ML- and Semiparametric Estimation in Logistic Models with Incomplete Covariate Data

VANESSA DIDELEZ

Department of Statistics, University College London,

Gower Street, London WC1 6BT, vanessa@stats.ucl.ac.uk

Abstract: ML–estimation of regression parameters with incomplete covariate information usually requires a distributional assumption regarding the concerned covariates that implies a source of misspecification. Semiparametric procedures avoid such assumptions at the expense of efficiency. In this paper a simulation study with small sample size is carried out to get an idea of the performance of the ML–estimator under misspecification and to compare it with the semiparametric procedures when the former is based on a correct assumption. The results show that there is only little gain by correct parametric assumptions, which does not justify the possibly large bias when the assumptions are not met. Additionally, the easily computed estimator proposed by CAIN and BRESLOW (1988) appears to be nearly semiparametric efficient.

Keywords: Logistic regression; Maximum likelihood; EM algorithm; Missing covariates; Missing data; Semiparametric efficiency.

Acknowledgements: The author would like to thank Iris Pigeot and Werner Vach for helpful comments. Financial support by the Deutsche Forschungsgemeinschaft is also gratefully acknowledged.

Short title: ML- and Semiparametric Estimation with Incomplete Covariates.

1 Introduction

The problem of coping with incomplete information in the covariates when estimating a regression parameter is common in applied work. A simple solution is given by the complete case analysis where all incomplete cases are discarded. The resulting complete case estimator, however, is obviously inefficient and not generally consistent under the missing at random assumption (MAR). This assumption excludes that the observability of a variable depends on its unobserved value but it does allow dependence on the values of the observed variables (RUBIN, 1976). Another standard method to cope with missings consists in imputing the unobserved values and then treating the data set as if it was complete. Besides requiring an appropriate model for generating the imputed values, this method obviously does not take the additional uncertainty due to the missing values into account and generates misleading variance estimations. Instead it seems in general more reasonable to use multiple imputation as proposed by RUBIN (1987) which is facilitated by simulation methods, e.g. Markov chain Monte Carlo, developed in the past decades (SCHAFER, 1997). However, in the context of regression analysis multiple imputation as well as proper ML–estimation via the EM–algorithm necessitate a distributional assumption concerning the covariates. Other approaches avoid the assumption of any distribution concerning the covariates and are therefore often termed 'semiparametric'. This paper aims at comparing these two strategies, assuming a possibly false covariate distribution versus avoiding such an assumption, in order to give an impression on the sensitivity to false distributional assumptions as well as to the loss of information when a semiparametric method is chosen. We further focus on the simple task of non–Bayesian estimation of the regression parameters and therefore omit multiple imputation. For a discussion of the latter method in the context of misspecified or semiparametric models and the associated problem of estimating the variance we refer to ROBINS and WANG (2000).

Full parametric procedures have for instance been proposed by LITTLE (1992),

BLACKHURST and SCHLUCHTER (1989), IBRAHIM (1990), and IBRAHIM and WEISBERG (1992). Typically, the conditional distribution of the incomplete covariate given a subset of, or all, the other covariates and the response variable is specified. The resulting ML–estimator is asymptotically efficient if the assumptions are correct. However, misspecification is likely to occur when restrictive assumptions are inevitable as for example when one of the involved variables is continuous. In such a situation, LITTLE (1992) and IBRAHIM and WEISBERG (1992) use a Gaussian covariate distribution. However, there is no apparent reason why this standard assumption should be correct and, so far, nothing is known about its 'robustness' against misspecification. In our simulation study we therefore investigate the behaviour of a ML–estimator assuming a Gaussian covariate distribution while the true distribution is Student or χ^2.

Semiparametric procedures have been intensively investigated in the last years. For the situation of two–stage case–control studies BRESLOW and CAIN (1988) propose a pseudoconditional likelihood approach yielding a consistent estimator under the MAR assumption. It has been shown (CAIN and BRESLOW, 1988; VACH and ILLI, 1997) that in the situation of a logistic regression model this turns out to be a simple modification of the complete case estimator. Another approach uses the empirical distribution or nonparametric kernel estimates to estimate the unknown distribution. This has been proposed by PEPE and FLEMMING (1991) and CARROLL and WAND (1991) in the context of mismeasured covariates but the resulting estimators for the regression parameter are only consistent if the missing mechanism is MAR and does not depend on the response variable. REILLY and PEPE (1995) apply the same idea in order to estimate the score function for incomplete observations. Their so–called mean score estimator is consistent under MAR. ROBINS et al. (1994) and ROBINS et al. (1995) address the performance of such semiparametric estimators by considering the lower variance bound of any regular semiparametric estimator. The estimator that attains this variance bound, however, usually depends on the unknown covariate distribution. For rather general

cases they describe adaptive semiparametric efficient estimators which are feasible without this knowledge. The simulation study conducted here gives an idea of the gain in efficiency of this estimator as compared to the mean score and the pseudo-conditional likelihood methods. In addition, the simulation results allow contrasting the performance of the semiparametric efficient estimator to the ML–estimator with correct assumption about the covariate distribution for finite sample size.

The outline of the paper is as follows. We restrict ourselves to the situation of a logistic regression model, which is mostly used in practice to model the influence of one ore more explanatory variables on a binary response. The situation of a binary outcome with discrete as well as continuous covariates often arises in biomedical applications e.g. when recovery from a disease is considered. In Section 2 we describe this model and the missing data situation to which we apply the different estimators that are, in turn, presented in Section 3. The considered estimators are the complete case, the ML–, the Breslow–and–Cain, the mean score, and the semiparametric efficient estimator. These are motivated for the special situation of a logistic regression restricting the presentation to the essentials since the general situation is treated in the literature mentioned above. The simulation designs are described in Section 4, the results of the simulation study in Section 5. Finally, we discuss the obtained results.

2 The Model

We compare the different approaches to estimating a regression parameter along the special case of a logistic regression. Let Y denote a binary response variable, X_1 a completely observed binary covariate and X_2 an incompletely observed continuous covariate. The logistic regression model is given by the assumption that

$$\Pr(Y = 1 | X_1 = x_1, X_2 = x_2; \beta) = \frac{\exp\{\beta^\top x^*\}}{1 + \exp\{\beta^\top x^*\}},$$

where $x^* = (1, x_1, x_2)^\top$, and $\beta^\top = (\beta_0, \beta_1, \beta_2)$ is the parameter vector to be estimated. For ease of notation we also write $\Pr(y|x_1, x_2; \beta)$ instead of $\Pr(Y = y|X_1 = x_1, X_2 = x_2; \beta)$.

