Biometrical Journal **44** (2002) 2, 227–239

Estimating Incidence Rates from Population-Based Case-Control Studies in the Presence of Nonrespondents

PATRICK G. ARBOGAST[1], D. Y. LIN[2], DAVID S. SISCOVICK[3,4,5], and STEPHEN M. SCHWARTZ[3,4]

[1] Division of Biostatistics, Department of Preventive Medicine, Vanderbilt University, Nashville, U.S.A.
[2] Department of Biostatistics, University of North Carolina, Chapel Hill, U.S.A.
[3] Cardiovascular Health Research Unit, Seattle, U.S.A.
[4] Department of Epidemiology, University of Washington, Seattle, U.S.A.
[5] Department of Medicine, University of Washington, Seattle, U.S.A.

Summary

In population-based case-control studies, it is of great public-health importance to estimate the disease incidence rates associated with different levels of risk factors. This estimation is complicated by the fact that in such studies the selection probabilities for the cases and controls are unequal. A further complication arises when the subjects who are selected into the study do not participate (i.e. become nonrespondents) and nonrespondents differ systematically from respondents. In this paper, we show how to account for unequal selection probabilities as well as differential nonresponses in the incidence estimation. We use two logistic models, one relating the disease incidence rate to the risk factors, and one modelling the predictors that affect the nonresponse probability. After estimating the regression parameters in the nonresponse model, we estimate the regression parameters in the disease incidence model by a weighted estimating function that weights a respondent's contribution to the likelihood score function by the inverse of the product of his/her selection probability and his/her model-predicted response probability. The resulting estimators of the regression parameters and the corresponding estimators of the incidence rates are shown to be consistent and asymptotically normal with easily estimated variances. Simulation results demonstrate that the asymptotic approximations are adequate for practical use and that failure to adjust for nonresponses could result in severe biases. An illustration with data from a cardiovascular study that motivated this work is presented.

Key words: Absolute risk; Disease incidence; Horvitz-Thompson estimator; Logistic regression; Missing data; Nonresponses.

1. Introduction

Case-control studies are widely used to investigate the association of putative risk factors with the incidence of rare diseases, such as cancer, stroke and myocardial infarction (BRESLOW and DAY, 1980, ch. 1). In a population-based case-control study, separate samples are taken of cases (i.e. diseased individuals) and controls (i.e. disease-free individuals) from a defined population, and the information on the exposure of interest and other risk factors is collected. The statistical analysis

 0323-3847/02/0203-0227 $ 17.50+.50/0

of these data is commonly based on the logistic model, which relates the probability of developing the disease to the exposure of interest and other risk factors. If one ignores the unequal selection probabilities of the case-control design and proceeds as if the observations came from a random sample of the entire population, the standard maximum likelihood method will provide valid inferences for the slope parameters (i.e. log odds ratios), but not for the intercept term (PRENTICE and PYKE, 1979).

The slope parameters of the logistic model estimated by the standard maximum likelihood method are useful measures of association between risk factors and disease incidence. It is also relevant in public health to estimate the absolute risk and the risk difference (i.e. the excess risk of disease due to exposure). As mentioned previously, the standard maximum likelihood method does not yield a valid estimator of the intercept term in the logistic model for case-control data. Therefore, this method cannot be used to estimate the absolute risk of disease or risk difference.

The difficulty in estimating the intercept term from the case-control data is caused by the fact that cases and controls are selected with unequal probabilities and consequently the study sample is not a random sample from the whole population. To account for the unequal selection probabilities between cases and controls, SCOTT and WILD (1986) suggested weighting an individual's contribution to the likelihood score function by the inverse of his/her selection probability. They showed that the resulting estimators of the intercept and slopes are consistent and asymptotically normal. Furthermore, maximum likelihood estimators of the intercept and slopes can be obtained by including a fixed offset equal to the logarithm of the ratio of the sampling fractions.

In this paper, we provide an extension of the Scott-Wild estimating function to adjust for nonresponses. In case-control studies that seek to obtain exposure information directly (e.g. using interviews) from the subjects who are eligible and selected into the study, some subjects may die prior to recruitment while others may refuse to participate. These subjects are called nonrespondents. In many studies, nonrespondents differ systematically from respondents. Consequently, failure to adjust for nonresponses could result in biased estimation of the regression parameters and incidence rates.

To remove the biases caused by nonresponses, we propose to measure the key predictors affecting nonresponses from both the respondents and nonrespondents. For example, in a study that conducts in-person interviews with the respondents, the information about the nonrespondents may be obtained via telephone interviews, mail questionnaires, or medical records. Given such data, we relate the probability of response to the measured predictors through a logistic model. We then incorporate the probabilities of response estimated from this model into the Scott-Wild type weighted estimating function for the disease incidence model. This two-step estimating procedure leads to consistent and asymptotically normal estimators of all the regression parameters (including the intercept and slopes) in

the disease incidence model. These results can then be used to estimate the incidence rates associated with various levels of exposure and other risk factors.

The proposed methodology is described in the next section. In Section 3, we report the results of our simulation studies. In Section 4, we provide an illustration with data taken from the Women's Cardiovascular Health Study (WCHS), which is a stratified population-based case-control study investigating the relationship between oral contraceptive use and incidence of stroke and myocardial infarction (SCHWARTZ et al., 1997). There were a considerable number of nonrespondents in the WCHS Study, and the information on the nonrespondents was collected through telephone interviews.

2. Methods

Let Y be the disease indicator, taking the value 1 for the case and 0 for the control. Also, let $\boldsymbol{X} = (1, X_1, \ldots, X_{p-1})'$ be a $p \times 1$ vector of covariates. The logistic model specifies that

$$\Pr(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}'\boldsymbol{x}}}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$ is a $p \times 1$ vector of unknown regression parameters. Write $p_1(\boldsymbol{x}; \boldsymbol{\beta}) = \Pr(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\beta})$ and $p_0(\boldsymbol{x}; \boldsymbol{\beta}) = 1 - p_1(\boldsymbol{x}; \boldsymbol{\beta})$. We wish to estimate $\boldsymbol{\beta}$ as well as $p_1(\boldsymbol{x}; \boldsymbol{\beta})$ for various values of $\boldsymbol{x}$.

