Ziegler, Kastner:

The Effect of Misspecified Response Probabilities on Parameter Estimates from Weighted Estimating Equations

Sonderforschungsbereich 386, Paper 200 (2000)

Online unter: http://epub.ub.uni-muenchen.de/

The Effect of Misspecified Response Probabilities on Parameter Estimates from Weighted Estimating Equations

Andreas Ziegler [1] Christian Kastner [2]

May 20, 2000

Abstract

Inference for the marginal mean using longitudinal data with monotone drop-outs in the response can be drawn with the weighted estimating equations (WEE; Robins, Rotnitzky and Zhao, 1995). Estimation proceeds in two steps. In the first step, a generalised linear model is usually applied to estimate response probabilities. In the second step, parameters of the mean structure are estimated by weighting a response inversely proportional to its estimated observation probability. The parameter estimates of the WEE are asymptotically normal and semiparametric efficient under suitable regularity conditions that include the correct specification of the model for the response probabilities. In this paper, we investigate the effect of misspecifying a) the parameters used to estimate the response probabilities and b) the link function for the response probabilities in a simulation study. We demonstrate that a slightly misspecified model for the response probabilities has an unimportant effect on the parameter estimates of the marginal mean from the WEE. We furthermore show that the choice of the link function has a negligible effect on the estimates of the marginal mean from the WEE. Our results are in line with classical findings for generalised linear models and for generalised estimating equations. Theoretical work should be added to our simulations that allow a quantification of the bias introduced by a misspecification of the model for the response probabilities.

Keywords: Correlated Data Analysis, Generalised Estimating Equations, Horvitz-Thompson Estimation, Marginal Models, Missing Data, Weighted Estimating Equations

1 Introduction

Several approaches for the analysis of the marginal mean using longitudinal data have been proposed. Probably, the most popular among these are the Generalised Estimating Equations (GEE; Liang and Zeger, 1986). The term "generalised" indicates that the association between the responses is modeled in addition to the mean structure which is of primary interest. If the association between the responses is not modeled but taken into account, the corresponding Estimating Equations (EE) are termed Independence Estimating Equations (IEE).

In analogy to other standard or advanced statistical methods, the GEE and the IEE were designed for complete data. However, many studies suffer from missing or incomplete data so that statistical analyses become more complicated. Approaches that ignore systematical differences between complete and incomplete clusters may be biased (Little and Schenker, 1995). In this paper we focus on item non-response in dependent variables (in the y) and assume monotone missing data patterns; that is, once a subject leaves the study, it will never return. Explanatory variables of interest are assumed to be completely observed.

One approach for solving the IEE in presence of missing dependent data received considerable attention (Robins et al., 1995; Robins and Rotnitzky, 1995). The basic idea of this approach is to weight observations inversely proportional to their respective response probabilities. The resulting estimators belong to the class of Horvitz-Thompson estimators. The corresponding estimating equations are termed Weighted Estimating Equations (WEE) and may be applied to data missing at random in Laird's (1988) sense. One disadvantage of the WEE was its unavailability in a standard computer package as noted by Carlin,

1 Medical Centre for Methodology and Health Research, Institute of Medical Biometry and Epidemiology, Philipps–University of Marburg, Bunsenstr. 3, 35033 Marburg, Germany, ziegler@mailer.uni-marburg.de

2 Institute of Statistics, LMU München, Ludwigstr. 33, 80539 München, Germany, kchris@stat.uni-muenchen.de

Wolfe, Coffey and Patton (1999). However, it has recently been implemented in the program MAREG (Kastner, Fieger and Heumann, 1997) which is freely available from the Web. One advantage of the WEE are their nice statistical properties: Thus, parameter estimates of the WEE are asymptotically normal and semiparametric efficient under suitable regularity conditions. The regularity conditions include the correct specification of the model for the response probabilities.

For this purpose we investigate a) the effect of misspecifying the parameters used to estimate the response probabilities and b) the effect of misspecifying the link function to model the response probabilities in a simulation study. The outline of this paper is as follows. In section 2 the IEE are derived assuming complete observations. The WEE of Robins et al. (1995) for a monotone missing data pattern are introduced in section 3. The results of our simulation study are discussed in section 4.

2 The Independence Estimating Equations

Let $\mathbf{y}_i$ be a vector of responses from n clusters with T observations for the ith cluster. The covariates $\mathbf{x}_{it}$ for each response y_{it} are summarized to the $\mathbf{X}_i = (\mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{iT})'$. The method may be easily extended to unequal cluster sizes T_i. The pairs $(\mathbf{y}_i, \mathbf{X}_i)$ are assumed to be independently identically distributed. We focus on marginal models for the mean and treat the association within a cluster as nuisance.

For independent observations, the well-known generalised linear model (GLM) allows flexibility in modeling mean and variance structures. In GLM, the mean structure is given by

$$E(y_{it}|\mathbf{x}_{it}) = \mu_{it} = g(\mathbf{x}'_{it}\boldsymbol{\beta}), \tag{1}$$

where g is a non-linear response function and $\boldsymbol{\beta}$ is the $p \times 1$ parameter vector of interest. If y_{it} is a binary variable, the connection between y_{it} and $\mathbf{x}_{it}$ may be established e.g. via the logit, the probit or the compound loglog link. In this situation, the variance function is usually chosen as the binomial variance function $\mu_{it}(1 - \mu_{it})$.

