On protected estimation of an odds ratio model with missing binary exposure and confounders

BY E. J. TCHETGEN TCHETGEN

Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.
 etchetge@hsph.harvard.edu

AND A. ROTNITZKY

Department of Economics, Di Tella University, Sáenz Valiente 1010, Buenos Aires, Argentina
 arotnitzky@utdt.edu

SUMMARY

We describe an estimator of the parameter indexing a model for the conditional odds ratio between a binary exposure and a binary outcome given a high-dimensional vector of confounders, when the exposure and a subset of the confounders are missing, not necessarily simultaneously, in a subsample. We argue that a recently proposed estimator restricted to complete-cases confers more protection to model misspecification than existing ones in the sense that the set of data laws under which it is consistent strictly contains each set of data laws under which each of the previous estimators are consistent.

Some key words: Inverse probability weighted; Logistic regression; Missing at random; Model misspecification.

1. INTRODUCTION

A common aim of epidemiologic observational studies is to evaluate the causal effect of a dichotomous point exposure A on the risk of a disease outcome encoded in a binary variable Y . In an effort to account for confounding bias in such nonexperimental studies, investigators routinely collect and adjust for in data analysis, a large number of confounding factors encoded in a vector L . A common target of inference is a finite-dimensional parameter ψ indexing a model $\text{OR}(L; \psi)$ for the conditional log-odds ratio function:

$$\text{OR}(L) = \log \left\{ \frac{\text{pr}(Y = 1 \mid A = 1, L) \text{pr}(Y = 0 \mid A = 0, L)}{\text{pr}(Y = 0 \mid A = 1, L) \text{pr}(Y = 1 \mid A = 0, L)} \right\}, \quad (1)$$

where $\text{OR}(\cdot; \cdot)$ is a known function and the true value of ψ is unknown. Under the assumption of no unmeasured confounders $\text{OR}(L)$ quantifies, on the log-odds ratio scale, the causal effect of A on Y . The function $\text{OR}(\cdot)$ is attractive as it is invariant to alterations in the marginal distributions of (Y, L) or of (A, L) . As a consequence, ψ can be consistently estimated from data (Y_i, A_i, L_i) ($i = 1, \dots, n$) collected under a variety of commonly employed epidemiological designs, in particular, simple random sample designs and case-control designs unmatched or matched on some components of L .

In this note, we consider the estimation of ψ when Y and a subset L_{obs} of the confounders L are always observed but A and the remaining components L_{mis} of L are not fully observed in all study participants, perhaps only A missing in some, only some random subset of L_{mis} missing in others and both missing in a third subsample. We examine inference under the assumptions that:

$$\text{pr}(R = 1 \mid A, L_{\text{mis}}, L_{\text{obs}}, Y) = \text{pr}(R = 1 \mid L_{\text{obs}}, Y) \quad (2)$$

and $\text{pr}(R = 1 \mid L_{\text{obs}}, Y) > 0$, where $R = 1$ if A and L_{mis} are jointly observed, and $R = 0$ otherwise. When A and L_{mis} can only be either both observed or both missing simultaneously, (2) is a particular

instance of the assumption that the data are missing at random (Little & Rubin, 2002) which is routinely made in the literature on missing covariate data (Robins et al., 1994; Zhao et al., 1996; Lipsitz et al., 1998; Little & Rubin, 2002; Parzen et al., 2002). For concreteness, we assume that (R_i, Y_i, A_i, L_i) ($i = 1, \dots, n$) are independent and identically distributed. Our discussion equally applies when the data are collected from a case-control design unmatched or matched on some of the confounders L . We argue that an estimator of ψ recently proposed by Tchetgen Tchetgen et al. (2010), computed just from the complete-case units, i.e., those for which A and L_{mis} are observed, is more robust than currently available estimators in the sense that the set of data laws under which it is consistent strictly contains each set of data laws under which each of the previous estimators is consistent.