The considered missing situation can be described as follows. Let R be an indicator variable indicating whether X_2 is observable ($R = 1$) or not ($R = 0$). The missing mechanism is assumed to satisfy the MAR assumption, that is $\Pr(R = 1|y, x_1, x_2) = \Pr(R = 1|y, x_1)$ for all y, x_1, x_2. These conditional probabilities for complete observations will be denoted by q_{yx_1}, $y, x_1 \in \{0, 1\}$, and are assumed to be bounded away from zero. The MAR assumption implies that unobserved values of X_2 have the same conditional distribution as the observed values. The likelihood generating (Y, X_1, X_2, R) reads as

$$\begin{aligned} L(\beta, \theta) &= f(x_1|\alpha) f(r|y, x_1; \gamma) \left\{\Pr(y|x_1, x_2; \beta) f(x_2|x_1; \xi)\right\}^r \\ &\quad \left\{\int \Pr(y|x_1, z; \beta) f(z|x_1; \xi)\, dz\right\}^{1-r}, \end{aligned} \tag{1}$$

where $\theta = (\alpha, \gamma, \xi)$ and f is used as generic symbol for a density. The parameters α, γ and ξ refer to the marginal distribution of X_1, to the conditional distribution of R given Y and X_1 which is Bernoulli with probabilities q_{yx_1}, and to the conditional distribution of X_2 given X_1, respectively. Maximising (1) in β is obviously not feasible without knowledge of $f(\cdot|x_1; \xi)$ whereas knowledge of $f(\cdot|\alpha)$ and the missing mechanism is not required as long as the latter is MAR. The parametric approach that will be proposed in the next section consists in specifying $f(\cdot|x_1; \xi)$ up to the unknown parameter ξ, which is assumed to be finite, and then maximising (1) simultaneously in β and ξ. The semiparametric approach views θ as an 'infinite dimensional' parameter with values in the set of the corresponding densities.

Assume now that (Y^i, X_1^i, X_2^i, R^i), $i = 1, \ldots, N$, is an independent sample of (Y, X_1, X_2, R). With $\mathcal{V} = \{i|r^i = 1\}$, the observable data is given by $\{(y^i, x_1^i, x_2^i, r^i)|i \in \mathcal{V}\} \cup \{(y^i, x_1^i, r^i)|i \in \overline{\mathcal{V}}\}$ where $\overline{\mathcal{V}} = \{1, \ldots, N\} \backslash \mathcal{V}$. The empirical response rates $\hat{q}_{yx_1}$ are given as the proportion of sample units with values y, x_1 and an unobserved X_2 among all those with values y and x_1. These rates can be regarded as estimates of

q_{yx_1}, $y, x_1 \in \{0,1\}$. Note that this straightforward estimation is only possible if Y and X_1 are discrete.

3 The Estimators

3.1 Complete case analysis

The complete case analysis consists in applying complete data methods to the reduced data set $\{(y^i, x_1^i, x_2^i) | i \in \mathcal{V}\}$, i.e. it maximizes

$$L^{CC}(\beta) = \prod_{i \in \mathcal{V}} \Pr(y^i | x_1^i, x_2^i; \beta).$$

The resulting estimator will be denoted by $\hat{\beta}^{CC}$. As shown by VACH and BLETTNER (1991) it is consistent if the missingness is conditionally independent of Y given X_1 and X_2 but it may be biased under MAR (see also ZHAO et al., 1996). Obviously the complete case estimator is in general not efficient since it ignores the information in $\{(y^i, x_1^i) | i \in \overline{\mathcal{V}}\}$.

3.2 ML–estimation

Following IBRAHIM and WEISBERG (1992) the considered ML–estimator is computed under the assumption that the conditional distribution of X_2 given X_1 is Gaussian. This is parametrised as follows: Let $\mu_x = \mathrm{E}(X_2 | X_1 = x)$, $x \in \{0,1\}$, denote the means depending on X_1, and σ^2 the variance, which is independent of X_1, i.e. we have in (1) that $\xi = (\mu_0, \mu_1, \sigma^2)$. The likelihood to be maximised is given by

$$L^{ML}(\beta, \xi) \;=\; \prod_{i \in \mathcal{V}} \left[\Pr(y^i | x_1^i, x_2^i; \beta) f(x_2^i | x_1^i; \xi) \right] \prod_{j \in \overline{\mathcal{V}}} \left[\int \Pr(y^j | x_1^j, z; \beta) f(z | x_1^j; \xi)\, dz \right],$$

where $f(\cdot | x_1; \xi)$ is the density of the Gaussian distribution with parameter $\xi = (\mu_{x_1}, \sigma^2)$. In general, maximisation of $L^{ML}(\beta, \xi)$ has to be carried out numerically due to the integration in the second product. This can partly be simplified by using

the EM algorithm (DEMPSTER et al., 1977), which is easy to apply when the considered model is an exponential family. In our special case, the joint conditional distribution of Y and X_2 given X_1 constitutes an exponential family as one can easily check. Still, the E–step involves numerical integration in order to compute the expectations with respect to the distribution of X_2 given Y and X_1 with density

$$f(x_2|y,x_1;\xi,\beta) = \frac{\Pr(y|x_1,x_2;\beta)f(x_2|x_1;\xi)}{\int \Pr(y|x_1,z;\beta)f(z|x_1;\xi)\,dz}. \tag{2}$$

In our simulation the denominator is approximated by a 10 point Gaussian quadrature which is sufficiently exact according to IBRAHIM and WEISBERG (1992).

3.3 Semiparametric estimation

In this section we first present some specific semiparametric estimators that leave the unknown distributions in (1) completely unrestricted and that are consistent under the MAR assumption. Their relation to the parametric ML–estimator is also addressed. Thereafter, a general class of semiparametric estimators is introduced containing the semiparametric efficient estimator.

3.3.1 Corrected complete case estimator

The complete case estimator may be biased under the MAR assumption. By considering the bias factor VACH and ILLI (1997) show that in the special case of a logistic regression model a simple correction is given by

$$\begin{aligned}
\hat{\beta}_0^{CCC} &= \hat{\beta}_0^{CC} + \log\frac{\hat{q}_{00}}{\hat{q}_{10}},\\
\hat{\beta}_1^{CCC} &= \hat{\beta}_1^{CC} + \log\frac{\hat{q}_{10}\hat{q}_{01}}{\hat{q}_{00}\hat{q}_{11}},\\
\hat{\beta}_2^{CCC} &= \hat{\beta}_2^{CC}.
\end{aligned} \tag{3}$$

Note that this estimator utilises the incomplete observations since the correction terms use $\hat{q}_{yx_1}$ and therefore the additional knowledge about the frequencies $N(y,x_1)$. CAIN and BRESLOW (1988) derive $\hat{\beta}^{CCC}$ as a special case of a pseudoconditional

likelihood approach in a more general setting where they prove the asymptotic normality (BRESLOW and CAIN, 1988).