An estimator $\hat{\boldsymbol{\beta}}_{IEE}$ is the solution of the IEE

$$\mathbf{u}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{D}'_i \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{2}$$

Here, $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}'$ is the diagonal matrix of first derivatives, $\boldsymbol{\mu}_i = \boldsymbol{\mu}_i(\boldsymbol{\beta}) = g(\mathbf{X}_i\boldsymbol{\beta})$ is the vector of the mean structure, and $\mathbf{V}_i$ is the diagonal matrix of the variances $\mathbf{V}_i = \mathrm{diag}(v_{it})$. In general, (2) are solved iteratively by a FISHER-scoring algorithm. The true variance matrix $Cov(\mathbf{y}_i|\mathbf{X}_i) = \boldsymbol{\Omega}_i \neq \mathbf{V}_i$ is not diagonal for correlated observations. Therefore, Zeger, Liang and Self (1985) proposed to use the sandwich information matrix

$$\widehat{V(\hat{\boldsymbol{\beta}}_{IEE})} = \left(\sum_{i=1}^{n} \hat{\mathbf{D}}'_i \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1} \left(\sum_{i=1}^{n} \hat{\mathbf{D}}'_i \hat{\mathbf{V}}_i^{-1} \hat{\boldsymbol{\Omega}}_i \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right) \left(\sum_{i=1}^{n} \hat{\mathbf{D}}'_i \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1} \tag{3}$$

with $\hat{\boldsymbol{\Omega}}_i = (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)'$. Eq. (3) yields consistent estimates of $V(\hat{\boldsymbol{\beta}}_{IEE})$, even if the dependent variables within a cluster are correlated.

3 The Weighted Estimating Equations

In the presence of missing response data, the IEE need not be an appropriate tool for the analysis of clustered data (s. e.g. Ziegler, Kastner and Chang-Claude, 2000). Weighting the observed data is one general tool for dealing with missing response data. The weights are the inverse response probabilities. In many applications, the missing data mechanism, thus the weights, can be explained by surrogate variables $\mathbf{z}_{it}$ that are observed in addition to the response variables y_{it} and covariates X_i. These surrogate variables need not be of the investigator's interest for the mean structure and are possibly only collected, if y_{it} is observed. Let $\mathbf{w}_{i0} = (\mathrm{vec}(\mathbf{X}_i)', y_{i0}, \mathbf{z}'_{i0})'$ be comprised of covariates $\mathbf{X}_i$ and the observations of y_{i0} and $\mathbf{z}_{i0}$ prior to follow-up. We assume that y_{i0} and $\mathbf{z}_{i0}$ are completely observed. Thus, the explanatory variables are assumed to be either fixed or independent of the response variables. Furthermore, we set $\mathbf{w}_{it} = (y_{it}, \mathbf{z}'_{it})'$ for $t = 1, \ldots, T$. Bars are used to indicate variables including the whole history except the current observation so that $\bar{\mathbf{w}}_{it} = (\mathbf{w}'_{i0}, \mathbf{w}'_{i1}, \ldots, \mathbf{w}'_{i(t-1)})'$.

r_{it} denotes the missing data indicator, such that $r_{it} = 1$, if the pair $(y_{it}, \mathbf{z}_{it})$ is observed and $r_{it} = 0$, if $(y_{it}, \mathbf{z}_{it})$ is missing. We assume a monotone missing data pattern so that $r_{i(t+1)} = 0$, if $r_{it} = 0$ for any t. Thus, $r_{iT} = 1$ indicates that the data of cluster i are completely observed. We assume that the data are missing at random (MAR) in the sense of Laird (1988). This implies that the response probability at time t only depends on observations prior to t. We do, however, not assume that the data are missing completely at random (MCAR; Laird, 1988). This would imply that the probability for a response at time t may depend on the explanatory variables $\mathbf{X}_i$ but not on the history $\mathbf{w}_{it}$ observed up to t. We assume that the probability λ_{it} to remain in the study is bounded away from 0:

$$\lambda_{it} = P(r_{it} = 1 | r_{i(t-1)} = 1, \bar{\mathbf{w}}_{it}) > \delta > 0\,. \tag{4}$$

The response probabilities $\lambda_{it}(\boldsymbol{\gamma}) = \lambda_{it}(\boldsymbol{\gamma}|r_{i(t-1)}, \bar{\mathbf{w}}_{it})$ may depend on an additional parameter $\boldsymbol{\gamma}$ that is modeled as a function of the history up to t and the observation status at $t-1$ as the missing is monotone. If the response probabilities $\lambda_{it}(\boldsymbol{\gamma})$ are unknown, $\boldsymbol{\gamma}$ needs to be estimated. If no observation is missing at a specific time point t, λ_{it} need not be estimated and $\lambda_{it} = 1$. If at least one observation is missing and present, respectively, for every time point, an estimate $\hat{\boldsymbol{\gamma}}$ can be obtained by maximizing the partial Likelihood function $L(\boldsymbol{\gamma}) = \prod_{i=1}^{n} L_i(\boldsymbol{\gamma})$, where

$$L_i(\boldsymbol{\gamma}) = \prod_{t=1}^{T} \left(\lambda_{it}(\boldsymbol{\gamma})^{r_{it}} [1 - \lambda_{it}(\boldsymbol{\gamma})]^{1-r_{it}} \right)^{r_{i(t-1)}}. \tag{5}$$

If the data are MAR, the product $\pi_{it} = \pi_{it}(\boldsymbol{\gamma}) = \lambda_{i1}(\boldsymbol{\gamma}) \cdot \ldots \cdot \lambda_{it}(\boldsymbol{\gamma})$ of the response probabilities including time t may be interpreted as the conditional probability of observing cluster i at time t given the entirely observed history $\bar{\mathbf{w}}_{it}$. In order to formulate the WEE, these conditional probabilities multiplied are collected together with their observational status in a $T \times T$ diagonal matrix $\boldsymbol{\Pi}_i = \boldsymbol{\Pi}_i(\boldsymbol{\gamma})$ with elements r_{it}/π_{it}. The multiplication by the actual observational status ensures that data with missing response do not contribute to the WEE. An estimator $\hat{\boldsymbol{\beta}}_{WEE}$ is the solution of the WEE

$$\mathbf{u}(\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} \boldsymbol{\Pi}_i(\hat{\boldsymbol{\gamma}}) \boldsymbol{\epsilon}_i = \mathbf{0}, \tag{6}$$

Robins et al. (1995, Appendix A) have shown that $\hat{\boldsymbol{\beta}}_{WEE}$ is asymptotically normal with mean $\boldsymbol{\beta}$ under suitable regularity conditions. A strongly consistent estimator of the variance is given e.g. by Robins et al. (1995). The reader should note that the usually applied robust variance matrix of eq. (3) need not yield positive estimates of the variance matrix Robins and Rotnitzky (1995). Robins and Rotnitzky (1995) have furthermore shown that prior knowledge concerning the response probabilities does not provide additional information, if (i) the mean structure is correctly specified, (ii) the data are MAR and (iii) the response probabilities are greater than 0. This result implies that the WEE may also be applied without loss of power even if the data are MCAR. This has been illustrated previously by Ziegler et al. (2000).