Suppose that stratified case-control sampling is taken from a finite source population of N subjects. The source population is regarded as a random sample of size N from an infinite population. Let N_{lj} be the total number of subjects in the source population who belong to the lth $(l = 0, 1)$ disease category and jth $(j = 1, \ldots, J)$ stratum. The information about the N_{lj}'s is often available from disease registries and official population statistics. For the lth disease category and jth stratum, n_{lj} subjects are drawn from the N_{lj} subjects in the subpopulation by simple random sampling without replacement.

Let $\boldsymbol{x}_{ljk}$ be the value of the covariate vector $\boldsymbol{X}$ for the kth subject in the lth disease category and jth stratum. For notational convenience, the subjects in the lth disease category and jth stratum are ordered such that the first n_{lj} subjects in the subpopulation correspond to those selected into the sample. If the sampled subjects were drawn randomly from the whole population, then the likelihood score function for $\boldsymbol{\beta}$ would be

$$\sum_{l=0}^{1} \sum_{j=1}^{J} \sum_{k=1}^{n_{lj}} (-1)^{l+1} \{1 - p_l(\boldsymbol{x}_{ljk}; \boldsymbol{\beta})\} \boldsymbol{x}_{ljk}. \tag{2}$$

In stratified case-control studies, subjects are sampled conditionally on their disease status and on their values of the stratification variables so that the resulting

sample is not a random sample from the whole population. Consequently, the use of estimating function (2) would not yield a consistent estimator of the intercept term β_0, although the estimators of the slope parameters $(\beta_1, \ldots, \beta_{p-1})'$ based on (2) are valid if one modifies model (1) to allow stratum-specific intercepts (PRENTICE and PYKE, 1979).

To obtain a consistent estimator of β_0, one needs to account for the unequal selection probabilities of the (stratified) case-control sampling scheme. The selection probability for the subjects in the lth disease category and jth stratum is $\pi_{lj} = n_{lj}/N_{lj}$. Using the HORVITZ-THOMPSON (1952) approach, we weight the elements in (2) by their inversed selection probabilities to yield the following estimating function for $\boldsymbol{\beta}$:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{l=0}^{1} \sum_{j=1}^{J} \pi_{lj}^{-1} \sum_{k=1}^{n_{lj}} (-1)^{l+1} \{1 - p_l(\boldsymbol{x}_{ljk}; \boldsymbol{\beta})\} \boldsymbol{x}_{ljk} . \tag{3}$$

For unstratified case-control studies, $\boldsymbol{U}(\boldsymbol{\beta}) = \mathbf{0}$ reduces to equation (4) of SCOTT and WILD (1986).

As mentioned in Section 1, there often exist nonrespondents among the selected subjects. If nonresponses do not occur in a purely random fashion, then the respondents will not be representative of all selected subjects. Consequently, failure to account for nonresponses would result in invalid inferences. We may remove the biases due to nonresponses by using the Horvitz-Thompson idea again if the probabilities of response can be determined.

Let $\boldsymbol{Z}$ be a set of predictors for nonresponses. Suppose that the measurements of $\boldsymbol{Z}$ are available on all or a random subset of the subjects who are selected into the study. These measurements may be obtained through interviews, mail questionnaires, or medical records, as mentioned in Section 1. Let ξ indicate, by the values 1 versus 0, on whether the subject is a respondent or nonrespondent. Suppose that ξ and $\boldsymbol{Z}$ are related through the logistic model

$$\Pr(\xi = 1 \mid \boldsymbol{Z} = z; \boldsymbol{\gamma}) = \frac{e^{\gamma' z}}{1 + e^{\gamma' z}} , \tag{4}$$

where $\boldsymbol{\gamma}$ is a set of unknown regression parameters. It is assumed that ξ is independent of $\boldsymbol{X}$ and Y given $\boldsymbol{Z}$. Let $q(z; \boldsymbol{\gamma}) = \Pr(\xi = 1 \mid \boldsymbol{Z} = z; \boldsymbol{\gamma})$. We can estimate $\boldsymbol{\gamma}$ by the standard maximum likelihood method. The likelihood score function for $\boldsymbol{\gamma}$ is

$$\boldsymbol{S}(\boldsymbol{\gamma}) = \sum_{l=0}^{1} \sum_{j=1}^{J} \sum_{k=1}^{n_{lj}} \psi_{ljk} \{\xi_{ljk} - q(z_{ljk}; \boldsymbol{\gamma})\} z_{ljk} , \tag{5}$$

where ψ_{ljk} indicates, by the values 1 versus 0, on whether or not z_{ljk} is measured. Let $\hat{\boldsymbol{\gamma}}$ be the solution to the equation $\{\boldsymbol{S}(\boldsymbol{\gamma}) = \mathbf{0}\}$. It follows from the standard likelihood theory that $\hat{\boldsymbol{\gamma}}$ is consistent and asymptotically normal with covariance matrix estimator $\hat{\boldsymbol{\Omega}}^{-1}(\hat{\boldsymbol{\gamma}})$, where $\hat{\boldsymbol{\Omega}}(\boldsymbol{\gamma}) = -\partial \boldsymbol{S}(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}'$. Given $\hat{\boldsymbol{\gamma}}$, the probability of response conditional on $\boldsymbol{Z} = z$ is estimated by $q(z; \hat{\boldsymbol{\gamma}})$.

By incorporating $q(z;\hat{\boldsymbol{\gamma}})$ into (3), we obtain the following estimating function for $\boldsymbol{\beta}$:

$$\boldsymbol{U}(\boldsymbol{\beta};\hat{\boldsymbol{\gamma}}) = \sum_{l=0}^{1}\sum_{j=1}^{J}\sum_{k=1}^{n_{lj}} \frac{(-1)^{l+1}\,\xi_{ljk}\psi_{ljk}\{1-p_l(\boldsymbol{x}_{ljk};\,\boldsymbol{\beta})\}\,\boldsymbol{x}_{ljk}}{\pi_{lj}q(\boldsymbol{z}_{ljk};\,\hat{\boldsymbol{\gamma}})}\,. \tag{6}$$

Note that $\boldsymbol{U}(\boldsymbol{\beta};\,\hat{\boldsymbol{\gamma}})$ includes only those subjects who are respondents and whose predictors of nonresponses are measured and that the contributions from those subjects are weighted inversely by their selection probabilities and their estimated response probabilities based on model (4).

