A Nonparametric Method for Dealing With Mismeasured Covariate Data

MARGARET SULLIVAN PEPE and THOMAS R. FLEMING*

Mismeasurement of covariate data is a frequent problem in statistical data analysis. However, when true and mismeasured data are obtained for a subsample of the observations, it is possible to estimate the parameters relating the outcome to the covariate of interest. Maximum likelihood methods that rely on parametric models for the mismeasurement have not met with much success. Realistic models for the mismeasurement process are difficult to construct; the form of the likelihood is often intractable and, more important, such methods are not robust to model misspecification. We propose an easily implemented method that is nonparametric with respect to the mismeasurement process and that is applicable when mismeasurement is due to the problem of missing data, errors in variables, or use of imperfect surrogate covariates. Specifically, denote the outcome variable by Y, the covariate data subject to mismeasurement by X, and the remaining covariates, including perhaps surrogates or mismeasured values of X, by Z. We consider a general regression model of the form $P_\beta(Y \mid X, Z)$. Suppose data regarding Y, X, and Z are available for a validation sample V, a random subsample of the total sample, whereas data regarding only Y and Z are available for the remainder, the nonvalidation sample $\bar{V}$. We propose to base inference on the estimated likelihood for β, $\hat{L}(\beta) = \Pi_{i \in V} P_\beta(Y_i \mid X_i, Z_i) \Pi_{j \in \bar{V}} \hat{P}_\beta(Y_j \mid Z_j)$, where $\hat{P}_\beta(Y_j \mid Z_j)$ is estimated empirically using the validation sample covariate data. Asymptotic results are derived for the case in which the surrogate or mismeasured covariates are categorical. The asymptotic variance of the estimated score involves not only the second derivative of the log estimated likelihood but also a term that captures the variability induced by estimating the nonvalidation sample likelihood. An example and a small simulation study demonstrate that this method may be of value for the missing covariate data and covariate measurement error problems.

KEY WORDS: Errors in variables; Missing data; Surrogate variables.

1. INTRODUCTION

Frequently components of a covariate vector U may be mismeasured. Specifically, there may be some measurement error associated with determining U (the errors in variables problem), or some components of the covariate vector may be missing (the missing covariate data problem). Alternatively, one might choose to measure some imperfect surrogate for a component of U because it is easier or less expensive to ascertain than the true covariate of interest.

Let X denote the true values of the components of the covariate vector U that are subject to mismeasurement in any of the senses mentioned above. Let Z denote the remaining components of U and the mismeasured values of X. Thus, for example, if $U = (U_1, U_2)$, with U_1 missing or mismeasured and U_3 a surrogate or mismeasured value for U_1, then $X = U_1$ and $Z = (U_2, U_3)$. Let the probability function for an outcome random variable Y, as a function of X and Z, be parameterized by β and denoted by $P_\beta(Y \mid X, Z)$. We suppose that all sample observations have data regarding Y and Z available but that only a random subsample, called the validation sample, has data regarding X also available. Since the conditional distribution of Y given Z is

$$P_\beta(Y \mid Z) = \int P_\beta(Y \mid x, Z)\, dP(x \mid Z),$$

inference regarding β is generally impossible without such a validation sample, through which the conditional distribution of X given Z can be identified. If the conditional distribution functions $P(X \mid Z)$ were completely known, then the likelihood for β, assuming iid observations, is

$$L(\beta) = \prod_{i \in V} P_\beta(Y_i \mid X_i, Z_i) \prod_{j \in \bar{V}} P_\beta(Y_j \mid Z_j),$$

where V and $\bar{V}$ are the set of indices for observations in the validation and nonvalidation samples, respectively. Our proposal is to estimate $P(X \mid Z)$ empirically from the validation sample and, in turn, to estimate the likelihood components $P_\beta(Y \mid Z)$ for nonvalidation sample members.

