Biased Estimation of Adjusted Odds Ratios From Incomplete Covariate Data Due to Violation of the Missing at Random Assumption

WERNER VACH[1] and SABINA ILLI[2]

[1] Center for Data Analysis and Model Building, University of Freiburg, Germany and
Institute of Medical Biometry and Informatics, University of Freiburg, Germany
[2] Institute of Statistics, Ludwig-Maximilians-University Munich, Germany

Summary

We investigate the possible bias due to an erroneous missing at random assumption if adjusted odds ratios are estimated from incomplete covariate data using the maximum likelihood principle. A relation between complete case estimates and maximum likelihood estimates allows us to identify situations where the bias vanishes. Numerical computations demonstrate that the bias is most serious if the degree of the violation of the missing at random assumption depends on the value of the outcome variable or of the observed covariate. Implications for the analysis of prospective and retrospective studies are given.

Key words: Adjusted odds ratio; Biased estimation; Case-control study; Complete case analysis; Maximum likelihood estimation; Missing at random; Missing value; Logistic model.

1. Introduction

The analysis of incomplete data is a challenge of the daily work of applied statisticians. The restriction to units with complete data is the standard approach of most statistical software packages. Such a complete case analysis, however, is wasteful of information. Recently the efficient analysis of regression models based on incomplete covariate data gathered a lot of attention, the book of VACH (1994) and the paper of ROBINS ET AL. (1994) present an overview. However, these sophisticated approaches to handle missing values rely on the missing at random (MAR) assumption, which excludes dependence of the observability of a covariate on its unobserved value.

In practical applications this assumption is often highly questionable. This is especially true if covariate data are collected by interviews or questionnaires, such that missing values can be due to an active refusal of subjects. Such a refusal may depend on the true value of a covariate, for example if one asks for alcohol consumption, sexual behaviour or income. Using documents like hospital records as a source for data collection similar problems occur. Strange symptoms or unusual treatments are usually well documented when they are present, but their absence results often only in a gap in the documents. In view of these problems one may argue that it is better to use methods not relying on the MAR assumption, e.g. a complete case analysis. However, in the analysis of case-control studies a complete-case analysis can result in biased estimates, and hence in this setting the use of advanced methods is even necessary to achieve consistent estimates.

In this paper we investigate the bias due to a violation of the MAR assumption in the special case of logistic regression which aims to estimate adjusted odds ratios. We restrict ourselves to the case of two categorical covariates where only the second is affected by missing values. In Section 2 we introduce some basic notations and in Section 3 some aspects of missing value mechanisms are discussed. Section 4 investigates the bias of complete case estimates, and in Section 5 we show, how the maximum likelihood (ML) estimate based on incomplete data is related to the complete case estimates. This allows us to investigate in Section 6 the possible asymptotic bias of the ML estimate, and some numerical results are presented, too. Section 7 investigates some alternative semiparametric procedures and in Section 8 we consider implications for the analysis of prospective and retrospective studies with incomplete covariate data. A general discussion finishes the paper.

2. Notation

Let Y be a binary outcome variable and X_1 and X_2 two categorical covariates with J and K categories, respectively. If all variables are observable the corresponding contingency table has the cell probabilities $p_{ijk} := P(Y = i, X_1 = j, X_2 = k)$. Within the k-th stratum of X_2 the odds ratio between the j-th category of X_1 and the first category is

$$\phi_{jk} := \frac{p_{1jk}p_{01k}}{p_{0jk}p_{11k}}$$

and within the j-th stratum of X_1 the odds ratio between the k-th category of X_2 and the first category is

$$\psi_{jk} := \frac{p_{1jk}p_{0j1}}{p_{0jk}p_{1j1}}$$

where we can replace p_{ijk} also by the conditional probabilities $p_{i|jk} := P(Y = i|X_1 = j, X_2 = k)$. In estimating adjusted odds ratios we assume that the odds ratios are constant over different strata, i.e.

$$\phi_j :\equiv \phi_{jk} \text{ for all } k \quad \text{ and } \quad \psi_k :\equiv \psi_{jk} \text{ for all } j$$

This can be equivalently expressed as

$$\frac{p_{1|jk}}{p_{0|jk}} = \tau\phi_j\psi_k$$

with an additional parameter τ. In logistic regression the logarithms of the odds ratios are considered as parameters; i.e.

$$\log\frac{p_{1|jk}}{p_{0|jk}} = \beta_0 + \beta_{1j} + \beta_{2k}$$

with $\exp(\beta_0) = \tau$, $exp(\beta_{1j}) = \phi_j$ and $\exp(\beta_{2k}) = \psi_k$, such that $p_{i|jk}(\beta) = \Lambda(\beta_0 + \beta_{1j} + \beta_{2k})^i(1 - \Lambda(\beta_0 + \beta_{1j} + \beta_{2k}))^{1-i}$ with $\Lambda(t) := (1 + \exp(-t))^{-1}$.

In this paper we consider the additional difficulty that the second covariate is unobservable for some subjects. The observability of X_2 is indicated by the binary random variable R_2, such that we observe instead of X_2 the random variable Z_2 with

$$Z_2 := \begin{cases} X_2 & \text{if } R_2 = 1 \\ ? & \text{otherwise} \end{cases} .$$

The observation of n independent realizations of (Y, X_1, Z_2) can be summarized in a $2 \times J \times (K+1)$ contingency table n_{ijk}, where the $(K+1)$-th category corresponds to a missing value for X_2. The cell probabilities of this table are determined by the original p_{ijk} and the response probabilities

$$q_{ijk} := P(R_2 = 1|Y = i, X_1 = j, X_2 = k) .$$

A special role will be played by the observable response rates $\widehat{Q}_{ij} := n_{ij+}/n_{ij.}$ with $n_{ij+} := n_{ij1} + \ldots + n_{ijK}$ and $n_{ij.} := n_{ij1} + \ldots + n_{ijK+1}$. Here $\widehat{Q}_{ij}$ is an estimate for $Q_{ij} := P(R_2 = 1|Y = i, X_1 = j) = \sum_k q_{ijk}P(X_2 = k|Y = i, X_1 = j)$.