Given $\hat{\boldsymbol{\gamma}}$, we use the Newton-Raphson algorithm to solve the equation $\{\boldsymbol{U}(\boldsymbol{\beta};\hat{\boldsymbol{\gamma}}) = \boldsymbol{0}\}$. Denote the resulting estimator by $\hat{\boldsymbol{\beta}}$. Let $\hat{\boldsymbol{A}}(\boldsymbol{\beta};\,\boldsymbol{\gamma}) = -\partial\boldsymbol{U}(\boldsymbol{\beta};\,\boldsymbol{\gamma})/\partial\boldsymbol{\beta}'$, $\hat{\boldsymbol{H}}(\boldsymbol{\beta};\,\boldsymbol{\gamma}) = -\partial\boldsymbol{U}(\boldsymbol{\beta};\,\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}'$, and

$$\hat{\boldsymbol{C}}(\boldsymbol{\beta};\,\boldsymbol{\gamma}) = \sum_{l=0}^{1}\sum_{j=1}^{J}\pi_{lj}^{-2}\sum_{k=1}^{n_{lj}}\left[\frac{\xi_{ljk}\psi_{ljk}\{1-p_l(\boldsymbol{x}_{ljk};\,\boldsymbol{\beta})\}^2\,\boldsymbol{x}_{ljk}^{\otimes 2}}{q^2(\boldsymbol{z}_{ljk};\,\boldsymbol{\gamma})}\right] - \sum_{l=0}^{1}\sum_{j=1}^{J}\frac{1-\pi_{lj}}{n_{lj}\pi_{lj}^2}\left[\sum_{k=1}^{n_{lj}}\frac{\xi_{ljk}\psi_{ljk}\{1-p_l(\boldsymbol{x}_{ljk};\,\boldsymbol{\beta})\}\,\boldsymbol{x}_{ljk}}{q(\boldsymbol{z}_{ljk};\,\boldsymbol{\gamma})}\right]^{\otimes 2},$$

where $a^{\otimes 2} = aa'$. We show in the Appendix that $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal with covariance matrix estimator $\hat{\boldsymbol{V}} \equiv \hat{\boldsymbol{A}}^{-1}(\hat{\boldsymbol{\beta}};\hat{\boldsymbol{\gamma}})\,\hat{\boldsymbol{B}}(\hat{\boldsymbol{\beta}};\hat{\boldsymbol{\gamma}})\,\hat{\boldsymbol{A}}^{-1}(\hat{\boldsymbol{\beta}};\hat{\boldsymbol{\gamma}})$, where $\hat{\boldsymbol{B}}(\boldsymbol{\beta};\,\boldsymbol{\gamma}) = \hat{\boldsymbol{C}}(\boldsymbol{\beta};\,\boldsymbol{\gamma}) - \hat{\boldsymbol{H}}(\boldsymbol{\beta};\,\boldsymbol{\gamma})\,\hat{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\gamma})\,\hat{\boldsymbol{H}}(\boldsymbol{\beta};\,\boldsymbol{\gamma})'$.

In this paper, we assume that simple random sampling without replacement is used to select subjects from each stratum so that the stratum sample sizes are fixed. An alternative sampling scheme is to select subjects by independent Bernoulli processes so that the stratum sample sizes are random. Under the latter sampling scheme, the covariance matrix for $\hat{\boldsymbol{\beta}}$ remains the same except that the second term in $\hat{\boldsymbol{C}}$ vanishes. Thus, simple random sampling reduces the variability of the estimator as compared to independent Bernoulli sampling. Another interesting phenomenon is that the second term in $\hat{\boldsymbol{B}}$ will vanish if $q(\boldsymbol{z}_{ljk};\,\hat{\boldsymbol{\gamma}})$ in (6) is replaced by $q(\boldsymbol{z}_{ljk};\,\boldsymbol{\gamma}_0)$, which means that it would be more efficient to estimate the response probabilities from the data even if they were known. Though this improved efficiency seems counterintuitive, this phenomenon has been studied before (e.g. ROBINS et al., 1994; WANG et al., 1997).

If the study is unstratified and there are no nonrespondents, then our parameter estimator $\hat{\boldsymbol{\beta}}$ reduces to that of SCOTT and WILD (1986) while our covariance matrix estimator differs slightly from theirs in that the factor $1-\pi_{lj}$ in the second term of $\hat{\boldsymbol{C}}$ is 1 in Scott and Wild's expression. The reason for this minor discrepancy is that Scott and Wild assumed an infinite source population, i.e. $N \to \infty$, whereas we assume a finite source population. In many population-based case-control studies, the majority of the cases in the population are selected into the study so that the factor $1-\pi_{1j}$ will be closer to 0 than to 1. Thus, the covariance matrix estimator given here is more accurate.

Once $\hat{\boldsymbol{\beta}}$ is obtained, we estimate $p_1(\boldsymbol{x}; \boldsymbol{\beta})$ by

$$p_1(\boldsymbol{x}; \hat{\boldsymbol{\beta}}) = \frac{e^{\hat{\boldsymbol{\beta}}'\boldsymbol{x}}}{1 + e^{\hat{\boldsymbol{\beta}}'\boldsymbol{x}}} .$$

For brevity, write $p_{\boldsymbol{x}} = p_1(\boldsymbol{x}; \boldsymbol{\beta})$ and $\hat{p}_{\boldsymbol{x}} = p_1(\boldsymbol{x}; \hat{\boldsymbol{\beta}})$. By the δ-method, $\hat{p}_{\boldsymbol{x}}$ is asymptotically normal with mean $p_{\boldsymbol{x}}$ and with variance estimator $\hat{p}_{\boldsymbol{x}}^2(1 - \hat{p}_{\boldsymbol{x}})^2\, \boldsymbol{x}'\hat{\boldsymbol{V}}\boldsymbol{x}$. On the other hand, the difference between the incidence rates $p_{\boldsymbol{x}_{(1)}}$ and $p_{\boldsymbol{x}_{(2)}}$ associated with $\boldsymbol{x}_{(1)}$ and $\boldsymbol{x}_{(2)}$ is estimated by $\hat{p}_{\boldsymbol{x}_{(1)}} - \hat{p}_{\boldsymbol{x}_{(2)}}$. Again by the δ-method, $\hat{p}_{\boldsymbol{x}_{(1)}} - \hat{p}_{\boldsymbol{x}_{(2)}}$ is asymptotically normal with mean $p_{\boldsymbol{x}_{(1)}} - p_{\boldsymbol{x}_{(2)}}$ and variance estimator

$$\{\hat{p}_{\boldsymbol{x}_{(1)}}(1 - \hat{p}_{\boldsymbol{x}_{(1)}})\, \boldsymbol{x}_{(1)} - \hat{p}_{\boldsymbol{x}_{(2)}}(1 - \hat{p}_{\boldsymbol{x}_{(2)}})\, \boldsymbol{x}_{(2)}\}'$$
$$\times \hat{\boldsymbol{V}}\{\hat{p}_{\boldsymbol{x}_{(1)}}(1 - \hat{p}_{\boldsymbol{x}_{(1)}})\, \boldsymbol{x}_{(1)} - \hat{p}_{\boldsymbol{x}_{(2)}}(1 - \hat{p}_{\boldsymbol{x}_{(2)}})\, \boldsymbol{x}_{(2)}\} .$$