4 A Simulation Study

In the first part of our simulation study, we investigate the properties of the WEE estimator for misspecified response probabilities using binary and continuous dependent variables. We use a study design that may be used in a randomized clinical trial. The simulation proceeds as follows.

1. The complete data set is generated without missing observations. The response depends on an intercept and a dummy-coded cluster-constant dichotomous variable. We use a balanced design so that the proportion of '1's is 50%. There is no difference between treatment groups at $t = 1$. At time points 2 and 3, the treatment effect is most pronounced being 1 and 2 for the binary and the continuous response, respectively. At $t = 4$, the treatment effect decreases by $\frac{1}{2}$ compared with $t = 2, 3$. We simulate 100 clusters with $t = 1, \ldots, 4$ each. Throughout the simulations, the number of replicates is 1000. $t = 1$ is used as baseline and assumed to be always observed. As the missing mechanism depends on y_{t-2}, it is necessary for y_{i2} to be always observed, too. The association structure is exchangeable. We use smaller associations for the binary response, since the correlation for multivariate binary data is restricted (s. e.g. Ziegler, Kastner and Blettner, 1998).

The continuous response variables are generated using a multivariate normally distributed variable with a pre-specified correlation structure. The binary response variables are generated using the log-linear representation since higher order marginal moments are restricted in general. However, in contrast

to Fitzmaurice and Laird (1993) who have used conditional second order moments, we generate the second-order moments marginally. Only third and fourth-order moments are conditional.

2. Observations are deleted from the complete simulated data set using pre-specified missing data mechanisms. Missing data are generated using an MAR process. The complete data generation process is described in full detail in Kastner (2000). The WEE are solved by MAREG (Kastner et al., 1997).

Tables 1 and 2 display the simulation results for the binary and the continuous response variable, respectively. FD denotes the complete (full) data. CC is the estimate if only complete clusters are used in the analysis. Thus, clusters are omitted from the analysis, if at least one observation in a cluster is missing. AC is the available case estimator, i.e. all observations are included in the analysis. Finally, WEE is the weighting estimating equations estimator. ME is the arithmetic mean of the parameter estimates from the 1000 replicates, and SD is the arithmetic mean of the estimated standard deviation of the parameter estimates. Finally, CI is the proportion of simulations in which the 95% confidence interval for the specific parameter covers the true parameter.

Table 1: Simulation results for a four-dimensional binary response and a binary treatment variable ($\beta_{\text{treat},1} = 0, \beta_{\text{treat},2} = 1, \beta_{\text{treat},3} = 1, \beta_{\text{treat},4} = 0.5$). Response probabilities either depend on y_{t-2} (denoted by ^{2}) or on y_{t-1} (denoted by ^{1}). The model for the response probability either includes y_{t-2} (denoted by $t-2$) or y_{t-1} (denoted by $t-1$).

			$\beta_{\text{treat},1}$				$\beta_{\text{treat},2}$				$\beta_{\text{treat},3}$				$\beta_{\text{treat},4}$			
Model	ρ		FD	CC	AC	WEE	FD	CC	AC	WEE	FD	CC	AC	WEE	FD	CC	AC	WEE
^{2} t-2	0.1	ME	0.001	0.052	0.001	0.001	1.024	1.447	1.024	1.024	1.029	1.472	1.075	1.042	0.520	0.591	0.591	0.534
		SD	0.286	0.355	0.286	0.283	0.325	0.459	0.325	0.322	0.326	0.464	0.368	0.355	0.296	0.372	0.372	0.364
		CI	0.971	0.952	0.971	0.955	0.943	0.902	0.943	0.943	0.950	0.900	0.960	0.942	0.927	0.948	0.948	0.925
^{2} t-2	0.5	ME	-0.008	0.293	-0.008	-0.008	1.018	1.608	1.018	1.018	1.022	1.608	1.206	1.033	0.508	0.825	0.825	0.527
		SD	0.286	0.359	0.286	0.282	0.325	0.486	0.325	0.321	0.326	0.486	0.382	0.376	0.295	0.387	0.387	0.378
		CI	0.964	0.877	0.964	0.943	0.941	0.837	0.941	0.941	0.933	0.824	0.929	0.941	0.949	0.905	0.905	0.962
^{2} t-1	0.1	ME	0.007	0.069	0.007	0.007	1.045	1.477	1.045	1.045	1.014	1.458	1.058	1.026	0.516	0.585	0.585	0.525
		SD	0.286	0.355	0.286	0.283	0.327	0.464	0.327	0.324	0.324	0.462	0.366	0.354	0.295	0.371	0.371	0.364
		CI	0.971	0.956	0.971	0.954	0.946	0.898	0.946	0.946	0.961	0.895	0.971	0.941	0.948	0.962	0.962	0.937
^{2} t-1	0.5	ME	-0.002	0.276	-0.002	-0.002	1.036	1.614	1.036	1.036	1.037	1.599	1.221	1.049	0.511	0.819	0.819	0.528
		SD	0.286	0.357	0.286	0.282	0.326	0.484	0.326	0.322	0.326	0.480	0.382	0.376	0.295	0.386	0.386	0.376
		CI	0.958	0.888	0.958	0.940	0.946	0.843	0.946	0.945	0.953	0.837	0.942	0.957	0.952	0.907	0.907	0.947
^{1} t-2	0.5	ME	-0.009	0.279	-0.009	-0.009	1.044	1.617	1.044	1.044	1.017	1.573	1.195	1.024	0.522	0.829	0.829	0.528
		SD	0.286	0.358	0.286	0.282	0.327	0.484	0.327	0.323	0.325	0.478	0.380	0.373	0.296	0.386	0.386	0.377
		CI	0.971	0.900	0.971	0.950	0.949	0.825	0.949	0.949	0.943	0.857	0.956	0.956	0.943	0.909	0.909	0.959