To estimate $P(X \mid Z)$ on the basis of the validation sample observations, note that some components of Z may be uninformative with respect to the distribution of X. For example, in the context of bone marrow transplantation, let X be serum levels of the drug, cyclosporine, at the end of the first week posttransplant. Then the dose of cyclosporine received might be informative with respect to X, but the degree of HLA antigen mismatch between the recipient and donor would not be. Let S denote the informative components of Z, in the sense that $P(X \mid Z) = P(X \mid S)$ almost surely. We assume that S is categorical. The empirical estimates of $P(X \mid Z)$ are

$$\hat{P}(X \mid Z) = \hat{P}(X \mid S) = \frac{\sum_{i \in V} I(X_i \le X, S_i = S)}{\sum_{i \in V} I(S_i = S)},$$

where $I(\ )$ is the indicator function. This yields an unbiased

* Margaret Sullivan Pepe is Assistant Member, Fred Hutchinson Cancer Research Center, Division of Public Health Sciences, 1124 Columbia Street, Seattle, WA 98104. Thomas R. Fleming is Professor, Department of Biostatistics, University of Washington, Seattle, WA 98195. This research was supported by NIH Grants GM-24472 and CA-32693.

© 1991 American Statistical Association
Journal of the American Statistical Association
March 1991, Vol. 86, No. 413, Theory and Methods

estimate of $P_\beta(Y_j \mid Z_j)$ for a nonvalidation sample member j,

$$\hat{P}_\beta(Y_j \mid Z_j) = \frac{\sum_{i\in V} P_\beta(Y_j \mid X_i, Z_j)\, I(S_i = S_j)}{\sum_{i\in V} I(S_i = S_j)}$$

$$= \int P_\beta(Y_j \mid x, Z_j)\, d\hat{P}(x \mid Z_j),$$

assuming at least one validation sample member i with covariate value $S_i = S_j$. We will show that inference for β can be based on the estimated likelihood

$$\hat{L}(\beta) = \prod_{i\in V} P_\beta(Y_i \mid X_i, Z_i) \prod_{j\in \bar{V}} \hat{P}_\beta(Y_j \mid Z_j).$$

An alternative approach to the mismeasured covariate data problem is to postulate a model for the mismeasurement $P_\theta(X \mid Z)$ and to maximize the full likelihood with respect to β and θ simultaneously. Pepe, Self, and Prentice (1989) used this approach for inference in the Cox model. Unfortunately, realistic models for the mismeasurement $P_\theta(X \mid Z)$ may be difficult to construct. Moreover, inference requiring specification of a parametric model, $P_\theta(X \mid Z)$, can be rather sensitive to the choice of mismeasurement model (Carroll Spiegelman, Lan, Gailey and Abbott 1984, Pepe et al. 1989). There are also technical difficulties in implementing maximum likelihood methods in this setting. Specifically, nonvalidation sample observations contribute terms of the form

$$P_{\beta,\theta}(Y \mid Z) = \int P_\beta(Y \mid x, Z)\, dP_\theta(x \mid Z),$$

which is rarely a tractable analytic expression. The integral usually must be evaluated numerically.

The estimated likelihood approach presented here circumvents this technical problem. Perhaps more important, by being nonparametric with respect to $P(X \mid Z)$, it circumvents the lack of robustness of the maximum likelihood approach to misspecification of the form of $P_\theta(X \mid Z)$, at least in large samples. The method can, in fact, be interpreted as one of the new semiparametric methods that are currently in vogue. It parameterizes only the relation of interest, $P_\beta(Y \mid X, Z)$, and allows the nuisance mismeasurement process to be completely arbitrary.

In Section 2, we discuss asymptotic distribution theory for the estimator β that maximizes the estimated likelihood. An illustrative example in Section 3 is followed by the results of a small simulation study.

2. ASYMPTOTIC RESULTS

We assume that the validation sample, of size n^V, is a simple random sample from the total sample of size n. If the fraction of the total sample in the validation subset is nonnegligible, specifically, if n^V/n approaches a limit, $\rho^V > 0$, as $n \rightarrow \infty$, then, in the appendix, it is shown that under regularity conditions and in a neighborhood of the true β, the solution $\hat{\beta}$ to the estimated score equation

$$\frac{d}{d\beta} \log \hat{L}(\beta) = 0$$

is consistent for β. Moreover, the estimated score function $n^{-1/2}\, d \log L(\beta)/d\beta$ is asymptotically normal, with mean 0 and variance–covariance matrix

$$(1 - \rho^V) E\left[-\frac{d^2}{d\beta^2} \log P_\beta(Y \mid Z)\right] + \rho^V E\left[-\frac{d^2}{d\beta^2} \log P_\beta(Y \mid X,Z)\right]$$
$$+ \frac{(1-\rho^V)^2}{\rho^V} \operatorname{var}\left\{E\left[\frac{d}{d\beta} \log P_\beta(Y \mid Z) | X,S\right]\right\}$$
$$\equiv I(\beta) + \frac{(1-\rho^V)^2}{\rho^V} \Sigma(\beta).$$