3.3.2 Mean score estimator

As shown by ROBINS et al. (1995), the contribution of an incomplete observation to the total score function is given by the derivation of the logarithm of (1) with respect to β, which can be written as

$$E\left(\frac{\partial}{\partial\beta}\log\Pr(Y|X_1,X_2;\beta)\middle|\, Y=y, X_1=x_1\right)$$
$$=\int\left(\frac{\partial}{\partial\beta}\log\left[\Pr(y|x_1,x_2;\beta)\right]\right)f(x_2|y,x_1;\beta,\xi)\,dx_2 \tag{4}$$

evaluated at the unknown true conditional density $f(\cdot|y,x_1;\beta,\xi)$. Under the MAR assumption a consistent estimate of $f(\cdot|y,x_1;\beta,\xi)$ can be based on the complete cases. REILLY and PEPE (1995) choose the empirical conditional distribution leading to

$$\sum_{i\in\mathcal{V}(y,x_1)}\frac{1}{V(y,x_1)}\frac{\partial}{\partial\beta}\log\Pr(y|x_1,x_2^i;\beta) \tag{5}$$

as an estimator for (4), where $V(y,x_1)=\#\{i\in\mathcal{V}|y^i=y,x_1^i=x_1\}$ and $\mathcal{V}(y,x_1)=\{i\in\mathcal{V}|y^i=y\wedge x_1^i=x_1\}$. As shown by the authors, replacing the unknown contribution of an incomplete observation to the total score function by (5) leads to a weighted sum of the contributions of the complete cases which motivates the name of the mean score method. The estimated total score function is thus given by

$$\sum_{i\in\mathcal{V}}\left(\frac{N(y^i,x_1^i)}{V(y^i,x_1^i)}\right)\frac{\partial}{\partial\beta}\log\Pr(y^i|x_1^i,x_2^i;\beta),$$

where $N(y,x_1)=\#\{i\in\{1,\ldots,N\}|y^i=y,x_1^i=x_1\}$. Computation of the corresponding estimator $\hat{\beta}^{RS}$ as root of the above expression is straightforward. REILLY and PEPE (1995) show that it is consistent and asymptotically normal.
PEPE and FLEMING (1991) and CARROLL and WAND (1991) pursue a similar idea. Note that expression (4) can be rewritten as

$$\int\left(\frac{\partial}{\partial\beta}\log\left[\Pr(y|x_1,x_2;\beta)\right]\right)\frac{\Pr(y|x_1,x_2;\beta)f(x_2|x_1;\xi)}{\int\Pr(y|x_1,z;\beta)f(z|x_1;\xi)dz}\,dx_2. \tag{6}$$

The authors propose to substitute $f(\cdot|x_1;\xi)$ in (6) by a nonparametric density estimator. Since this estimator has to be based on the complete cases it is only consistent if the missing mechanism is MAR and, additionally, does not depend on the response variable. It follows that the resulting estimator of β, too, is only consistent under this more restrictive condition. A detailed discussion can be found in ROBINS et al. (1995).

Note that the contribution of an incomplete observation given by (4) or (6) is identical to the expectation of the loglikelihood for a complete observation with respect to the conditional distribution of X_2 given Y and X_1 (cf. equation (2)). The idea of REILLY and PEPE (1995) and PEPE and FLEMING (1991) can therefore be viewed as approximation of the ML–estimation by estimating the E–step and performing only one iteration of the EM algorithm.

3.3.3 Semiparametric efficient estimation

ROBINS et al. (1994) propose a class of semiparametric estimators which depend on two functions: With K denoting the dimension of the regression parameter the first one, $h : \mathbb{R}^2 \to \mathbb{R}^K$, is a function of the covariates and the second one, $\varphi : \mathbb{R}^2 \to \mathbb{R}^K$, is a function of the completely observed variables. The associated estimator $\hat{\beta}(h,\varphi)$ is given as solution of the following equation system

$$\sum_{i=1}^{N}\left(\frac{m^i h(x_1^i,x_2^i)\varepsilon^i(\beta)}{q_{y^i x_1^i}} - \frac{(m^i - q_{y^i x_1^i})\varphi(y^i,x_1^i)}{q_{y^i x_1^i}}\right) = 0, \tag{7}$$

where $\varepsilon^i(\beta) = y^i - \mathrm{E}(Y|x_1^i,x_2^i;\beta)$. The above equation can be regarded as a corrected and weighted estimation equation where the first term depends on the complete and the second on the incomplete cases. Under regularity conditions and under MAR $\hat{\beta}(h,\varphi)$ is consistent and asymptotically normal. If the unknown missing mechanism $q_{y^i x_1^i}$ in (7) is replaced by $\hat{q}_{y^i x_1^i}$ the resulting estimator of β will be denoted by $\widehat{\hat{\beta}}(h,\varphi)$. Further, if $\varphi = 0$ the incomplete observations do not contribute to equation (7). Thus, $\widehat{\hat{\beta}}(h,0)$ can be regarded as a pseudo complete case estimator since it utilises the incomplete observations only to estimate the response rates $\hat{q}_{yx_1}$.

The main interest of ROBINS et al. (1994) concerns the derivation of an estimator which is semiparametric efficient. They show that the proposed class contains an estimator $\hat{\beta}(h_{eff}, \varphi_{eff})$ that attains the lower variance bound with the functions h_{eff} and φ_{eff} given as follows. The first is the solution of the functional equation

$$h(x_1, x_2) = t(x_1, x_2)\Big[\frac{\partial}{\partial\beta}\mu^0_{x_1,x_2} + E\left\{(q^{-1}_{YX_1} - 1)E\left(h(X_1, X_2)\varepsilon(\beta^0)|Y, X_1 = x_1\right)\varepsilon(\beta^0)|x_1, x_2\right\}\Big] \tag{8}$$

with $\mu^0_{x_1,x_2} = \mathrm{E}(Y|x_1, x_2; \beta^0)$, $t(x_1, x_2) = \{\mathrm{E}(\varepsilon(\beta^0)^2/q_{YX_1}|X_1 = x_1, X_2 = x_2)\}^{-1}$ and β^0 as true value of the regression parameter. The function φ^h that minimizes the asymptotic variance of $\hat{\beta}(h, \varphi^h)$ for a general function h is given as conditional expectation

$$\varphi^h(y, x_1) = \mathrm{E}(h(X_1, X_2)\varepsilon(\beta)|Y = y, X_1 = x_1). \tag{9}$$

It follows that $\varphi_{eff} = \varphi^{h_{eff}}$. The authors further show that $\hat{\beta}(h, \varphi^h)$ is asymptotically equivalent to the pseudo complete case estimator $\widehat{\hat{\beta}}(h, 0)$ for any choice of the function h.