2. DIMENSION REDUCING ESTIMATION STRATEGIES

The right-hand side of (1) and the baseline log-odds function

$$\text{odds}Y(L) = \log \left\{ \frac{\text{pr}(Y = 1 \mid A = 0, L)}{\text{pr}(Y = 0 \mid A = 0, L)} \right\} \quad (3)$$

determine the conditional probability function $\text{pr}(Y = 1 \mid A, L)$. In the absence of missing data, to ensure maximal robustness against model misspecification, it is desirable to estimate ψ under a model that assumes just (1) rather than under a model that parameterizes the regression function $\text{pr}(Y = 1 \mid A, L)$ and consequently that parameterizes the right-hand sides of both (1) and (3). Tchetgen Tchetgen et al. (2010) showed that due to the curse of dimensionality this is not feasible when, as is often the case in applications to render the assumption of no unmeasured confounders plausible, L is a vector with several components. The situation is aggravated when (A, L_{mis}) are missing by happenstance, as then the conditional missingness probability function $\text{pr}(R = 1 \mid L_{\text{obs}}, Y)$ is also unknown. The following dimension reduction strategies appear to be available. One strategy is to assume (a) a parametric model $\text{odds}A(L; \alpha)$ for

$$\text{odds}A(L) = \log \left\{ \frac{\text{pr}(A = 1 \mid Y = 0, L)}{\text{pr}(A = 0 \mid Y = 0, L)} \right\}. \quad (4)$$

Model (1) coupled with model $\text{odds}A(L; \alpha)$ for (4) renders the parametric retrospective logistic regression model for A on Y and L ,

$$\log \left\{ \frac{\text{pr}(A = 1 \mid Y, L; \psi, \alpha)}{\text{pr}(A = 0 \mid Y, L; \psi, \alpha)} \right\} = \text{odds}A(L; \alpha) + \text{OR}(L; \psi)Y. \quad (5)$$

The problem then reduces to estimation of a parametric logistic regression function with incomplete observations of the outcomes A and/or a subset L_{mis} of the covariates, when the probability of observing both A and L_{mis} depends only on the observed model covariates Y and L_{obs} . Under such conditions, it is well known that the standard logistic regression estimator of ψ , say $\tilde{\psi}_{\text{retro}}$, restricted to the complete-case units, i.e., the units with no missing data, is consistent and asymptotically normal. In fact, when $L_{\text{obs}} = L$ so that $L_{\text{mis}} = \emptyset$, $\tilde{\psi}_{\text{retro}}$ is semiparametric efficient under the model that assumes (5) and condition (2) on the missing data mechanism, so under just these assumptions, no information is lost in large samples by discarding incomplete units (Robins et al., 1994).

An alternative strategy would be to assume (b) a parametric model $\text{odds}Y(L; \eta)$ for $\text{odds}Y(L)$, which, in conjunction with model (1), renders the parametric logistic regression model for Y on A and L ,

$$\log \left\{ \frac{\text{pr}(Y = 1 \mid A, L; \psi, \eta)}{\text{pr}(Y = 0 \mid A, L; \psi, \eta)} \right\} = \text{odds}Y(L; \eta) + \text{OR}(L; \psi)A. \quad (6)$$

The problem then reduces to estimation under a logistic regression model with a subset of the covariates incompletely observed and probability of complete observations that depends at most on the outcome and on the always observed covariates. However, when L is high-dimensional, this dimension-reducing strategy is insufficient to avoid the curse of dimensionality and further model restrictions are needed. One option

is to postulate, in addition to (b), a parametric model $f(A, L_{\text{mis}} | L_{\text{obs}}; \nu)$ for $f(A, L_{\text{mis}} | L_{\text{obs}})$ and to conduct maximum likelihood estimation of the model parameters (ψ, η, ν) . This strategy, however, leads to even more restrictions than the sole restriction (a) required by the retrospective logistic regression $\tilde{\psi}_{\text{retro}}$ because it leads not only to a parametric model for $f(A | Y, L)$ but for the joint law $f(Y, A, L_{\text{mis}} | L_{\text{obs}})$.