3. Missing Value Mechanisms

The properties of any statistical method to handle such incomplete data depend on the unknown response probabilities q_{ijk}, which cannot be estimated observing only Y, X_1 and Z_2. Hence assumptions on the missing value mechanism are necessary to insure desired statistical properties. Of central importance is the MAR assumption introduced by RUBIN (1976), which excludes a dependence of response probabilities on unobserved values. In our setting, the MAR assumption reads

$$q_{ijk} \equiv q_{ij} \ ,$$

and is equivalent to $P(X_2 = k|Y = i, X_1 = j, R_2 = 1) = P(X_2 = k|Y = i, X_1 = j, R_2 = 0)$ for all i, j, k; i.e., it allows us to assume that unobserved values of X_2 have the same conditional distribution as the observed values. Hence this assumption is the key to an adequate handling of missing values.

In this paper we focus on the possible bias due to an erroneous MAR assumption. The theoretical results will depend on the assumption that the response rates can be decomposed to $q_{ijk} = q_{jk}q_i$, $q_{ijk} = q_{ik}q_j$, or $q_{ijk} = q_{ij}q_k$. In practice it will be difficult to identify situations, where such a decomposition holds without that one of the factors is equal to one. But in many applications assumptions like $q_{ijk} \equiv q_{jk}$ or $q_{ijk} \equiv q_{ik}$ can be justified by the design of a study, and these are special cases of the above decompositions. This will be discussed in more detail in Section 8.

4. Complete case analysis

In a complete case analysis all subjects with incomplete covariate data are omitted in the analysis. Hence estimation is based on the analysis of a $2 \times J \times K$ contingency table with cell probabilities

$$\begin{aligned} p^{CC}_{ijk} &:= P(Y = i, X_1 = j, X_2 = k|R_2 = 1) \\ &= q_{ijk}p_{ijk}/P(R_2 = 1) \end{aligned}$$

and hence with odds ratios

$$\phi^{CC}_{jk} := \phi_j f^{\phi}_{jk} \text{ with } f^{\phi}_{jk} := \frac{q_{1jk}q_{01k}}{q_{0jk}q_{11k}} \quad \text{and}$$

$$\psi^{CC}_{jk} := \psi_k f^{\psi}_{jk} \text{ with } f^{\psi}_{jk} := \frac{q_{1jk}q_{0j1}}{q_{0jk}q_{1j1}} \quad .$$

In general, ϕ^{CC}_{jk} may depend on k and ψ^{CC}_{jk} may depend on j; hence the essential assumption for the estimation of adjusted odds ratios is violated. However, if the response rates can be decomposed into two factors, i.e. if

$q_{ijk} = q_{jk} q_i$, $q_{ijk} = q_{ik} q_j$, or $q_{ijk} = q_{ij} q_k$, then the odds ratios are constant and hence coincide with the stochastic limits of the estimates $\hat{\phi}_j^{CC}$ and $\hat{\psi}_k^{CC}$ from a complete case analysis. The resulting asymptotic bias factors ϕ_j^{CC}/ϕ_j and ψ_k^{CC}/ψ_k are summarized in the following table:

condition	ϕ_j^{CC}/ϕ_j	ψ_k^{CC}/ψ_k
$q_{ijk} \equiv q_{ij} q_k$	$\frac{q_{1j} q_{01}}{q_{0j} q_{11}}$	1
$q_{ijk} \equiv q_{ik} q_j$	1	$\frac{q_{1k} q_{01}}{q_{0k} q_{11}}$
$q_{ijk} \equiv q_{jk} q_i$	1	1

The third condition of this table was previously identified by Glynn and Laird (1983) as the essential condition to assure consistent estimation for a logistic regression analysis based on complete cases. For any regression model the condition $q_{ijk} \equiv q_{jk}$ implies consistency of the complete case regression estimates, because the selection process only changes the distribution of the covariates, but not the regression model. Logistic regression allows additionally the factor q_i, because a selection depending only on the outcome only changes the intercept, which is the essential argument in using logistic regression in the analysis of case-control studies (Breslow and Day 1980). Further aspects of biased estimation in a complete case analysis are discussed by Vach and Blettner 1991 and Vach (1994, pp. 18-20).

5. ML estimation under the MAR assumption

Obviously a complete case analysis is not an efficient method of estimating ϕ_j, because the neglected subjects carry information on the relation between Y and X_1. The MAR assumption allows us to use the information from these subjects. Let us first note that under the MAR assumption the complete case estimates may be biased. As now q_{ij} and Q_{ij} coincide, the bias factor depends only on estimable quantities. A first idea is to correct for this bias, i.e. to consider the corrected complete case estimates

$$\widehat{\phi}_j^{CCC} := \hat{\phi}_j^{CC} / \widehat{f}_j^{\phi} \text{ with } \widehat{f}_j^{\phi} := \frac{\widehat{Q}_{1j} \widehat{Q}_{01}}{\widehat{Q}_{0j} \widehat{Q}_{11}} \,,$$

$$\widehat{\psi}_k^{CCC} := \hat{\psi}_k^{CC} \text{ and } \hat{\tau}^{CCC} := \hat{\tau}^{CC} / \frac{\widehat{Q}_{11}}{\widehat{Q}_{01}}$$

This estimate was first considered by White (1982) for the case of a binary X_1. The same estimate was considered by Cain and Breslow (1988) as a special case of a conditional maximum likelihood estimate in a more general setting (Breslow and Cain 1988). The motivation outlined above was first presented by Vach and Blettner (1991).