In order to get closed expressions for φ_{eff} and h_{eff} we can make use of the fact that Y is discrete. Let $\mathcal{Y}$ denote the finite set of possible realisations of Y. Then taking expectation of (9) with respect to the conditional distribution of X_2 given Y and X_1 leads to a closed expression for each $\varphi_{eff}(y_0, \cdot)$, $y_0 \in \mathcal{Y}$. These are imputed in (8) to get h_{eff}. Both functions obviously still depend on the unspecified distribution of X_2 given Y and X_1, on the true value β^0 of the regression parameter, and on the missing mechanism q_{yx_1}. Estimators $\hat{h}_{eff}$ and $\hat{\varphi}_{eff}$ with the property that $\widehat{\hat{\beta}}(\hat{h}_{eff}, \hat{\varphi}_{eff})$ is asymptotically equivalent to $\hat{\beta}(h_{eff}, \varphi_{eff})$ can for example be obtained in the following way. The conditional distribution of X_2 given Y and X_1 is estimated by the corresponding empirical one and the unknown β^0 can be replaced by any consistent estimator even an inefficient one. In our simulation study we choose the mean score estimator since it is easy to compute. Finally, $q_{y^i x_1^i}$ is replaced by $\hat{q}_{y^i x_1^i}$. ROBINS et al. (1994) show the desired asymptotic equivalence of the resulting estimator $\hat{\beta}^{eff}$

to the semiparametric efficient estimator.

Note that the semiparametric estimators proposed in the previous sections are elements of the class just defined. If we choose $h = h^F_{eff}$ as the optimal function for complete data and if $q_{yx_1} \equiv q$ where q is a constant then $\hat{\beta}^{CC} = \hat{\beta}(h^F_{eff}, 0)$. But in contrast to the complete case estimator, $\widehat{\hat{\beta}}(h^F_{eff}, 0)$ is consistent for general MAR mechanisms and identical to the mean score estimator. Furthermore, one can find a function h^{BC} such that $\widehat{\hat{\beta}}(h^{BC}, 0)$ is asymptotically equivalent to the estimator proposed by BRESLOW and CAIN (1988), which in our case is the corrected complete case estimator. But neither $\widehat{\hat{\beta}}(h^F_{eff}, 0)$ nor $\widehat{\hat{\beta}}(h^{BC}, 0)$ are in general semiparametric efficient.

4 Simulation Designs

The simulation study presented here compares the proposed estimators for small sample size. A similar study has been carried out by ROBINS et al. (1994) with a large sample size (N=2000) and without including the ML–estimator. Other studies (ZHAO and LIPSITZ, 1992; VACH, 1994) consider only discrete covariates where the problem of misspecification, which is of special interest here, does not occur.

The different simulation designs are given by varying the type of missing mechanism, the type of the conditional distribution of X_2 given X_1, the dependence between these covariates, and the regression parameter. The chosen missing mechanisms can be read off Table 1 and are all MAR–mechanisms. The first mechanism means missing completely at random (MCAR) since the missingness is independent of Y and X_1. The second depends only on X_1 (MDX) and the third only on Y (MDY). Consequently, MDXY means that the mechanism depends on both, Y and X_1. Note that the MCAR mechanism leads to a greater over all missing rate than the other mechanisms, which has to be taken into account when interpreting the results.

The conditional distribution of X_2 given X_1 is either Gaussian, or $t(6)$ representing

a symmetric but heavy–tailed distribution, or $\chi^2(2)$ representing a non–symmetric distribution. All distributions are rescaled so as to have variance equal to one. These choices of covariate distributions are rather meant as archetypes than as being realistic. With respect to the dependence between the covariates we consider two choices for $\mu_x = \mathrm{E}(X_2|X_1 = x)$ implemented by shifting the above distributions. In the case $\mu_0 = \mu_1 = 0$ the covariates are independent, in the case $\mu_0 = -1, \mu_1 = 1$ they are dependent.

Table 1: The missing mechanisms and the corresponding probabilities q_{yx_1}.

	q_{00}	q_{10}	q_{01}	q_{11}
MCAR	0.3	0.3	0.3	0.3
MDX	0.8	0.8	0.3	0.3
MDY	0.8	0.3	0.8	0.3
MDXY	0.8	0.3	0.3	0.8

To keep the number of parameter constellations limited we let β_2 take the values $\{-1.5, 0, 1.5\}$ whereas β_0 and β_1 are kept fixed as $\beta_0 = 0$ and $\beta_1 = 1$. The covariate X_1 follows a Bernoulli distribution with $\Pr(X_1 = 1) = 0.5$. The sample size is chosen to be $N = 200$ (before generating the missings). For each of the resulting 72 designs 1000 samples are generated using Turbo Pascal 7.0.

5 Results

In order to compare the estimators we compute the estimated relative mean squared errors which are the ratios of the Monte Carlo mean squared error of the semiparametric efficient estimator and that of the respective other estimator. This will simply be called relative MSE, i.e. all relative MSEs are relative to the semiparametric efficient estimator. A relative MSE larger than one thus means that the considered estimator is better, w.r.t. the MSE, than the semiparametric efficient estimator and

we would expect such findings only for the ML–estimator when the covariate distribution is correctly specified. Note that the relative MSE is no absolute measure: In some situations both estimators can be bad. Since the semiparametric efficient estimator is used as a reference we report its observed bias first (Section 5.1.1). ROBINS et al. (1994) consider the estimated relative efficiencies, i.e. the ratio of the Monte Carlo variances, instead of the relative MSE. This is not sensible, here, as the sample size is considerably smaller and hence bias is not negligible. For the same reason we additionally compute the means of the observed biases which we will simply call bias. A negative bias indicates that the estimator tends to underestimate the true value.

5.1 Comparison of the semiparametric efficient estimator and the ML–estimator

We first discuss the bias of the semiparametric efficient and the ML–estimator and then the relative MSE of the latter one. The former is asymptotically unbiased but there can be considerable deviations for finite samples. The latter is not expected to be unbiased when the distributional assumptions are false. We therefore distinguish the cases where the specification of the covariate distribution is correct or false.

5.1.1 Bias of the semiparametric efficient estimator

The bias of the semiparametric efficient estimator can be read off Table 2. In case that X_2 has no influence, i.e. $\beta_2 = 0$, the bias of all three components $\hat{\beta}_0^{eff}$, $\hat{\beta}_1^{eff}$ and $\hat{\beta}_2^{eff}$ is in general negligible.

