27

Simple Methods, Direct Likelihood, and Weighted Generalized Estimating Equations

27.1 Introduction

Commonly used methods to analyze incomplete longitudinal data include complete case analysis (CC) and last observation carried forward (LOCF). However, such methods rest on strong assumptions, including missing completely at random (MCAR). Such assumptions are too strong to generally hold. Over the past decades, a number of full longitudinal data analysis methods have become available, such as the linear, generalized linear, and non-linear mixed modeling frameworks, and the likelihood-based models of Chapters 6 and 7, that are valid under the much weaker missing at random (MAR) assumption. Such methods are useful, even if the scientific question is in terms of a single time point, e.g., the last planned measurement occasion in a clinical trial. The validity of such a method rests on the use of maximum likelihood, under which the missing data mechanism is ignorable as soon as MAR applies. Specific attention needs to be devoted to generalized estimating equations, given their non-likelihood status.

In many clinical trial and other settings, the standard methodology used to analyze incomplete longitudinal data is based on such methods as *last observation carried forward* (LOCF), *complete case analysis* (CC), or simple forms of imputation. This is often done without questioning the possible influence of these assumptions on the final results, even though several authors have written about this topic. A relatively early account is given in Heyting, Tolboom, and Essers (1992). Mallinckrodt *et al* (2003ab) and Lavori, Dawson, and Shera (1995) propose direct-likelihood and multiple-

imputation methods, respectively, to deal with incomplete longitudinal data. Siddiqui and Ali (1998) compare direct-likelihood and LOCF methods.

It is unfortunate that such a strong emphasis is placed on methods like LOCF and CC in clinical trial settings, as they are based on strong and unrealistic assumptions. Even the strong MCAR assumption does not suffice to guarantee that an LOCF analysis is valid. In contrast, under the less restrictive assumption of MAR, valid inference can be obtained through a likelihood-based analysis without modeling the dropout process. One can then use linear or generalized linear mixed models (Verbeke and Molenberghs 2000, see also Chapter 4 in this volume), without additional complication or effort. We will argue that such an analysis is more likely to be valid, and even easier to implement than LOCF and CC analyses.

In Section 27.2, the status of longitudinal and non-longitudinal data analysis is briefly discussed in the context of incomplete longitudinal sequences. Section 27.3 reviews simple methods, with emphasis on CC and LOCF, and then goes on to advocate direct likelihood as an important and viable alternative. The bias that occurs in CC and LOCF is studied analytically, in the context of a specific and simple model, is studied in Section 27.4. The specific situation of generalized estimating equations is the topic of Section 27.5. The concepts developed in this chapter are then exemplified using a depression clinical trial (Section 27.6), the Age Related Macular Degeneration study (Section 27.7), which was introduced in Section 2.9 and analyzed before in Section 24.4, and finally the analgesic trial (Section 27.8), which has been analyzed before in Chapter 17 and Section 18.4.

27.2 Longitudinal Analysis or Not?

In principle, one should start by considering the density of the full data (26.2), but by the very nature of the missing data problem, parts of the outcome vector $\boldsymbol{Y}_i$ may be left unobserved, and hence one has to focus on the observed data only, i.e., $\boldsymbol{Y}_i^o$ and $\boldsymbol{R}_i$. Of course, when ignorability applies (Section 26.2.3), one can further ignore the missing data itself. As stated in the introduction, one often sees much simpler analyses, which often overlook the important issues altogether.

Whatever the perspective taken, it usually belongs to one of two possible views for the measurement model on the one hand and a philosophy adopted for the missingness model on the other hand. We will describe these in turn.

Model for measurements. A choice has to be made regarding the modeling approach to the measurements. Several views are possible.

View 1. One can choose to analyze the entire longitudinal profile, irrespective of whether interest focuses on the entire profile (e.g., difference in slope between groups) or on a specific time point (e.g., the last planned occasion). In the latter case, one would make inferences about such an occasion using the posited model.

View 2. One states the scientific question in terms of the outcome at a well-defined point in time. Several choices are possible:

View 2a. The scientific question is defined in terms of the *last planned occasion*. In this case, one can either accept the dropout as it is or use one or other strategy (e.g., imputation) to incorporate the missing outcomes.

View 2b. One can choose to define the question and the corresponding analysis in terms of the *last observed measurement*.

Although Views 1 and 2a necessitate reflection on the missing data mechanism, View 2b avoids the missing data problem because the question is couched completely in terms of observed measurements. Thus, under View 2b, an LOCF analysis might be acceptable, provided it matched the scientific goals, but is then better described as a last observation analysis because nothing is carried forward. Such an analysis should properly be combined with an analysis of time to dropout, perhaps in a survival analysis framework. Of course, an investigator should reflect very carefully on whether View 2b represents a relevant and meaningful scientific question (Shih and Quan 1997).

Method for handling missingness. A choice has to be made regarding the modeling approach for the missingness process. Under certain assumptions this process can be ignored (e.g., a likelihood-based ignorable analysis). Some simple methods, such as a complete case analysis and LOCF, do not explicitly address the missingness process either.

The measurement model will depend on whether or not a full longitudinal analysis is done. When the focus is on the last observed measurement or on the last measurement occasion only, one typically opts for classical two- or multi-group comparisons (t-test, Wilcoxon, etc.). When a longitudinal analysis is deemed necessary, the choice depends on the nature of the outcome. Options include the linear and generalized linear mixed models, generalized estimating equations, etc.

27.3 Simple Methods

We will briefly review a number of relatively simple methods that still are commonly used. For the validity of many of these methods, MCAR is required. For others, such as LOCF, MCAR is necessary but not sufficient.

The focus will be on the complete case method, for which data are removed, and on imputation strategies, where data are filled in. Regarding imputation, one distinguishes between single and multiple imputation. In the first case, a single value is substituted for every 'hole' in the dataset and the resulting dataset is analyzed as if it represented the true complete data. Multiple imputation acknowledges the uncertainty stemming from filling in missing values rather than observing them (Rubin 1987, Schafer 1997). LOCF will be discussed within the context of imputation strategies, although LOCF can be placed in other frameworks as well.

