cross-sectional and longitudinal effects of aging,

$$\mathrm{H}_0\colon (\beta_3 - \beta_7) = (\beta_4 - \beta_8) = (\beta_5 - \beta_9) = (\beta_6 - \beta_{10}) = 0,$$

produces a (multivariate) Wald statistic, $W^2 = 2.06$, with 4 df, ($p > 0.70$). This suggests that the results for aging presented in Table 13.2 are not confounded by cohort effects.

Next we consider a marginal logistic regression model without the gender × age interactions. Specifically, we consider the following model for the log odds of obesity:

$$\log\left\{\frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)}\right\} = \beta_1 + \beta_2\,\mathrm{Gender}_i + \beta_3\,\mathrm{Age}_{ij} + \beta_4\,\mathrm{Age}_{ij}^2,$$

while retaining the same assumptions about the variances and pairwise log odds ratios. The estimated regression coefficients (and pairwise log odds ratios) for this model are presented in Table 13.3. The estimated effect of age^2 is significant at the 0.05 level and these results provide evidence that the log odds of obesity increases from 6 to 12 years, levels off between age 12 to age 14, and declines between 14 to 18 years. Although the rates of obesity are significantly higher for girls at all ages, the patterns of change in the rates of obesity over time do not depend on gender. To translate these results on to a more interpretable scale, we can estimate the probability of obesity at each age for boys and girls,

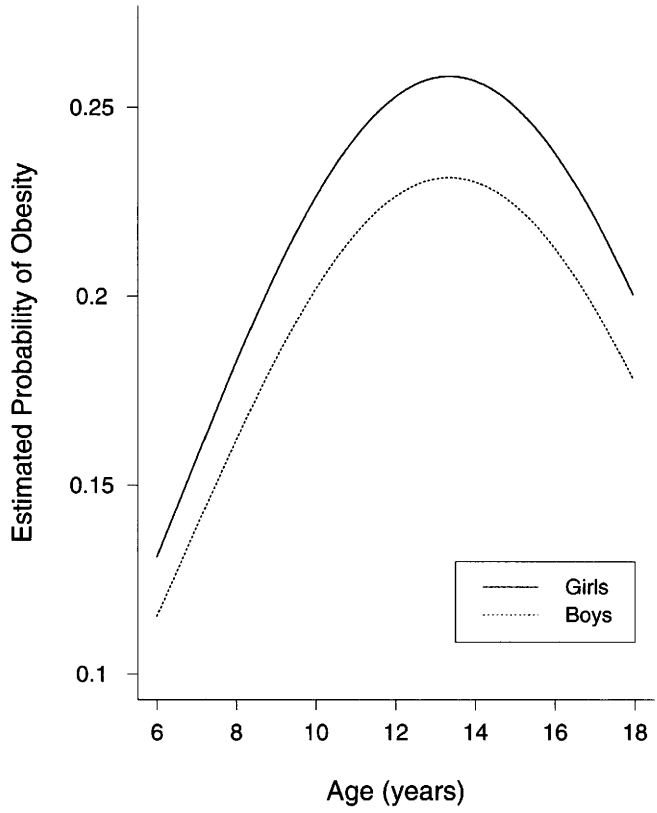

Fig. 13.1 Estimated probability of obesity versus age for boys and girls in the Muscatine Coronary Risk Factor study.

$$\widehat{\mu}_{ij} = \frac{e^{\widehat{\beta}_1 + \widehat{\beta}_2 \,\mathrm{Gender}_i + \widehat{\beta}_3 \,\mathrm{Age}_{ij} + \widehat{\beta}_4 \,\mathrm{Age}_{ij}^2}}{1 + e^{\widehat{\beta}_1 + \widehat{\beta}_2 \,\mathrm{Gender}_i + \widehat{\beta}_3 \,\mathrm{Age}_{ij} + \widehat{\beta}_4 \,\mathrm{Age}_{ij}^2}}.$$

For example, the estimated probability of obesity for boys at ages 6, 10, 14, and 18 is 0.12, 0.20, 0.23, and 0.18, respectively; for girls, the estimated probability of obesity at ages 6, 10, 14, and 18 is 0.13, 0.22, 0.26, and 0.20, respectively (see Figure 13.1). Note that with the logistic model, an additive effect of gender does not translate into a constant difference over time in the probability of obesity. Potential confounding of these trends by cohort effects can be examined by including linear and quadratic effects of both mean age (where averaging is over time) and current age minus mean age in the logistic model. A test of equality of the cross-sectional and longitudinal effects of aging produces a (multivariate) Wald statistic, $W^2 = 1.74$,

Table 13.4 Parameter estimates and standard errors from marginal logistic regression model for the obesity data from the Muscatine Coronary Risk Factor study, omitting gender × age and gender × age^2 interactions.

Variable	Estimate	SE[a]	Z
Intercept	−1.2270	0.0477	−25.72
Gender	0.1445	0.0627	2.31
Age	0.0416	0.0091	4.58
Age^2	−0.0156	0.0023	−6.77
α_1	3.0684	0.0957	32.07
α_2	2.5929	0.1353	19.17

[a] SE based on "sandwich" variance estimator.

with 2 df, ($p > 0.40$), suggesting that the results for aging presented in Table 13.3 are not confounded by cohort effects.

For illustrative purposes, we consider the same model for the log odds of obesity,

$$\log\left\{\frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)}\right\} = \beta_1 + \beta_2\,\text{Gender}_i + \beta_3\,\text{Age}_{ij} + \beta_4\,\text{Age}_{ij}^2,$$

and retain the same assumptions about the variances, but assume a Toeplitz pattern for the log odds ratios:

$$\log \text{OR}(Y_{ij}, Y_{ij-1}) = \alpha_1,$$
$$\log \text{OR}(Y_{ij}, Y_{ij-2}) = \alpha_2.$$

The estimated parameters for this model are displayed in Table 13.4, and the estimates of the regression parameters (and their standard errors) are very similar to those reported in Table 13.3. The estimates of the within-subject log odds ratios display a characteristic decreasing time-dependence as the time separation increases. Finally, since the Toeplitz pattern for the within-subject log odds ratios is nested within the unstructured pairwise log odds ratio model, it is possible to assess the goodness of fit of the Toeplitz model. That is, by appropriately reparameterizing the unstructured pairwise log odds ratio model,

$$\log \text{OR}(Y_{i1}, Y_{i2}) = \alpha_1,$$
$$\log \text{OR}(Y_{i1}, Y_{i3}) = \alpha_2,$$
$$\log \text{OR}(Y_{i2}, Y_{i3}) = \alpha_1 + \alpha_3,$$

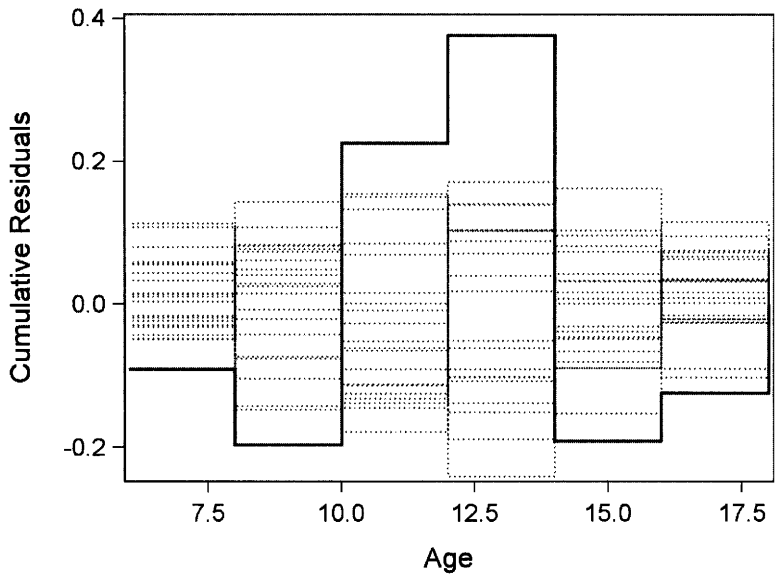

Fig. 13.2 Plot of observed cumulative sum of residuals versus age (solid curve) and 20 simulated realizations (dotted curves) from the null distribution assuming a correctly specified model for the log odds of obesity (with quadratic trend for age).

a 1-degree-of-freedom goodness-of-fit test (based on a Wald test) for the Toeplitz model can be constructed. The test of the null hypothesis, H_0: $\alpha_3 = 0$, produces a Wald $Z = -0.99$ ($p > 0.30$), indicating that the Toeplitz pattern is defensible for these data.