We suggest to construct the confidence intervals for incidence rates based on the log transformation, which not only ensures that the lower confidence limits will be positive but also improves the small-sample coverage of the confidence intervals. Based on the log transformation, the 95% confidence interval for $p_{\boldsymbol{x}}$ is $\hat{p}_{\boldsymbol{x}}\, e^{\pm 1.96(1-\hat{p}_{\boldsymbol{x}})\,(\boldsymbol{x}'\hat{\boldsymbol{V}}\boldsymbol{x})^{1/2}}$.

3. Simulation Studies

We conducted a series of simulation studies to evaluate the performance of the methods described in the previous section. We generated disease incidence from the logistic model

$$\text{logit}\,\{\Pr(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\beta})\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 , \tag{7}$$

where $\beta_0 = -7.9$, $\beta_1 = 0.5$, $\beta_2 = 1.0$, X_1 is standard normal, and X_2 is Bernoulli with 0.2 success probability if $X_1 < 0$ and with 0.5 success probability if $X_1 \geq 0$. The dependence of the success probability of X_2 on X_1 creates a confounding effect of X_1 on X_2. Under this model, approximately 0.1% of the subjects are cases (i.e. diseased). We used two strata defined by $X_1 < 0.5$ versus $X_1 \geq 0.5$. We let $N = 1{,}000{,}000$ and $n_{0j} = n_{1j} = 50$, 100, 200 or 300, and drew cases and controls randomly from their respective strata. This sampling scheme mimics that of the WCHS Study (SCHWARTZ et al., 1997) in that the exposure of interest (i.e. oral contraceptive use) is dichotomous while the major confounder (i.e. age) is continuous and the study is stratified on the discrete version of the major confounder (i.e. age group). We generated nonrespondents through the model

$$\text{logit}\,\{\Pr(\xi = 1 \mid \boldsymbol{Z} = \boldsymbol{z}; \boldsymbol{\gamma})\} = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_1 z_2 ,$$

where $Z_1 = Y$, $Z_2 = X_2$, $\gamma_0 = 0.75$, $\gamma_1 = \gamma_2 = 0$ and $\gamma_3 = 0.75$. Under this model, the response rate for controls is approximately 70% and does not depend on X_2, whereas the response rate for cases is roughly 85% when $X_2 = 1$ and 70% when $X_2 = 0$.

Table 1 summarizes the simulation results for the estimation of the regression parameters $(\beta_0, \beta_1, \beta_2)'$ and incidence rates $p_1^{(0)}$ and $p_1^{(1)}$, the latter being $p_1(\boldsymbol{x}; \boldsymbol{\beta})$

Table 1
Summary statistics for the simulation studies

		Not Adjusting for Nonresponses				Adjusting for Nonresponses			
n_{ij}	Para.	Mean of Est.	SE of Est.	Mean of SEE	Cov. Prob.	Mean of Est.	SE of Est.	Mean of SEE	Cov. Prob.
50	β_0	−7.92	0.22	0.23	0.96	−7.91	0.17	0.17	0.96
	β_1	0.51	0.19	0.18	0.95	0.51	0.18	0.17	0.96
	β_2	1.22	0.38	0.38	0.92	1.02	0.32	0.32	0.96
	$p_1^{(0)}$	3.72	0.82	0.83	0.96	3.71	0.61	0.62	0.96
	$p_1^{(1)}$	21.06	4.85	4.55	0.86	17.21	3.07	2.93	0.95
100	β_0	−7.91	0.16	0.16	0.95	−7.91	0.12	0.12	0.95
	β_1	0.51	0.13	0.13	0.95	0.51	0.12	0.12	0.95
	β_2	1.20	0.26	0.26	0.90	1.01	0.22	0.22	0.96
	$p_1^{(0)}$	3.72	0.60	0.59	0.95	3.71	0.46	0.45	0.95
	$p_1^{(1)}$	20.47	3.21	3.10	0.76	16.90	2.08	2.04	0.95
200	β_0	−7.90	0.12	0.12	0.95	−7.91	0.09	0.09	0.95
	β_1	0.51	0.09	0.09	0.95	0.50	0.09	0.09	0.95
	β_2	1.19	0.19	0.18	0.82	1.01	0.16	0.16	0.95
	$p_1^{(0)}$	3.71	0.43	0.43	0.95	3.70	0.34	0.33	0.95
	$p_1^{(1)}$	20.24	2.23	2.22	0.59	16.76	1.49	1.50	0.95
300	β_0	−7.90	0.10	0.10	0.95	−7.90	0.07	0.08	0.96
	β_1	0.50	0.08	0.08	0.95	0.50	0.07	0.07	0.95
	β_2	1.19	0.15	0.15	0.75	1.01	0.13	0.13	0.94
	$p_1^{(0)}$	3.73	0.37	0.36	0.95	3.72	0.28	0.29	0.96
	$p_1^{(1)}$	20.25	1.86	1.87	0.44	16.77	1.26	1.28	0.96

Note: SE and SEE stand for standard error and standard error estimate, and Cov. Prob. stands for the coverage probability of the 95% confidence interval. Estimates and standard errors for $p_1^{(0)}$ and $p_1^{(1)}$ are per 10,000 people.

evaluated at $x_1 = x_2 = 0$ and $x_1 = x_2 = 1$, respectively. Under model (7), we have $p_1^{(0)} = 3.706$ and $p_1^{(1)} = 16.59$ per 10,000 people. Each entry in Table 1 is based on 1000 simulation samples. With adjustment for nonresponses, the estimators of the regression parameters and incidence rates as well as their standard error estimators are virtually unbiased, and the corresponding confidence intervals have proper coverage probabilities. By contrast, without adjustment of nonresponses, the estimators are biased and the confidence intervals have poor coverages, especially with respect to β_2 and $p_1^{(1)}$.