The case $\beta_2 \neq 0$ is more serious especially concerning the estimation of β_2 when the covariate distribution is χ^2. Here, the bias of $\hat{\beta}_2$ is for $\mu_0 = \mu_1$ often, and for $\mu_0 \neq \mu_1$ and any covariate distribution nearly always in absolute value larger than 0.1, and even larger for the χ^2 distribution. The bias of $\hat{\beta}_0$ and $\hat{\beta}_1$ is for $\beta_2 \neq 0$ and $\mu_0 = \mu_1$ still relatively small but serious deviations occur for $\hat{\beta}_1$ if $\mu_0 \neq \mu_1$. In general, the

bias given a t–distribution is only slightly larger than for a Gaussian covariate.

Table 2: Bias of the semiparametric efficient estimator.

covariable distrib. = Gaussian		$\mu_0 = \mu_1 = 0$			$\mu_0 = -1, \mu_1 = 1$		
		$\hat{\beta}_0^{eff}$	$\hat{\beta}_1^{eff}$	$\hat{\beta}_2^{eff}$	$\hat{\beta}_0^{eff}$	$\hat{\beta}_1^{eff}$	$\hat{\beta}_2^{eff}$
MCAR, $\beta_2 =$	-1.5	*-0.02*	*0.08*	*-0.13*	*-0.07*	*0.21*	*-0.18*
	0	*-0.01*	*0.03*	*-0.01*	*-0.02*	*0.07*	*-0.02*
	1.5	*-0.00*	*0.08*	*0.19*	*0.07*	*-0.07*	*0.24*
MDX, $\beta_2 =$	-1.5	*-0.00*	*0.05*	*-0.08*	*-0.02*	*0.11*	*-0.10*
	0	*-0.01*	*0.02*	*-0.01*	*-0.01*	*0.03*	*-0.00*
	1.5	*-0.02*	*0.07*	*0.09*	*0.01*	*-0.00*	*0.10*
MDY, $\beta_2 =$	-1.5	*-0.02*	*0.05*	*-0.09*	*-0.05*	*0.11*	*-0.09*
	0	*-0.01*	*0.02*	*0.00*	*-0.02*	*0.06*	*-0.01*
	1.5	*-0.03*	*0.04*	*0.08*	*0.01*	*-0.02*	*0.13*
MDXY, $\beta_2 =$	-1.5	*-0.03*	*0.08*	*-0.09*	*-0.05*	*0.13*	*-0.08*
	0	*-0.02*	*0.04*	*-0.01*	*-0.01*	*0.03*	*0.01*
	1.5	*-0.01*	*0.07*	*0.08*	*0.02*	*0.03*	*0.14*

covariable distrib. = χ^2		$\mu_0 = \mu_1 = 0$			$\mu_0 = -1, \mu_1 = 1$		
		$\hat{\beta}_0^{eff}$	$\hat{\beta}_1^{eff}$	$\hat{\beta}_2^{eff}$	$\hat{\beta}_0^{eff}$	$\hat{\beta}_1^{eff}$	$\hat{\beta}_2^{eff}$
MCAR, $\beta_2 =$	-1.5	*-0.05*	*0.06*	*-0.24*	*-0.16*	*0.32*	*-0.25*
	0	*-0.00*	*0.04*	*0.02*	*0.03*	*-0.00*	*0.02*
	1.5	*0.06*	*0.04*	*0.23*	*0.20*	*-0.33*	*0.31*
MDX, $\beta_2 =$	-1.5	*-0.04*	*0.08*	*-0.14*	*-0.06*	*0.12*	*-0.09*
	0	*0.00*	*0.02*	*0.00*	*0.01*	*-0.00*	*-0.00*
	1.5	*0.04*	*0.02*	*0.14*	*0.06*	*-0.08*	*0.10*
MDY, $\beta_2 =$	-1.5	*-0.03*	*0.04*	*-0.08*	*-0.12*	*0.20*	*-0.14*
	0	*-0.01*	*0.03*	*0.00*	*-0.01*	*0.01*	*-0.01*
	1.5	*0.02*	*0.04*	*0.11*	*-0.00*	*-0.04*	*0.14*
MDXY, $\beta_2 =$	-1.5	*-0.03*	*0.08*	*-0.12*	*-0.15*	*0.25*	*-0.15*
	0	*-0.01*	*0.04*	*-0.01*	*-0.01*	*0.04*	*-0.00*
	1.5	*0.01*	*0.05*	*0.15*	*0.00*	*-0.01*	*0.12*

covariable distrib. = student		$\mu_0 = \mu_1 = 0$			$\mu_0 = -1, \mu_1 = 1$		
		$\hat{\beta}_0^{eff}$	$\hat{\beta}_1^{eff}$	$\hat{\beta}_2^{eff}$	$\hat{\beta}_0^{eff}$	$\hat{\beta}_1^{eff}$	$\hat{\beta}_2^{eff}$
MCAR, $\beta_2 =$	-1.5	*-0.02*	*0.09*	*-0.22*	*-0.08*	*0.24*	*-0.18*
	0	*-0.01*	*0.03*	*-0.00*	*0.00*	*0.04*	*-0.00*
	1.5	*-0.01*	*0.08*	*0.17*	*0.11*	*-0.11*	*0.28*
MDX, $\beta_2 =$	-1.5	*0.01*	*0.01*	*-0.08*	*-0.05*	*0.15*	*-0.11*
	0	*-0.01*	*0.04*	*-0.01*	*0.01*	*0.00*	*0.01*
	1.5	*-0.00*	*0.04*	*0.08*	*0.02*	*-0.00*	*0.08*
MDY, $\beta_2 =$	-1.5	*-0.02*	*0.04*	*-0.08*	*-0.07*	*0.13*	*-0.12*
	0	*0.00*	*0.01*	*-0.01*	*-0.01*	*0.03*	*-0.01*
	1.5	*-0.01*	*0.01*	*0.08*	*0.01*	*-0.02*	*0.11*
MDXY, $\beta_2 =$	-1.5	*-0.03*	*0.09*	*-0.08*	*-0.06*	*0.17*	*-0.11*
	0	*-0.02*	*0.05*	*-0.01*	*-0.03*	*0.10*	*-0.01*
	1.5	*-0.03*	*0.10*	*0.11*	*0.02*	*0.00*	*0.15*

An additional aspect concerns the direction of the bias. The estimation of β_0 has nearly always a negative bias. In contrast to this, the bias of $\hat{\beta}_1^{eff}$ is in general positive. The direction of the bias of $\hat{\beta}_2^{eff}$ depends on the true value: it is negative for $\beta_2 = -1.5$ and positive for $\beta_2 = 1.5$.

5.1.2 Bias of the ML–estimator

As can be seen from Table 3, the bias of the ML–estimator is very similar to the one of the semiparametric efficient estimator for the Gaussian covariate distribution.