Finally, we consider the use of cumulative sums of residuals (see Sections 10.4 and 13.3) to assess the adequacy of the model for the probability of obesity. The results presented so far have assumed that the log odds of obesity changes curvilinearly with age, allowing the rates of obesity to increase from 6 to 12 years, level off between age 12 to age 14, and decline between 14 to 18 years (see Figure 13.1). Specifically, the curvilinear trend in the log odds of obesity is assumed to be a quadratic function of age. Figure 13.2 shows a plot of the observed cumulative sum of the residuals (solid curve), with respect to age for the quadratic trend model. On the vertical axis is the cumulative sum of residuals; the horizontal axis denotes age (in years). Superimposed on the graph are 20 simulated realizations (dotted curves) of the cumulative sum from the null distribution under the assumption that the model for the log odds of obesity is correctly specified. The realizations of the cumulative sum under the null are computer simulated from the appropriate Gaussian mean-zero process. By comparing the observed cumulative sum to the 20 different realizations under the null, it is possible to determine whether any apparent trend is systematic or due to chance fluctuations; a more formal comparison is made using the p-value for the supremum test based on 10,000 simulated realizations from the null distribution.

Table 13.5 Parameter estimates and standard errors from marginal logistic regression model for the obesity data from the Muscatine Coronary Risk Factor study, assuming a cubic trend for age.

Variable	Estimate	SE[a]	Z
Intercept	−1.2228	0.0477	−25.65
Gender	0.1457	0.0627	2.33
Age	0.0078	0.0144	0.54
Age^2	−0.0166	0.0024	−6.99
Age^3	0.0018	0.0006	3.01
α_{12}	3.1501	0.1290	24.42
α_{13}	2.6135	0.1353	19.32
α_{23}	2.9933	0.1231	24.31

[a]SE based on "sandwich" variance estimator.

From Figure 13.2 it would appear that the observed cumulative sum displays a systematic pattern. In particular, the observed cumulative sum is far too large between 10 and 14 years; it also appears to be too small before 10 years and after 14 years. This suggests that the assumed functional form for age, a quadratic trend, may not be adequate. This graphical assessment of fit can be complemented by a numerical assessment. The maximum absolute value of the observed cumulative sum is 0.376. The supremum test yields a p-value of 0.0039, based on 10,000 simulated realizations of the process under the null. That is, out of 10,000 simulated realizations under the null hypothesis that a quadratic trend is adequate, only 39 had a maximum absolute value that exceeded 0.376. Thus both the graphical and numerical results suggest that the functional form for age may be inappropriate.

Next we consider a refinement to the model to allow for a cubic trend in the log odds as a function of age. Specifically, we consider the following model for the log odds of obesity:

$$\log\left\{\frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)}\right\} = \beta_1 + \beta_2\,\text{Gender}_i + \beta_3\,\text{Age}_{ij} + \beta_4\,\text{Age}_{ij}^2 + \beta_5\,\text{Age}_{ij}^3,$$

while retaining the same assumptions about the variances and pairwise log odds ratios as before. The estimated regression coefficients (and pairwise log odds ratios) for this model are presented in Table 13.5. The estimated effect of age^3 is significant at the 0.05 level ($Z = 3.01, p < 0.003$). These results provide evidence that the log odds of obesity depart from a quadratic trend in age. To translate these results on to

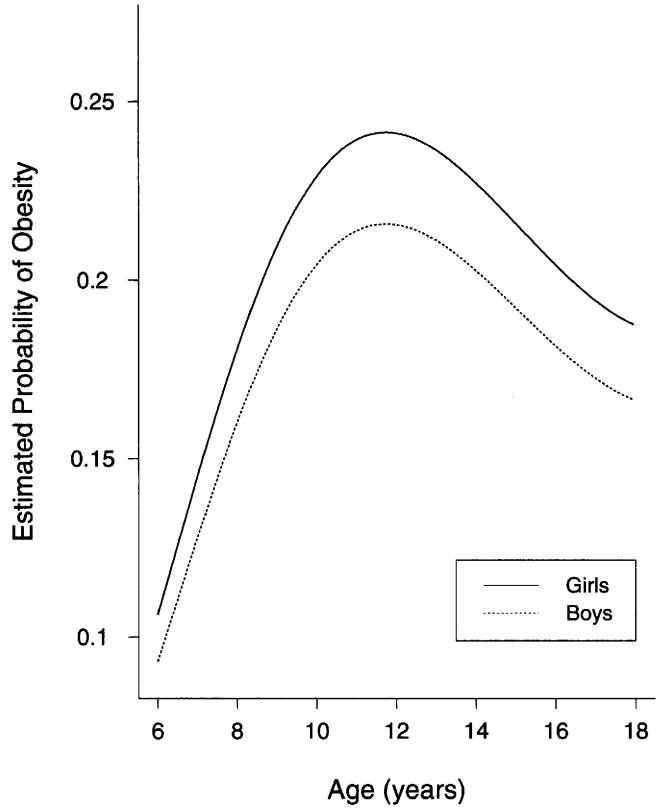

Fig. 13.3 Estimated probability of obesity versus age for boys and girls in the Muscatine Coronary Risk Factor study.

a more interpretable scale, we can plot the estimated probability of obesity at each age for boys and girls (see Figure 13.3). From Figure 13.3 it is clear that the rates of obesity increase sharply from 6 to 12 years but then decline, albeit at a slower rate, thereafter. Although the rates of obesity are significantly higher for girls at all ages, the overall pattern of change in the log odds of obesity does not depend on gender (a 3 df test of interaction with gender yields a chi-square statistic of 0.82, $p > 0.80$). Overall, these results suggest that the rates of obesity level off earlier and decline at a somewhat less steep rate than was indicated by the quadratic trend model (compare Figures 13.1 and 13.3).

We can assess the adequacy of this revised model using cumulative sums of residuals. Figure 13.4 shows a plot of the observed cumulative sum of the residuals, with respect to age; superimposed on the graph are 20 realizations from the Gaussian mean-zero null distribution. This plot suggests there is no systematic trend in the observed

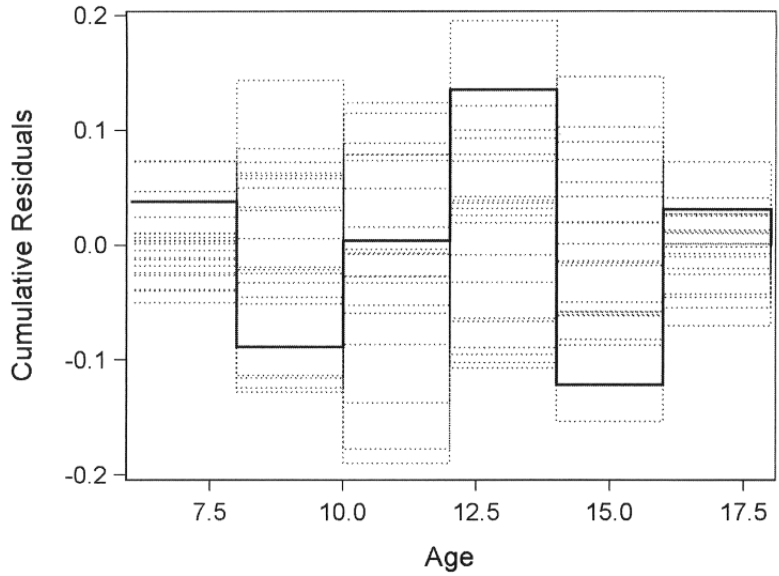

Fig. 13.4 Plot of observed cumulative sum of residuals versus age (solid curve) and 20 simulated realizations (dotted curves) from the null distribution assuming a correctly specified model for the log odds of obesity (with cubic trend for age).

curve. This is confirmed by a numerical assessment. The maximum absolute value of the observed cumulative sum is 0.135, with corresponding p-value for the supremum test equal to 0.346. With the inclusion of a cubic trend for age, both the graphical and numerical diagnostics no longer provide evidence of model misspecification.