Alternatively, one might consider assuming (b) and (c) a parametric model $\text{pr}(R = 1 | L_{\text{obs}}, Y; \lambda)$ for $\text{pr}(R = 1 | L_{\text{obs}}, Y)$, thus arriving at the problem of estimating the logistic regression model (6) under a parametric model for the probability of complete observations. In such a setting, the weighted logistic regression estimator of ψ , say $\tilde{\psi}_{\text{ipw}}$, in model (6) restricted to the complete-case units and weights equal to $\hat{\pi}^{-1} = \text{pr}(R = 1 | L_{\text{obs}}, Y; \hat{\lambda})^{-1}$, where $\hat{\lambda}$ is the maximum likelihood estimator of λ , is consistent and asymptotically normal (Robins et al., 1994). This estimator, known as the inverse probability weighted estimator, was used, for example, in Moore et al. (2009), in a setting in which only A was missing.

It is possible to construct a so-called augmented inverse probability weighted estimator of ψ , say $\tilde{\psi}_{\text{aipw}}$, which is more robust than $\tilde{\psi}_{\text{ipw}}$. Like $\tilde{\psi}_{\text{ipw}}$, the estimator, $\tilde{\psi}_{\text{aipw}}$ is consistent if (b) and (c) hold, but unlike $\tilde{\psi}_{\text{ipw}}$, it is also consistent even if (c) fails provided (a), (b) and (d) a parametric model $f(L_{\text{mis}} | L_{\text{obs}}, Y; \tau)$ for $f(L_{\text{mis}} | L_{\text{obs}}, Y)$ hold. To compute $\tilde{\psi}_{\text{aipw}}$ one first computes, using only the complete-case units, the maximum likelihood estimators $\hat{\tau}$ and $(\hat{\alpha}_{\text{retro}}, \tilde{\psi}_{\text{retro}})$ of τ and (α, ψ) under models (d) and (6), respectively. The estimator $\tilde{\psi}_{\text{aipw}}$ of ψ is the subvector of $(\psi^T, \tilde{\eta}^T)^T$ solving

$$E_n \left[R \hat{\pi}^{-1} S(\psi, \eta) - (\hat{\pi}^{-1} R - 1) \hat{E}\{S(\psi, \eta) | Y, L_{\text{obs}}\} \right] = 0,$$

where $S(\psi, \eta) = (1, A, L')\{Y - \text{pr}(Y = 1 | A, L; \psi, \eta)\}$, $\hat{E}(\cdot | Y, L_{\text{obs}})$ is the expectation with respect to $\hat{f}(A, L_{\text{mis}} | L_{\text{obs}}, Y) = f(A | L, Y; \hat{\alpha}_{\text{retro}}, \tilde{\psi}_{\text{retro}})f(L_{\text{mis}} | L_{\text{obs}}, Y; \hat{\tau})$ and here and throughout $E_n(\cdot)$ denotes the empirical mean operator. Tchetgen Tchetgen (2009) developed a simple algorithm for implementing this estimator when L is fully observed.

In summary, of the dimension reducing procedures examined so far, the two which are consistent under the weakest modelling assumptions are:

- (i) the estimator $\tilde{\psi}_{\text{retro}}$, which requires that model (a) be correct;
- (ii) the estimator $\tilde{\psi}_{\text{aipw}}$, which requires that either models (b) and (c) or models (a) and (b) and (d) be correct.