A **complete case analysis** includes only those cases for which all measurements were recorded. This method has obvious advantages. It is simple to describe and almost any software can be used because there are no missing data. Unfortunately, the method suffers from severe drawbacks. First, there is nearly always a substantial loss of information. For example, suppose there are 20 measurements, with 10% of missing data on each measurement. Suppose, further, that missingness on the different measurements is independent; then, the estimated percentage of incomplete observations is as high as 87%. The impact on precision and power may be dramatic. Even though the reduction of the number of complete cases will be less severe in settings where the missingness indicators are correlated, this loss of information will usually militate against a complete case analysis. Second, severe bias can result when the missingness mechanism is MAR but not MCAR. Indeed, should an estimator be consistent in the complete data problem, then the derived complete case analysis is consistent only if the missingness process is MCAR. A CC analysis can be conducted when Views 1 and 2 of Section 27.2 are adopted. It is obviously not a reasonable choice with View 2b.

An alternative way to obtain a data set on which complete data methods can be used is to fill in rather than delete (Little and Rubin 1987). Concern has been raised regarding imputation strategies. Dempster and Rubin (1983) write: "The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases." For example, Little and Rubin (1987) show that the application of imputation could be considered acceptable in a linear model with one fixed effect and one error term, but that it is generally not acceptable for hierarchical models, split-plot designs, repeated measures with a complicated error structure, random-effects, and mixed-effects models.

Thus, the user of imputation strategies faces several dangers. First, the imputation model could be wrong and, hence, the point estimates biased. Second, even for a correct imputation model, the uncertainty resulting from missingness is ignored. Indeed, even when one is reasonably sure about the

mean value the unknown observation *would have had*, the actual stochastic realization, depending on both the mean and error structures, is still unknown. In addition, most methods require the MCAR assumption to hold while some even require additional and often unrealistically strong assumptions.

A method that has received considerable attention (Siddiqui and Ali 1998, Mallinckrodt *et al* 2003ab) is **last observation carried forward** (LOCF). In the LOCF method, whenever a value is missing, the last observed value is substituted. The technique can be applied to both monotone and non-monotone missing data. It is typically applied in settings where incompleteness is due to attrition.

LOCF can, but not necessarily has to, be regarded as an imputation strategy, depending on which of the views of Section 27.2 is taken. The choice of viewpoint has a number of consequences. First, when the problem is approached from a missing data standpoint, one has to think it plausible that subjects' measurements do not change from the moment of dropout onwards (or during the period they are unobserved in the case of intermittent missingness). In a clinical trial setting, one might believe that the response profile *changes* as soon as a patient goes off treatment and even that it would flatten. However, the constant profile assumption is even stronger. Second, LOCF shares with other single imputation methods that it artificially increases the amount of information in the data, by treating imputed and actually observed values on an equal footing. This is especially true if a longitudinal view is taken. Verbeke and Molenberghs (1997, Chapter 5) have shown that all features of a linear mixed model (group difference, evolution over time, variance structure, correlation structure, random effects structure, ...) can be affected.

Thus, scientific questions with which LOCF is compatible will be those that are phrased in terms of the last obtained measurement (View 2b). Whether or not such questions are sensible should be the subject of scientific debate, which is quite different from a *post hoc* rationale behind the use of LOCF. Likewise, it can be of interest to model the complete cases separately and to make inferences about them. In such cases, a CC analysis is of course the only reasonable way forward. This is fundamentally different from treating a CC analysis as one that can answer questions about the randomized population as a whole.

We will briefly describe two other imputation methods. The idea behind **unconditional mean imputation** (Little and Rubin 1987) is to replace a missing value with the average of the observed values on the same variable over the other subjects. Thus, the term *unconditional* refers to the fact that one does not use (i.e., condition on) information on the subject for which an imputation is generated. It is clear that this method is developed primarily for continuous data and its application to binary outcomes would be problematic. Because values are imputed that are unrelated to a subject's other measurements, all aspects of a model, such as a linear

mixed model, are typically distorted (Verbeke and Molenberghs 1997). In this sense, unconditional mean imputation can be as damaging as LOCF.

Buck's method or **conditional mean imputation** (Buck 1960, Little and Rubin 1987) is similar in complexity to mean imputation. Consider, for example, a single multivariate normal sample. The first step is to estimate the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ from the complete cases, assuming that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For a subject with missing components, the regression of the missing components ($\boldsymbol{Y}_i^m$) on the observed ones ($\boldsymbol{y}_i^o$) is

$$\boldsymbol{Y}_i^m | \boldsymbol{y}_i^o \sim N(\boldsymbol{\mu}^m + \boldsymbol{\Sigma}^{mo}(\boldsymbol{\Sigma}^{oo})^{-1}(\boldsymbol{y}_i^o - \boldsymbol{\mu}_i^o), \boldsymbol{\Sigma}^{mm} - \boldsymbol{\Sigma}^{mo}(\boldsymbol{\Sigma}^{oo})^{-1}\boldsymbol{\Sigma}^{om}).$$

The second step calculates the conditional mean from the regression of the missing components on the observed components, and substitutes the conditional mean for the corresponding missing values. In this way, "vertical" information (estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) is combined with "horizontal" information ($\boldsymbol{y}_i^o$). Buck (1960) showed that under mild conditions, the method is valid under MCAR mechanisms. Little and Rubin (1987) added that the method is also valid under certain MAR mechanisms. Even though the distribution of the observed components is allowed to differ between complete and incomplete observations, it is very important that the regression of the missing components on the observed ones is constant across missingness patterns. Again, this method shares with other single imputation strategies that, although point estimation may be consistent, the precision will be overestimated. There is a connection between *the concept* of conditional mean imputation and a likelihood-based ignorable analysis, in the sense that the latter analysis produces expectations for the missing observations that are formally equal to those obtained under conditional mean imputation. However, in likelihood-based ignorable analyses, no explicit imputation takes place, hence the amount of information in the data is not overestimated and important model elements, such as mean structure and variance components, are not distorted.

Historically, an important motivation behind the simpler methods was their simplicity. Currently, with the availability of commercial software tools such as, for example, the SAS procedures MIXED, GLIMMIX, and NLMIXED and the SPlus and R nlme libraries, this motivation no longer applies. Arguably, an MAR analysis is the preferred choice. Of course, the correctness of an MAR analysis is in its own right never completely verifiable. Purely resorting to MNAR analyses (Chapters 29 and 30) is not satisfactory either since important sensitivity issues (Chapter 31) then arise. See also Verbeke and Molenberghs (2000).