4. Application to WCHS

This work was motivated by the previously mentioned WCHS Study (SCHWARTZ et al., 1997), which is a population-based case-control study investigating the relationship between oral contraceptive (OC) use and incidence of stroke and myocar-

dial infarction. For our application, only stroke incidence was considered. The study was conducted by the Cardiovascular Health Research Unit at the University of Washington in Seattle, Washington, USA, and sponsored by the National Institute of Child Health and Human Development. Eligible cases and controls were selected from women 18–44 years old residing in the King, Pierce and Snohomish counties in the State of Washington between July 1, 1991 and February 28, 1995.

The stroke cases were women diagnosed with a first fatal or non-fatal stroke and not having a prior history of major cardiovascular diseases. Controls were sampled using random digit telephone dialing and also excluded women with a past history of major cardiovascular diseases. The study sought to recruit every eligible available case while controls were randomly sampled from 5 age groups. The selection probabilities for the controls were 4.33×10^{-5}, 7.88×10^{-5}, 12.45×10^{-5}, 34.88×10^{-5} and 58.35×10^{-5} for age groups 18–24, 25–29, 30–34, 35–39 and 40–44, respectively. Participation involved an extensive in-person interview eliciting history of oral contraceptive use, reproductive history, demographic information and cardiovascular risk factors.

The study involved 891 subjects: 242 cases and 649 controls. The nonresponse rates were 29% for the cases and 25% for the controls. A telephone questionnaire was administered on the nonrespondents. The telephone questionnaire included two dozen questions from the in-person interview questionnaire which were considered the key predictors for nonresponses.

We first used model (4) to relate the probability of response to the following predictors: case-control status, age, OC use, and the interactions of case-control status with age and OC use. These predictors were chosen because they had the most appreciable effects on the nonresponse probability. The parameter estimates for case-

Table 2

Estimation of regression parameters in the incidence model for the WCHS study, adjusting and not adjusting for nonresponses

	Unadjusted		Adjusted	
Parameter	Estimate	Stand. Error	Estimate	Stand. Error
Intercept	−9.09	0.26	−8.93	0.28
Age	0.12	0.02	0.12	0.02
Current OC use	−0.18	0.31	−0.31	0.37
Former OC use	−0.70	0.26	−0.84	0.29
Current smoking	1.30	0.20	1.31	0.26
Former smoking	0.29	0.24	0.26	0.30
Treated hypertension	1.71	0.24	1.71	0.42
Treated diabetes	0.99	0.47	0.84	0.82
BMI	0.02	0.02	0.02	0.02
African American	0.94	0.29	0.99	0.47
Other non-white	0.33	0.28	0.31	0.34

Table 3
Estimation of incidence rates for the WCHS study, adjusting and not adjusting for nonresponses

Risk Factors	Unadjusted			Adjusted		
	Est.	SE	95% CI	Est.	SE	95% CI
Baseline	11.3	2.9	(6.8, 18.8)	13.2	3.7	(7.6, 22.9)
25 years old	2.1	0.7	(1.1, 4.0)	2.4	0.9	(1.1, 5.1)
44 years old	20.7	5.7	(12.1, 35.6)	24.3	7.4	(13.4, 44.1)
Current OC user	9.5	2.3	(5.9, 15.2)	9.7	2.9	(5.4, 17.4)
Past OC user	5.6	1.0	(4.0, 8.0)	5.7	1.1	(3.8, 8.4)
Current smoker	41.4	11.8	(23.7, 72.4)	49.0	16.1	(25.8, 93.4)
Past smoker	15.2	3.9	(9.3, 25.0)	17.1	5.4	(9.1, 31.9)
Rx for hypertension	62.7	19.4	(34.2, 114.9)	72.9	34.5	(28.8, 184.1)
Rx for diabetes	30.6	12.3	(13.9, 67.2)	30.5	24.1	(6.5, 143.7)
BMI = 15	9.0	2.9	(4.8, 17.0)	10.5	3.8	(5.1, 21.5)
BMI = 53	21.3	9.9	(8.5, 53.2)	24.9	14.8	(7.8, 80.1)
African American	28.9	11.7	(13.1, 63.8)	35.5	19.5	(12.1, 104.0)
Other non-white	15.8	4.6	(8.9, 27.8)	18.0	6.0	(9.3, 34.6)

Note: The estimates and confidence intervals are per 100,000 people. Baseline refers to a 39 year old white female who never used OC's, never smoked, was never treated for hypertension or diabetes, and with a BMI value of 25. For all other entries in the table, the risk factors take the same values as the baseline except for the specified value of the particular risk factor.

control status, age, current OC use and former OC use were 2.70, 0.05, 0.70 and 0.37, respectively, and the parameter estimates for the interactions of case-control status with age, current and former OC use were −0.08, 0.29 and 0.78, respectively.

We then used model (1) to relate stroke incidence to OC use, age, smoking status, treatment for hypertension, treatment for diabetes, body mass index (BMI) and race. The regression results, both adjusting and not adjusting for nonresponses, are shown in Table 2. The corresponding results for the estimation of incidence rates are given in Table 3. Because the nonresponse rates were similar between cases and controls and did not depend too much on the measured predictors, adjustment of nonresponses only modestly improved the estimates of the regression parameters and incidence rates in this particular study.

5. Discussion

This paper provides a model-based approach for estimating the regression parameters of the logistic model and the corresponding incidence rates which properly adjusts for the differential selection probabilities of the stratified case-control sampling scheme as well as nonresponses. Case-control studies, especially those that are population-based, have played and will continue to play a major role in the

epidemiologic investigation of rare diseases. Practically all case-control studies in which data are collected directly from study subjects have nonrespondents. If nonresponses are not completely at random, then failure to account for nonresponses could result in biased estimators of regression parameters and incidence rates.