In summary, the results indicate that the rates of obesity are significantly higher for girls at all ages. However, the overall pattern of change in the log odds of obesity does not depend on gender. For both boys and girls the rates of obesity increase sharply from 6 to 12 years but then decline, albeit at a slower rate, thereafter.

Clinical Trial of Antibiotics for Leprosy

Next we consider count data from a placebo-controlled clinical trial of 30 patients with leprosy at the Eversley Childs Sanitarium in the Philippines (Snedecor and Cochran, 1967). Participants in the study were randomized to either of two antibiotics (denoted treatment drug A and B) or to a placebo (denoted treatment drug C). Prior to receiving treatment, baseline data on the number of leprosy bacilli at six sites of the body where the bacilli tend to congregate were recorded for each patient. After several months of treatment, the number of leprosy bacilli at six sites of the body were recorded a second time. The outcome variable is the total count of the number of leprosy bacilli at the six sites.

Table 13.6 Mean count of leprosy bacilli at six sites of the body (and variance) pre- and post-treatment.

Treatment Group	Baseline	Post-Treatment
Drug A (Antibiotic)	9.3	5.3
	(22.7)	(21.6)
Drug B (Antibiotic)	10.0	6.1
	(27.6)	(37.9)
Drug C (Placebo)	12.9	12.3
	(15.7)	(51.1)

Before proceeding with the analysis, a feature of these data should be noted. These data display substantially greater variability than that predicted by the mean under a Poisson distribution assumption. The mean number of bacilli and the variance are displayed in Table 13.6. Although the sample sizes are relatively small, these descriptive statistics reveal that the variances are substantially greater than the means. As a result a Poisson assumption for the variance, with $\text{Var}(Y_{ij}) = \mu_{ij}$, is not appropriate for these data. Instead, we consider

$$\text{Var}(Y_{ij}) = \phi\, \mu_{ij},$$

where it is assumed that $\phi > 1$.

In this study, the question of main scientific interest is whether treatment with antibiotics (drugs A and B) reduces the abundance of leprosy bacilli at the six sites of the body when compared to placebo (drug C). To address this question we can compare the changes, from baseline to follow-up, in the average count of leprosy bacilli in the three treatment groups. This can be expressed in the following marginal model for the expected counts of leprosy bacilli

$$\log E(Y_{ij}) = \log \mu_{ij} = \beta_1 + \beta_2 \,\text{Time}_{ij} + \beta_3 \,\text{Time}_{ij} \times \text{Trt}_{1i} + \beta_4 \,\text{Time}_{ij} \times \text{Trt}_{2i},$$

where Y_{ij} is the count of the number of leprosy bacilli for the i^{th} patient in the j^{th} period of observation ($j = 1, 2$). The variables Trt_1 and Trt_2 are indicator variables for drugs A and B respectively, with $\text{Trt}_1 = 1$ if a patient was randomized to drug A and $\text{Trt}_1 = 0$ otherwise, and $\text{Trt}_2 = 1$ if a patient was randomized to drug B and $\text{Trt}_2 = 0$ otherwise. The binary variable, Time, denotes the baseline and post-treatment follow-up periods, with $\text{Time} = 0$ for the baseline period (period 1) and $\text{Time} = 1$ for the post-treatment follow-up period (period 2). Because patients were

Table 13.7 Parameters of the marginal log-linear regression model for the leprosy bacilli data.

Treatment Group	Period	$\log(\mu_{ij})$
Drug A (Antibiotic)	Baseline	β_1
	Follow-up	$\beta_1 + \beta_2 + \beta_3$
Drug B (Antibiotic)	Baseline	β_1
	Follow-up	$\beta_1 + \beta_2 + \beta_4$
Drug C (Placebo)	Baseline	β_1
	Follow-up	$\beta_1 + \beta_2$

randomized to one of the three treatments, the model does not include main effects of treatment (since the mean count of the number of leprosy bacilli at baseline can be assumed to be equal in the three treatment groups). To complete the specification of the model, we must make assumptions about the variances of the counts and the within-subject association among the repeated counts. Because of the discernible *overdispersion* in these data (relative to Poisson variability), we assume that the variance of Y_{ij} is given by

$$\text{Var}(Y_{ij}) = \phi\, \mu_{ij},$$

where ϕ can be thought of as an overdispersion factor. Finally, the within-subject association is accounted for by assuming a common correlation,

$$\text{Corr}(Y_{i1}, Y_{i2}) = \alpha.$$

In this marginal model for the expected number of leprosy bacilli, all of the covariates are dichotomous and the log-linear regression parameters can be given interpretations in terms of (log) rate ratios. In Table 13.7 we summarize the interpretation of β in terms of the log expected counts in the three groups at baseline and during post-treatment follow-up. So, for example, the expected count of leprosy bacilli at the six sites of the body at baseline in the placebo group (drug C) is e^{β_1}, while the expected count during the follow-up period is $e^{\beta_1+\beta_2}$. Thus e^{β_2} is the rate ratio of leprosy bacilli, comparing the follow-up period to baseline, in the placebo group (drug C). Similarly $e^{\beta_2+\beta_3}$ is the rate ratio of leprosy bacilli, comparing the follow-up period to baseline, in the group randomized to drug A. Finally, $e^{\beta_2+\beta_4}$ is the rate ratio of leprosy bacilli, comparing the follow-up period to baseline, in the group randomized to drug B.

As a result a direct comparison of the three treatment groups in terms of changes in the expected rates of leprosy bacilli is expressible in terms of β_3 and β_4. That is, β_3 and β_4 represents the difference between the changes in the log expected rates, comparing

Table 13.8 Parameter estimates and standard errors (based on sandwich variance estimator) from marginal log-linear regression model for the leprosy bacilli data.

Variable	Estimate	SE	Z
Intercept	2.3734	0.0801	29.62
Time	−0.0138	0.1573	−0.09
Time × Trt_1	−0.5406	0.2186	−2.47
Time × Trt_2	−0.4791	0.2279	−2.10

Note: Estimated scale or dispersion parameter: $\widehat{\phi} = 3.45$. Estimated working correlation: $\widehat{\alpha} = 0.797$.

drug A and B to the placebo (drug C). For example, a value of $\beta_3 < 0$ indicates a greater reduction in the rate of bacilli from baseline in the group randomized to drug A (when compared to the placebo group).

The estimated regression coefficients, obtained using the GEE approach, are displayed in Table 13.8 (with standard errors based on the "sandwich" estimator). A test of the null hypothesis, H_0: $\beta_3 = \beta_4 = 0$, produces a (multivariate) Wald statistic, $W^2 = 6.99$, with 2 degrees of freedom ($p < 0.05$). This indicates that treatment with antibiotics significantly reduces the abundance of leprosy bacilli at the six sites of the body. A test of the null hypothesis that both antibiotics are equally effective, H_0: $\beta_3 = \beta_4$, produces a Wald statistic, $W^2 = 0.08$, with 1 degree of freedom ($p > 0.7$). Thus we cannot reject the null hypothesis that the two antibiotics are equally effective in reducing the number of leprosy bacilli. To obtain a common estimate of the log rate ratio, comparing both antibiotics (drugs A and B) to placebo, we can fit the reduced model

$$\log E(Y_{ij}) = \log \mu_{ij} = \beta_1 + \beta_2 \,\text{Time}_{ij} + \beta_3 \,\text{Time}_{ij} \times \text{Trt}_i,$$

where the variable Trt is an indicator variable for antibiotics, with $\text{Trt} = 1$ if a patient was randomized to either drug A or B and $\text{Trt} = 0$ otherwise. We retain the same assumptions about the variance and correlation as before.

The estimated regression coefficients are displayed in Table 13.9 (with standard errors based on the "sandwich" estimator). The common estimate of the log rate ratio, comparing post-treatment rates of bacilli in the antibiotics group (drugs A and B) to placebo, is −0.5141. Thus the rate ratio is 0.60 (or $e^{-0.5141}$), with 95% confidence interval, 0.41 to 0.88, indicating that treatment with antibiotics significantly reduces the average number of bacilli when compared to placebo. In the placebo group, there is a non-significant reduction in the average number of bacilli of approximately 1% (or $[1 - e^{-0.0108}] \times 100\%$), while in the antibiotics group there is a significant reduction of approximately 40% (or $[1 - e^{-0.0108-0.5141}] \times 100\%$).