If not only the removal of bias but also efficiency is our goal, estimation by the ML principle is a straightforward choice. However, besides the MAR assumption application of the ML principle requires a specification of a parametric family of distributions for the conditional distribution of X_2 given X_1 (Ibrahim 1990, Vach and Schumacher 1993). As both covariates are categorical in our setting, we can use the conditional probabilities

$$\pi_{k|j} := P(X_2 = k | X_1 = j)$$

directly. Hence the ML estimates for (β, π) result from maximizing

$$L(\beta,\pi) := \prod_{i,j} \left\{ \prod_k \left(p_{i|jk}(\beta)\pi_{k|j} \right)^{n_{ijk}} \right\} \left(\sum_k p_{i|jk}(\beta)\pi_{k|j} \right)^{n_{ij?}} .$$

Using preliminary results of Weinberg and Wacholder (1993), we show in the Appendix that the resulting estimates are identical to the corrected complete case estimates. This allows a simple investigation of the asymptotic bias in the next section.

Standard theory for maximum likelihood estimation ensure that the ML estimates of β are consistent and efficient (e.g. Lehmann 1983, p. 430), if the MAR assumption is valid.

6. Asymptotic bias of ML estimates under violation of the MAR assumption

As $\widehat{\psi}_k^{CCC} = \widehat{\psi}_k^{CC}$, the results on the asymptotic bias of $\widehat{\psi}_k^{CC}$ apply also to the ML estimates of ψ_k. Hence we can restrict ourselves to the asymptotic bias in estimating ϕ_j.

Whenever the complete case estimate $\widehat{\phi}_j^{CC}$ is consistent, the asymptotic bias factor of the ML estimate, i.e. the ratio between the stochastic limit of $\widehat{\phi}_j^{ML}$ and ϕ_j is equal to F_j^{-1} with

$$F_j := \frac{Q_{1j}Q_{01}}{Q_{0j}Q_{11}} .$$

In the general case, F_j^{-1} has to be multiplied additionally by the bias factor of the complete case estimate. This allows someone to identify two important situations, where the asymptotic bias vanishes: First, if the second covariate has no influence or, second, if $\phi_j = 1$ and the covariates are independent. However, we need additional assumptions on the missing value mechanism.

Lemma 1: *If $\psi_k \equiv 1$ and one of the following conditions holds, then the ML estimate for ϕ_j is consistent. The three conditions are:*

$$
\begin{array}{rl}
i) & q_{ijk} = q_{jk} q_i \\
ii) & q_{ijk} = q_{ij} q_k \\
iii) & q_{ijk} = q_{ik} q_j \wedge X_1 \text{ and } X_2 \text{ are independent}
\end{array}
$$

Proof: $\psi_k \equiv 1$ implies $P(X_2 = k|Y = i, X_1 = j) \equiv \pi_{k|j}$. Hence $Q_{ij} = \sum_k q_{ijk}\pi_{k|j}$. Now $i)$ implies $Q_{ij} = q_i \sum_k q_{jk}\pi_{k|j}$; hence $F_j = 1$. $ii)$ implies $Q_{ij} = q_{ij} \sum_k q_k \pi_{k|j}$, hence $F_j = \frac{q_{1j}q_{01}}{q_{0j}q_{11}}$, which coincides with the asymptotic bias factor of the complete case estimate. $iii)$ implies $\pi_{k|j} \equiv \pi_k$ and further $Q_{ij} = q_j \sum_k q_{ik}\pi_k$; hence $F_j = 1$. □

Lemma 2: *If $\phi_j \equiv 1, q_{ijk} = q_{ij}q_k$ and if X_1 and X_2 are independent, then the ML estimate of ϕ_j is consistent.*

Proof: $\phi_j \equiv 1$ and independence of X_1 and X_2 imply $P(X_2 = k|Y = i, X_1 = j) \equiv P(X_2 = k|Y = i)$. Hence $Q_{ij} = q_{ij} \sum_k q_k P(X_2 = k|Y = i)$ and F_j coincides with the asymptotic bias factor of the complete case estimate. □

Under the conditions of Lemma 2, but with $\phi_j \neq 1$, it does not hold that the bias is always toward 1. Even if additionally $q_{ijk} \equiv q_k$ and X_1 is a balanced dichotomous covariate, there exist constellations with $\phi_j > 1$ and an asymptotic bias factor larger than 1.

These theoretical results are directly of no great practical value, because none of the conditions is likely to be satisfied completely in practice. However, they may indicate the main factors with influence on the asymptotic bias: The size of the effect of X_2, the degree of dependence between X_1 and X_2, and the special type of the violation of the MAR assumption. To validate these factors, we compute the numerical value of the asymptotic bias for a variety of parameter constellations in the setting of two dichotomous covariates. For each choice of response rates we compute the maximal absolute bias in estimating a true β_{12} of 1.0 varying $P(X_1 = 1), P(X_2 = 1)$ and $P(Y = 1)$ between 0.2 and 0.8 by a step width of 0.1 and considering all possible combinations. For the dependence of X_1 and X_2 we investigate odds ratios of 1.0, 3.0 and 9.0 and with respect to the influence of X_2 we consider the cases $\beta_{22} = 1.0$ and $\beta_{22} = 2.0$.