ROBINS, ROTNITZKY, and ZHAO (1994), among others, showed how to handle data missing by happenstance, including nonresponses, when the original sample consists of i.i.d. observations from an infinite population. Although Robins et al. regarded the case-control design itself as an example of missing data in their Sections 6.3 and 6.4, they did not allow the subjects selected into the case-control sample to have nonresponses by happenstance. Our work extends the existing literature on case-control designs by allowing differential nonresponses and also extends the standard literature on missing data by allowing the original observations to be sampled nonindependently from a finite population rather than independently from an infinite population. Although it would be possible to produce semiparametric efficient estimators under the conditions considered in this paper using the approach of Robins et al., the resulting estimators would be much more difficult to implement than the ones given here.

To use the proposed methods or any other potential methods for handling differential nonresponses, it is necessary to measure the predictors for nonresponses. One can use telephone interviews, as done in the WCHS Study, mail questionnaires or medical records to collect the relevant information from nonrespondents. If one anticipates differential nonresponses and wishes to remove the induced biases in the analysis, then one must incorporate a means of collecting the needed information from nonrespondents into their study design. This was done in the planning stage of the WCHS Study.

Appendix. Derivation of the Asymptotic Results

Because the case-control sample is a biased sample from a finite population, the standard asymptotic techniques based on random sampling from an infinite population are not sufficient for establishing the asymptotic properties of $\hat{\boldsymbol{\beta}}$. We will appeal to some results from survey sampling theory, especially the variance formula and central limit theorem for sampling from a finite population.

For $l = 0, 1$, $j = 1, \ldots, J$ and $k = 1, \ldots, N_{lj}$, let ϕ_{ljk} indicate, by the values 1 versus 0, on whether or not the kth subject of the lth disease category and jth stratum is selected into the case-control sample. Then (5) and (6) can be written as

$$\boldsymbol{S}(\boldsymbol{\gamma}) = \sum_{l=0}^{1} \sum_{j=1}^{J} \sum_{k=1}^{N_{lj}} \phi_{ljk} \psi_{ljk} \{\xi_{ljk} - q(\boldsymbol{z}_{ljk}; \boldsymbol{\gamma})\} \boldsymbol{z}_{ljk} ,$$

$$\boldsymbol{U}(\boldsymbol{\beta}; \hat{\boldsymbol{\gamma}}) = \sum_{l=0}^{1} \sum_{j=1}^{J} \sum_{k=1}^{N_{lj}} \frac{\phi_{ljk}(-1)^{l+1} \xi_{ljk} \psi_{ljk} \{1 - p_l(\boldsymbol{x}_{ljk}; \boldsymbol{\beta})\} \boldsymbol{x}_{ljk}}{\pi_{lj} q(\boldsymbol{z}_{ljk}; \hat{\boldsymbol{\gamma}})} .$$

Let E denote the expectation, and let $\mathcal{F}$ denote all the random variables except the selection indicators ϕ_{ljk}'s. By definitions, $\Pr(\phi_{ljk} = 1 \mid \mathcal{F}) = \pi_{lj}$ and $\Pr(\xi_{ljk} = 1 \mid \mathcal{F}) = q(\boldsymbol{z}_{ljk}; \boldsymbol{\gamma})$. Thus,

$$\begin{aligned} E\{\boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{\gamma})\} &= E[E\{\boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{\gamma}) \mid \mathcal{F}\}] \\ &= \Pr(\psi = 1)\, E\left[\sum_{l=0}^{1}\sum_{j=1}^{J}\sum_{k=1}^{N_{lj}} (-1)^{l+1} \{1 - p_l(\boldsymbol{x}_{ljk}; \boldsymbol{\beta})\}\, \boldsymbol{x}_{ljk}\right]. \end{aligned}$$

Since the term inside the squared-bracket in the above display is the population score function, we have $E\{\boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{\gamma})\} = \boldsymbol{0}$.

By the law of large numbers and the consistency of $\hat{\boldsymbol{\gamma}}$, $N^{-1}\boldsymbol{U}(\boldsymbol{\beta}; \hat{\boldsymbol{\gamma}})$ converges in probability to $\bar{\boldsymbol{u}}(\boldsymbol{\beta}) \equiv \lim_{N\to\infty} N^{-1}E\{\boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{\gamma})\}$, which is zero. Recall that $\hat{\boldsymbol{A}}(\boldsymbol{\beta}; \boldsymbol{\gamma}) = -\partial \boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{\gamma})/\partial \boldsymbol{\beta}'$, i.e.

$$\hat{\boldsymbol{A}}(\boldsymbol{\beta}; \boldsymbol{\gamma}) = \sum_{l=0}^{1}\sum_{j=1}^{J}\sum_{k=1}^{N_{lj}} \frac{\phi_{ljk}\xi_{ljk}\psi_{ljk} p_0(\boldsymbol{x}_{ljk}; \boldsymbol{\beta})\, p_1(\boldsymbol{x}_{ljk}; \boldsymbol{\beta})\, \boldsymbol{x}_{ljk}^{\otimes 2}}{\pi_{lj} q(\boldsymbol{z}_{ljk}; \boldsymbol{\gamma})},$$

which is positive semidefinite. Again by the law of large numbers and the consistency of $\hat{\boldsymbol{\gamma}}$, $N^{-1}\hat{\boldsymbol{A}}(\boldsymbol{\beta}; \hat{\boldsymbol{\gamma}})$ converges in probability to $\boldsymbol{A}(\boldsymbol{\beta}) \equiv -\partial \bar{\boldsymbol{u}}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$. Assume that $\boldsymbol{A}(\boldsymbol{\beta})$ is nonsingular, which implies that $\boldsymbol{A}(\boldsymbol{\beta})$ is positive definite. It then follows from convex analysis that $\hat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}$.