27.4 Bias in LOCF, CC, and Ignorable Likelihood Methods

It is often quoted that LOCF or CC, though problematic for parameter estimation, produce randomization-valid hypothesis testing, but this is questionable. First, in a CC analysis partially observed data are selected out, with probabilities that may depend on post-randomization outcomes, thereby undermining any randomization justification. Second, if the focus is on one particular time point, e.g., the last one scheduled, then LOCF plugs in data. Such imputations, apart from artificially inflating the information content, may deviate in complicated ways from the underlying data. In contrast, a likelihood-based MAR analysis uses all available data, with the need for neither deletion nor imputation, which suggests that a likelihood-based MAR analysis would usually be the preferred one for testing as well. Third, although the size of a randomization-based LOCF test may reach its nominal size under the null hypothesis of no difference in treatment profiles, there will be other regions of the alternative space where the power of the LOCF test procedure is equal to its size, which is completely unacceptable.

Using the simple but insightful setting of two repeated follow-up measures, the first of which is always observed while the second can be missing, we establish some properties of the LOCF and CC estimation procedures under different missing data mechanisms, against the background of an MAR process operating. In this way, we bring LOCF and CC within a general framework that makes clear their relationships with more formal modeling approaches, enabling us to make a coherent comparison among the different approaches. The use of a moderate amount of algebra leads to some interesting conclusions.

It is most convenient to consider continuous outcomes, although similar arguments hold for non-Gaussian outcomes as well. Let us assume each subject i is to be measured on two occasions $t_i = 0, 1$. Subjects are randomized to one of two treatment arms: $T_i = 0$ for the standard arm and $T_i = 1$ for the experimental arm. The probability of an observation being observed on the second occasion ($D_i = 2$) is p_0 and p_1 for treatment groups 0 and 1, respectively. We can write the means of the observations in the two dropout groups as follows:

$$\text{dropouts } D_i = 1 \quad : \quad \beta_0 + \beta_1 T_i + \beta_2 t_i + \beta_3 T_i t_i, \tag{27.1}$$

$$\text{completers } D_i = 2 \quad : \quad \gamma_0 + \gamma_1 T_i + \gamma_2 t_i + \gamma_3 T_i t_i. \tag{27.2}$$

The true underlying population treatment difference at time $t_i = 1$, as determined from (27.1)–(27.2), is equal to:

$$\begin{aligned} \Delta_{\text{true}} \quad = \quad & p_1(\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3) + (1 - p_1)(\beta_0 + \beta_1 + \beta_2 + \beta_3) \\ & -[p_0(\gamma_0 + \gamma_2) + (1 - p_0)(\beta_0 + \beta_2)]. \end{aligned} \tag{27.3}$$

If we use LOCF, the expectation of the corresponding estimator equals:

$$\begin{aligned}\Delta_{\text{LOCF}} &= p_1(\gamma_0+\gamma_1+\gamma_2+\gamma_3)+(1-p_1)(\beta_0+\beta_1)\\ &\quad -[p_0(\gamma_0+\gamma_2)+(1-p_0)\beta_0].\end{aligned} \tag{27.4}$$

Alternatively, if we use CC, the above expression changes to:

$$\Delta_{\text{CC}} = \gamma_1+\gamma_3. \tag{27.5}$$

Hence, in general, both procedures yield biased estimators.

We will now consider the special but important cases where the true missing data mechanisms are MCAR and MAR, respectively. Each of these will impose particular constraints on the β and γ parameters in Model (27.1)–(27.2). Under MCAR, the β parameters are equal to their γ counterparts and (27.3) simplifies to

$$\Delta_{\text{MCAR,true}} = \beta_1+\beta_3 \equiv \gamma_1+\gamma_3. \tag{27.6}$$

Suppose we apply the LOCF procedure in this setting, the expectation of the resulting estimator then simplifies to:

$$\Delta_{\text{MCAR,LOCF}} = \beta_1+(p_1-p_0)\beta_2+p_1\beta_3. \tag{27.7}$$

The bias is given by the difference between (27.6) and (27.7):

$$B_{\text{MCAR,LOCF}} = (p_1-p_0)\beta_2-(1-p_1)\beta_3. \tag{27.8}$$

While of a simple form, we can learn several things from this expression by focusing on each of the terms in turn. First, suppose $\beta_3 = 0$ and $\beta_2 \neq 0$, implying that there is no differential treatment effect between the two measurement occasions, but there is an overall time trend. Then, the bias can go in either direction depending on the sign of $p_1 - p_0$ and the sign of β_2. Note that $p_1 = p_0$ only in the special case that the dropout rate is the same in both treatment arms. Whether or not this is the case has no impact on the status of the dropout mechanism (it is MCAR in either case, even though in the second case dropout is treatment-arm dependent), but is potentially very important for the bias implied by LOCF. Second, suppose $\beta_3 \neq 0$ and $\beta_2 = 0$. Again, the bias can go in either direction depending on the sign of β_3, i.e., depending on whether the treatment effect at the second occasion is larger or smaller than the treatment effect at the first occasion. In conclusion, even under the strong assumption of MCAR, we see that the bias in the LOCF estimator typically does not vanish and, even more importantly, the bias can be positive or negative and can even induce an apparent treatment effect when one does not exist.

In contrast, as can be seen from (27.5) and (27.6), the CC analysis is unbiased.

Let us now turn to the MAR case. In this setting, the constraint implied by the MAR structure of the dropout mechanism is that the conditional distribution of the second observation given the first is the same in both dropout groups (Molenberghs *et al* 1998). Based on this result, the expectation of the second observation in the standard arm of the dropout group is

$$E(Y_{i2}|D_i = 1, T_i = 0) = \gamma_0 + \gamma_2 + \sigma(\beta_0 - \gamma_0), \tag{27.9}$$

where $\sigma = \sigma_{21}\sigma_{11}^{-1}$, σ_{11} is the variance of the first observation in the fully observed group and σ_{12} is the corresponding covariance between the pair of observations. Similarly, in the experimental group we obtain

$$E(Y_{i2}|D_i = 1, T_i = 1) = \gamma_0 + \gamma_1 + \gamma_2 + \gamma_3 + \sigma(\beta_0 + \beta_1 - \gamma_0 - \gamma_1). \tag{27.10}$$

The true underlying population treatment difference (27.3) then becomes

$$\begin{aligned} \Delta_{\text{MAR,true}} &= \gamma_1 + \gamma_3 + \sigma[(1 - p_1)(\beta_0 + \beta_1 - \gamma_0 - \gamma_1) \\ &\quad -(1 - p_0)(\beta_0 - \gamma_0)]. \end{aligned} \tag{27.11}$$

In this case, the bias in the LOCF estimator can be written as:

$$\begin{aligned} B_{\text{MAR,LOCF}} &= p_1(\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3) + (1 - p_1)(\beta_0 + \beta_1) \\ &\quad -p_0(\gamma_0 + \gamma_2) - (1 - p_0)\beta_0 - \gamma_1 - \gamma_3 \\ &\quad -\sigma[(1 - p_1)(\beta_0 + \beta_1 - \gamma_0 - \gamma_1) \\ &\quad -(1 - p_0)(\beta_0 - \gamma_0)]. \end{aligned} \tag{27.12}$$

Again, although involving more complicated relationships, it is clear that the bias can go in either direction, thus contradicting the claim often put forward that the bias in LOCF leads to conservative conclusions. Further, it is far from clear what conditions need to be imposed in this setting for the corresponding estimator to be either unbiased or conservative.