Table 2: Simulation results for a four-dimensional normally distributed response and a binary treatment variable ($\beta_{\text{treat},1} = 0, \beta_{\text{treat},2} = 2, \beta_{\text{treat},3} = 2, \beta_{\text{treat},4} = 1$). Response probabilities either depend on y_{t-2} (denoted by ^{2}) or on y_{t-1} (denoted by ^{1}). The model for the response probability either includes y_{t-2} (denoted by $t-2$) or y_{t-1} (denoted by $t-1$).

			$\beta_{\text{treat},1}$				$\beta_{\text{treat},2}$				$\beta_{\text{treat},3}$				$\beta_{\text{treat},4}$			
Model	ρ		FD	CC	AC	WEE	FD	CC	AC	WEE	FD	CC	AC	WEE	FD	CC	AC	WEE
^{2} t-2	0.5	ME	-0.003	0.473	-0.003	-0.003	1.996	2.264	1.996	1.996	1.999	2.247	2.231	2.029	0.995	1.247	1.247	1.032
		SD	0.139	0.145	0.139	0.138	0.140	0.166	0.140	0.138	0.140	0.170	0.169	0.189	0.140	0.169	0.169	0.195
		CI	0.952	0.106	0.952	0.950	0.945	0.643	0.945	0.942	0.943	0.694	0.731	0.858	0.944	0.680	0.680	0.849
^{2} t-2	0.9	ME	0.005	0.482	0.005	0.005	2.007	2.443	2.007	2.007	2.005	2.440	2.426	2.069	1.004	1.437	1.437	1.074
		SD	0.139	0.143	0.139	0.138	0.140	0.150	0.140	0.138	0.140	0.150	0.151	0.183	0.140	0.151	0.151	0.186
		CI	0.942	0.085	0.942	0.940	0.939	0.179	0.939	0.939	0.945	0.179	0.190	0.774	0.946	0.176	0.176	0.755
^{2} t-1	0.5	ME	-0.003	0.478	-0.003	-0.003	1.999	2.269	1.999	1.999	2.008	2.256	2.243	2.176	1.003	1.252	1.252	1.125
		SD	0.139	0.145	0.139	0.121	0.140	0.167	0.140	0.137	0.140	0.170	0.169	0.161	0.140	0.170	0.170	0.172
		CI	0.951	0.092	0.951	0.913	0.942	0.640	0.942	0.939	0.946	0.661	0.685	0.782	0.937	0.672	0.672	0.857
^{2} t-2	0.9	ME	-0.006	0.481	-0.006	-0.006	1.992	2.436	1.992	1.992	1.995	2.436	2.422	2.215	0.990	1.432	1.432	0.989
		SD	0.140	0.145	0.140	0.137	0.139	0.150	0.139	0.138	0.140	0.151	0.152	0.153	0.140	0.151	0.151	0.182
		CI	0.939	0.083	0.939	0.937	0.934	0.188	0.934	0.933	0.941	0.187	0.218	0.620	0.944	0.199	0.199	0.813
^{1} t-2	0.5	ME	0.002	0.316	0.002	0.002	1.997	2.464	1.997	1.997	1.996	2.464	2.177	2.114	0.996	1.309	1.309	1.237
		SD	0.140	0.177	0.140	0.136	0.140	0.156	0.140	0.122	0.140	0.156	0.158	0.141	0.139	0.175	0.175	0.176
		CI	0.930	0.553	0.930	0.925	0.922	0.167	0.922	0.886	0.945	0.157	0.776	0.828	0.935	0.558	0.558	0.692

The first two blocks within tables 1 and 2 show the results for the correctly specified model of the response probabilities. It is seen that the CC and the AC estimators do not yield consistent parameter estimates, since the response data are not MCAR but MAR. The bias increases with the correlation of the responses and, generally, with the proportion of missing values. In contrast to the CC and the AC estimator, the WEE produces consistent parameter estimates. As expected, the standard deviation (SD) of the WEE is greater than that of the FD due to the presence of missing data.

The consistency of the WEE requires the correct specification of the model for the response probabilities. It is interesting to note that the bias turns out to be negligible in the models considered in our simulations if the true missing data process depends on y_{t-2} but y_{t-1} is used to model the response probabilities (tables 1 and 2, blocks 3 and 4). Unexpectedly, there is no trend for an increased bias if the correlation of the responses decreases. The findings also hold if the true missing data process depends on y_{t-1} but y_{t-2} is used in the model of the response probabilities. There seems to be a trend for a bias in the estimate of $\beta_{\text{treat},4}$ for the continuous response (table 2). This trend, however, is not seen for other parameters and in other simulated models (s. table 1; results from other simulated models not shown). Compared with standard approaches the WEE are, however, preferable even if the model for the response probabilities is slightly misspecified (tables 1 and 2).

Tables 1, 2, 3 and 4 show the results for misspecified response probabilities. The endogenous variable used to model the response probabilities is misspecified in tables 1 and 2. Tables 3 and 4, however, display results for both response probabilities estimated without covariates and response probabilities estimated without lagged explanatory variables of the mean model. The simulation to set up tables 3 and 4 proceeds as above with the exception that we use trivariate binary and continuous responses, respectively, instead of four-dimensional dependent variables. The model without X leads to the same conclusions as the misspecified models in tables 1 and 2. In the case that the response probabilities are estimated without any y, i.e. we assume MCAR, the WEE are nearly as biased as CC and AV.