3. THE PROPOSED ESTIMATOR

We now argue that an estimator of ψ recently derived in Tchetgen Tchetgen et al. (2010) computed just from the complete-case units confers more robustness to model misspecification than the estimators $\tilde{\psi}_{\text{retro}}$ and $\tilde{\psi}_{\text{aipw}}$. To compute this estimator, one first computes the standard logistic regression estimators $\hat{\alpha}$ of α and $\hat{\eta}$ of η in models (a) and (b) using only data from complete-case units. One then discards the separate estimators of ψ obtained from the fit of each model and instead, one computes the desired estimator, say $\hat{\psi}$, by solving the estimating equations $E_n\{U(\psi; \hat{\eta}, \hat{\alpha})\} = 0$ where

$$U(\psi; \eta, \alpha) = R \frac{\partial \text{OR}(L; \psi)}{\partial \psi} \{A - d(L; \psi, \eta, \alpha)\} [Y - \text{expit}\{\text{odds}Y(L; \eta) + \text{OR}(L; \psi)A\}]$$

with

$$d(L; \psi, \eta, \alpha) = 1 + \frac{\exp\{\text{OR}(L; \psi)\} \times [1 + \exp\{\text{OR}(L; \psi) + \text{odds}Y(L; \eta)\}]}{\exp\{\text{odds}A(L; \alpha)\} \times [1 + \exp\{\text{odds}Y(L; \eta)\}]}.$$

Solving the estimating equation may appear computationally challenging, but this is not the case as Tchetgen Tchetgen & Rotnitzky (2011) provide an iterative algorithm that implements $\hat{\psi}$ with standard logistic regression software. The algorithm is reproduced in the online Supplementary Material for completeness. These authors showed that in the absence of missing data, $\hat{\psi}$ is consistent and asymptotically normal for ψ provided either $\text{odds}A(L; \alpha)$ is a correct model for $\text{odds}A(L)$ or, $\text{odds}Y(L; \eta)$ is a correct model for $\text{odds}Y(L)$, but not necessarily both assertions hold. This result immediately implies that with

missing data, $\hat{\psi}$ is an estimator of the parameter ψ indexing a model $\text{OR}(L; \psi)$ for the complete-case log-OR function:

$$\text{OR}_{\text{cc}}(L) = \log \left\{ \frac{\text{pr}(Y = 1 \mid A = 1, L, R = 1) \times \text{pr}(Y = 0 \mid A = 0, L, R = 1)}{\text{pr}(Y = 1 \mid A = 0, L, R = 1) \times \text{pr}(Y = 0 \mid A = 1, L, R = 1)} \right\},$$

which is consistent and asymptotically normal for ψ provided either $\text{odds}A(L; \alpha)$ is a correct model for $\text{odds}_{\text{cc}}A(L)$ or, $\text{odds}Y(L; \eta)$ is a correct model for $\text{odds}_{\text{cc}}Y(L)$, but not necessarily both, where

$$\text{odds}A_{\text{cc}}(L) = \log \frac{\text{pr}(A = 1 \mid Y = 0, L, R = 1)}{\text{pr}(A = 0 \mid Y = 0, L, R = 1)}, \quad \text{odds}Y_{\text{cc}}(L) = \log \frac{\text{pr}(Y = 1 \mid A = 0, L, R = 1)}{\text{pr}(Y = 0 \mid A = 0, L, R = 1)}$$

are the complete-case baseline log-odds functions.

Under (2), $\text{OR}_{\text{cc}}(L) = \text{OR}(L)$ because the full-data law $f(A, Y, L)$ differs from the complete-case data law $f(A, Y, L \mid R = 1)$ only in that $f(Y, L_{\text{obs}})$ is distinct from $f(Y, L_{\text{obs}} \mid R = 1)$ and, as noted earlier, $\text{OR}(L)$ remains unchanged under departures from the marginal law of (Y, L) . Thus, a model $\text{OR}(L; \psi)$ for $\text{OR}_{\text{cc}}(L)$ is also a model for $\text{OR}(L)$. Furthermore, because $\text{odds}A_{\text{cc}}(L)$ depends solely on $\text{pr}(A = 1 \mid Y = 0, L, R = 1)$ and this is equal to $\text{pr}(A = 1 \mid Y = 0, L)$ under (2), then $\text{odds}A_{\text{cc}}(L) = \text{odds}A(L)$ and thus model $\text{odds}A(L; \alpha)$ is also a model for $\text{odds}A(L)$.