The bias in the CC estimator case takes the form:

$$B_{\text{MAR,CC}} = -\sigma[(1 - p_1)(\beta_0 + \beta_1 - \gamma_0 - \gamma_1) - (1 - p_0)(\beta_0 - \gamma_0)]. \tag{27.13}$$

Even though this expression is simpler than in the LOCF case, it is still true that the bias can operate in either direction.

Thus, in all cases, LOCF typically produces bias of which the direction and magnitude depend on the true but unknown treatment effects. Hence, caution is needed when using this method. In contrast, an ignorable likelihood based analysis, as outlined in Section 27.3, provides a consistent estimator of the true treatment difference at the second occasion under both MCAR and MAR. Although this is an assumption, it is rather a mild one in contrast to the stringent conditions required to justify the LOCF method, even when the qualitative features of the bias are considered more important than the quantitative ones. Note that the LOCF method is not valid even under the strong MCAR condition, whereas the CC approach is valid under MCAR.

27.5 Weighted Generalized Estimating Equations

In the previous sections, in particular in the last one, it was shown that direct likelihood is a method of choice, due to the ease with which it can be implemented and the validity under MAR.

For categorical outcomes, as we have seen before, the GEE approach could be adopted. However, as Liang and Zeger (1986) pointed out, inferences with the GEE are valid only under the strong assumption that the data are missing completely at random (MCAR). To allow the data to be missing at random (MAR), Robins, Rotnitzky, and Zhao (1995) proposed a class of weighted estimating equations. These can be viewed as an extension of generalized estimating equations.

The idea of weighted generalized estimating equations (WGEE) is to weight each subject's measurements in the GEEs by the inverse probability that a subject drops out at that particular measurement occasion. Such a weight can be calculated as

$$\begin{aligned}\nu_{ij} \equiv P(D_i = j) &= \prod_{k=2}^{j-1}[1 - P(R_{ik} = 0 | R_{i2} = \ldots = R_{i,k-1} = 1)] \times \\ &\quad P(R_{ij} = 0 | R_{i2} = \ldots = R_{i,j-1} = 1)^{I\{j \leq n_i\}} \quad (27.14)\end{aligned}$$

if dropout occurs by time j or we reach the end of the measurement sequence, and

$$\nu_{ij} \equiv P(D_i = j) = \prod_{k=2}^{j}[1 - P(R_{ik} = 0 | R_{i2} = \ldots = R_{i,k-1} = 1)] \quad (27.15)$$

otherwise.

Recall that we partitioned $\boldsymbol{Y}_i$ into the unobserved components $\boldsymbol{Y}_i^m$ and the observed components $\boldsymbol{Y}_i^o$. Similarly, we can make the exact same partition of $\boldsymbol{\mu_i}$ into $\boldsymbol{\mu_i}^m$ and $\boldsymbol{\mu_i}^o$. In the weighted GEE approach, which is proposed to reduce possible bias of $\widehat{\boldsymbol{\beta}}$, the score equations to be solved are:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} W_i \, \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}'} (A_i^{1/2} R_i A_i^{1/2})^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu_i}) \;=\; \mathbf{0},$$

where W_i is a diagonal matrix with the elements of $\boldsymbol{\nu}_i$ along the diagonal, or

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{d=2}^{n_i+1} \frac{I(D_i = d)}{\nu_{id}} \, \frac{\partial \boldsymbol{\mu_i}(d)}{\partial \boldsymbol{\beta}'} (A_i^{1/2} R_i A_i^{1/2})^{-1}(d) (\boldsymbol{y}_i(d) - \boldsymbol{\mu_i}(d)) \;=\; \mathbf{0},$$

where $\boldsymbol{y}_i(d)$ and $\boldsymbol{\mu_i}(d)$ are the first $d-1$ elements of $\boldsymbol{y}_i$ and $\boldsymbol{\mu_i}$, respectively. We define

$$\frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}'}(d)$$

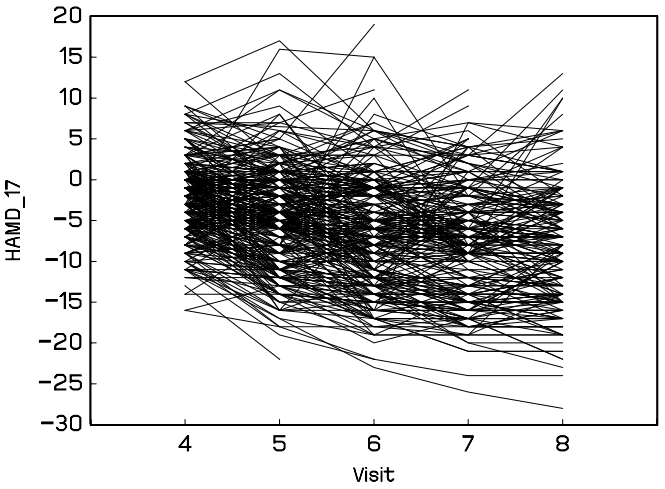

FIGURE 27.1. *Depression Trial. Individual profiles.*

and $(A_i^{1/2} R_i A_i^{1/2})^{-1}(d)$ analogously, in line with the definition of Robins, Rotnitzky and Zhao (1995).

Thus, not only likelihood methods but also appropriately adapted generalized estimating equations can be used with ease, under MAR. Both can be adapted to the MNAR setting as well (Chapters 29 and 30). Although it is beneficial to have both of these tools in one's toolkit, it is also important to realize that both 'schools' have strong supporters. An important discussion of these issues is given in Davidian, Tsiatis, and Leon (2005). Lipsitz *et al* (2001) studied bias in weighted estimating equations.