Our simulations clearly demonstrate that the WEE are consistent if the missing data process is MAR and if the model for the response probabilities is correctly specified. The CC and the AC estimator yield biased parameter estimates in these situations. The WEE may yield biased estimates if the model for the response probabilities is misspecified. If, however, the misspecification is negligible, the WEE may serve as an appropriate working model. A theoretical solution that quantifies the bias of the parameter estimates of the mean structure is required for the situation of a misspecified model for the response probability.

Table 3: Simulation results for a trivariate binary response and a binary treatment variable ($\beta_{\text{treat},1} = 0, \beta_{\text{treat},2} = 1, \beta_{\text{treat},3} = 0.5$). The association structure is exchangeable with $\rho = 0.5$. Response probabilities depend on y_{t-1}. Results depend on the model for estimating response probabilities.

Model		$\beta_{\text{treat},1}$				$\beta_{\text{treat},2}$				$\beta_{\text{treat},3}$			
		FD	CC	AC	WEE	FD	CC	AC	WEE	FD	CC	AC	WEE
without X	mean	-0.005	0.506	-0.005	-0.005	1.026	1.619	1.234	1.070	0.544	0.918	0.918	0.657
	std.dev	0.286	0.381	0.286	0.283	0.325	0.506	0.400	0.396	0.297	0.410	0.410	0.404
	CI	0.970	0.748	0.970	0.946	0.949	0.847	0.949	0.962	0.945	0.873	0.873	0.938
without y	mean	-0.004	0.495	-0.004	-0.004	1.029	1.653	1.259	1.259	0.552	0.910	0.910	0.910
	std.dev	0.286	0.380	0.286	0.286	0.326	0.511	0.403	0.403	0.297	0.409	0.409	0.409
	CI	0.974	0.736	0.974	0.974	0.940	0.836	0.949	0.949	0.954	0.859	0.859	0.859

Table 4: Simulation results for a trivariate continuous response and a binary treatment variable ($\beta_{\text{treat},1} = 0, \beta_{\text{treat},2} = 2, \beta_{\text{treat},3} = 1$). The association structure is exchangeable with $\rho = 0.5$. Response probabilities depend on y_{t-1}. Results depend on the model for estimating response probabilities.

Model		$\beta_{\text{treat},1}$				$\beta_{\text{treat},2}$				$\beta_{\text{treat},3}$			
		FD	CC	AC	WEE	FD	CC	AC	WEE	FD	CC	AC	WEE
MAR without X	mean	0.008	0.447	0.008	0.008	1.999	2.240	2.209	2.091	1.007	1.236	1.236	1.113
	std.dev	0.139	0.143	0.139	0.138	0.138	0.162	0.164	0.177	0.139	0.165	0.165	0.180
	CI	0.942	0.127	0.942	0.940	0.933	0.679	0.753	0.856	0.944	0.684	0.684	0.826
MCAR	mean	-0.000	0.477	-0.000	-0.000	1.992	2.262	2.225	2.225	1.000	1.245	1.245	1.245
	std.dev	0.140	0.146	0.140	0.140	0.139	0.165	0.168	0.168	0.139	0.170	0.170	0.170
	CI	0.934	0.095	0.934	0.934	0.947	0.632	0.710	0.710	0.947	0.673	0.673	0.673

In the second part of our simulation study we analyze the effect of misspecifying the link function for the response probability model. The simulation proceeds as above. We use trivariate binary responses and trivariate continuous responses, respectively. Subject to variation were the sample size (number of clusters n), the correlation ρ of the association structure, the proportion of missing values (%miss) and the link function used to generate the missing values (Link).

Table 5 shows the results for a trivariate binary response and a binary treatment variable with $(\beta_{\text{treat},1} = 0, \beta_{\text{treat},2} = 2, \beta_{\text{treat},3} = 1)$. The association structure is exchangeable with $\rho = 0.5$. The model for the response probabilities is correctly specified. The logit link function and the compound loglog link function are used to generate the missing data. The proportion of missing values in the treatment group (%miss) varies between 6% and 18%. The sample consists of either 100 or 1000 clusters. Obviously, parameter estimates, standard deviations and coverage probabilities differ only slighty between the logit, probit and the compund loglog link functions that are used to solve the WEE. These results hold for missing data generated using the logit and the compund loglog link, respectively. Nevertheless, we have to point out that the data for $\beta_{\text{treat},2}$ are slightly biased even for the complete data (FD).

Table 6 shows the simulation results for a trivariate binary response, a binary treatment variable with $(\beta_{\text{treat},1} = 0, \beta_{\text{treat},2} = 1, \beta_{\text{treat},3} = 0.5)$ and a continuous covariate $(\beta_{\text{cont}} = 1)$. The association structure is exchangeable with $\rho = 0.5$. The model for the response probabilities is correctly specified and missing data are generated using the logit link function. The proportion of missing values in the treatment group (%miss) varies between 4% and 86%. The sample consists either of 100 or 1000 clusters. Analogously to table 5, the results are very similar for the different link functions. It is interesting to note that the treatment effect at time 2 can be estimated appropriately from a total of 100 clusters even if 62% and 86% of the data in the treatment group are missing at times 2 and 3, respectively.

Table 7 shows the simulation results for a normally distributed response and a binary treatment variable $(\beta_{\text{treat},1} = 0, \beta_{\text{treat},2} = 2, \beta_{\text{treat},3} = 1)$. Table 8 displays simulation results for similar models. However, a continuous covariate $(\beta_{\text{cont}} = 1)$ is included in the model in addition to the binary treatment variable. In both simulation settings, the association structure is exchangeable. The model for the response probabilities is correctly specified. The number of clusters is 1000. Missing data are generated using the logit link function. Results depend on the proportion of missing values in the treatment group (%miss) and the correlation of the responses (ρ).

The simulation results for normally distributed responses are similar to those of the binary dependent variables. Thus, there is no apparent difference between the three applied link functions. However, the reader should note that the coverage probabilities for $\beta_{\text{treat},2}$ decrease if a) the proportion of missing values increases and b) the correlation increases. These results are as expected since the precision decreases with a higher proportion of missing values. Furthermore, the additional information gained from multiple observations within a cluster is low for high response correlations.