We thus arrive at the conclusion that under (2), $\hat{\psi}$ is consistent for ψ if one of Conditions 1 and 2 below holds, but not necessarily both.

Condition 1. Model (a) is correctly specified.

Condition 2. A parametric model for $\text{odds}Y_{\text{cc}}(L)$ is correctly specified.

The identity

$$\log \left\{ \frac{\text{pr}(Y = 1 \mid A = 0, L, R = 1)}{\text{pr}(Y = 0 \mid A = 0, L, R = 1)} \right\} = \log \left\{ \frac{\text{pr}(Y = 1 \mid A = 0, L)}{\text{pr}(Y = 0 \mid A = 0, L)} \right\} + \log \left\{ \frac{\text{pr}(R = 1 \mid Y = 1, L)}{\text{pr}(R = 1 \mid Y = 0, L)} \right\}$$

implies that a parametric model for the full-data baseline log-odds $\text{odds}Y(L)$ and another for $\text{pr}(R = 1 \mid L, Y) = \text{pr}(R = 1 \mid L_{\text{obs}}, Y)$ determine a parametric model for $\text{odds}Y_{\text{cc}}(L)$. Consequently Condition 2 is met, in particular, if Condition 3 is met and a parametric model for $\text{odds}Y_{\text{cc}}(L)$ is constructed combining (b) and (c) via the identity (7).

Condition 3. Models (b) and (c) are correctly specified.

In summary,

(iii) consistency of $\hat{\psi}$ requires that either model (a) or models (b) and (c) be correct.

Contrasting (iii) with (i) and (ii), we observe that the conditions for consistency of $\hat{\psi}$ are met if and only if either the conditions for consistency of $\tilde{\psi}_{\text{retro}}$ or those for consistency of $\tilde{\psi}_{\text{aipw}}$ are met, but not necessarily both. We thus conclude that $\hat{\psi}$ confers more protection to model misspecification than each individual estimator $\tilde{\psi}_{\text{retro}}$ and $\tilde{\psi}_{\text{aipw}}$. This in turn implies that $\hat{\psi}$ confers more protection to model misspecification than the prospective parametric likelihood-based estimator and the estimator $\tilde{\psi}_{\text{ipw}}$ discussed in § 2, as consistency of the former requires even more stringent conditions than those for consistency of $\tilde{\psi}_{\text{retro}}$ and consistency of the latter requires even more stringent conditions than those for consistency of $\tilde{\psi}_{\text{aipw}}$.

4. SIMULATION STUDY

We compared $\hat{\psi}$ with $\tilde{\psi}_{\text{retro}}$ and $\tilde{\psi}_{\text{aipw}}$ in a simulation study in which A is sometimes missing but $L_{\text{obs}} = L$ is always observed. We simulated 500 random samples of (Y, RA, R, L)

Table 1. *Simulation results*

Line	Model		Estimator	
	Correct		$\tilde{\psi}_{\text{retro}}$	$\tilde{\psi}_{\text{aipw}}$
				$\hat{\psi}$
1	(a), (b), (c)	Bias	-0.001	0.0003
		Variance	0.0285	0.029
2	(a), (b)	Bias	-0.001	0.00005
		Variance	0.028	0.031
3	(a), (c)	Bias	-0.001	0.242
		Variance	0.028	0.029
4	(b), (c)	Bias	0.184	-0.043
		Variance	0.024	0.028
5	(a)	Bias	-0.001	-0.240
		Variance	0.028	0.028
6	(b)	Bias	0.184	-0.028
		Variance	0.024	0.033
7	(c)	Bias	0.184	0.210
		Variance	0.024	0.031
8	—	Bias	0.184	0.215
		Variance	0.024	0.026

(a) model for odds $A(L)$; (b) model for odds $Y(L)$; (c) model for $\text{pr}(R = 1 \mid Y, L_{\text{obs}})$.