27.6 The Depression Trial

We will illustrate various methods discussed in this chapter by means of a clinical trial in depression, analyzed before by Molenberghs *et al* (2004), Jansen *et al* (2005), Dmitrienko *et al* (2005, Chapter 5), and Molenberghs *et al* (2005).

27.6.1 The Data

The depression trial data come from a clinical trial including 342 patients with post-baseline data. The Hamilton Depression Rating Scale ($HAMD_{17}$) is used to measure the depression status of the patients. For each patient, a baseline assessment is available.

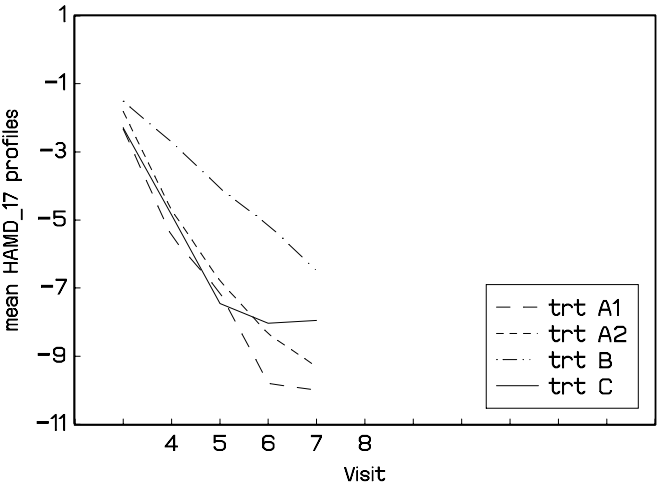

FIGURE 27.2. *Depression Trial. Mean profiles per treatment arm.*

For blinding purposes, therapies are coded as A1 for primary dose of experimental drug, A2 for secondary dose of experimental drug, and B and C for non-experimental drugs. Individual profiles and mean profiles of the changes from baseline in $HAMD_{17}$ scores per treatment arm are shown in Figures 27.1 and 27.2 respectively.

The contrast of primary interest is between A1 and C. Emphasis is on the difference between arms at the end of the study. A graphical representation of the dropout, per arm, is given in Figure 27.3. Part of the depression data set is given below. Therapies A1, A2, B, and C are denoted as treatment 1, 2, 3, and 4 respectively. Dots represent unobserved measurements.

We will focus on the analysis of the binary outcome, defined as 1 if the $HAMD_{17}$ score is larger than 7, and 0 otherwise. These analyses are in line with Jansen *et al* (2004), Dmitrienko *et al* (2005, Chapter 5), and Molenberghs *et al* (2005).

The primary null hypothesis will be tested using both GEE and WGEE, as well as GLMM. We include the fixed categorical effects of treatment, visit, and treatment-by-visit interaction, as well as the continuous, fixed covariates of baseline score and baseline score-by-visit interaction. A random intercept will be included when considering the random-effect models.

Analyses will be implemented using the SAS procedures GENMOD, GLIMMIX, and NLMIXED.

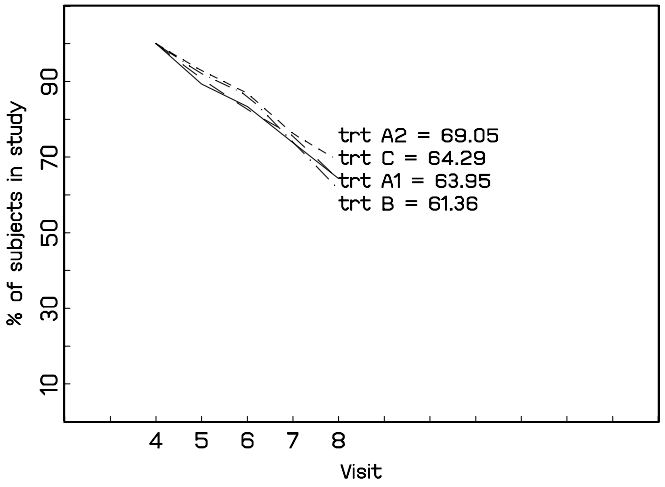

FIGURE 27.3. *Depression Trial. Evolution of dropout per treatment arm.*

27.6.2 Marginal Models

First, let us consider the GEE approach. Although we can consider both empirically corrected and model-based standard errors (Chapter 8), it is sensible to confine inferences to the empirically corrected ones. Several contrasts are of interest as well. The first one to test for treatment effect at the endpoint, the second one for the average treatment effect over the course of the study. Depending on the primary and secondary scientific questions, more of these can be considered. Both standard GEE (Section 8.2) as well as linearization-based GEE (Section 8.8) are considered. It will allow us to assess similarities and differences in this context, knowing how closely they agree from, for example, Chapter 8.

Of course, given the incomplete nature of the data, it is careful to consider weighted generalized estimating equations, unless one has strong belief that the MCAR assumption holds. This implies that weights have to be constructed, based on the probability to drop out at a given time, given the patient is still in the study, given his or her past measurements, and given covariates. We restrict attention to the previous outcome and treatment indicator. The resulting model is of a standard logistic regression or probit regression type, and can be easily fitted using standard logistic regression software, such as the SAS procedures GENMOD and LOGISTIC. The code is exemplified in Section 32.5. The result of fitting this logistic regression did not reveal strong evidence for a dependence on the previous outcome (estimate -0.097, s.e. 0.351), nor on the treatment allocation (estimate 0.065, s.e. 0.314).

TABLE 27.1. *Depression Trial. Results of marginal models: Parameter estimates (model-based standard errors; empirically corrected standard errors) for standard unweighted and weighted GEE (denoted GEE and WGEE, respectively) and the linearization based method (interaction terms are not shown).*

Effect	GEE	WGEE	Linearization
Intercept	-1.22 (0.77;0.79)	-0.56 (0.63;0.91)	-1.23 (0.75;0.79)
Treatment	-0.71 (0.38;0.38)	-0.91 (0.32;0.41)	-0.67 (0.37;0.38)
Visit 4	0.43 (1.05;1.22)	-0.15 (0.85;1.90)	0.45 (1.05;1.22)
Visit 5	-0.45 (0.91;1.23)	-0.23 (0.68;1.54)	-0.47 (0.92;1.23)
Visit 6	0.06 (0.86;1.03)	0.15 (0.69;1.13)	0.05 (0.86;1.03)
Visit 7	-0.25 (0.89;0.91)	-0.27 (0.78;0.89)	-0.25 (0.89;0.91)
Baseline	0.08 (0.04;0.04)	0.06 (0.03;0.05)	0.08 (0.04;0.04)

Results of fitting the standard GEE as well as weighted GEE, combined with the results of the linearization-based method, are presented in Table 27.1. Apart from treatment allocation, the effect of baseline value and indicators for time at visits 4, 5, 6, and 7 were included into the model. Further, the interactions between treatment and visit and between baseline and visit were included in the model.