Table 3: Relative MSE and bias of the ML–estimator with incomplete data assuming a Gaussian covariate distribution (bias in brackets).

covariable distrib. = Gaussian		$\mu_0 = \mu_1 = 0$			$\mu_0 = -1, \mu_1 = 1$		
		$\hat{\beta}_0^{ML}$	$\hat{\beta}_1^{ML}$	$\hat{\beta}_2^{ML}$	$\hat{\beta}_0^{ML}$	$\hat{\beta}_1^{ML}$	$\hat{\beta}_2^{ML}$
MCAR, $\beta_2 =$	-1.5	1.05 *(-0.02)*	1.05 *(0.07)*	1.04 *(-0.13)*	1.05 *(0.19)*	1.04 *(-0.06)*	1.03 *(-0.17)*
	0	1.00 *(-0.01)*	1.00 *(0.03)*	1.00 *(-0.01)*	1.01 *(-0.02)*	1.01 *(0.07)*	1.00 *(-0.02)*
	1.5	1.04 *(-0.00)*	1.05 *(0.08)*	1.02 *(0.19)*	1.06 *(0.06)*	1.05 *(-0.07)*	1.07 *(0.22)*
MDX, $\beta_2 =$	-1.5	1.01 *(-0.00)*	1.04 *(0.05)*	1.01 *(-0.08)*	1.00 *(-0.02)*	1.01 *(0.10)*	1.01 *(-0.10)*
	0	1.00 *(-0.01)*	1.00 *(0.02)*	1.00 *(-0.01)*	1.00 *(-0.01)*	1.00 *(0.03)*	1.00 *(-0.00)*
	1.5	1.00 *(-0.02)*	1.03 *(0.07)*	1.01 *(0.09)*	1.01 *(0.01)*	1.03 *(0.01)*	1.04 *(0.10)*
MDY, $\beta_2 =$	-1.5	1.07 *(-0.02)*	1.07 *(0.04)*	1.10 *(-0.09)*	1.07 *(-0.04)*	1.04 *(0.10)*	1.05 *(-0.09)*
	0	1.00 *(-0.01)*	0.99 *(0.03)*	0.99 *(0.00)*	1.01 *(-0.02)*	1.01 *(0.06)*	1.00 *(-0.01)*
	1.5	1.05 *(-0.02)*	1.06 *(0.04)*	1.04 *(0.08)*	1.03 *(0.02)*	1.04 *(-0.00)*	1.10 *(0.12)*
MDXY, $\beta_2 =$	-1.5	1.06 *(-0.02)*	1.07 *(0.06)*	1.07 *(-0.08)*	1.06 *(-0.03)*	1.07 *(0.10)*	1.03 *(-0.07)*
	0	1.00 *(-0.02)*	1.00 *(0.04)*	0.99 *(-0.01)*	1.02 *(0.00)*	1.02 *(0.02)*	1.01 *(0.01)*
	1.5	1.04 *(-0.01)*	1.04 *(0.06)*	1.04 *(0.08)*	1.06 *(0.03)*	1.07 *(0.01)*	1.16 *(0.12)*

Table 4 shows the bias of the ML–estimator in the situations where the distributional assumptions are wrong, i.e. for the χ^2 and Student covariate distribution. Here, we observe only a small bias whenever $\beta_2 = 0$. As should be expected, the wrong assumption about the covariate distribution does not appear to affect the consistency of the ML–estimator when this covariate has no influence.

If $\beta_2 \neq 0$ and X_2 follows the χ^2 distribution the bias is clearly affected. Especially the estimation of β_2 in the presence of a missing mechanism that depends on the response variable (MDY and MDXY) appears to be distinctly biased. The bias when estimating β_0 and β_1 is also quite large in these situations, especially when $\mu_0 \neq \mu_1$. For the MCAR and MDX mechanisms we observe no such severe bias although it is

sometimes over 0.1 for the estimation of β_1 and β_2. The largest observed absolute bias among the designs with χ^2 distribution is 0.85

If the covariate distribution is Student we can observe the same bias pattern as for the Gaussian covariate distribution with a slight general tendency to extreme values and a clear tendency to extreme values for the estimation of β_1 and β_2 in the special case of $\mu_0 = \mu_1$, $\beta_2 \neq 0$, and a MDXY missing mechanism. With this last exception, the results are also similar to those of the semiparametric efficient estimator for the Student distribution. The largest observed absolute bias among the designs with Student covariate distribution is 0.27.

Table 4: Relative MSE and bias of the ML–estimator with incomplete data falsely assuming a Gaussian covariate distribution.

covariable distrib. $= \chi^2$		$\mu_0 = \mu_1 = 0$			$\mu_0 = -1, \mu_1 = 1$		
		$\hat{\beta}_0^{ML}$	$\hat{\beta}_1^{ML}$	$\hat{\beta}_2^{ML}$	$\hat{\beta}_0^{ML}$	$\hat{\beta}_1^{ML}$	$\hat{\beta}_2^{ML}$
MCAR, $\beta_2 =$	-1.5	1.02 *(0.09)*	1.05 *(0.07)*	1.36 *(-0.12)*	1.53 *(0.08)*	1.36 *(0.08)*	1.23 *(-0.15)*
	0	0.98 *(-0.00)*	0.98 *(0.05)*	0.98 *(0.03)*	1.03 *(0.03)*	1.02 *(-0.01)*	0.99 *(0.02)*
	1.5	1.09 *(-0.07)*	0.72 *(0.12)*	1.20 *(0.17)*	1.60 *(-0.05)*	1.53 *(0.15)*	1.62 *(0.15)*
MDX, $\beta_2 =$	-1.5	1.05 *(0.00)*	0.87 *(0.19)*	1.17 *(-0.09)*	1.11 *(0.01)*	1.02 *(0.09)*	1.09 *(-0.06)*
	0	1.00 *(-0.00)*	1.00 *(0.02)*	1.00 *(0.00)*	1.01 *(0.01)*	1.01 *(-0.00)*	1.01 *(-0.00)*
	1.5	1.03 *(0.00)*	0.82 *(0.01)*	1.08 *(0.12)*	1.17 *(-0.02)*	1.05 *(0.19)*	1.16 *(0.05)*
MDY, $\beta_2 =$	-1.5	1.05 *(0.06)*	0.98 *(0.04)*	0.79 *(0.44)*	1.11 *(0.46)*	1.02 *(-0.78)*	1.33 *(0.30)*
	0	0.99 *(-0.01)*	0.99 *(0.04)*	0.95 *(0.03)*	0.98 *(0.02)*	0.94 *(-0.04)*	0.93 *(0.02)*
	1.5	0.79 *(-0.03)*	0.86 *(0.08)*	0.25 *(0.77)*	0.89 *(0.20)*	0.60 *(-0.38)*	1.03 *(0.23)*
MDXY, $\beta_2 =$	-1.5	0.59 *(-0.08)*	0.85 *(0.19)*	0.98 *(-0.15)*	1.28 *(0.13)*	1.17 *(0.09)*	1.14 *(-0.08)*
	0	0.95 *(-0.02)*	0.89 *(0.06)*	0.66 *(0.03)*	0.80 *(0.03)*	0.85 *(-0.02)*	0.66 *(0.04)*
	1.5	0.72 *(-0.13)*	0.18 *(0.69)*	0.29 *(0.74)*	0.48 *(0.51)*	0.64 *(-0.57)*	0.33 *(0.85)*