of size 2000 according to $L_1 \sim N(0, 0.9^2)$, $L_2 \mid L_1 \sim N(L_1^2/2, 0.9^2)$, $f(Y, A \mid L) \propto \exp(-0.5Y + 0.5YL_1 + 0.4YL_2 + 0.7YL_1L_2 + AY + 0.8A - 0.4AL_1 + 0.5AL_2 + 0.6AL_1L_2)$ and $R \mid (Y, A, L) \sim \text{Ber}[\{1 + \exp(0.3L_1 + 0.3L_2Y + 0.4YL_1L_2)\}^{-1}]$. Table 1 reports the results for the estimators $\hat{\psi}$, $\tilde{\psi}_{\text{retro}}$ and $\tilde{\psi}_{\text{aipw}}$ in model $\text{OR}(L; \psi) = \psi$ computed under all eight combinations of correct or incorrect models for odds $A(L)$, odds $Y(L)$ and $\text{pr}(R = 1 \mid L, Y)$, labelled (a), (b) and (c), respectively. Correct models used odds $A(L; \alpha) = \alpha_0 + \alpha_1L_1 + \alpha_2L_2 + \alpha_3L_1L_2$, odds $Y(L; \eta) = \eta_0 + \eta_1L_1 + \eta_2L_2 + \eta_3L_1L_2$ and $\text{pr}(R = 1 \mid L, Y; \lambda) = \{1 + \exp(-\lambda_0 - \lambda_1L_1 - \lambda_2L_2 - \lambda_3YL_2 - \lambda_4YL_1L_2)\}^{-1}$. Incorrect models ignored the terms in L_1L_2 . According to theory, $\hat{\psi}$ is consistent in lines 1–5, $\tilde{\psi}_{\text{retro}}$ is consistent in lines 1, 2, 3 and 5 and $\tilde{\psi}_{\text{aipw}}$ is consistent in lines 1, 2 and 4. The Monte Carlo bias of each estimator reported in the corresponding lines is small, but the biases of $\tilde{\psi}_{\text{retro}}$ in line 4 and that of $\tilde{\psi}_{\text{aipw}}$ in lines 3 and 5 are substantial, thus illustrating the robustness advantage of $\hat{\psi}$. According to theory, none of the three estimators is consistent in lines 6–8, a fact reflected in the substantial Monte Carlo bias observed for all three estimators in these lines, with the exception of the bias of estimators $\tilde{\psi}_{\text{aipw}}$ and $\hat{\psi}$ in line 6. This exception is not supported by theory and is thus likely particular to our data generating process.

As indicated in § 2, according to theory, when odds $A(L)$ is correctly modelled, as in lines 1, 2, 3 and 5, $\tilde{\psi}_{\text{retro}}$ is indeed asymptotically efficient, in spite of being based only on complete-case units. A comparison of the Monte Carlo standard errors of $\tilde{\psi}_{\text{retro}}$ and $\tilde{\psi}_{\text{aipw}}$ in lines 1 and 2, corresponding to cases in which both are consistent, confirms this result. A comparison of the Monte Carlo standard errors of $\tilde{\psi}_{\text{retro}}$ and $\hat{\psi}$ in lines 1, 2, 3 and 5, the cases in which both are consistent, further indicates that the new complete-case estimator $\hat{\psi}$ does not incur in any sizeable efficiency loss. Furthermore, a comparison of the standard errors of $\hat{\psi}$ and $\tilde{\psi}_{\text{aipw}}$ in line 4, a situation in which both are consistent, additionally illustrates the fact that even though $\tilde{\psi}_{\text{aipw}}$ uses data from incomplete-case units it can be less efficient than $\hat{\psi}$ when the model for odds $A(L)$ is incorrectly specified.