We first consider constellations where the response rates do not depend on the outcome variable (Table 1). For all constellations we observe that the bias depends on the degree of correlation between X_1 and X_2 and on the size of β_{22}. If the response rates do not depend on X_1, that is $q_{ijk} \equiv q_k$, a small violation of the MAR assumption results in a small bias, even if β_{22}

and the degree of dependence between X_1 and X_2 is large, but increasing the degree of violation the bias may become unacceptably large. A similar picture is shown in the third and fourth row, where the response rates depend on X_1, but still can be factorized, that is $q_{ijk} \equiv q_j q_k$. However, if this does not hold, a small violation can induce a large bias even if the covariates are independent, which is demonstrated in the fifth row, where the ratio of the response probabilities between $X_2 = 2$ and $X_2 = 1$ depend on X_1. Even if the covariates are independent and balanced, a dependence of the degree of violation on the first covariate is a major source of bias, which is shown in Figure 1.

Second we consider constellations where the response rates depend on the outcome variable, but not on the first covariate (Table 2). Again the influence of the degree of dependence of the covariates and the size of β_{22} is obvious. In the first two rows the response rates can be factorized, that is $q_{ijk} \equiv q_i q_k$, and the results agree with the corresponding results in Table 1, which can be also shown using the results above. From the third row we conclude that a dependence of the degree of violation on the outcome can

q_{i11}	q_{i21}	q_{i12}	q_{i22}	$OR(X_1, X_2)$	$\beta_{22} = 1$	$\beta_{22} = 2$
				1.0	0.006	0.021
0.8		0.6		4.0	0.025	0.043
				9.0	0.047	0.085
				1.0	0.021	0.070
0.8		0.3		4.0	0.083	0.143
				9.0	0.155	0.278
				1.0	0.009	0.030
0.9	0.75	0.6	0.5	3.0	0.035	0.060
				9.0	0.066	0.118
				1.0	0.015	0.050
0.8	0.4	0.4	0.2	3.0	0.060	0.103
				9.0	0.112	0.201
				1.0	0.103	0.201
0.8	0.8	0.6	0.4	3.0	0.111	0.202
				9.0	0.138	0.253

Table 1: Maximal absolute bias in estimating $\beta_{12} = 1.0$ for selected response probabilities not depending on the outcome variable

be a source of bias, however independence of the covariates seems to limit this bias. This combined influence of the dependence of the covariates and of the dependence of the degree of violation on the outcome variable is further demonstrated in Figure 2 for the case of balanced covariates and a balanced outcome variable.

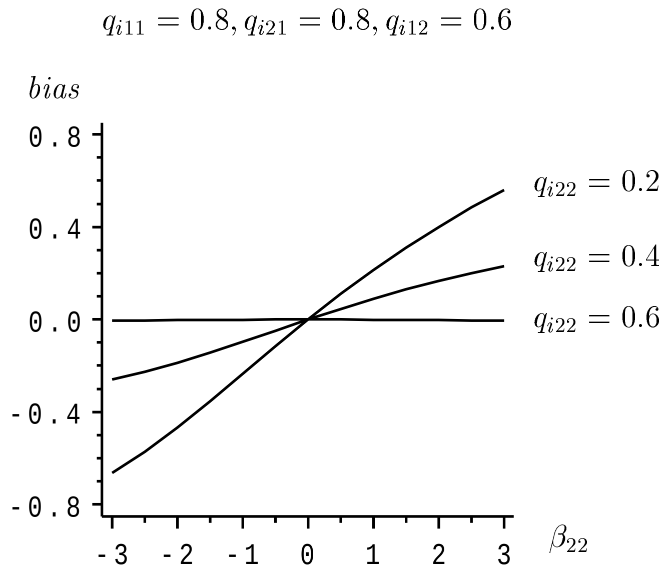

Figure 1: Asymptotic bias of the ML estimate $\widehat{\beta}_{12}^{ML}$ in dependence of β_{22} and q_{i22}, if $\beta_{12} = 1.0, P(X_1 = 1) = P(X_2 = 1) = P(Y = 1) = 0.5$ and $OR(X_1, X_2) = 1.0$

q_{1j1}	q_{0j1}	q_{1j2}	q_{0j2}	$OR(X_1, X_2)$	$\beta_{22} = 1$	$\beta_{22} = 2$
				1.0	0.009	0.030
0.9	0.75	0.6	0.5	3.0	0.035	0.060
				9.0	0.066	0.118
				1.0	0.015	0.050
0.8	0.4	0.4	0.2	3.0	0.060	0.103
				9.0	0.112	0.201
				1.0	0.030	0.067
0.8	0.8	0.6	0.4	3.0	0.119	0.133
				9.0	0.235	0.272

Table 2: Maximal absolute bias in estimating $\beta_{12} = 1.0$ for selected response probabilities depending on the outcome variable.