Although GEE and its linearization based version produce very similar results, in line with earlier observations, there are differences with the weighted version, in parameter estimates as well as standard errors. The difference in standard errors (often, but not always, larger under WGEE) are explained by the fact that additional sources of uncertainty, due to missingness, are taken into account. The resulting inferences can be different. For example, the treatment effect parameter is non-significant with GEE ($p = 0.0633$ with standard GEE and $p = 0.1184$ with the linearized version) while a significant difference is found under the correct WGEE analysis ($p = 0.0268$). Also, the difference is marked for treatment effect at endpoint: $p = 0.0658$ with standard GEE and $p = 0.0631$ with the linearized version, while a significant difference is found under the correct WGEE analysis ($p = 0.0289$).

Thus, one may fail to detect such important effects as treatment differences when GEE is used rather than the, admittedly, somewhat more laborious WGEE.

27.6.3 Random-effects Models

Because the generalized linear mixed model is typically fitted using maximum likelihood, based on numerical integration or data approximations (Chapter 14), standard fitting algorithms can be used, without modification, provided the MAR assumption and the mild regularity conditions

TABLE 27.2. *Depression Trial. Results of random-effects model fitting. Parameter estimates (standard errors) for GLMM with adaptive Gaussian quadrature (Num. int.) and penalized-quasi likelihood methods (PQL) (interaction terms are not shown).*

Effect	PQL	Num. int.
Intercept	-1.70 (1.06)	-2.31 (1.34)
Treatment	-0.84 (0.55)	-1.20 (0.72)
Visit 4	0.66 (1.48)	0.64 (1.75)
Visit 5	-0.44 (1.29)	-0.78 (1.51)
Visit 6	0.17 (1.22)	0.19 (1.41)
Visit 7	-0.23 (1.25)	-0.27 (1.43)
Baseline	0.10 (0.06)	0.15 (0.07)
R.I. var.	2.53 (0.53)	5.71 (1.53)

for ignorability are fulfilled, as presented in Section 26.2.3. Dmitrienko *et al* (2005, Chapter 5) and Molenberghs *et al* (2005) have indicated that also here the choice between adaptive and non-adaptive quadrature, the number of quadrature points, and the choice between quasi-Newton and Newton-Raphson, has a noticeable impact on the results, where adaptive quadrature and Newton-Raphson iteration produce the most reliable results, with no difference in the parameter estimates and standard errors observed, whether 10, 20, or 50 quadrature points are used. These results are contrasted with PQL based estimates in Table 27.2.

Once again, there are considerable differences between both approaches, and the PQL estimates are rather close to the GEE estimates. This indicates that, though the method is in principle likelihood based, the poverty of the approximation jeopardizes its validity under MAR even more than when data are complete and, if at all possible, the numerical integration method ought to be the preferred one. Turning to the treatment effect, the treatment effect at endpoint is not significant in either of the analyses, but the difference in p-value is noticeable: $p = 0.0954$ for numerical integration and $p = 0.1286$ with PQL.

27.7 Age Related Macular Degeneration Trial

In Section 24.4 we considered a longitudinal analysis, jointly for the binary and continuous outcomes at 4, 12, 24, and 52 weeks, for the ARMD study introduced in Section 2.9. Results were reported in Table 24.7. All analyses done in Section 24.4 were based on 190 subjects with complete information at weeks 24 and 52. However, the total number of subjects equals 240, meaning that a substantial portion of the data is subject to missingness.

TABLE 27.3. *Age Related Macular Degeneration Trial. Overview of missingness patterns and the frequencies with which they occur. 'O' indicates observed and 'M' indicates missing.*

Measurement occasion					
4 wks	12 wks	24 wks	52 wks	Number	%
Completers					
O	O	O	O	188	78.33
Dropouts					
O	O	O	M	24	10.00
O	O	M	M	8	3.33
O	M	M	M	6	2.50
M	M	M	M	6	2.50
Non-monotone missingness					
O	O	M	O	4	1.67
O	M	M	O	1	0.42
M	O	O	O	2	0.83
M	O	M	M	1	0.42

Both intermittent missingness as well as dropout occurs. An overview is given in Table 27.3.

Thus, 78.33% of the profiles are complete, while 18.33% exhibit monotone missingness. Out of the latter group, 2.5% or 6 subjects have no follow-up measurements. The remaining 3.33%, representing 8 subjects, have intermittent missing values. Although the group of dropouts is of considerable magnitude, the ones with intermittent missingness is much smaller. Nevertheless, it is cautious to include all into the analyses. This is certainly possible for direct likelihood analyses and for standard GEE, but WGEE is more complicated in this respect. One solution is to monotonize the missingness patterns by means of multiple imputation (Section 28.2) and then conduct WGEE.

In the analysis of Section 24.4, 190 'completers' were used, even though Table 27.3 shows there are 188 completers only. However, the analyses in Section 24.4 were done on subjects with measurements at weeks 24 and 52. The table shows that these can come from either profile 'OOOO,' the completers, but also from 'MOOO,' thus amounting to $188 + 2 = 190$ subjects.