Summing up, the second part of our simulations shows that the use of the link function to model the response probabilities is expected to have a negligible effect in most circumstances.

Table 5: Simulation results for a trivariate binary response and a binary treatment variable ($\beta_{\text{treat},1} = 0, \beta_{\text{treat},2} = 2, \beta_{\text{treat},3} = 1$). The association structure is exchangeable with $\rho = 0.5$. The model for the response probabilities is correctly specified. Results depend on the link function used to generate the missing data (Link), the proportion of missing values in the treatment group (%miss) and the number of clusters (n).

				$\beta_{\text{treat},1}$				$\beta_{\text{treat},2}$				$\beta_{\text{treat},3}$			
Link	%miss	n		FD	logit	probit	loglog	FD	logit	probit	loglog	FD	logit	probit	loglog
logit	6	100	ME	0.079	0.079	0.079	0.079	1.892	1.899	1.889	1.894	1.118	1.120	1.105	1.113
			SD	0.286	0.285	0.283	0.283	0.431	0.436	0.438	0.439	0.333	0.323	0.343	0.343
			CI	0.970	0.967	0.945	0.945	0.927	0.931	0.929	0.935	0.954	0.946	0.963	0.965
logit	6	1 000	ME	0.082	0.082	0.082	0.082	1.816	1.816	1.804	1.811	1.084	1.083	1.065	1.074
			SD	0.090	0.090	0.089	0.089	0.129	0.132	0.132	0.132	0.103	0.106	0.106	0.106
			CI	0.829	0.829	0.829	0.829	0.657	0.686	0.657	0.671	0.871	0.896	0.923	0.908
logit	16	100	ME	0.090	0.090	0.090	0.090	1.866	1.868	1.861	1.865	1.117	1.119	1.107	1.114
			SD	0.286	0.284	0.283	0.284	0.425	0.461	0.463	0.463	0.333	0.370	0.377	0.377
			CI	0.951	0.932	0.926	0.929	0.912	0.913	0.911	0.913	0.947	0.958	0.967	0.966
logit	16	1 000	ME	0.085	0.085	0.085	0.085	1.818	1.821	1.813	1.817	1.087	1.091	1.077	1.085
			SD	0.090	0.090	0.090	0.090	0.129	0.142	0.142	0.142	0.103	0.116	0.116	0.116
			CI	0.835	0.835	0.835	0.835	0.674	0.739	0.719	0.728	0.860	0.877	0.898	0.889
loglog	18	100	ME	0.083	0.083	0.083	0.083	1.886	1.888	1.894	1.891	1.116	1.110	1.119	1.113
			SD	0.286	0.285	0.284	0.284	0.430	0.469	0.472	0.472	0.334	0.371	0.382	0.381
			CI	0.950	0.935	0.924	0.925	0.915	0.912	0.917	0.915	0.925	0.927	0.942	0.942
loglog	18	1 000	ME	0.086	0.086	0.086	0.086	1.823	1.816	1.825	1.819	1.092	1.077	1.091	1.082
			SD	0.090	0.101	0.090	0.090	0.130	0.142	0.143	0.142	0.103	0.117	0.118	0.117
			CI	0.837	0.837	0.837	0.837	0.682	0.720	0.735	0.723	0.847	0.914	0.900	0.910

Table 6: Simulation results for a trivariate binary response, a binary treatment variable ($\beta_{\text{treat},1} = 0, \beta_{\text{treat},2} = 1, \beta_{\text{treat},3} = 0.5$) and a continuous covariate ($\beta_{\text{cont}} = 1$). The association structure is exchangeable with $\rho = 0.5$. The model for the response probabilities is correctly specified. Missing data are generated using the logit link function. Results depend on the proportion of missing values in the treatment group (%miss) and the number of clusters (n).

			$\beta_{\text{treat},1}$				$\beta_{\text{treat},2}$				$\beta_{\text{treat},3}$				β_{cont}			
%miss	n		FD	logit	probit	loglog	FD	logit	probit	loglog	FD	logit	probit	loglog	FD	logit	probit	loglog
4	100	ME	0.005	0.003	-0.001	0.001	1.017	1.014	1.005	1.010	0.602	0.602	0.591	0.597	1.022	1.027	1.036	1.031
		SD	0.371	0.363	0.370	0.369	0.433	0.435	0.435	0.435	0.398	0.404	0.406	0.405	0.536	0.506	0.538	0.534
		CI	0.951	0.944	0.948	0.948	0.939	0.937	0.938	0.937	0.948	0.947	0.949	0.949	0.954	0.915	0.946	0.944
4	1 000	ME	-0.002	-0.003	-0.009	-0.005	1.002	1.001	0.989	0.996	0.588	0.588	0.572	0.580	1.008	1.009	1.023	1.015
		SD	0.116	0.117	0.117	0.117	0.134	0.136	0.136	0.136	0.124	0.126	0.127	0.127	0.168	0.172	0.173	0.172
		CI	0.941	0.944	0.941	0.944	0.950	0.949	0.949	0.950	0.884	0.882	0.907	0.895	0.949	0.951	0.950	0.947
14	100	ME	0.012	0.014	0.009	0.011	1.042	1.043	1.031	1.038	0.621	0.621	0.606	0.615	1.031	1.028	1.039	1.033
		SD	0.372	0.373	0.375	0.375	0.437	0.460	0.462	0.461	0.400	0.434	0.437	0.436	0.538	0.549	0.557	0.556
		CI	0.944	0.937	0.941	0.938	0.945	0.944	0.947	0.947	0.950	0.949	0.953	0.951	0.946	0.936	0.940	0.941
14	1 000	ME	-0.002	-0.002	-0.007	-0.004	1.003	1.004	0.992	0.999	0.586	0.587	0.570	0.579	0.999	0.999	1.011	1.004
		SD	0.116	0.118	0.118	0.118	0.134	0.142	0.142	0.142	0.124	0.135	0.135	0.135	0.168	0.176	0.176	0.176
		CI	0.956	0.961	0.960	0.961	0.954	0.957	0.959	0.955	0.920	0.910	0.931	0.924	0.960	0.965	0.965	0.965
86	100	ME	-0.033	-0.053	-0.061	-0.053	0.967	0.943	0.908	0.943	0.566	-0.183	-0.231	-0.183	1.029	1.076	1.094	1.077
		SD	0.371	0.397	0.397	0.397	0.427	0.639	0.643	0.639	0.396	0.924	0.933	0.924	0.536	0.644	0.643	0.644
		CI	0.955	0.958	0.956	0.958	0.948	0.950	0.947	0.950	0.950	0.939	0.941	0.938	0.937	0.939	0.938	0.938
86	1 000	ME	-0.001	0.001	-0.006	0.001	0.999	1.002	0.970	1.003	0.589	0.607	0.561	0.608	1.003	0.998	1.016	0.998
		SD	0.116	0.126	0.125	0.126	0.134	0.200	0.200	0.200	0.124	0.294	0.294	0.294	0.168	0.204	0.203	0.204
		CI	0.955	0.952	0.952	0.951	0.951	0.951	0.941	0.952	0.898	0.950	0.955	0.950	0.964	0.958	0.955	0.958