Note that the first component of the variance, $I(\beta)$, is the expected information based on $L(\beta)$, the likelihood for the observed data if $P(X \mid Z)$ was completely known. The second term is therefore the variance induced by estimating the likelihood components for nonvalidation sample members.

This asymptotic normality result, which is derived in the appendix, follows from the representation of the estimated score:

$$\frac{1}{\sqrt{n}} \frac{d}{d\beta} \log \hat{L}(\beta) = \frac{1}{\sqrt{n}} \frac{d}{d\beta} \log L(\beta) + \frac{1}{\sqrt{n}} \frac{(1-\rho^V)}{\rho^V} \times \sum_{i\in V} \overline{W}_{X_i,S_i}(\beta) + O_p(1/\sqrt{n}),$$

where

$$\overline{W}_{X_i,S_i}(\beta) = \frac{\sum_{j\in\bar{V}} \left\{ \frac{dP_\beta(Y_j \mid X_i, Z_j)/d\beta}{P_\beta(Y_j \mid Z_j)} - \frac{dP_\beta(Y_j \mid Z_j)/d\beta}{[P_\beta(Y_j \mid Z_j)]^2} \times P_\beta(Y_j \mid X_i, Z_j) \right\} I(S_j = S_i)}{\sum_{j\in\bar{V}} I(S_j = S_i)}.$$

$\overline{W}_{X_i,S_i}$ is, essentially, the contribution of a validation sample member i to the difference between the true and estimated scores for the nonvalidation sample observations. Conditioning on the nonvalidation sample data and on the covariate data S for the validation sample, the $\{\overline{W}_{X_i,S_i}(\beta),\ i \in V\}$ are independent, with variance–covariance matrix converging to $\Sigma(\beta)$ with probability 1. Therefore, the sample variance–covariance of $\{\overline{W}_{X_i,S_i}(\beta),\ i \in V\}$ is consistent for $\Sigma(\beta)$, and by substituting estimates for the unknown quantities in $\overline{W}_{X_i,S_i}(\beta)$, a consistent estimator for $\Sigma(\beta)$ is the sample variance–covariance of $\{\hat{\overline{W}}_{X_i,S_i}(\beta),\ i \in V\}$,

$$\hat{\Sigma}(\beta) = \widehat{\operatorname{var}}\{\hat{\overline{W}}_{X_i,S_i}(\beta),\ i \in V\},$$

$$\bar{\hat{W}}_{X_i,S_i}(\beta) = \frac{\sum_{j\in\bar{V}} \left\{ \dfrac{dP_\beta(Y_j \mid X_i, Z_j)/d\beta}{\hat{P}_\beta(Y_j \mid Z_j)} - \dfrac{d\hat{P}_\beta(Y_j \mid Z_j)/d\beta}{[\hat{P}_\beta(Y_j \mid Z_j)]^2} \times P_\beta(Y_j \mid X_i, Z_j) \right\} I(S_j = S_i)}{\sum_{j\in\bar{V}} I(S_j = S_i).}$$

It follows that a consistent estimator of the variance–covariance matrix of the estimated score function is

$$\hat{I}(\hat{\beta}) + \frac{(1 - \hat{\rho}^V)^2}{\hat{\rho}^V} \hat{\Sigma}(\hat{\beta}) \quad \text{where} \quad \hat{I}(\beta) = -\frac{d^2}{d\beta^2} \log \hat{L}(\beta)$$

and $\hat{\rho}^V = n^V/n$.

A Taylor series expansion provides asymptotic distribution theory for $\hat{\beta}$, the maximum estimated likelihood estimate. Large sample properties are summarized in the following result.