Table 13.9 Parameter estimates and standard errors (based on sandwich variance estimator) from marginal log-linear regression model for the leprosy bacilli data, with common estimate of the effect of antibiotics.

Variable	Estimate	SE	Z
Intercept	2.3734	0.0801	29.62
Time	−0.0108	0.1572	−0.07
Time × Trt	−0.5141	0.1966	−2.62

Note: Estimated scale or dispersion parameter: $\widehat{\phi} = 3.41$.

Table 13.10 Parameter estimates and model-based standard errors from marginal log-linear regression model for the leprosy bacilli data, with common estimate of the effect of antibiotics.

Variable	Estimate	SE	Z
Intercept	2.3734	0.0557	42.59
Time	−0.0108	0.0619	−0.17
Time × Trt	−0.5141	0.0832	−6.18

Note: Fixed scale or dispersion parameter: $\phi = 1$.

Finally, the estimated pairwise correlation is relatively large (approximately 0.8), suggesting that there may be substantial heterogeneity among patients in their disease severity. Of note, the estimated scale parameter is approximately 3.4, revealing overdispersion relative to that predicted by Poisson variability. Recall from Section 13.2 that in addition to correcting for misspecification of the within-subject association, the "sandwich" estimator of the standard errors also corrects for any misspecification of the variance, including overdispersion. Therefore standard errors based on the "sandwich" estimator are automatically adjusted for potential overdispersion; it is not necessary to include an additional parameter, ϕ, to account for overdispersion. It is instructive to compare the standard errors in Table 13.9 with the corresponding model-based standard errors when ϕ is fixed at 1 (Poisson variance assumption) and when ϕ is estimated from the data (allowing for overdispersion by a constant factor $\phi > 0$). Results for the former are presented in Table 13.10; results for the latter are presented in Table 13.11. First, note that regardless of whether ϕ is fixed at 1 or estimated as an additional parameter, the estimates of the regression parameters, $\widehat{\beta}$, are the same. Second, the model-based standard errors in Table 13.10 are discernibly

Table 13.11 Parameter estimates and model-based standard errors from marginal log-linear regression model for the leprosy bacilli data, with common estimate of the effect of antibiotics.

Variable	Estimate	SE	Z
Intercept	2.3734	0.1028	23.08
Time	−0.0108	0.1142	−0.09
Time × Trt	−0.5141	0.1536	−3.35

Note: Estimated scale or dispersion parameter: $\widehat{\phi} = 3.41$.

smaller than those in Table 13.9, reflecting the fact that they are based on the assumption of Poisson variability. Third, the model-based standard errors in Table 13.11 are larger than those reported in Table 13.10 by a constant factor of 1.845 (or $\sqrt{3.41}$). Although the standard errors in Tables 13.11 and 13.9 have both made adjustments for overdispersion, the former corrections are based on a constant multiple of the standard errors obtained under a Poisson variance assumption whereas the latter corrections allow for more general departures from Poisson variability. For example, standard errors based on the "sandwich" estimator make corrections for overdispersion that may be differential by treatment group. Finally, we note that overdispersion relative to Poisson variation can also be accounted for by assuming that the counts have negative binomial variance. Assuming negative binomial variance for the counts of leprosy bacilli (and a log link function) yields estimates of the regression parameters and standard errors that are qualitatively very similar to those in Table 13.11. Recall that the negative binomial variance allows for overdispersion by assuming the variance increases as a *quadratic* function of the mean (see Section 11.5). However, as with the use of a constant scale factor (which implicitly assumes the variance increases as a *linear* function of the mean), this correction for overdispersion does not allow for more general departures from Poisson variability. In contrast, corrections based on the "sandwich" variance estimator allow for *any* departures from Poisson variability when the conditions required for its use are met (see Section 13.2).

Arthritis Clinical Trial

The final example is from a longitudinal clinical trial comparing auranofin therapy (3 mg of oral gold, twice daily) and placebo for the treatment of rheumatoid arthritis (Bombardier et al., 1986). In this six-month, randomized, double-blind trial, 303 patients with classic or definite rheumatoid arthritis were randomized to one of the two treatment groups and followed over time. The outcome variable of interest is a global impression scale (Arthritis Categorical Scale) measured at baseline (month 0), month 2, month 4, and month 6. This is a self-assessment of a patient's current arthritis, mea-

sured on a five-level ordinal scale: (1) very good, (2) good, (3) fair, (4) poor, and (5) very poor. Baseline data on this outcome variable are available for 303 of the patients who participated in this trial; follow-up data at 6 months are available for 294 patients.

The goal of the analysis is to assess changes in the odds of a more favorable response over the duration of the study, and to determine whether treatment with auranofin has an influence on these changes. Letting Y_{ij} denote the ordinal response for the i^{th} subject at the j^{th} occasion, we assume that the log odds of a more favorable response at each occasion follows the proportional odds model

$$\log\left\{\frac{\Pr(Y_{ij} \leq k)}{\Pr(Y_{ij} > k)}\right\} = \alpha_k + \beta_1 \,\mathrm{Trt}_i + \beta_2 \sqrt{\mathrm{Month}_{ij}} + \beta_3 \,\mathrm{Trt}_i \times \sqrt{\mathrm{Month}_{ij}},$$

where Month_{ij} is the timing of the measurement, in months, for the i^{th} subject at the j^{th} occasion, $\mathrm{Trt}_i = 1$ if the i^{th} subject is randomized to auranofin, and $\mathrm{Trt}_i = 0$ if randomized to placebo. This specifies the first component of a marginal model, the model for the mean response. Specifically, the model assumes that the log odds of a favorable response changes linearly with (square-root transformed) time, but the slopes over time are allowed to differ between the two treatment groups. Preliminary analyses indicated that changes in the log odds of a more favorable response were approximately linear in square-root transformed time; a 2 df Wald test of non-linearity (quadratic trend) yielded $W^2 = 0.18, p > 0.90$. The second component of the marginal model, the variance of the multinomial response, is completely determined by the model for the mean response. For the third component, we initially make a "working independence" assumption for the within-subject association among the repeated ordinal responses and rely on the empirical variance estimator for making valid inferences.

The GEE estimates in Table 13.12 indicate that the trajectories for the log odds over time are significantly different for patients treated with placebo versus patients treated with auranofin ($Z = 2.66, p < 0.01$). Specifically, patients treated with auranofin have a significantly greater increase in the odds of a more favorable response over the duration of the study. Relative to baseline the odds of a more favorable response at month 6 has increased by a factor of 1.84 (or $e^{0.2481 \times \sqrt{6}}$) for the placebo group but by a factor of 3.27 (or $e^{(0.2481+0.2354) \times \sqrt{6}}$) for the auranofin group. At the completion of the trial, patients treated with auranofin are approximately twice (or $e^{0.2354 \times \sqrt{6}} = 1.78$) as likely to have a more favorable response when compared to patients treated with placebo. Not surprisingly, due to the randomization, $\widehat{\beta}_1 \approx 0$ indicating that the two groups have a similar log odds of a favorable response at baseline (month 0).

Finally, for illustrative purposes we re-fit the model with an unstructured pattern for the within-subject association among the repeated ordinal responses. In general, specification of a "working covariance" for ordinal responses (other than "working independence") is very challenging because it requires the specification and estimation of a large number of parameters. With n repeated measures of a K-level ordinal response, there are $(K-1)^2 \times n \times (n-1)/2$ pairwise parameters. In our example with four repeated measures of a five-level ordinal response, an unstructured pattern

Table 13.12 GEE estimates and standard errors (empirical) from the proportional odds model for the arthritis clinical trial data.

Variable	Estimate	SE	Z
α_1	−3.1902	0.1994	−16.00
α_2	−1.2042	0.1523	−7.91
α_3	0.5736	0.1464	3.92
α_4	2.4770	0.1995	12.42
Trt	0.0714	0.1975	0.36
$\sqrt{\text{Month}}$	0.2481	0.0613	4.05
Trt × $\sqrt{\text{Month}}$	0.2354	0.0883	2.66

has 96 pairwise association parameters that require estimation. Fitting a model with an unstructured pattern yielded estimates and standard errors very similar to those reported in Table 13.12. For the effect of main interest, the Trt × $\sqrt{\text{Month}}$ interaction, the analysis with unstructured pattern for the within-subject association yields $\widehat{\beta}_3 = 0.2377$ (model-based SE = 0.0869, empirical SE = 0.0877). This estimate of β_3 is similar to the estimate reported in Table 13.12 ($\widehat{\beta}_3 = 0.2354$) under a "working independence" assumption for the within-subject association. In addition the empirical SE for $\widehat{\beta}_3$ reported in Table 13.12 (empirical SE = 0.0883) is very similar to both the model-based and empirical standard errors obtained from the analysis with unstructured pattern for the within-subject association. The fact that the empirical standard errors are so similar under these two "working" assumptions for the within subject association suggests that there has been negligible loss of efficiency from basing the analysis on a "working independence" assumption; however, we caution that this cannot be expected in general.