covariable distrib. = student		$\mu_0 = \mu_1 = 0$			$\mu_0 = -1, \mu_1 = 1$		
		$\hat{\beta}_0^{ML}$	$\hat{\beta}_1^{ML}$	$\hat{\beta}_2^{ML}$	$\hat{\beta}_0^{ML}$	$\hat{\beta}_1^{ML}$	$\hat{\beta}_2^{ML}$
MCAR, $\beta_2 =$	-1.5	0.97 *(-0.02)*	0.97 *(0.11)*	1.03 *(-0.21)*	1.03 *(-0.05)*	1.00 *(0.18)*	0.95 *(-0.17)*
	0	0.99 *(-0.01)*	1.00 *(0.03)*	1.00 *(-0.00)*	1.00 *(0.00)*	1.01 *(0.04)*	1.00 *(-0.00)*
	1.5	0.97 *(-0.01)*	0.98 *(0.09)*	1.03 *(0.16)*	1.12 *(0.06)*	1.11 *(-0.03)*	1.07 *(0.26)*
MDX, $\beta_2 =$	-1.5	0.99 *(0.01)*	0.98 *(0.03)*	1.01 *(-0.08)*	1.02 *(-0.04)*	1.02 *(0.12)*	1.01 *(-0.10)*
	0	1.00 *(-0.01)*	1.00 *(0.04)*	1.00 *(-0.01)*	1.00 *(0.01)*	1.00 *(0.01)*	1.00 *(0.01)*
	1.5	0.99 *(-0.00)*	0.96 *(0.06)*	1.00 *(0.08)*	1.01 *(0.01)*	1.06 *(0.05)*	1.01 *(0.08)*
MDY, $\beta_2 =$	-1.5	0.94 *(-0.04)*	0.98 *(0.05)*	1.02 *(-0.07)*	0.98 *(-0.05)*	1.05 *(0.08)*	1.05 *(-0.11)*
	0	1.00 *(-0.00)*	0.99 *(0.01)*	0.99 *(-0.01)*	0.99 *(-0.01)*	1.00 *(0.03)*	0.99 *(-0.01)*
	1.5	0.97 *(-0.04)*	1.01 *(0.02)*	1.04 *(0.07)*	1.07 *(-0.01)*	1.06 *(0.05)*	1.10 *(0.11)*
MDXY, $\beta_2 =$	-1.5	0.89 *(-0.07)*	0.78 *(0.19)*	0.84 *(-0.15)*	1.03 *(-0.01)*	1.07 *(0.07)*	1.15 *(-0.05)*
	0	0.99 *(-0.02)*	0.98 *(0.05)*	0.94 *(-0.01)*	0.98 *(-0.02)*	0.98 *(0.09)*	0.96 *(-0.01)*
	1.5	0.86 *(-0.07)*	0.76 *(0.21)*	0.85 *(0.18)*	1.00 *(0.05)*	1.03 *(-0.06)*	0.80 *(0.27)*

Note that the estimation of β_2 is always biased away from zero, i.e. the absolute effect is overestimated. The same phenomenon can be observed for all other methods. This seems to illustrate that consistency is merely an asymptotic property.

5.1.3 Relative MSE of the ML–estimator with correct assumptions

The results presented in Table 3 further allow a direct comparison of the semiparametric efficient and the parametric efficient estimation. Since the bias is similar for both estimators the relative MSE essentially reflects the gain in efficiency due to the additional parametric assumption.

At first, one can say that the results of both estimators are nearly equal for $\beta_2 = 0$, i.e. when the incompletely observed covariate has no effect on the response. Furthermore, the gain in efficiency is generally only modest for the MCAR and MDX missing mechanisms. If $\beta_2 \neq 0$ and the missing mechanism is MDY or MDXY we can observe that in more than half of the designs the relative MSE is greater than 1.05 reaching the maxima of 1.10 and 1.16, respectively, for $\beta_2 = 1.5$. The missing mechanisms depending on the response variable may therefore be those where the ML–estimator truly outperforms the semiparametric efficient one.

5.1.4 Relative MSE of the ML–estimator with wrong assumptions

Despite the wrong distributional assumption there are some situations where the ML–estimator performs almost as good as the semiparametric efficient one with respect to the relative MSE. This is the case when $\beta_2 = 0$ for both covariate distributions and all missing mechanism except MDXY while at the same time the bias is always very small as we have seen above. Thus, in these situations the ML–estimator seems neither inconsistent nor inefficient. For the MDXY designs and the χ^2 distribution, however, a serious loss in efficiency of the ML–estimator can be observed while the bias is still very small.

If $\beta_2 \neq 0$ and X_2 is distributed according to the χ^2 distribution the ML–estimator performs fairly well for the MCAR and MDX missing mechanisms. Taking the bias

into account, it follows that the good results are mainly due to a small variance of the ML–estimator. But if the missing mechanism additionally depends on the response variable the results indicate a serious deficiency of the ML approach. The smallest observed relative MSE amounts to 0.18 and occurs in the MDXY designs. It is not surprising that, in contrast, the ML–estimator performs nearly as well for the Student as for the Gaussian covariate distribution. The designs with a small loss in efficiency are given when the covariates are independent and the missing mechanism is not MCAR. It reaches a minimal MSE of 0.76 for the MDXY mechanism. If, in contrast, $\mu_0 \neq \mu_1$ the relative MSE is not worse for the Student than for the Gaussian covariate distribution. This may suggest that the ML–estimator is still appropriate for dependent covariates because it makes a correct assumption about the dependence structure although the distributional assumption is wrong.

5.2 Performance of the semiparametric estimators

In this section, we discuss the performance of the complete case, the corrected complete case, and the mean score estimators compared with the semiparametric efficient one. The results of the simulation study are not given in details.

5.2.1 The complete case estimator

For the designs where the complete case estimator is inconsistent we get that the bias of $\hat{\beta}_0^{CC}$ is always less than -1 for both missing mechanisms that depend on the response variable whereas $\hat{\beta}_1^{CC}$ is distinctly biased only for the MDXY mechanism showing a bias of typically more than 2. But even when $\hat{\beta}^{CC}$ is consistent the relative MSE is severely affected by discarding the incomplete cases, it often takes values between 0.55 and 0.8.