Result

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, I^{-1}(\beta) + \frac{(1 - \rho^V)^2}{\rho^V} I^{-1}(\beta)\Sigma(\beta)I^{-1}(\beta)\right)$$

and a consistent estimator of the variance–covariance matrix is

$$\hat{I}^{-1}(\hat{\beta}) + \frac{(1 - \hat{\rho}^V)^2}{\hat{\rho}^V} \hat{I}^{-1}(\hat{\beta})\hat{\Sigma}(\hat{\beta})\hat{I}^{-1}(\hat{\beta}).$$

The estimate $\hat{\beta}$ can be evaluated by implementing the usual Newton–Raphson procedure, the score and information matrix being "estimated" at each iteration. Therefore, the method can be easily incorporated into many routine analyses by including this one extra step at each iteration. The extra variability term $\hat{\Sigma}(\hat{\beta})$ is calculated only after the final iteration.

3. EXAMPLES AND SIMULATIONS

Example 1: Missing Covariate Data

Most statistical packages will drop observations with incomplete covariate data from the analysis. This can result in a much reduced effective sample size and serious loss of efficiency when a substantial proportion of observations is missing data. An alternative strategy is to impute the missing values with predicted values from a regression of the missing component on the available components and then to use the completed data set for inference. Such simple imputation methods, however, are known to introduce bias into the estimate itself and into its variance estimate. Imputation methods also require specification of a model for the mismeasurement process $P_\theta(X \mid Z)$.

In this example, we illustrate the alternative approach to handling missing covariate data based on the estimated likelihood method and contrast the results to those obtained when observations with incomplete data are dropped from the analysis. The data concern 97 patients given bone marrow transplants from female sibling donors who were matched for genes in the major histocompatibility complex. Patients were evaluated for acute graft versus host disease, the binary outcome of interest in this study. Factors thought to be predictive of the risk of acute graft versus host disease are previous donor pregnancies, patient age, isolation in a laminar airflow room during the immediate posttransplant period, and the acute graft versus host disease prophylactic treatment received. The primary focus in the study was on the effect of prophylactic regimen. Donor pregnancy status was missing for 31 patients, due to a lack of effort among some physicians in ascertaining a complete medical history for the donor. Identifying the mismeasured covariate X with donor pregnancy status and the surrogate S with patient age, the likelihood, score, and information components for those with missing donor pregnancy data were estimated empirically using the validation sample. The validation sample here consisted of those 66 observations with complete covariate data.

The results of the estimated likelihood analysis using a logistic regression model are displayed in Table 1. There was a substantial gain in efficiency by using the estimated likelihood approach, which included the 31 observations with missing donor pregnancy data, over the maximum likelihood analysis with those 31 observations excluded. The estimated standard error of the coefficient for prophylaxis decreased from .87 to .52. The validation sample in this analysis was large, composing 68% of the sample. As a result, the extra variability introduced into the prophylaxis coefficient by estimating the likelihood was negligible. Specifically, the extra variability component of the variance of β, $[(1 - \rho^V)^2/n\rho^V]\ I^{-1}(\beta)\Sigma(\beta)I^{-1}(\beta)$, was estimated to be $.33 \times 10^{-5}$, which compares with the inverse information $I^{-1}(\beta)$ component of the variance of β, estimated as .276.

Interestingly, in this analysis, there was no decrease in the standard error of the estimated regression coefficient for donor pregnancy. It might be that patient age was a rather poor surrogate for donor pregnancy status in this study. We could have omitted the donor pregnancy covariate from the model and performed the maximum likelihood analysis for the reduced model using all 97 observations. This would not be satisfactory, however, since it is known that the omission of an important covariate can bias the estimates of the remaining covariates in a model (Gail, Wieand, and Pantadosi 1984, Anderson 1989). Moreover, the coefficient estimated from the estimated likelihood analysis can be interpreted in the context of the richer model, which includes donor pregnancy status.