Analogous to the analysis presented in Section 27.6, and inspired by the model for the binary data reported in Table 24.4, we compare analyses performed on the completers only (CC), on the LOCF imputed data, as well as on the observed data. In all cases, standard GEE, and linearization-based GEE will be considered. For the observed, partially incomplete data,

GEE is supplemented with WGEE. Further, a random-intercepts GLMM is considered, based on both PQL and numerical integration. The GEE analyses are reported in Table 27.4 and the random-effects models in Table 27.6. In all cases, we use the logit link. For GEE, a working exchangeable correlation matrix is considered. The model has four intercepts and four treatment effects. The advantage of having separate treatment effects at each time is that particular attention can be given at the treatment effect assessment at the last planned measurement occasion, i.e., after one year. From Table 27.4 it is clear that there is very little difference between the standard GEE and linearization-based GEE results. This is undoubtedly the case for CC, LOCF, and unweighted GEE on the observed data. For these three cases, also the model-based and empirically corrected standard errors agree extremely well. This is due to the unstructured nature of the full time by treatment mean structure. However, we do observe differences in the WGEE analyses. Not only are the parameter estimates mildly different between the two GEE versions, there is a dramatic difference between the model-based and empirically corrected standard errors. This is entirely due to the weighting scheme. The weights were not calibrated to add up to the total sample size, which is reflected in the model-based standard errors. In the linearization case, part of the effect is captured as overdispersion. This can be seen from adding the parameters σ^2 and τ^2. In all other analyses, the sum is close to one, as it should be when there is no residual overdispersion, but in the last column these add up to 3.14. Nevertheless, the two sets of empirically corrected standard errors agree very closely, which is reassuring.

When comparing parameter estimates across CC, LOCF, and observed data analyses, it is clear that LOCF has the effect of artificially increasing the correlation between measurements. The effect is mild in this case. The parameter estimates of the observed-data GEE are close to the LOCF results for earlier time points and close to CC for later time points. This is to be expected, as at the start of the study the LOCF and observed populations are virtually the same, with the same holding between CC and observed populations near the end of the study. Note also that the treatment effect under LOCF, especially at 12 weeks and after 1 year, is biased downward in comparison to the GEE analyses. To properly use the information in the missingness process, WGEE can be used. To this end, a logistic regression for dropout, given covariates and previous outcomes, needs to be fitted. Parameter estimates and standard errors are given in Table 27.5. Intermittent missingness will be ignored. Covariates of importance are treatment assignment, the level of lesions at baseline (a four-point categorical variable, for which three dummies are needed), and time at which dropout occurs. For the latter covariates, there are three levels, since dropout can occur at times 2, 3, or 4. Hence, two dummy variables are included. Finally, the previous outcome does not have a significant impact, but will be kept in the model nevertheless. In spite of there being

TABLE 27.4. *Age Related Macular Degeneration Trial. Parameter estimates (model-based standard errors; empirically corrected standard errors) for the marginal models: standard and linearization-based GEE on the CC and LOCF population, and on the observed data. In the latter case, also WGEE is used.*

Effect	Par.	CC	LOCF	Observed data	
				Unweighted	WGEE
Standard GEE					
Int.4	β_{11}	-1.01(0.24;0.24)	-0.87(0.20;0.21)	-0.87(0.21;0.21)	-0.98(0.10;0.44)
Int.12	β_{21}	-0.89(0.24;0.24)	-0.97(0.21;0.21)	-1.01(0.21;0.21)	-1.78(0.15;0.38)
Int.24	β_{31}	-1.13(0.25;0.25)	-1.05(0.21;0.21)	-1.07(0.22;0.22)	-1.11(0.15;0.33)
Int.52	β_{41}	-1.64(0.29;0.29)	-1.51(0.24;0.24)	-1.71(0.29;0.29)	-1.72(0.25;0.39)
Tr.4	β_{12}	0.40(0.32;0.32)	0.22(0.28;0.28)	0.22(0.28;0.28)	0.80(0.15;0.67)
Tr.12	β_{22}	0.49(0.31;0.31)	0.55(0.28;0.28)	0.61(0.29;0.29)	1.87(0.19;0.61)
Tr.24	β_{32}	0.48(0.33;0.33)	0.42(0.29;0.29)	0.44(0.30;0.30)	0.73(0.20;0.52)
Tr.52	β_{42}	0.40(0.38;0.38)	0.34(0.32;0.32)	0.44(0.37;0.37)	0.74(0.31;0.52)
Corr.	ρ	0.39	0.44	0.39	0.33
Linearization-based GEE					
Int.4	β_{11}	-1.01(0.24;0.24)	-0.87(0.21;0.21)	-0.87(0.21;0.21)	-0.98(0.18;0.44)
Int.12	β_{21}	-0.89(0.24;0.24)	-0.97(0.21;0.21)	-1.01(0.22;0.21)	-1.78(0.26;0.42)
Int.24	β_{31}	-1.13(0.25;0.25)	-1.05(0.21;0.21)	-1.07(0.23;0.22)	-1.19(0.25;0.38)
Int.52	β_{41}	-1.64(0.29;0.29)	-1.51(0.24;0.24)	-1.71(0.29;0.29)	-1.81(0.39;0.48)
Tr.4	β_{12}	0.40(0.32;0.32)	0.22(0.28;0.28)	0.22(0.29;0.29)	0.80(0.26;0.67)
Tr.12	β_{22}	0.49(0.31;0.31)	0.55(0.28;0.28)	0.61(0.28;0.29)	1.85(0.32;0.64)
Tr.24	β_{32}	0.48(0.33;0.33)	0.42(0.29;0.29)	0.44(0.30;0.30)	0.98(0.33;0.60)
Tr.52	β_{42}	0.40(0.38;0.38)	0.34(0.32;0.32)	0.44(0.37;0.37)	0.97(0.49;0.65)
	σ^2	0.62	0.57	0.62	1.29
	τ^2	0.39	0.44	0.39	1.85
Corr.	ρ	0.39	0.44	0.39	0.59

no strong evidence for MAR, the results between GEE and WGEE differ quite a bit. It is noteworthy that at 12 weeks, a treatment effect is observed with WGEE which goes unnoticed with the other marginal analyses. This finding is mildly confirmed by the random-intercept model, when the data as observed are used.

The results for the random-intercept models are given in Table 27.6. We observe the usual downward bias in the PQL *versus* numerical integration analysis, as well as the usual relationship between the marginal parameters of Table 27.4 and their random-effects counterparts. Note also that the random-intercepts variance is largest under LOCF, underscoring again that this method artificially increases the association between measurements on the same subject. In this case, unlike for the marginal models, LOCF and in fact also CC, slightly to considerably overestimates the treatment effect at certain times, in particular at 4 and 24 weeks.