To derive the asymptotic distribution for $\hat{\boldsymbol{\beta}}$, we need to study the behavior of $\boldsymbol{U}(\boldsymbol{\beta}; \hat{\boldsymbol{\gamma}})$. By Taylor series expansions, $\boldsymbol{U}(\boldsymbol{\beta}; \hat{\boldsymbol{\gamma}}) = \boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{\gamma}) - \hat{\boldsymbol{H}}(\boldsymbol{\beta}; \boldsymbol{\gamma}^{\dagger})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ and $\boldsymbol{S}(\boldsymbol{\gamma}) = \hat{\boldsymbol{\Omega}}(\boldsymbol{\gamma}^{\ddagger})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}^{\dagger}$ and $\boldsymbol{\gamma}^{\ddagger}$ are on the line segment between $\hat{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$. Thus,

$$\boldsymbol{U}(\boldsymbol{\beta}; \hat{\boldsymbol{\gamma}}) = \boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{\gamma}) - \hat{\boldsymbol{H}}(\boldsymbol{\beta}; \boldsymbol{\gamma}^{\dagger})\, \hat{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\gamma}^{\ddagger})\, \boldsymbol{S}(\boldsymbol{\gamma}).$$

Note that

$$\begin{aligned} &\hat{\boldsymbol{H}}(\boldsymbol{\beta}; \boldsymbol{\gamma}) \\ &= \sum_{l=0}^{1}\sum_{j=1}^{J}\sum_{k=1}^{N_{lj}} \frac{\phi_{ljk}(-1)^{l+1}\, \xi_{ljk}\psi_{ljk}\{1 - p_l(\boldsymbol{x}_{ljk}; \boldsymbol{\beta})\}\, \{1 - q(\boldsymbol{z}_{ljk}; \boldsymbol{\gamma})\}\, \boldsymbol{x}_{ljk}\boldsymbol{z}_{ljk}'}{\pi_{lj} q(\boldsymbol{z}_{ljk}; \boldsymbol{\gamma})}, \end{aligned}$$

and

$$\hat{\boldsymbol{\Omega}}(\boldsymbol{\gamma}) = \sum_{l=0}^{1}\sum_{j=1}^{J}\sum_{k=1}^{N_{lj}} \phi_{ljk}\psi_{ljk} q(\boldsymbol{z}_{ljk}; \boldsymbol{\gamma})\, \{1 - q(\boldsymbol{z}_{ljk}; \boldsymbol{\gamma})\}\, \boldsymbol{z}_{ljk}^{\otimes 2}.$$

By the law of large numbers and the consistency of $\hat{\boldsymbol{\gamma}}$, $N^{-1}\hat{\boldsymbol{H}}(\boldsymbol{\beta}; \boldsymbol{\gamma}^{\dagger})$ and $N^{-1}\hat{\boldsymbol{\Omega}}(\boldsymbol{\gamma}^{\ddagger})$ converge to well-defined limits, say $\boldsymbol{H}$ and $\boldsymbol{\Omega}$. Therefore,

$$N^{-1/2}\boldsymbol{U}(\boldsymbol{\beta}; \hat{\boldsymbol{\gamma}}) = N^{-1/2}\boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{\gamma}) - \boldsymbol{H}\boldsymbol{\Omega}^{-1}N^{-1/2}\boldsymbol{S}(\boldsymbol{\gamma}) + \boldsymbol{o}_p(1). \tag{A.1}$$

Let us make the decomposition: $\boldsymbol{U}(\boldsymbol{\beta};\,\boldsymbol{\gamma}) = \boldsymbol{U}^F(\boldsymbol{\beta};\,\boldsymbol{\gamma}) + \boldsymbol{U}^D(\boldsymbol{\beta};\,\boldsymbol{\gamma})$, where

$$\boldsymbol{U}^F(\boldsymbol{\beta};\,\boldsymbol{\gamma}) = \sum_{l=0}^{1}\sum_{j=1}^{J}\sum_{k=1}^{N_{lj}} \frac{(-1)^{l+1}\,\xi_{ljk}\psi_{ljk}\{1 - p_l(\boldsymbol{x}_{ljk};\,\boldsymbol{\beta})\}\,\boldsymbol{x}_{ljk}}{q(\boldsymbol{z}_{ljk};\,\boldsymbol{\gamma})}\,,$$

$$\boldsymbol{U}^D(\boldsymbol{\beta};\,\boldsymbol{\gamma}) = \sum_{l=0}^{1}\sum_{j=1}^{J}\sum_{k=1}^{N_{lj}} \frac{(\phi_{ljk} - \pi_{lj})\,(-1)^{l+1}\,\xi_{ljk}\psi_{ljk}\{1 - p_l(\boldsymbol{x}_{ljk};\,\boldsymbol{\beta})\}\,\boldsymbol{x}_{ljk}}{\pi_{lj}q(\boldsymbol{z}_{ljk};\,\boldsymbol{\gamma})}\,.$$

By the multivariate central limit theorem, $N^{-1/2}\boldsymbol{U}^F(\boldsymbol{\beta};\,\boldsymbol{\gamma})$ converges in distribution to a zero-mean normal random vector with covariance matrix

$$\boldsymbol{C}^F = \lim_{N\to\infty} N^{-1}\sum_{l=0}^{1}\sum_{j=1}^{J}\sum_{k=1}^{N_{lj}} \frac{\xi_{ljk}\psi_{ljk}\{1 - p_l(\boldsymbol{x}_{ljk};\,\boldsymbol{\beta})\}^2\,\boldsymbol{x}_{ljk}^{\otimes 2}}{q^2(\boldsymbol{z}_{ljk};\,\boldsymbol{\gamma})}\,.$$

By the Wald-Wolfowitz-Noether-Hajek central limit theorem for sampling without replacement from a finite population (see Cochran, 1977, pp. 39–40), conditionally on $\mathcal{F}$, $N^{-1/2}\boldsymbol{U}^D(\boldsymbol{\beta};\,\boldsymbol{\gamma})$ converges in distribution to a zero-mean normal random vector with covariance matrix

$$\boldsymbol{C}^D = \lim_{N\to\infty} \sum_{l=0}^{1}\sum_{j=1}^{J} \frac{(n_{lj}/N)\,(1 - \pi_{lj})}{\pi_{lj}^2} \left(N_{lj}^{-1}\sum_{k=1}^{N_{lj}} \frac{\xi_{ljk}\psi_{ljk}\{1 - p_l(\boldsymbol{x}_{ljk};\,\boldsymbol{\beta})\}^2\,\boldsymbol{x}_{ljk}^{\otimes 2}}{q^2(\boldsymbol{z}_{ljk};\,\boldsymbol{\gamma})} - \left[N_{lj}^{-1}\sum_{k=1}^{N_{lj}} \frac{\xi_{ljk}\psi_{ljk}\{1 - p_l(\boldsymbol{x}_{ljk};\,\boldsymbol{\beta})\}\,\boldsymbol{x}_{ljk}}{q(\boldsymbol{z}_{ljk};\,\boldsymbol{\gamma})} \right]^{\otimes 2} \right).$$