As suggested by a referee, we performed a small simulation study to validate the use of the large sample distribution results in this example. Covariate data for patient age, laminar airflow room, and prophylaxis, as they are dichotomized in Table 1, were generated independently using their observed marginal distributions in this data set. Donor pregnancy status was then generated using the observed marginal distributions within the age category defined for that observation. Finally, the outcome was generated using a logistic regression model with coefficients equal to those estimated in the estimated likelihood analysis of Table 1. Thirty-one observations were chosen at random to constitute the subset of observations with missing donor pregnancy data. Coefficients were then estimated using the estimated-likelihood analysis. On the basis of 500 simula-

Table 1. Logistic Regression Models for Acute Graft Versus Host Disease

	Log odds ratio (standard error)			
Factor	*Estimated likelihood analysis (n = 97)*		*Observations with unknown donor pregnancy data excluded (n = 66)*	
Donor pregnancy (yes/no)	1.26	(.66)	1.27	(.67)
Patient age (30–40 years/20–29 years)	−.03	(.52)	.01	(.67)
Laminar airflow room (yes/no)	−.58	(.57)	−.42	(.77)
Prophylaxis (MTX + CSP/other)*	−1.36	(.52)	−1.59	(.87)
Constant	−.69	(.54)	−.47	(.79)

*MTX + CSP = a combination of methotrexate and cyclosporine; other = regimes used other than the combination of methotrexate and cyclosporine.

tions, the coverages of the 90% confidence intervals implied by the asymptotic results were 88%, 88%, 91%, 90%, and 93% for the coefficients associated with donor pregnancy status, patient age, laminar airflow room, prophylaxis, and the intercept, respectively. We also performed a similar simulation study for the classical logistic regression analysis of Table 1, using only the 66 observations for which donor pregnancy data were known. Coverages of the 90% confidence intervals implied by the asymptotic results were 92%, 86%, 91%, 90%, and 90% in this case.

Example 2: Errors in Covariates—A Simulation Study

A simulation study, using the mathematical programming environment of GAUSS 2.0 (1988), was performed to investigate the adequacy of the asymptotic results in small samples. For each of n observations, a normally distributed covariate, $X \sim N(0,1)$, and a binary outcome Y from the logistic probability function,

$$P(Y = 1 \mid X = x) = e^{\alpha+\beta x}/(1 + e^{\alpha+\beta x}),$$

were generated. Independent additive normal error, $\varepsilon \sim N(0,\sigma^2)$, was generated, and a binary surrogate covariate was defined by

$$Z = I(X + \varepsilon > 0).$$

In the bone marrow transplantation setting, this model might be appropriate for, say, the risk of developing acute graft versus host disease (Y) as a function of the logarithm of serum cyclosporine levels (X) achieved at the end of the first-week posttransplant. An indicator (Z) of whether an inaccurate measurement of serum cyclosporine is above some threshold value might be readily available for all subjects, whereas the more expensive, but accurately measured, covariate of interest (X) might be available only for a small subgroup.

Using the data $\{(Y_j, Z_j), j \in \bar{V}; (Y_i, Z_i, X_i), i \in V\}$, where V was chosen as a simple random sample from the total, the maximum estimated-likelihood estimates were obtained. The results in Table 2 cover a range of covariate effects parameterized by β, measurement error effects parameterized by σ, and rather small validation sample size fractions ρ^V.

When the covariate X has no effect ($\beta = 0$), it appears

*Table 2. Simulation Study Results for Logistic Regression, $P(Y = 1/X = x) = e^{\alpha+\beta x}/(1 + e^{\alpha+\beta x})$ Where $X \sim N(0,1)$ and $Z = I(X + \varepsilon > 0)$, $\varepsilon \sim N(0,\sigma^2)$**