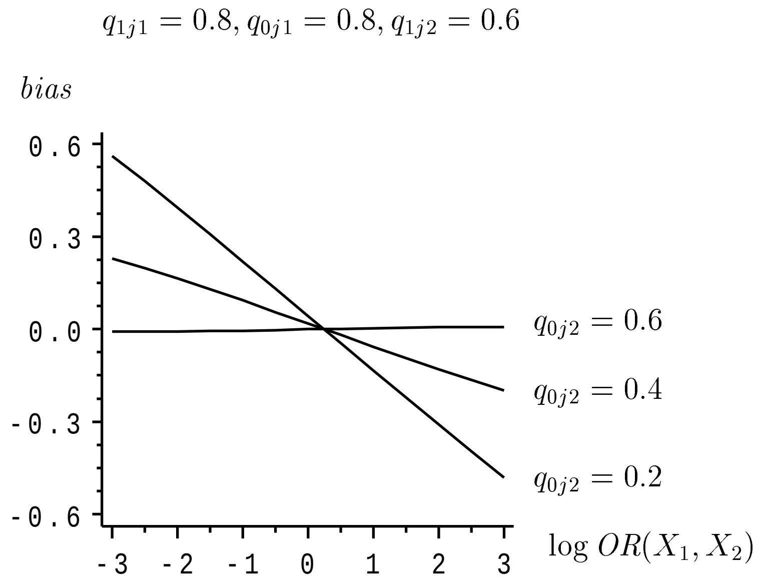

Figure 2: Asymptotic bias of the ML estimate $\widehat{\beta}_{12}^{ML}$ in dependence of $\log OR(X_1, X_2)$ and q_{0j2}, if $\beta_{12} = 1.0, \beta_{22} = 1.0, P(X_1 = 1) = P(X_2 = 1) = P(Y = 1) = 0.5$

7. Semiparametric methods

In the case of two categorical covariates the ML principle provides an appropriate tool to achieve efficient estimates. If the covariates are continuous, the necessity to specify parametric families for conditional distributions among the covariates prevents its application in practice. Semiparametric approaches avoid this problem and have been considered by several authors (PEPE and FLEMING 1991, CARROLL and WAND 1991, FLANDERS and GREENLAND 1991, REILLY and PEPE 1994, ROBINS ET AL. 1994). In our setting two of these approaches result in rather simple and intuitive methods, which have been shown to be less efficient than ML estimation (VACH 1994). Hence it may be worth to investigate, whether the loss of efficiency under the MAR assumption may be counterbalanced by a smaller bias under violation of the MAR assumption.

The approaches of PEPE and FLEMING (1991) and CARROLL and WAND (1991) reduce in our setting to the maximization of the likelihood $L(\beta, \hat{\pi})$ where $\hat{\pi}$ is a consistent estimate for π. VACH and SCHUMACHER (1993) showed that the estimate

$$\hat{\pi}^{VS}_{k|j} := \frac{n_{1jk}/\hat{Q}_{1j} + n_{0jk}/\hat{Q}_{0j}}{n_{.j.}}$$

is consistent under the MAR assumption and does not require additional assumptions as do the original proposals. The estimate achieved by maximizing $L(\beta, \hat{\pi})$ can be regarded as a pseudo maximum likelihood estimate $\hat{\beta}^{PML}$. In the Appendix we show that under the assumption $q_{ijk} \equiv q_{jk}$ the estimates $\hat{\pi}^{VS}_{k|j}$ and $\hat{\pi}^{ML}_{k|j}$ have the same stochastic limit. This implies that $\hat{\beta}^{ML}$ and $\hat{\beta}^{PML}$ have the same asymptotic bias.

The approach of REILLY and PEPE (1994) reduces in our setting to the analysis of a $2 \times J \times K$ contingency table with estimated entries

$$\hat{n}_{ijk} := n_{ijk} + n_{ij?}\frac{n_{ijk}}{n_{ij+}} = \frac{n_{ijk}}{\hat{Q}_{ij}} .$$

Due to the appealing interpretation VACH and BLETTNER (1991) called this method "Filling". Under the assumption $q_{ijk} = q_i q_{jk}$ the stochastic limit $\tilde{p}_{ijk}$ of $\hat{n}_{ijk}/n$ is equal to $p_{ijk}q_i q_{jk}/Q_{ij}$; hence the corresponding odds ratios satisfy $\tilde{\phi}_{jk} = \phi_j/F_j$ and $\tilde{\psi}_{jk} = \psi_k$. This implies that estimates from the filled table have the same asymptotic bias as the ML estimates. Computations of the asymptotic bias for response rates not satisfying $q_{ijk} \equiv q_{jk}$ or $q_{ijk} \equiv q_i q_{jk}$, respectively, show only slight differences between the three estimates.

Hence these methods provide no alternative to reduce the sensitivity against violation of the MAR assumption.

8. Implications for the analysis of prospective and retrospective studies

So far we have identified some major sources of bias due to a violation of the MAR assumption. Specific constellations of the response probabilities are one source. As these probabilities are unknown and cannot be estimated from the available data, we have to rely on a-priori assumptions. However, the design of a study and the conceptual meaning of covariates may allow such assumptions.

In many studies we first collect data on the covariates and later on data on the outcome variable representing an event happening after finishing collection of covariate data. This prospective measurement of the outcome variable is typical for controlled clinical trials, where all covariates are measured at baseline. In this setting we can usually exclude a dependence of the response rates on the outcome variable, i.e. $q_{ijk} \equiv q_{jk}$ holds. This implies that complete case estimates are consistent. If we decide to use ML estimation under the MAR assumption to improve efficiency, a large bias due to a violation of the MAR assumption can easily be identified, because then the ML estimates differ distinctly from the complete case estimates. However, if the difference is small, we do not know, whether this indicates bias or whether it is just the necessary correction to improve efficiency. The next point is to check whether we can additionally assume $q_{jk} \equiv q_k$, as this would reduce the risk of a serious bias. There are situations where it is obvious that we cannot exclude such a dependence, e.g. if X_1 is age or sex and X_2 is a question on sexual behaviour. However, often the conceptual context of the covariates allows to exclude such a dependence. This is especially true if X_1 is a randomized treatment. In any case one should look at the size of the effect of X_2 and the degree of dependence between X_1 and X_2. Whereas the first is estimated in a consistent manner even if the MAR assumption is violated, estimates of the latter may be biased too.