5.2.2 The corrected complete case estimator

Surprisingly, the corrected complete case estimator produces results nearly identical to the semiparametric efficient one. The relative MSEs are almost always between 0.99 and 1.00, exceptions arising only for the non–Gaussian covariate distributions when the missing mechanism depends on the response variable and $\beta_2 \neq 0$. But even then the relative MSE is at least 0.98. Concerning the bias we can observe the same pattern as for the semiparametric efficient estimator with a slight tendency to a greater bias of $\hat{\beta}_2^{CCC}$ for the missing mechanisms that depend on the response variable.

5.2.3 The mean score estimator

The mean score estimator is clearly dominated by the semiparametric efficient estimator. The relative MSE is almost always definitely smaller than 1.00. The worst result is a relative MSE of 0.65 but in most cases it is still at least 0.8 and even greater than 0.9 for the MCAR missing mechanism. The main difficulty seems to concern the estimation in the MDX situation especially for $\mu_0 \neq \mu_1$. Here, the relative MSEs are roughly about 0.8.

Although the results are similar for the different covariate distributions, it can be observed that in case of a non-MCAR mechanism, $\beta_2 = 0$, and $\mu_0 \neq \mu_1$ the relative MSE of all three components is always greater for the Gaussian covariate distribution than for the others. Note that the performance of the mean score estimator is essentially the same in the case of a discrete covariate X_2 (cf. VACH, 1994, p. 34).

6 Discussion

The main result of the simulation study concerns the performance of the ML–estimator compared to the semiparametric efficient one proposed by ROBINS et al. (1994). On the one hand, we have seen that in the situation of a correct as-

sumption about the covariate distribution and rather small sample size the gain in efficiency by ML–estimation is only modest. On the other hand, this parametric approach can lead to serious bias if the assumed covariate distribution is 'far away' from the true one, where 'far away' means χ^2 instead of Gaussian. The Student distribution is in contrast similar enough to the Gaussian for the bias of the ML–estimator to be negligible, at least for a sample size of 200. However, simulations with a sample size of 1000, which are not reported here, show a more serious bias of the ML–estimator given a Student covariate distribution. In contrast, the performance of the semiparametric efficient estimator appears to be satisfying also for finite sample size notwithstanding that efficiency is an asymptotic property. Strictly speaking, these results of course only apply for the specific situations considered in the simulation study. As conclusion we propose that if one doubts the appropriateness of the Gaussian distribution in a specific application one may consider semiparametric efficient estimation as a reasonable alternative, in particular if the missing mechanism is far from being completely at random.

Another interesting result has been obtained for the corrected complete case estimator. It strengthens the conjecture that in the special case of a logistic regression where all variables except the incomplete one are discrete the estimator proposed by BRESLOW and CAIN (1988) is nearly semiparametric efficient. An analytic proof of this property is not known to the author. Moreover, we have to restrict this conjecture to the logistic regression model since ROBINS et al. (1994) show that $\hat{\beta}^{CCC}$ is not semiparametric efficient in general regression models.
The remaining estimators, complete case and mean score, meet the expectation of being biased or inefficient so that they are not recommendable despite their simple computation.

The possibility of misspecifying the missing mechanism has not been addressed so far but has to be taken into account. All the discussed semiparametric approaches

require an estimation of the observation probabilities given by $\hat{q}_{yx_1}$. For continuous Y or X_1, however, there is no such straightforward estimation procedure. Instead, a model for the missing mechanism has to be assumed. As shown by ZHAO et al. (1996) the correctness of this model is crucial in assuring the consistency of the semiparametric estimators.

References

BLACKHURST, D.W. and M.D. SCHLUCHTER (1989), Logistic regression with a partially observed covariate, *Communications in Statistics – Simulation and Computation* **18**, 163-177.

BRESLOW, N.E. and K.C. CAIN (1988), Logistic regression for two–stage case–control data, *Biometrika* **75**, 11-20.

CAIN, K.C. and N.E. BRESLOW (1988), Logistic regression analysis and efficient design for two–stage studies, *American Journal of Epidemiology* **128**, 1198-1206.

CARROLL, R.J. and M.P. WAND (1991), Semiparametric estimation in logistic measurement error models, *Journal of the Royal Statistical Society B* **53**, 573-585.

DEMPSTER, A.P., N.M. LAIRD and D.B. RUBIN (1977), Maximum likelihood estimation from incomplete data via the EM–algorithm, *Journal of the Royal Statistical Society B* **39**, 1-38.

IBRAHIM, J.G. (1990), Incomplete data in generalised linear models, *Journal of the American Statistical Association* **85** 765-769.

IBRAHIM, J.G. and S. WEISBERG (1992), Incomplete data in generalised linear models with continuous covariates, *Australian Journal of Statistics* **34**, 461-470.

LITTLE, J.A. (1992), Regression with missing X's: A review, *Journal of the American Statistical Association* **87**, 1227-1237.

PEPE, M.S. and T.R. FLEMING (1991), A nonparametric method for dealing with mismeasured covariate data, *Journal of the American Statistical Association* **86**, 108-113.

REILLY, M. and M. PEPE (1995), A mean score method for missing and auxiliary covariate data in regression models, *Biometrika* **82**, 299-314.

ROBINS, J.M., A. ROTNITZKY and L.P. ZHAO (1994), Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association* **89**, 846-866.

ROBINS, J.M., F. HSIEH and W. NEWEY (1995), Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates, *Journal of the Royal Statistical Society B* **57**, 409-424.

ROBINS, J.M., N. WANG (2000), Inference for imputation estimators, *Biometrika* **87**, 113-124.

RUBIN, D.B. (1976), Inference and missing data, *Biometrika* **63**, 581-592.

RUBIN, D.B. (1987), *Multiple imputation for nonresponse in surveys*. Wiley, New York

SCHAFER, J.L. (1997), *Analysis of incomplete multivariate data*. Chapman and Hall, London.

VACH, W. (1994), *Logistic regression with missing values in the covariates*. Springer, New York.

VACH, W. and M. BLETTNER (1991), Biased estimation of the odds ratio in case–control studies due to the use of ad–hoc methods of correcting for missing

values for confounding variables. *American Journal of Epidemiology* **134**, 895-907.

VACH, W. and S. ILLI (1997), Biased estimation of adjusted odds ratios from incomplete covariate data due to the violation of the missing at random assumption, *Biometrical Journal* **39**, 13 –28.

ZHAO, L.P. and S. LIPSITZ (1992), Design and analysis of two–stage designs, *Statistics in Medicine* **11**, 769-782.

ZHAO, L.P., S. LIPSITZ and D. LEW (1996), Regression analysis with missing covariate data using estimating equations, *Biometrics* **52**, 1165-1182.