13.5 MARGINAL MODELS AND TIME-VARYING COVARIATES

In this section[†] we return to an implicit assumption in marginal models that was highlighted in the previous chapter at the end of Section 12.2. Marginal models assume that the conditional mean of the j^{th} response, given $X_{i1}, ..., X_{in_i}$, depends

[†]This section examines in detail the implicit assumption in marginal models concerning time-varying covariates. This assumption has important implications for estimation and interpretation of time-varying covariate effects. Although the content of this section is somewhat challenging, and can be omitted on first reading without loss of continuity, we strongly encourage the reader to return to this section.

only on X_{ij}. As we will see, this assumption has some important ramifications for time-varying covariates in marginal models. Recall that the vector of covariates at the j^{th} occasion, X_{ij}, includes two main types of covariates: covariates whose values do not change throughout the duration of the study and covariates whose values change over time. The former are referred to as time-invariant or between-subject covariates (e.g., gender and fixed experimental treatments), while the latter are referred to as time-varying or within-subject covariates (e.g., time since baseline, current smoking status, and environmental exposures). In this section we consider estimation and aspects of interpretation of time-varying covariates in marginal models. We want to emphasize at the outset that the issues discussed here in the context of marginal models apply equally to the linear models for longitudinal data in Part II of the book.

When considering time-varying covariates, we can distinguish covariates that vary systematically over time but are fixed by design of the study and covariates that vary randomly over time. An example of a time-varying covariate that is fixed by design is a treatment group indicator in a crossover trial. Another example, and one that is commonly encountered in a longitudinal study, is time since baseline (when the measurement occasions are fixed by the study design). Covariates that vary randomly over time are often referred to as *stochastic*, that is, values of the covariate at any occasion cannot be precisely predicted since they are governed by a random mechanism. An example of a time-varying covariate that is stochastic is current blood glucose level. In an observational study of diabetics, participants' blood sugar levels can vary randomly over the duration of the study. Additional examples include current smoking status or cumulative pack years, blood pressure, cholesterol level, fat intake, and exposure to environmental pollutants. As we will later see, when a covariate is both time-varying and stochastic, new issues arise concerning the interpretation and estimation of regression parameters in marginal models for longitudinal data.

Marginal models for the mean response described in this and earlier chapters can be specified as

$$g(\mu_i) = g\{E(Y_i|X_i)\} = X_i\beta, \tag{13.4}$$

for some known link function $g(\cdot)$. This use of vector and matrix notation implies that the model for the mean at each occasion is given by

$$g(\mu_{ij}) = g\{E(Y_{ij}|X_i)\} = X'_{ij}\beta, \quad (j = 1, ..., n_i).$$

However, what is often overlooked is the implicit assumption that the conditional mean of the j^{th} response, given $X_{i1}, ..., X_{in_i}$, depends only on X_{ij}

$$E(Y_{ij}|X_i) = E(Y_{ij}|X_{i1}, ..., X_{in_i}) = E(Y_{ij}|X_{ij}). \tag{13.5}$$

With time-invariant covariates, this assumption necessarily holds since $X_{ij} = X_{ik}$ for all occasions $k \neq j$. Also, with time-varying covariates that are fixed by design of the study (e.g., treatment group indicator in a crossover trial), the assumption also holds since values of the covariates at any occasion are determined a priori by study design and in a manner completely unrelated to the longitudinal response. However, when a covariate is time-varying and stochastic (13.5) may not necessarily hold. For

example, the assumption will be violated when the current value of Y_{ij}, given X_{ij}, predicts the subsequent value of X_{ij+1}. In that case

$$E(Y_{ij}|X_{ij}, X_{ij+1}) \neq E(Y_{ij}|X_{ij}),$$

and X_{ij+1} is said to confound the relationship between Y_{ij} and X_{ij}. In general, when (13.5) does not hold, then preceding and/or subsequent values of the time-varying covariate confound the relationship between Y_{ij} and X_{ij}; this can lead to biased estimates of β in the marginal model given by (13.4).

To fix ideas, consider a longitudinal study designed to examine the effects of physical exercise on reducing blood glucose levels in patients with type 2 diabetes mellitus. We let X_{ij} denote the cumulative amount of physical activity at the j^{th} occasion and Y_{ij} denote a measure of blood glucose. The goal of the study is to determine the relationship between Y_{ij} and X_{ij}. Next suppose that subjects with elevated blood glucose levels at the j^{th} occasion subsequently increase their level of physical activity, while subjects with the same cumulative amount of physical activity at the j^{th} occasion, but with normal blood glucose levels, continue to maintain their usual level of physical activity. Then the assumption that

$$E(Y_{ij}|X_{i1}, ..., X_{in_i}) = E(Y_{ij}|X_{ij})$$

does not hold and the relationship between Y_{ij} and X_{ij} is confounded by X_{ij+1}. In particular, the strength of the relationship between Y_{ij} and X_{ij} will be underestimated if subjects with elevated blood glucose levels subsequently increase their amount of physical activity.

In general, when a covariate is time-varying and stochastic, much greater care is needed in modeling its relationship to the response variable. It is important to assess the assumption made in (13.5), namely that the conditional mean of Y_{ij}, given the entire time-varying covariate profile $X_{i1}, ..., X_{in_i}$, depends only on the covariate value at the j^{th} occasion, X_{ij}. Note, however, that X_{ij} can be defined in terms of functions of the explanatory variables measured at or preceding the j^{th} occasion (e.g., cumulative exposure at the j^{th} occasion). When (13.5) is violated the relationship between the mean of Y_{ij} and X_{ij}, expressed in terms of β, will be confounded by preceding and/or subsequent values of the covariate and misleading inferences about β can result.

Finally, even when (13.5) holds, there can be problems with the *interpretation* of regression parameters relating the mean response to stochastically time-varying covariates. In particular, the regression parameters β in (13.4) may not have the implied causal interpretation. For example, the model given by (13.4) may correctly specify the relationship between mean blood glucose level and physical activity at the last measurement occasion, since at the last occasion X_{in_i}, the cumulative amount of physical activity, is a function of the entire time-varying covariate profile. However, even though (13.5) holds, the regression parameters β may not have the implied causal interpretation without making additional assumptions. To see why, let us consider a simplified version of the example discussed earlier.

Suppose that a group of diabetics are measured at two occasions. Let Y_{i1} and Y_{i2} denote the blood glucose levels at baseline and follow-up, and X_{i1} and X_{i2} denote

measures of physical activity at the two occasions. Suppose that it is of interest to determine the association between the cumulative amount of physical activity, $X_i^* = X_{i1} + X_{i2}$, and blood glucose level at the completion of the study, Y_{i2}. The following model is assumed:

$$E(Y_{i2}|X_i^*) = \beta_1 + \beta_2 X_i^*,$$

where, for ease of exposition, an identity link function is assumed. In this model β_2 appears to have interpretation as the effect of a unit increase in the cumulative amount of physical activity on the mean blood glucose level at follow-up, since

$$E(Y_{i2}|X_i^* = x + 1) - E(Y_{i2}|X_i^* = x) = \beta_2.$$

However, because X_{ij} is time-varying and stochastic, this interpretation of β_2 rests on the validity of *either* of the following two assumptions: (1) Y_{i2} is not predicted by Y_{i1}, given X_{i1} and X_{i2}, or (2) X_{i2} is not predicted by Y_{i1}, given X_{i1}. In particular, if neither of these assumptions holds, Y_{i1} "confounds" the relationship between Y_{i2} and X_i^* and β_2 does not have the desired causal interpretation. We loosely use the term "confounding" to emphasize that Y_{i1} obscures or distorts the association of real scientific interest between Y_{i2} and X_i^*. Strictly speaking, Y_{i1} can be considered both a "confounder" and an "intermediate variable" on the causal path between Y_{i2} and X_i^*. When Y_{i1} is both a confounder and an intermediate variable, standard methods of adjustment for confounding no longer apply (since Y_{i1} is predicted by X_{i1}, and so should not be adjusted for, but also predicts X_{i2}, and so should be adjusted for in the analysis of the association between Y_{i2} and X_i^*). Instead, advanced statistical methods for causal inference (e.g., marginal structural models and structural nested models; see references at end of chapter) are required when neither assumption 1 or 2 holds. However, a discussion of statistical methods for causal inference is beyond the scope of this chapter; some references to the statistical literature on this topic appear at the end of the chapter.