Table 7: Simulation results for a normally distributed response and a binary treatment variable ($\beta_{\text{treat},1} = 0, \beta_{\text{treat},2} = 2, \beta_{\text{treat},3} = 1$). The association structure is exchangeable. The model for the response probabilities is correctly specified. The number of clusters is 1000. Missing data are generated using the logit link function. Results depend on the proportion of missing values in the treatment group (%miss) and the correlation of the responses (ρ).

			$\beta_{\text{treat},1}$				$\beta_{\text{treat},2}$				$\beta_{\text{treat},3}$			
%miss	ρ		FD	logit	probit	loglog	FD	logit	probit	loglog	FD	logit	probit	loglog
4	0.5	ME	0.003	0.003	0.003	0.003	2.003	2.003	2.003	2.003	1.004	1.004	1.004	1.004
		SD	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.046	0.046	0.046
		CI	0.950	0.950	0.950	0.950	0.948	0.956	0.952	0.954	0.945	0.945	0.946	0.945
4	0.9	ME	-0.001	-0.001	-0.001	-0.001	1.999	1.999	2.001	2.000	1.000	1.000	1.001	1.000
		SD	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045
		CI	0.960	0.959	0.959	0.959	0.962	0.961	0.961	0.961	0.955	0.958	0.958	0.959
12	0.5	ME	0.000	0.000	0.000	0.000	2.000	2.001	2.013	2.006	1.000	1.001	1.012	1.005
		SD	0.045	0.045	0.044	0.045	0.045	0.048	0.047	0.048	0.045	0.049	0.048	0.048
		CI	0.953	0.952	0.951	0.951	0.949	0.944	0.936	0.949	0.933	0.937	0.936	0.940
12	0.9	ME	-0.001	-0.001	-0.001	-0.001	1.999	2.001	2.024	2.010	0.999	1.002	1.023	1.010
		SD	0.044	0.044	0.044	0.044	0.044	0.049	0.046	0.047	0.044	0.050	0.047	0.048
		CI	0.935	0.935	0.935	0.935	0.947	0.947	0.912	0.938	0.938	0.947	0.914	0.939
86	0.5	ME	0.000	0.000	0.000	0.000	2.002	2.002	2.026	1.994	0.998	0.999	1.018	0.993
		SD	0.045	0.045	0.045	0.045	0.045	0.090	0.088	0.091	0.045	0.134	0.130	0.135
		CI	0.962	0.961	0.961	0.961	0.968	0.953	0.945	0.955	0.955	0.945	0.937	0.946

Table 8: Simulation results for a normally distributed response, a binary treatment variable ($\beta_{\text{treat},1} = 0, \beta_{\text{treat},2} = 2, \beta_{\text{treat},3} = 1$) and a continuous covariate ($\beta_{\text{cont}} = 1$). The association structure is exchangeable. The model for the response probabilities is correctly specified. The number of clusters is 1000. Missing data are generated using the logit link function. Results depend on the proportion of missing values in the treatment group (%miss) and the correlation of the responses (ρ).

			$\beta_{\text{treat},1}$				$\beta_{\text{treat},2}$				$\beta_{\text{treat},3}$				β_{cont}			
%miss	ρ		FD	logit	probit	loglog	FD	logit	probit	loglog	FD	logit	probit	loglog	FD	logit	probit	loglog
12	0.5	ME	-0.001	-0.001	-0.001	-0.000	1.999	1.998	2.009	2.004	1.001	1.001	1.010	1.006	1.000	1.000	1.000	0.999
		SD	0.056	0.057	0.056	0.057	0.056	0.059	0.059	0.059	0.056	0.060	0.059	0.060	0.081	0.082	0.082	0.082
		CI	0.941	0.942	0.940	0.943	0.962	0.956	0.952	0.955	0.944	0.955	0.951	0.952	0.936	0.935	0.940	0.936
66	0.5	ME	-0.002	-0.002	-0.008	-0.004	1.996	1.996	2.034	2.004	0.998	0.999	1.028	1.004	1.003	1.003	1.017	1.007
		SD	0.056	0.060	0.058	0.059	0.057	0.077	0.072	0.076	0.056	0.084	0.076	0.082	0.081	0.093	0.090	0.092
		CI	0.956	0.950	0.949	0.950	0.939	0.946	0.923	0.939	0.945	0.954	0.934	0.951	0.959	0.946	0.935	0.947
66	0.9	ME	-0.000	0.001	-0.011	-0.003	2.001	2.006	2.074	2.021	1.001	1.007	1.059	1.015	1.002	1.001	1.028	1.010
		SD	0.060	0.065	0.062	0.064	0.060	0.085	0.074	0.083	0.060	0.095	0.081	0.093	0.096	0.110	0.103	0.108
		CI	0.947	0.951	0.945	0.948	0.948	0.947	0.825	0.933	0.951	0.934	0.868	0.926	0.953	0.947	0.928	0.943
88	0.5	ME	-0.001	-0.001	-0.024	0.005	2.001	2.005	2.045	1.957	1.002	1.037	1.053	1.010	0.998	0.998	1.051	0.983
		SD	0.057	0.092	0.085	0.096	0.056	0.169	0.150	0.187	0.056	0.219	0.202	0.227	0.081	0.186	0.169	0.194
		CI	0.936	0.919	0.899	0.908	0.952	0.923	0.912	0.925	0.962	0.860	0.869	0.831	0.952	0.905	0.875	0.899