5. EXTENSIONS TO NONBINARY OUTCOMES AND/OR TREATMENTS

For the situations in which A and/or Y are not binary, a number of recent articles have focused attention on the estimation of the parameter ψ indexing an odds ratio model $\text{OR}(Y, A, L; \psi)$ for $\text{OR}(Y, A, L) = \log[\{f(Y \mid A, L)f(y_0 \mid a_0, L)\}/\{f(y_0 \mid A, L)f(Y \mid a_0, L)\}]$ where (y_0, a_0) is a user-specified point in the sample space (Chen, 2007; Osius, 2009; Tchetgen Tchetgen et al., 2010; Tchetgen Tchetgen, 2010). As

in the binary case, $OR(Y, A, L)$ is an attractive target of inference as it is invariant to alterations in the marginal laws of (Y, L) or of (A, L) . Tchetgen Tchetgen et al. (2010) describe estimators of ψ that are consistent and asymptotically normal provided one of the following two models is correctly specified, but not necessarily both.

Model 1. A model odds $A(A, L; \alpha)$ for odds $A(A, L) = \log\{f(A|y_0, L)/f(a_0|y_0, L)\}$.

Model 2. A model odds $Y(Y, L; \eta)$ for odds $Y(Y, L) = \log\{f(Y|a_0, L)/f(y_0|a_0, L)\}$.

Just as in the binary case, under the missing data conditions of this note, the estimators of Tchetgen Tchetgen et al. (2010) restricted to the complete-case units are consistent and asymptotically normal for ψ so long as odds $A(A, L; \alpha)$ is a correctly specified model for odds $A(A, L)$ or odds $Y(A, L; \eta)$ is now a correctly specified model for odds $Y_{cc}(Y, L) = \log\{f(Y|a_0, L, R=1)/f(y_0|a_0, L, R=1)\}$, but not necessarily both.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes an iterative algorithm.

ACKNOWLEDGEMENT

The authors were funded by grants from the U.S. National Institutes of Health. The authors wish to thank the editor and the reviewers for helpful comments. Andrea Rotnitzky is also affiliated with the Harvard School of Public Health.

REFERENCES

- CHEN, Y. H. (2007). A semi-parametric odds ratio model for measuring association. *Biometrics* **63**, 413–21.
- LIPSITZ, S., PARZEN, M. & EWELL, M. (1998). Inference using conditional logistic regression with missing covariates. *Biometrics* **54**, 295–303.
- LITTLE, R. & RUBIN, D. (2002). *Statistical Analysis with Missing Data*, 2nd ed. New York: John Wiley.
- MOORE, C. G., LIPSITZ, S. R., ADDY, C. L., HUSSEY, J. R., FITZMAURICE, G. & NATARAJAN, S. (2009). Logistic Regression with incomplete covariate data in complex survey sampling: application of re-weighted estimating equations. *Epidemiology* **20**, 382–90.
- OSIUS, G. (2009). Asymptotic inference for semiparametric association models. *Ann. Statist.* **37**, 459–89.
- PARZEN, M., LIPSITZ, S., IBRAHIM, J. & LIPSHULTZ, S. (2002). A weighted estimating equation for linear regression with missing covariate data. *Statist. Med.* **21**, 2421–36.
- ROBINS J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.
- TCHETGEN TCHETGEN, E. (2009). A simple implementation of doubly robust estimation in logistic regression with covariates missing at random. *Epidemiology* **20**, 391–4.
- TCHETGEN TCHETGEN, E., ROBINS, J. & ROTNITZKY, A. (2010). On doubly robust estimation in a semi-parametric odds ratio model. *Biometrika* **97**, 171–80.
- TCHETGEN TCHETGEN, E. & ROTNITZKY, A. (2011). Double-robust adjustment for confounding in cohort and case-control studies. *Statist. Med.* **30**, 335–47.
- TCHETGEN TCHETGEN, E. (2010). On the interpretation, robustness, and power of varieties of case-only tests of gene-environment interaction. *Am. J. Epidemiol.* **172**, 1335–8.
- ZHAO, L. P., LIPSITZ, S. & LEW, D. (1996). Regression analysis with missing covariate data using estimating equations. *Biometrics* **52**, 1165–82.

[Received July 2010. Revised March 2011]