Let us consider these two assumptions in context. In a longitudinal study, it is unlikely that assumption 1 would ever hold, since the repeated responses are usually positively correlated (given the covariates, X_{i1} and X_{i2}). Therefore the causal interpretation of β_2 usually rests on the validity of assumption 2. For example, assumption 2 would be violated if subjects with elevated blood glucose levels at baseline subsequently increase their level of physical activity, while subjects with the same amount of physical activity at baseline, but with normal blood glucose levels, continue to maintain their usual level of physical activity. When assumption 2 holds, the covariate is said to be *external* with respect to the response variable and β has the desired causal interpretation.

In summary, when a covariate is both time-varying and stochastic, we must consider the relationship between the response at any occasion, say Y_{ij}, and the subsequent value of the covariate, X_{ij+1}. A time-varying covariate is said to be *external* when the current and preceding values of the response at the j^{th} occasion $(Y_{i1}, ..., Y_{ij})$, given the current and preceding values of the time-varying covariate $(X_{i1}, ..., X_{ij})$, do not predict the subsequent value of X_{ij+1}. More formally, a time-varying covariate

is *external* (or sometimes referred to as *exogenous*) when

$$f(X_{ij+1}|X_{i1},...,X_{ij},Y_{i1},...,Y_{ij}) = f(X_{ij+1}|X_{i1},...,X_{ij}); \tag{13.6}$$

otherwise, the covariate is said to be *internal* (or *endogenous*). This generalizes assumption 2. Note that when a covariate is external,

$$E(Y_{ij}|X_i) = E(Y_{ij}|X_{i1},...,X_{in_i}) = E(Y_{ij}|X_{i1},...,X_{ij}),$$

which is a weaker assumption than (13.5). An example of an external covariate is air pollution in studies of children's lung function growth. The outdoor levels of air pollutants (e.g., ozone, fine suspended particulate matter, and sulfur dioxide) are time-varying and stochastic, but conditional on past values, future values are not predicted by the lung function responses of the study participants and (13.6) holds. Note, however, that children's personal exposure to air pollution would not be considered an external covariate if children with poor lung function growth subsequently altered their daily behavior (e.g., spending less time outdoors) to avoid exposure to high levels of air pollution. In principle, it is possible to examine the assumption that a time-varying covariate is *external* by considering regression models for the dependence of X_{ij} on $Y_{i1},...,Y_{ij-1}$ (or some known function(s) of $Y_{i1},...,Y_{ij-1}$) and $X_{i1},...,X_{ij-1}$ (or some known function(s) of $X_{i1},...,X_{ij-1}$). The absence of any relationships between X_{ij} and $Y_{i1},...,Y_{ij-1}$, given the preceding covariate profile, $X_{i1},...,X_{ij-1}$, provides support for the validity of the assumption that the covariate process is *external*.

In conclusion, when covariates are time-varying and stochastic the regression parameters do not necessarily have the implied causal interpretation even when (13.5) holds. The regression parameters can be given a causal interpretation only when it can be further assumed that the time-varying covariates are external with respect to the response variable (i.e., when (13.6) holds).

13.6 COMPUTING: GENERALIZED ESTIMATING EQUATIONS USING PROC GENMOD IN SAS

To fit marginal models using the generalized estimating equations approach, we can use an enhanced option for repeated measures data in the PROC GENMOD procedure in SAS. Although PROC GENMOD is primarily a procedure for fitting generalized linear models to a single response, the use of a REPEATED statement in PROC GENMOD allows for the fitting of marginal models to correlated responses using the GEE approach. In Chapter 15, Section 15.6, we describe how an alternative procedure in SAS, PROC GLIMMIX, can also fit marginal models using the GEE approach.

For example, to fit a marginal logistic regression model to longitudinal data from two groups, with the within-subject associations specified in terms of log odds ratios, we can use the illustrative SAS commands given in Table 13.13. Similarly, to fit a marginal log-linear regression model to longitudinal data from two groups, with the within-subject associations specified in terms of correlations, we can use the illustrative SAS commands given in Table 13.14. To fit a marginal proportional odds

Table 13.13 Illustrative commands for a marginal logistic regression, with within-subject associations specified in terms of log odds ratios, using PROC GENMOD in SAS.

```
PROC GENMOD DESCENDING;
    CLASS id group time;
    MODEL y=group time group*time / DIST=BINOMIAL LINK=LOGIT;
    REPEATED SUBJECT=id / WITHINSUBJECT=time LOGOR=FULLCLUST;
```

Table 13.14 Illustrative commands for a marginal log-linear regression, with within-subject associations specified in terms of correlations, using PROC GENMOD in SAS.

```
PROC GENMOD;
    CLASS  id  group time;
    MODEL  y=group  time  group*time / DIST=POISSON  LINK=LOG;
    REPEATED  SUBJECT=id / WITHINSUBJECT=time  TYPE=UN;
```

model to longitudinal ordinal data, under a "working independence" assumption for the within-subject association, we can use the illustrative SAS commands given in Table 13.15. To assess the adequacy of the functional form for the time trend in the marginal model for the mean based on cumulative sums of residuals, we can use the illustrative SAS commands given in Table 13.16. Next we describe the most salient parts of the command syntax required for fitting marginal models to longitudinal data using the GEE approach within PROC GENMOD in SAS.

PROC GENMOD <options>;

This statement calls the procedure GENMOD in SAS. It can also include an option for specifying the level of the response variable that is modeled. By default, the lower response level is modeled. For a binary response, coded (0,1), it is the probability that $Y = 0$ that is modeled. For an ordinal response, coded (1,2,...,K), the response categories are ordered from lowest to highest and the probabilities of the lower response levels are modeled. Use of the DESCENDING option reverses the default ordering of the response levels, resulting in the highest response level(s) being modeled (i.e., the probability that $Y = 1$ for binary data that are coded as 0 and 1).

CLASS variables;

The CLASS statement is used to define all variables that are to be regarded as categorical or factors. By default, this statement will create indicator variables for each factor using a reference group coding, with the last level (where "last"

Table 13.15 Illustrative commands for a marginal proportional odds regression, with a "working independence" assumption for the within-subject associations, using PROC GENMOD in SAS.

```
PROC GENMOD;
    CLASS id group;
    MODEL y=group time group*time / DIST=MULT LINK=CUMLOGIT;
    REPEATED SUBJECT=id / TYPE=IND;
```

here refers to the level with the largest alphanumeric value) regarded as the reference group. Different sort orders for the CLASS variables can be requested by the ORDER=<option> on the PROC GENMOD statement.

MODEL response = <effects> / <options>;
MODEL events/trials = <effects> / <options>;

The MODEL statement specifies the response variable and the covariate effects. The second form of the MODEL statement, with the `events/trials` syntax, allows the response to be in the form of a ratio of two variables (e.g., counts of the number of successes and the number of trials) and is used for binomial response data. The linear predictor can include both discrete (defined in the CLASS statement) and quantitative (excluded from the CLASS statement) covariates. By default, PROC GENMOD includes a column of 1's for the intercept in the model.

The option that ordinarily is used to specify the distribution of a single univariate response has a somewhat different role when fitting a marginal model using the GEE approach. The option DIST=*keyword* does not specify a distribution for the vector of correlated responses; instead, it specifies the default canonical link function and variance function that happen to be associated with particular exponential family distributions. For example, the option DIST=POISSON does not specify that the response vector (or even its separate components) has a Poisson distribution; instead, it specifies that the mean of the response vector is related to the covariates via a log link function (the canonical link for the Poisson distribution), and the mean and variance of the responses are related by $\text{Var}(Y) = E(Y) = \mu$ (i.e., the variance function is $v(\mu) = \mu$).