We should mention that even in such a prospective setting the assumption $q_{ijk} \equiv q_{jk}$ may be violated, if there is a latent variable with strong impact on the outcome variable and the missing value mechanism. In clinical trials such a variable may be a positive/negative attitude to clinical medicine in general or the patients expectation on the success of the therapy or unpleasant side effects.

If data on the outcome variable is collected in a retrospective manner, the assumption $q_{ijk} = q_{jk}$ is highly questionable. Especially in case-control studies different data collection procedures for cases and controls imply some dependence of the response probabilities on the outcome. Additionally, if the missing at random assumption is questionable different data collection procedures are likely to result in a different degree of the violation of the MAR assumption. Hence the situation $q_{ijk} \equiv q_{ik}$ can be regarded as typical for a case control study. Now if X_1 is the exposure of interest and X_2 is a potential confounder, i.e. if X_1 and X_2 are correlated and X_2 has an effect on Y, our numerical results suggest that the typical situation results in substantial bias! Additionally, any problem mentioned above for the prospective setting can occur also in the retrospective setting.

Note that in case-control studies with incomplete covariate data the use of the prospective logistic model can be justified (WACHOLDER and WEINBERG 1994, CARROLL ET AL. 1995).

9. Conclusions

A violation of the MAR assumption can result in a serious bias, if methods relying on this assumption are used to handle incomplete covariate data. We investigated this bias for the case of logistic regression with two categorical covariates, where only the second is affected by missing values. With respect to the estimation of the effect of the completely observed covariate, our investigations suggest that this bias is small if we have a pure violation in the sense that the response probabilities depend only on the true value of the covariate. An additional dependence on the first covariate or on the outcome variable can be a source of serious bias. Furthermore the degree of dependence between the covariates and the size of the effect of the second covariate have an impact on the bias.

Our results can be generalized in the way that X_2 can be a vector of categorical covariates and that the regression model includes interactions with X_1. If X_2 is continuous our results are also valid, if we consider ML estimation with arbitrary conditional distributions $\mathcal{D}(X_2|X_1 = j)$ putting mass only on the observed values of X_2. In a related framework COSSLETT (1981) considered estimates of this type.

The results of this paper allow some qualitative statements about the magnitude of a potential bias. In applications we need additional quantitative information about a possible bias, especially if we have some prior information on the kind and magnitude of the violation of the MAR assumption. VACH and BLETTNER (1995) provide a framework to estimate regression parameters under a specified non MAR mechanism and suggest a sensitivity analysis by investigating systematically the variation of the parameter estimates under specified violations. Their conclusions derived from some examples agree with the results of this paper: First, they observe that the estimates are not too sensitive against violations of the MAR assumption if the observed response rates are equal. Second, they observe in the analysis of a case-control study that estimates are highly sensitive against violations of the MAR assumption if the degree of violation differs between cases and controls.

Acknowledgements

The authors are indebted to Martin Schumacher and Clarice Weinberg for comments on a previous draft of this paper. The first author has been partially supported by the Deutsche Forschungsgemeinschaft.

Appendix

We have to show the identity of $\widehat{\beta}^{CCC}$ and $\widehat{\beta}^{ML}$. We will do this in a more general framework of regression models with

$$\frac{p_{1|jk}(\gamma,\theta)}{p_{0|jk}(\gamma,\theta)} = \gamma_j g_{jk}(\theta)$$

with a prespecified function g_{jk}. Our proof follows closely that of Weinberg and Wacholder (1993), who show a similar identity to justify the prospective analysis of case-control studies.

The ML estimate $(\widehat{\gamma}^{ML}, \widehat{\theta}^{ML}, \widehat{\pi}^{ML})$ maximizes

$$L(\gamma,\theta,\pi) := \prod_{i,j,k} p_{i|jk}(\gamma,\theta)^{n_{ijk}} \prod_{j,k} \pi_{k|j}^{n_{\cdot jk}} \prod_{i,j} p_{i|j}(\gamma,\theta,\pi)^{n_{ij?}}$$

$$\text{with } p_{i|j}(\gamma,\theta,\pi) := \sum_{k=1}^{K} p_{i|jk}(\gamma,\theta)\pi_{k|j}$$

The estimate $(\widehat{\gamma}^{CCC}, \widehat{\theta}^{CCC})$ can be expressed as the maximum of

$$L^*(\gamma,\theta) = \prod_{i,j,k} p^*_{i|jk}(\gamma,\theta)^{n_{ijk}} \text{ with } \frac{p^*_{1|jk}(\gamma,\theta)}{p^*_{0|jk}(\gamma,\theta)} = \gamma_j \frac{\widehat{Q}_{1j}}{\widehat{Q}_{0j}} g_{jk}(\theta)\ ,$$

i.e. of a likelihood with an appropriate offset.