The convergence of the distribution also holds unconditionally because $\boldsymbol{C}^D$ is a deterministic matrix that does not depend on the actual values of $(\boldsymbol{x}_{ljk},\,\boldsymbol{z}_{ljk},\,\xi_{ljk},\,\psi_{ljk})$ $(l = 0, 1;\ j = 1, \ldots, J;\ k = 1, \ldots, N_{lj})$. Since $E(\phi_{ljk} \mid \mathcal{F}) = \pi_{lj}$, it is easy to show that $N^{-1/2}\boldsymbol{U}^F(\boldsymbol{\beta};\,\boldsymbol{\gamma})$ and $N^{-1/2}\boldsymbol{U}^D(\boldsymbol{\beta};\,\boldsymbol{\gamma})$ are uncorrelated and thus asymptotically independent. Therefore, $N^{-1/2}\boldsymbol{U}(\boldsymbol{\beta};\,\boldsymbol{\gamma})$ converges in distribution to a zero-mean normal random vector with covariance matrix $\boldsymbol{C} = \boldsymbol{C}^F + \boldsymbol{C}^D$, which is the limit of

$$N^{-1}\sum_{l=0}^{1}\sum_{j=1}^{J}\sum_{k=1}^{N_{lj}} \frac{\xi_{ljk}\psi_{ljk}\{1 - p_l(\boldsymbol{x}_{ljk};\,\boldsymbol{\beta})\}^2\,\boldsymbol{x}_{ljk}^{\otimes 2}}{\pi_{lj}q^2(\boldsymbol{z}_{ljk};\,\boldsymbol{\gamma})} - N^{-1}\sum_{l=0}^{1}\sum_{j=1}^{J} \frac{1 - \pi_{lj}}{n_{lj}} \left[\sum_{k=1}^{N_{lj}} \frac{\xi_{ljk}\psi_{ljk}\{1 - p_l(\boldsymbol{x}_{ljk};\,\boldsymbol{\beta})\}\,\boldsymbol{x}_{ljk}}{q(\boldsymbol{z}_{ljk};\,\boldsymbol{\gamma})} \right]^{\otimes 2}.$$

Likewise, $N^{-1/2}\boldsymbol{U}(\boldsymbol{\beta};\,\boldsymbol{\gamma})$ and $N^{-1/2}\boldsymbol{S}(\boldsymbol{\gamma})$ are asymptotically joint normal, and the limiting covariance matrix between $N^{-1/2}\boldsymbol{U}(\boldsymbol{\beta};\,\boldsymbol{\gamma})$ and $N^{-1/2}\boldsymbol{S}(\boldsymbol{\gamma})$ is

$$\lim_{N\to\infty} N^{-1}\sum_{l=0}^{1}\sum_{j=1}^{J}\sum_{k=1}^{N_{lj}} \frac{\phi_{ljk}(-1)^{l+1}\,\xi_{ljk}\psi_{ljk}\{1 - p_l(\boldsymbol{x}_{ljk};\,\boldsymbol{\beta})\}\,\{1 - q(\boldsymbol{z}_{ljk};\,\boldsymbol{\gamma})\}\,\boldsymbol{x}_{ljk}\boldsymbol{z}_{ljk}'}{\pi_{lj}q(\boldsymbol{z}_{ljk};\,\boldsymbol{\gamma})},$$

which is exactly $\boldsymbol{H}$. Hence, it follows (A.1) that $N^{-1/2}\boldsymbol{U}(\boldsymbol{\beta}; \hat{\boldsymbol{\gamma}})$ converges in distribution to a zero-mean normal random vector with covariance matrix $\boldsymbol{B} = \boldsymbol{C} - \boldsymbol{H}\boldsymbol{\Omega}^{-1}\boldsymbol{H}'$.

Taking the Taylor series expansion of $\boldsymbol{U}(\hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\gamma}})$ at $\boldsymbol{U}(\boldsymbol{\beta}; \hat{\boldsymbol{\gamma}})$, we have

$$N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left\{N^{-1}\hat{\boldsymbol{A}}(\boldsymbol{\beta}^*; \hat{\boldsymbol{\gamma}})\right\}^{-1} N^{-1/2}\boldsymbol{U}(\boldsymbol{\beta}; \hat{\boldsymbol{\gamma}}),$$

where $\boldsymbol{\beta}^*$ is on the line segment between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. The asymptotic normality of $\boldsymbol{U}(\boldsymbol{\beta}; \hat{\boldsymbol{\gamma}})$, together with the consistency of $N^{-1}\hat{\boldsymbol{A}}(\boldsymbol{\beta}; \hat{\boldsymbol{\gamma}})$ and $\hat{\boldsymbol{\beta}}$, then implies that $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a zero-mean normal random vector with covariance matrix $\boldsymbol{V} = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}$. Replacing the unknown parameters in $\boldsymbol{V}$ by their respective sample estimators yield the consistent covariance matrix estimator given in Section 2.

References

Breslow, N. E. and Day, N. E., 1980: *Statistical Methods in Cancer Research, Volume 1, The Analysis of Case Control Studies*. Int'l. Agency for Research on Cancer, Lyon.

Cochran, W. G., 1977: *Sampling Techniques*. Wiley, New York.

Horvitz, D. G. and Thompson, D. J., 1952: A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Prentice, R. L. and Pyke, R., 1979: Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.

Robins, J. M., Rotnitzky, A., and Zhao, L., 1994: Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

Schwartz, S. M., Siscovick, D. S., Longstreth, W. T., Psaty, B. M., Beverly, R. K., Raghunathan, T. E., Lin, D. Y., and Koepsell, T. D., 1997: Low-dose oral contraceptive use and stroke in young women. *Annals of Internal Medicine* **127**, 596–603.

Scott, A. J. and Wild, C. J., 1986: Fitting logistic models under case-control or choice based sampling. *Journal of the Royal Statistical Society Series B* **48**, 170–182.

Wang, C. Y., Wang, S., Zhao, L., and Ou, S., 1997: Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association* **92**, 512–525.

Patrick G. Arbogast
Division of Biostatistics
Department of Preventive Medicine
Vanderbilt University
U-8201 Medical Center North
Nashville, TN 37232-2637
U.S.A.
E-mail: patrick.arbogast@mcmail.vanderbilt.edu.

Received, October 2000
Revised, May 2001
Accepted, September 2001