Note that PROC GENMOD also provides a wide choice of options for the inclusion of a dispersion parameter, ϕ. However, the scale parameter ϕ is assumed to be time-invariant. This restriction on the scale parameter is a limitation of the implementation of the GEE approach that makes it unappealing for analyzing longitudinal data when the response variable is continuous and the variance of the repeated measurements is not constant over the duration of the study (e.g., the variance of baseline measurements is often discernibly smaller than post-baseline measurements).

The LINK=*keyword* specifies the choice of built-in link function relating the mean response to the linear predictor. If the LINK=*keyword* is omitted, the default link function is the canonical link function associated with the particular exponential family distribution specified on DIST=*keyword*.

A final option often required when modeling count data is an offset. The OFFSET=*variable* specifies a variable to be used as an offset. For example, in modeling count data, the rate is often of more direct interest and the denominator for the counts or "population at risk" can be included as an offset. Note that this variable cannot be a CLASS variable, and it should not be included as one of the covariates listed on the MODEL statement.

REPEATED SUBJECT=subject-effect / <options>;

The REPEATED statement distinguishes the fitting of a generalized linear model for a single univariate responses via maximum likelihood from the fitting of a marginal model to a vector of correlated responses using the GEE approach. The REPEATED statement is used to specify the assumed structure of the within-subject association among the repeated measurements.

In particular, the REPEATED statement defines a variable that determines the clustering of observations within an individual. The latter is achieved by including a subject identifier, that distinguishes clusters of correlated responses, on the SUBJECT=*subject-effect*; this is not optional, a *subject-effect* must be included with the REPEATED statement and this variable must be listed in the CLASS statement. By including a subject identifier, pairs of observations with the same value of that variable are regarded as correlated (by virtue of arising from the same subject) while pairs of observations with distinct values are regarded as independent.

A useful option on the REPEATED statement is the WITHINSUBJECT=*within-subject effect*. With this option a variable denoting the "repeated effect" can be included and this identifies the order of the repeated measurements within subjects. In the context of longitudinal data, the "repeated effect" identifies the measurement occasions. While it is not always necessary to include this variable, failure to do so may have unforeseen consequences when there are vectors of repeated measures of different length and/or when the vector of responses are not in the same order for all subjects. To avoid any potential problems, this variable should be included on the REPEATED statement, whenever possible, to ensure that the within-subject association is estimated appropriately.

While the REPEATED statement in PROC GENMOD has a similar function to the REPEATED statement in PROC MIXED, the order in which the *subject-effect* and the *within subject-effect* appear in the REPEATED statement are reversed (for reasons perhaps best known only to the developers at SAS Institute).

By default, PROC GENMOD produces a table of regression parameter estimates, standard errors, and Z statistics. The standard errors and Z statistics are based on the empirical or "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ described in Section 13.2. Use of the REPEATED statement with the MODELSE option produces the corresponding table based on the "model-based" estimator of $\text{Cov}(\widehat{\beta})$.

Table 13.16 Illustrative commands for requesting model assessment of the functional form for the time trend based on cumulative sums of residuals, with a "working independence" assumption for the within-subject associations, using PROC GENMOD in SAS.

```
PROC GENMOD;
    CLASS id group;
    MODEL y=group time group*time / DIST=BINOMIAL LINK=LOGIT;
    REPEATED SUBJECT=id / TYPE=IND;
    ASSESS VAR(time) / RESAMPLES = 10000 SEED=7435865;
```

Finally, two additional options are used for specifying assumptions about the structure of the working correlation matrix or the log odds ratios (for binary responses only) among the repeated measurements. The TYPE=*correlation-structure* specifies the working correlation structure. PROC GENMOD provides a number of build-in correlation structures, including unstructured (UN), m-dependent (MDEP(m), where m is the order of dependence), first-order autoregressive (AR), and exchangeable (analogous to "compound symmetry") or equicorrelated (EXCH/CS). For ordinal responses with a multinomial distribution, PROC GENMOD currently only supports a "working independence" assumption (IND); indeed, one of the main challenges with extending the GEE approach to ordinal responses has been that the "working covariance" for ordinal responses (other than "working independence"), in general, requires the specification and estimation of a large number of nuisance parameters.

For binary responses only, the structure of the within-subject association among the responses can be specified in terms of log odds ratios using the LOGOR=*log odds ratio structure* option. For example, in Table 13.13, the LOGOR=FULLCLUST option estimates separate log odds ratios for all pairs of responses; this is analogous to an "unstructured" odds ratio pattern. PROC GENMOD also allows a very flexible regression structure for the log odds ratios. Note that either the TYPE=*correlation-structure* or the LOGOR=*log odds ratio structure* option should be specified, but not both. By default, a working independence structure is assumed.

ASSESS VAR=effect | LINK / <options>;

The ASSESS statement computes and plots statistics based on aggregates of residuals. Three types of aggregates are available: cumulative sums of residuals, moving sums of residuals, and lowess smoothed residuals; the default is cumulative sums. To create an analysis, either VAR=*effect* or LINK must be specified. VAR=*effect* requests that the functional form of a covariate be assessed by performing the analysis with respect to the variable identified by the effect. The effect must be specified in the MODEL statement and must be

a continuous variables. LINK requests the assessment of the link function by performing the analysis with respect to the linear predictor.
The WINDOW and LOESS options in the ASSESS statement requests model assessment based on moving sums of residuals and lowess smoothed residuals respectively.
An important option in the ASSESS statement is the RESAMPLES<=number> option; this specifies the number of paths used for computing the p-value for the supremum test (the default is 1,000 simulated paths). Another useful option is the SEED=number option; this specifies a seed for the normal random number generator used in creating simulated realizations of aggregates of residuals for plots and estimating p-values. Specifying a seed allows you to reproduce identical graphs and p-values from a later run of the procedure.

Of note, the initial output produced by PROC GENMOD is the standard output from a generalized linear model assuming that all observations are independent. The resulting estimates of the regression coefficients are used as initial values for the generalized estimating equations algorithm. However, the reader is cautioned that this initial output should be ignored. In particular, the reported value of the log-likelihood and various likelihood-based goodness of fit statistics should not be considered part of the GEE output.

13.7 FURTHER READING

Burton et al. (1998) provide an accessible introduction to generalized estimating equations. A more comprehensive description of generalized estimating equations can be found in Chapter 6 of the textbook by Myers et al. (2001); also see Chapter 11 of the textbook by Agresti (2002).

Bibliographic Notes

The early foundations for statistical methods for the analysis of repeated categorical responses can be traced to a general approach developed by Grizzle, Starmer, and Koch (1969); this approach became known as the GSK method. Koch et al. (1977) applied the GSK method to the analysis of repeated measurements. However, the application of the GSK method was limited to categorical covariates. The GEE approach overcame many of the limitations of the GSK method.

The theoretical foundation for the generalized estimating equations approach can be found in Godambe (1960) and Durbin (1960); also see Huber (1967, 1981) and White (1982). Liang and Zeger (1986) and Zeger and Liang (1986), in companion papers, proposed a class of generalized estimating equations for repeated measures and longitudinal data; see Liang and Zeger (1995) for a historical perspective on generalized estimating equations. Connections between the GEE approach and likelihood-based methods were made by Zhao et al. (1992), Fitzmaurice and Laird (1993), and Fitzmaurice et al. (1993).

The "sandwich" variance estimator was suggested by Cox (1961) and derived in Huber (1967), White (1982), Gourieroux et al. (1984), and Royall (1986); see Hinkley and Wang (1991) and Kauermann and Carroll (2001) for a discussion of properties of the "sandwich" variance estimator. For finite samples, simulation studies have shown that Wald tests using the sandwich estimator tend to be liberal, that is, have nominal p-values that are too small (see Lin and Wei, 1989; Emrich and Piedmonte, 1992; Gunsolley et al., 1995; Fay et al., 1998; Mancl and DeRouen, 2001; Fay and Graubard, 2001).

The implicit assumption in marginal regression models with time-varying covariates given by (13.5) is discussed in Fitzmaurice et al. (1993), Pepe and Anderson (1994), Robins et al. (1999), Pan et al. (2000), and Pan and Connett (2002); also see Schildcrout and Heagerty (2005) for a discussion of the bias–variance trade-off when the assumption does not hold.