In a first step, we show that for θ fixed, $\widehat{\gamma}^{ML}(\theta)$ and $\widehat{\gamma}^{CCC}(\theta)$ coincide. For complete data summarized in a $2 \times J \times K$ contingency table with entries $\widetilde{n}_{ijk}$ it can be shown that for θ fixed the ML estimate $\widetilde{\gamma}$ is uniquely determined by

$$\sum_k \widetilde{n}_{\cdot jk} p_{0|jk}(\widetilde{\gamma},\theta) = n_{0j\cdot} \quad \text{for all } j \tag{1}$$

and the ML estimate $\widetilde{\pi}$ satisfies

$$\widetilde{\pi}_{k|j} = \frac{\widetilde{n}_{\cdot jk}}{\widetilde{n}_{\cdot j\cdot}} \tag{2}$$

For incomplete data, the ML estimate $(\widehat{\gamma}, \widehat{\pi}) := (\widehat{\gamma}^{ML}(\theta), \widehat{\pi}^{ML}(\theta))$ is a fixed point of the EM algorithm (Dempster, Laird and Rubin 1977), hence $\widehat{\gamma}$ and $\widehat{\pi}$ satisfy (1) and (2) for the contingency table with entries

$$\widehat{n}_{ijk} := n_{ijk} + n_{ij?} P_{\widehat{\gamma},\widehat{\pi},\theta}(X_2 = k | Y = i, X_1 = j) \ ;$$

hence

$$\sum_k \widehat{n}_{\cdot jk} p_{0|jk}(\widehat{\gamma}, \theta) = \widehat{n}_{0j\cdot} = n_{0j\cdot} \ \text{ for all } j \ \text{ and} \tag{3}$$

$$\widehat{\pi}_{k|j} = \frac{\widehat{n}_{\cdot jk}}{\widehat{n}_{\cdot j\cdot}} = \frac{\widehat{n}_{\cdot jk}}{n_{\cdot j\cdot}}$$

This implies

$$p_{0|j}(\widehat{\gamma}, \theta, \widehat{\pi}) = \sum_k p_{0|jk}(\widehat{\gamma}, \theta) \widehat{\pi}_{k|j} = \frac{1}{n_{\cdot j\cdot}} \sum_k \widehat{n}_{\cdot jk} p_{0|jk}(\widehat{\gamma}, \theta) = \frac{n_{0j\cdot}}{n_{\cdot j\cdot}} \tag{4}$$

and

$$\widehat{\pi}_{k|j} := \frac{n_{\cdot jk}}{n_{\cdot j\cdot} H^{\widehat{Q}}_{jk}(\widehat{\gamma}, \theta)} \ \text{with } H^{Q}_{jk}(\gamma, \theta) := \sum_{i=0}^{1} Q_{ij} p_{i|jk}(\gamma, \theta) \ . \tag{5}$$

The latter follows from

$$\widehat{\pi}_{k|j} = \frac{\widehat{n}_{\cdot jk}}{n_{\cdot j\cdot}} = \frac{1}{n_{\cdot j\cdot}} \left(n_{\cdot jk} + \sum_{i=0}^{1} n_{ij?} \frac{p_{i|jk}(\widehat{\gamma}, \theta) \widehat{\pi}_{k|j}}{p_{i|j}(\widehat{\gamma}, \theta)} \right)$$

and hence with (4)

$$\begin{aligned}
\widehat{\pi}_{k|j} &= n_{\cdot jk} \left(n_{\cdot j\cdot} - \sum_{i=0}^{1} n_{ij?} \frac{n_{\cdot j\cdot}}{n_{ij\cdot}} p_{i|jk}(\widehat{\gamma}, \theta) \right)^{-1} \\
&= \frac{n_{\cdot jk}}{n_{\cdot j\cdot}} \left(1 - \sum_{i=0}^{1} (1 - \widehat{Q}_{ij}) p_{i|jk}(\widehat{\gamma}, \theta) \right)^{-1} .
\end{aligned}$$

Now (3) is equivalent to

$$\sum_k n_{\cdot jk} p^{*}_{0|jk}(\widehat{\gamma}, \theta) = n_{0j+} \ \text{ for all } j \tag{6}$$

because $\widehat{n}_{\cdot jk} = n_{\cdot j\cdot} \widehat{\pi}_{k|j} = n_{\cdot jk} / H^{\widehat{Q}}_{jk}(\widehat{\gamma}, \theta)$ and

$$\frac{p_{0|jk}(\widehat{\gamma}, \theta)}{H^{\widehat{Q}}_{jk}(\widehat{\gamma}, \theta)} = \frac{1}{1 + \widehat{\gamma}_j g_{jk}(\theta)} \left(\widehat{Q}_{1j} \frac{\widehat{\gamma}_j g_{jk}(\theta)}{1 + \widehat{\gamma}_j g_{jk}(\theta)} + \widehat{Q}_{0j} \frac{1}{1 + \widehat{\gamma}_j g_{jk}(\theta)} \right)^{-1}$$

$$= \frac{1}{\widehat{Q}_{0j} + \widehat{Q}_{1j}\widehat{\gamma}_j g_{jk}(\theta)} = \frac{1}{\widehat{Q}_{0j}} \frac{1}{1 + \widehat{\gamma}_j g_{jk}(\theta)\frac{\widehat{Q}_{1j}}{\widehat{Q}_{0j}}} = \frac{n_{0j\cdot}}{n_{0j+}} p^*_{0|jk}(\widehat{\gamma}, \theta) \ .$$