5 Discussion

The application of GEE to estimate clustered data has become increasingly popular in recent years. They have been implemented in several standard software packages (s. e.g. Ziegler and Grömping, 1998). However, they all rely on the assumption that missing data are only MCAR in Laird's (1988) sense. Imputation (Paik, 1997; Xie and Paik, 1997) or weighting approaches (Robins et al., 1995; Robins and Rotnitzky, 1995; Rotnitzky and Robins, 1995) that may be applied if missing data in the responses are MAR in Laird's (1988) sense have only received little attention. Recently, it has been stated that "one could possibly exploit recent developments in semi-parametric modeling approaches". However, it has been criticized that this possibility is not "available in an accessible form with current software" (Carlin et al., 1999). This important shortcoming of the WEE that weights observations inversely proportional to their respective response probabilities has now been solved. The WEE proposed by Robins and co-workers have been implemented in the stand-alone package MAREG (Kastner et al., 1997) which is freely available. They seem to be an interesting approach to estimate the marginal from clustered follow-up data. Its current implementation is, however, restricted to monotone missing data patterns. An extension to arbitrary missing data patterns is under construction.

In this paper we have investigated the effect of a slightly misspecified model for the response probabilities and a misspecified link function by simulations using the WEE for monotone missing data patterns.

Firstly, we have shown that the WEE yield consistent parameter estimates of the mean structure if missing responses are MAR, while the usually applied complete cluster (CC) or available case (AC) estimator failed to be consistent. Our results furthermore demonstrate that a slight misspecification of the model for the response probabilities will have a negligible effect on the parameter estimates of the mean structure. This result fits well within the classical findings for linear and generalised linear models (GLM, s. e.g. Greene, 1993).

Secondly, we have shown that the choice of the link function has a negligible effect on the parameter estimates of the mean structure. This result also agrees with classical findings for GLM and GEE (Park and Weisberg, 1998; Li and Duan, 1989). These authors have shown that FISHER consistent estimates of regression coefficients can be obtained even if the link function in the GLM is misspecified. Thus, we can expect that the choice of the link function for the response probabilities will generally have a trifling effect on the estimate of the response probabilities.

Theoretical work would be helpful that allows a quantification of the bias introduced by a misspecification of the model for the response probabilities. Further work should also investigate the use of alternatives to GLM for the modeling of response probabilities for WEE.

Acknowledgements

The work of C.K. was supported by the Deutsche Forschungsgemeinschaft.

References

Carlin, J. B., Wolfe, R., Coffey, C. and Patton, G. C. (1999). Analysis of binary outcomes in longitudinal studies using weighted estimating equations and discrete-time survival methods: Prevalence and incidence of smoking in an adolescent cohort, *Statistics in Medicine* **18**: 2655–2679.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses, *Biometrika* **80**: 141–151.

Greene, W. H. (1993). *Econometric Analysis*, 2 edn, Macmillan, New York.

Kastner, C. (2000). *Fehlende Werte bei korrelierten Beobachtungen*, Dissertation, Ludwig-Maximilians-Universität München.

Kastner, C., Fieger, A. and Heumann, C. (1997). MAREG and WinMAREG—a tool for marginal regression models, *Computational Statistics and Data Analysis* **24**: 235–241. URL: (http://www.stat.uni-muenchen.de/∼andreas/mareg/winmareg.html)

Laird, N. M. (1988). Missing data in longitudinal studies, *Statistics in Medicine* **7**: 305–315.

Li, K. C. and Duan, N. (1989). Regression analysis under link violation, *Annals of Statistics* **17**: 1009–1052.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**: 13–22.

Little, R. J. A. and Schenker, N. (1995). Missing data, *in* G. Arminger, C. C. Clogg and M. E. Sobel (eds), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Plenum, New York, pp. 39–75.

Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random, *Journal of the American Statistical Association* **92**: 1320–1329.

Park, C. and Weisberg, S. (1998). Fisher consistency of GEE models under link misspecification, *Computational Statistics and Data Analysis* **27**: 229–235.

Robins, J. M. and Rotnitzky, A. G. (1995). Semiparametric efficiency in multivariate regression models with missing data, *Journal of the American Statistical Association* **90**: 122–129.

Robins, J. M., Rotnitzky, A. G. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association* **90**: 106–120.

Rotnitzky, A. G. and Robins, J. M. (1995). Semiparametric estimation of models for means and covariances in the presence of missing data, *Scandinavian Journal of Statistics* **22**: 323–333.

Xie, F. and Paik, M. C. (1997). Multiple imputation methods for the missing covariates in generalized estimating equation, *Biometrics* **53**: 1538–1546.

Zeger, S. L., Liang, K.-Y. and Self, S. G. (1985). The analysis of binary longitudinal data with time-independent covariates, *Biometrika* **72**: 31–38.

Ziegler, A. and Grömping, U. (1998). The generalised estimating equations: A comparison of procedures available in commercial statistical software packages, *Biometrical Journal* **40**: 245–260.

Ziegler, A., Kastner, C. and Blettner, M. (1998). The generalised estimating equations: An annotated bibliography, *Biometrical Journal* **40**: 115–139.

Ziegler, A., Kastner, C. and Chang-Claude, J. (2000). Analysis of pregnancy and other factors on detection of HPV infection using weighted estimating equations for follow-up data, *SFB386 – Discussion paper 201*, Ludwig-Maximilians-Universität München.