A general discussion of methods for estimating the causal effect of time-varying covariates in marginal models for longitudinal data can be found in Robins et al. (1999) and the references therein; also see Chapter 23 of Fitzmaurice et al. (2009). Chapter 12 of Diggle et al. (2002) presents a useful summary of the key ideas.

Regression diagnostics for marginal models fit by generalized estimating equations were developed by Preisser and Qaqish (1996). They provide computational formulae for one-step approximations to deletion diagnostics for the influence of a single observation and for the influence of the set of correlated observations on a single individual (or cluster).

Problems

13.1 In a clinical trial of patients with respiratory illness, 111 patients from two different clinics were randomized to receive either placebo or an active treatment. Patients were examined at baseline and at four visits during treatment. At each examination, respiratory status (categorized as 1 = good, 0 = poor) was determined. These data are from Koch et al. (1990), and are reported in Davis (1991) and Stokes et al. (1995). The main objective of the analyses is to understand the joint effects of treatment and time on the probability that respiratory status is classified as good. It is also of interest to determine whether the effect of treatment is the same for patients from the two clinics.

The raw data are stored in an external file: `respir.dat`

Each row of the data set contains the following eight variables:

ID Clinic Treatment Y_0 Y_1 Y_2 Y_3 Y_4

Note: The respiratory status response variable Y_j is coded 1 = good, and 0 = poor, at the j^{th} occasion. The categorical (character) variable Treatment is coded A = Active drug, P = Placebo. The categorical variable Clinic is coded 1 = clinic 1, 2 = clinic 2.

13.1.1 Ignoring the clinic variable, consider a model for the log odds that respiratory status is classified as good, including the main effects of treatment and time (where time is regarded as a categorical variable with five levels), and their interaction.
Use generalized estimating equations (GEE), assuming separate pairwise log odds ratios (or separate pairwise correlations, if available software does not permit the within-subject association to be parameterized in terms of log odds ratios) among the five binary responses. Construct a test of the null hypothesis of no effect of treatment on *changes* in the log odds that respiratory status is classified as good based on the empirical standard errors.

13.1.2 What conclusions do you draw about the effect of treatment on changes in the log odds? Provide results that support your conclusions.

13.1.3 Patients in this trial were drawn from two separate clinics. Repeat the analysis for Problem 13.1.1, allowing the effects of treatment (and, possibly, time) to depend on clinic.

(a) Is the effect of treatment the same in the two clinics? Present results to support your conclusion.

(b) Find a parsimonious model that describes the effects of clinic, treatment, and time, on the log odds that respiratory status is classified as good. For the model selected, give a clear interpretation of the estimated regression parameters for the final model selected.

13.1.4 For the final model selected in Problem 13.1.3, construct a table of the estimated probabilities that respiratory status is classified as good as a function of both time and treatment group (and, possibly, clinic). What do you conclude from this table?

13.2 In a clinical trial of patients suffering from epileptic seizures (Thall and Vail, 1990), patients were randomized to receive either a placebo or the drug progabide, in addition to standard therapy. A baseline count of the number of epileptic seizures in an 8-week period prior to randomization was obtained. In addition counts of the number of epileptic seizures in each of four successive 2-week (post-baseline) treatment periods were obtained. The goal of the analysis is to make a comparison between the two treatment groups in terms of changes in the rates of epileptic seizures throughout the duration of the study.

The raw data are stored in an external file: `epilepsy.dat`

Each row of the data set contains the following eight variables:

ID Y_1 Y_2 Y_3 Y_4 Treatment Y_0 Age

Note: The response variable Y_0 is a baseline count of the number of epileptic seizures in an 8-week interval. The response variables Y_j are counts of the number of epileptic

seizures in the four successive 2-week (post-baseline) treatment intervals, for $j = 1, ..., 4$. The categorical variable Treatment is coded 1 = Progabide, 0 = Placebo. The variable Age is the age of each patient (in years) at baseline.

13.2.1 Consider a model for the log seizure rate that includes the main effects of treatment and time (where time is regarded as a categorical variable with five levels), and their interaction. Use generalized estimating equations (GEE), assuming separate pairwise correlations among the five responses. Construct a test of the null hypothesis of no effect of treatment on changes in the log seizure rate based on the empirical standard errors.

13.2.2 What conclusions do you draw about the effect of treatment on *changes* in the log seizure rate?

13.2.3 Construct a new variable, Ptime, where:
Ptime = 0 if baseline, and Ptime = 1 if post-baseline (any of the four successive 2-week intervals).
Repeat the analysis for Problem 13.2.1 using Ptime (instead of time as a categorical variable with five levels). Construct a test of the null hypothesis of no effect of treatment on changes in the log seizure rate based on the empirical standard errors.

13.2.4 From the results of the analysis for Problem 13.2.3, what conclusions do you draw about the effect of treatment on changes in the log seizure rate?

13.2.5 Patient 49 (ID = 49) is a potential outlier. This patient reported 151 seizures during the 8-week baseline interval and 302 (102+65+72+63) seizures during the four successive 2-week intervals. Repeat all of the analyses in Problems 13.2.1 to 13.2.4, excluding all of the repeated count data from patient 49. When the data from patient 49 are excluded, what conclusions do you draw about the effect of treatment on changes in the log seizure rate?

13.3 In a clinical trial of patients with insomnia (Francom, Chuang-Stein, and Landis, 1989), patients were randomized to receive either a hypnotic drug or placebo. An ordinal response, denoting patients' reported time (in minutes) to fall asleep after going to bed was recorded at baseline and after two weeks of treatment. The four-level ordinal response is coded 1: < 20 minutes, 2: 20–30 minutes, 3: 30–60 minutes, 4: > 60 minutes; these data are from Chapter 11 (Table 11.4) of Agresti (2002). The main objective of the analyses is to assess changes in the odds of a more favorable

response (shorter reported time to fall asleep) over the duration of the study, and to determine whether treatment with the hypnotic drug has an influence of these changes.

The raw data are stored in an external file: `insomnia.dat`

Each row of the data set contains the following four variables:

ID Trt Time Y

Note: The response variable Y is a four-level ordinal response denoting patients' reported time (in minutes) to fall asleep after going to bed (1: $<$ 20 minutes, 2: 20–30 minutes, 3: 30–60 minutes, 4: $>$ 60 minutes). The variable Time denotes baseline (Time = 0) and 2-week follow-up (Time = 1). The categorical treatment variable, Trt, is coded $1 =$ Hypnotic drug, $0 =$ Placebo. The variable Age is the age of each patient (in years) at baseline.

13.3.1 Consider a proportional odds model for the cumulative log odds of response that includes the main effects of treatment and time, and their interaction,

$$\log\left\{\frac{\Pr(Y_{ij} \le k)}{\Pr(Y_{ij} > k)}\right\} = \alpha_k + \beta_1 \,\mathrm{Trt}_i + \beta_2 \,\mathrm{Time}_{ij} + \beta_3 \,\mathrm{Trt}_i \times \mathrm{Time}_{ij}.$$

Fit the model using generalized estimating equations (GEE), with a "working independence" assumption for the within-subject association.

13.3.2 What is the interpretation of the estimate of β_2?

13.3.3 What is the interpretation of the estimate of β_3?

13.3.4 Construct a test of the null hypothesis of no effect of treatment on changes in the cumulative log odds of response based on the empirical standard errors. What conclusions do you draw about the effect of treatment?

13.3.5 Based on the results from Problem 13.3.1, estimate the odds ratio of a more favorable response at week 2 relative to baseline for patients receiving placebo.

13.3.6 Based on the results from Problem 13.3.1, estimate the odds ratio of a more favorable response at week 2 relative to baseline for patients receiving the hypnotic drug.

13.3.7 Based on the results from Problem 13.3.1, estimate the probability that patients receiving the hypnotic drug report falling asleep in less than 20 minutes (i.e, the probability of response level 1) at week 2.

13.3.8 Based on the results from Problem 13.3.1, estimate the probability that patients receiving the hypnotic drug report falling asleep in 30–60 minutes (i.e., the probability of response level 3) at week 2.
Hint: $\Pr(Y_{ij} = k) = \Pr(Y_{ij} \le k) - \Pr(Y_{ij} \le k - 1)$.