As $\widehat{\gamma}^{CCC}(\theta)$ is uniquely determined by (6), this finishes the first step. In the second step we show that the difference of the profile loglikelihood

$$\log L^*(\widehat{\gamma}(\theta), \theta) - \log L(\widehat{\gamma}(\theta), \theta, \widehat{\pi}(\theta)) \tag{7}$$

is independent of θ, which implies the coincidence of $\widehat{\theta}^{CCC}$ and $\widehat{\theta}^{ML}$. (7) is equal to

$$\sum_{i,j,k} n_{ijk} \log \frac{p^*_{i|jk}(\widehat{\gamma}(\theta), \theta)}{p_{i|jk}(\widehat{\gamma}(\theta), \theta)} - \sum_{j,k} n_{\cdot jk} \log \widehat{\pi}_{k|j}(\theta) - \sum_{i,j} n_{ij?} \log p_{i|j}(\widehat{\gamma}(\theta), \theta, \widehat{\pi}(\theta))$$

and $p^*_{i|jk}(\widehat{\gamma}(\theta), \theta)/p_{i|jk}(\widehat{\gamma}(\theta), \theta) = \widehat{Q}_{ij}/H^{\widehat{Q}}_{jk}(\widehat{\gamma}(\theta), \theta)$, which we have shown above for $i = 0$ and which can be shown similarly for $i = 1$. With $\widehat{\pi}_{k|j}(\theta) = \frac{n_{\cdot jk}}{n_{\cdot j\cdot}}/H^{\widehat{Q}}_{jk}(\widehat{\gamma}(\theta), \theta)$ and (4) we have shown that (7) does not depend on θ.

It remains to prove the results of Section 7. Under the assumption $q_{ijk} \equiv q_{jk}$ the stochastic limit of $\widehat{\pi}^{VS}_{k|j}$ is equal to

$$\frac{1}{p_{\cdot j\cdot}} \left(\frac{q_{jk} p_{1jk}}{Q_{1j}} + \frac{q_{jk} p_{0jk}}{Q_{0j}} \right) = q_{jk} \pi_{k|j} \left(\frac{p_{1|jk}}{Q_{1j}} + \frac{p_{0|jk}}{Q_{0j}} \right)$$

and by (5) the stochastic limit of $\widehat{\pi}^{ML}_{k|j}$ is equal to $q_{jk}\pi_{k|j} H^Q_{jk}(\gamma^*, \theta)^{-1}$ where γ^* is the stochastic limit of $\widehat{\gamma}^{ML}$, i.e. $\gamma^*_j := \gamma_j / \frac{Q_{1j}}{Q_{0j}}$. Now

$$\begin{aligned}
H^Q_{jk}(\gamma^*, \theta)^{-1} &= \left(Q_{1j} \frac{\gamma_j g_{jk}(\theta) \frac{Q_{0j}}{Q_{1j}}}{1 + \gamma_j g_{jk}(\theta) \frac{Q_{0j}}{Q_{1j}}} + Q_{0j} \frac{1}{1 + \gamma_j g_{jk}(\theta) \frac{Q_{0j}}{Q_{1j}}} \right)^{-1} \\
&= \frac{1 + \gamma_j g_{jk}(\theta) \frac{Q_{0j}}{Q_{1j}}}{Q_{0j}(1 + \gamma_j g_{jk}(\theta))} = \frac{1}{Q_{0j}} \frac{1}{1 + \gamma_j g_{jk}(\theta)} + \frac{1}{Q_{1j}} \frac{\gamma_j g_{jk}(\theta)}{1 + \gamma_j g_{jk}(\theta)} \\
&= \frac{p_{0|jk}}{Q_{0j}} + \frac{p_{1|jk}}{Q_{1j}} \ ;
\end{aligned}$$

hence the stochastic limits coincide.

References

Breslow, N.E. and Cain, K.C., 1988: Logistic regression for two-stage case-control data. Biometrika **75**, 11-20.

BRESLOW, N.E. and DAY, N.E. 1980: *Statistical methods in cancer research, Vol. 1 – The analysis of case-control studies.* IARC Scientific Publications No. 32. Lyon.

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

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

CARROLL, R.J., WANG, S. and WANG, C.Y., 1995: Prospective analysis of logistic case-control studies. Journal of the American Statistical Association **90**, 157-169.

COSSLETT, S.R. 1981. Efficient estimation of discrete choice models. In: *Structural analysis of discrete data with econometric applications*, Ed. Manski, C.F., McFadden, D., MIT Press, Cambridge MA. 51-111.

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

FLANDERS, W.D. and GREENLAND, S., 1991: Analytic methods for two-stage case-control studies and other stratified designs. Statistics in Medicine **10**, 739-747.

GLYNN, R.J. and LAIRD, N.M., 1983. Regression estimates and missing data: complete case analysis; Technical Report.

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

LEHMANN, E.L. 1983: *Theory of point estimation.* Wiley. New York.

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

REILLY, M. and PEPE, M., 1995: A Mean Score method for missing and auxiliary covariate data in regression models. Biometrika **82**, 299-314.

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

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

VACH, W. 1994: *Logistic regression with missing values in the covariates.* Lecture Notes in Statistics **86**. Springer. New York.

VACH, W. and BLETTNER, M., 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 BLETTNER, M., 1995: Logistic regression with incompletely observed categorical covariates – Investigating the sensitivity against violation of the missing at random assumption. Statistics in Medicine **14**, 1315-1329.

VACH, W. and SCHUMACHER, M., 1993: Logistic regression with incompletely observed categorical covariates – A comparison of three approaches. Biometrika **80**, 353-362.

WACHOLDER, S. and WEINBERG, C.R., 1994: Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling. Biometrics **50**, 350-357.

WEINBERG, C.R. and WACHOLDER, S, 1993: Prospective analysis of case-control data under general multiplicative-intercept risk models. Biometrika **80**, 461-465.

WHITE, J.E., 1982: A two-stage design for the study of the relationship between a rare exposure and a rare disease. American Journal of Epidemiology **115**, 119-128.