TABLE 27.5. *Age Related Macular Degeneration Trial. Parameter estimates (standard errors) for a logistic regression model to describe dropout.*

Effect	Parameter	Estimate (s.e.)
Intercept	ψ_0	0.14 (0.49)
Previous outcome	ψ_1	0.04 (0.38)
Treatment	ψ_2	-0.86 (0.37)
Lesion level 1	ψ_{31}	-1.85 (0.49)
Lesion level 2	ψ_{32}	-1.91 (0.52)
Lesion level 3	ψ_{33}	-2.80 (0.72)
Time 2	ψ_{41}	-1.75 (0.49)
Time 3	ψ_{42}	-1.38 (0.44)

27.8 The Analgesic Trial

The binary satisfaction outcome in the analgesic trial (Section 2.2) was given extensive treatment in Chapter 17 and its ordinal counterpart was studied in Section 18.4. An important feature of the data is that a subgroup of patients does not complete the study but rather leaves prior to the scheduled end of the trial. Out of the 491 patients available for analysis, 223 are complete, and there are 55, 54, and 63 dropouts after the third, second, and first visit, respectively. Further, 96 patients have no follow up measurements. Among these, 63 have intermediate missing values as well. To further illustrate the impact of missingness on generalized estimating equations, we will conduct an analysis on the monotone sequences, with both ordinary and weighted generalized estimating equations, using the same marginal model (17.2) as fitted in Chapter 17.

A logistic regression is built for the dropout indicator, in terms of the previous outcome (for which the ordinal version is used by means of 4 dummies), pain control assessment at baseline, physical functioning at baseline, and genetic disorder measured at baseline. All of these are significant and parameter estimates are given in Table 27.7. This implies that there is evidence against MCAR in favor of MAR. This is a stronger result than observed in Section 27.6.2 for the depression trial.

In agreement with the procedure outlined in Section 27.5 and as illustrated on the depression trial, the predicted probabilities from this logistic regression are then used to calculate the weights, to be used in weighted GEE. Parameter estimates and standard errors for these are presented in Table 27.8. Clearly, though the evidence against MCAR is strong, the effect of the method chosen is noticeable but not terribly strong. We also note the impact on the standard errors. Weighted analyses are typically less precise, but more correct, than unweighted ones. Correction for the missingness mechanism has the effect of reducing the magnitude of the pa-

TABLE 27.6. *Age Related Macular Degeneration Trial. Parameter estimates (standard errors) for the random-intercept models: PQL and numerical-integration based fits on the CC and LOCF population, and on the observed data (direct-likelihood).*

Effect	Parameter	CC	LOCF	Direct lik.
		PQL		
Int.4	β_{11}	-1.19(0.31)	-1.05(0.28)	-1.00(0.26)
Int.12	β_{21}	-1.05(0.31)	-1.18(0.28)	-1.19(0.28)
Int.24	β_{31}	-1.35(0.32)	-1.30(0.28)	-1.26(0.29)
Int.52	β_{41}	-1.97(0.36)	-1.89(0.31)	-2.02(0.35)
Trt.4	β_{12}	0.45(0.42)	0.24(0.39)	0.22(0.37)
Trt.12	β_{22}	0.58(0.41)	0.68(0.38)	0.71(0.37)
Trt.24	β_{32}	0.55(0.42)	0.50(0.39)	0.49(0.39)
Trt.52	β_{42}	0.44(0.47)	0.39(0.42)	0.46(0.46)
R.I. s.d.	τ	1.42(0.14)	1.53(0.13)	1.40(0.13)
R.I. var.	τ^2	2.03(0.39)	2.34(0.39)	1.95(0.35)
		Numerical integration		
Int.4	β_{11}	-1.73(0.42)	-1.63(0.39)	-1.50(0.36)
Int.12	β_{21}	-1.53(0.41)	-1.80(0.39)	-1.73(0.37)
Int.24	β_{31}	-1.93(0.43)	-1.96(0.40)	-1.83(0.39)
Int.52	β_{41}	-2.74(0.48)	-2.76(0.44)	-2.85(0.47)
Trt.4	β_{12}	0.64(0.54)	0.38(0.52)	0.34(0.48)
Trt.12	β_{22}	0.81(0.53)	0.98(0.52)	1.00(0.49)
Trt.24	β_{32}	0.77(0.55)	0.74(0.52)	0.69(0.50)
Trt.52	β_{42}	0.60(0.59)	0.57(0.56)	0.64(0.58)
R.I. s.d.	τ	2.19(0.27)	2.47(0.27)	2.20(0.25)
R.I. var.	τ^2	4.80(1.17)	6.08(1.32)	4.83(1.11)

rameter estimates. In both cases, unstructured working assumptions were used. There is a noticeable effect on the working correlation matrix as well. With GEE, we obtain

$$R_{\text{UN, GEE}} = \begin{pmatrix} 1 & 0.173 & 0.246 & 0.201 \\ & 1 & 0.177 & 0.113 \\ & & 1 & 0.456 \\ & & & 1 \end{pmatrix},$$

TABLE 27.7. *Analgesic Trial. Parameter estimates (standard errors) for a logistic regression model to describe dropout.*

Effect	Parameter	Estimate (s.e.)
Intercept	ψ_0	-1.80 (0.49)
Previous GSA= 1	ψ_{11}	-1.02 (0.41)
Previous GSA= 2	ψ_{12}	-1.04 (0.38)
Previous GSA= 3	ψ_{13}	-1.34 (0.37)
Previous GSA= 4	ψ_{14}	-0.26 (0.38)
Basel. PCA	ψ_2	0.25 (0.10)
Phys. func.	ψ_3	0.009 (0.004)
Genetic disfunc.	ψ_4	0.59 (0.24)

TABLE 27.8. *Analgesic Trial. Parameter estimates (empirically corrected standard errors) for standard GEE and weighted GEE (WGEE) fitted to the monotone sequences.*

Effect	Parameter	GEE	WGEE
Intercept	β_1	2.95 (0.47)	2.17 (0.69)
Time	β_2	-0.84 (0.33)	-0.44 (0.44)
Time^2	β_3	0.18 (0.07)	0.12 (0.09)
Basel. PCA	β_4	-0.24 (0.10)	-0.16 (0.13)

whereas the WGEE version is

$$R_{\text{UN, WGEE}} = \begin{pmatrix} 1 & 0.215 & 0.253 & 0.167 \\ & 1 & 0.196 & 0.113 \\ & & 1 & 0.409 \\ & & & 1 \end{pmatrix}.$$

Of course, in line with general warnings issued in Section 8.2, care should be taken with interpreting the working correlation structure. In principle, it is a set of nuisance parameters, merely included to obtain reasonably efficient GEE estimates.