				$\hat\beta$					$\hat\alpha$		
α	β	σ	ρ^V	Mean($\hat\beta$)	Median($\hat\beta$)	Var($\hat\beta$)	Mean (vâr($\hat\beta$))	Coverage of 90% confidence interval for $\hat\beta$	Mean($\hat\alpha$)	Var($\hat\alpha$)	Mean (vâr($\hat\alpha$))
						n = 200					
0	0	.25	.2	.006	.011	.033	.030	89%	−.004	.020	.021
0	0	.25	.1	.009	.012	.036	.034	91%	.009	.021	.021
0	0	1.00	.2	−.005	−.015	.053	.051	91%	.002	.020	.022
0	0	1.00	.1	−.006	−.006	.076	.073	91%	.007	.022	.024
0	.693	.25	.2	.703	.683	.053	.043	88%	.000	.027	.024
0	.693	.25	.1	.724	.693	.060	.050	91%	.002	.035	.025
0	.693	1.00	.2	.728	.692	.088	.079	87%	.006	.032	.026
0	.693	1.00	.1	.762	.671	.176	.146	88%	.007	.056	.035
						n = 100					
0	0	.25	.2	.012	−.002	.072	.067	91%	.000	.049	.043
0	0	1.00	.2	−.035	−.037	.118	.117	92%	.003	.048	.047
0	.693	.25	.2	.735	.722	.099	.093	90%	−.001	.049	.049
0	.693	1.00	.2	.779	.716	.225	.220	91%	.011	.082	.065
−1	0	.25	.2	−.013	.000	.087	.100	95%	−1.034	.058	.068
−1	0	1.00	.2	.004	−.002	.211	.213	94%	−1.047	.059	.090
−1	.693	.25	.2	.763	.717	.145	.157	93%	−1.037	.082	.101
−1	.693	1.00	.2	.737	.678	.240	.502	93%	−1.034	.103	.307

*Based on 500 simulations in each case.

that the asymptotic results for $\hat{\beta}$ approximate the small-sample distribution very closely. This was true in this example, even with validation sample sizes as small as 20. Moreover, the estimated-likelihood method is fully efficient when $\beta = 0$. This results from the fact that the extra variability component of the variance of $\hat{\beta}$, namely, $(1 - \rho^V)^2/\rho^V I^{-1}(\beta)\Sigma(\beta)I^{-1}(\beta)$, is zero at $\beta = 0$ since

$$\Sigma(0) = \operatorname{var}\left\{E\left[\frac{d}{d\beta}\log P_\beta(Y \mid Z)|X, Z\right]\right\}_{\beta=0},$$

and

$$E\left[\frac{d}{d_\beta}\log P_\beta(Y \mid Z)|X, Z\right]_{\beta=0} = E\left[\frac{d}{d_\beta}\log P_\beta(Y \mid Z)|Z\right]_{\beta=0} = 0.$$

Indeed, at $\beta = 0$, $\hat{P}_\beta(Y \mid Z) = P_\beta(Y \mid Z)$. It is not surprising therefore that the estimated-likelihood estimate is as efficient as the maximum likelihood estimate with $P(X \mid Z)$ known.

When the covariate X had a rather strong effect, with relative risk of 2.0 for $X = 1$ versus $X = -1$, there did appear to be some positive bias in the estimate $\hat{\beta}$. Moreover, the estimator $\{\hat{I}^{-1}(\hat{\beta}) + [(1 - \rho^V)^2/\rho^V]\, \hat{I}^{-1}(\hat{\beta})\Sigma(\hat{\beta})\hat{I}^{-1}(\hat{\beta})\}/n$ provided an underestimate of the true variance. On the other hand, the median estimate $\hat{\beta}$ was quite close to the true value, and the 90% confidence intervals appeared adequate, having a higher coverage probability than the average variance estimate would suggest in most cases. These observations suggest that the true distribution of β is somewhat skewed. We verified this in the simulation study. A few extremely large positive values may be a result of the form of the estimated score for nonvalidation sample members $[d\hat{P}(Y \mid Z)/d\beta]/\hat{P}(Y \mid Z)$, which is likely to have heavy tails in small samples. The distribution of the variance estimates for $\hat{\alpha}$ and $\hat{\beta}$ also appeared to be skewed, particularly in the asymmetric case with $\alpha = -1$, $\beta = .693$. The last line of Table 2 displays average variance estimates for $\hat{\alpha}$ and $\hat{\beta}$, .307 and .502, respectively, which were much larger than the corresponding median variance estimates, .096 and .204, respectively, in this case. Coverage of the 90% confidence intervals for both α and β, however, remained adequate.

4. CONCLUDING REMARKS

The empirical estimation of the likelihood presented here is a simple and a natural approach to the mismeasured covariate data problem when validation data are available in addition to the mismeasured covariate data.

Published research on the covariate measurement error problem has focused on methods that require a parametric specification of the form of the mismeasurement $P(X \mid Z)$. Indeed, the case of normally distributed covariates with normally distributed errors (Carroll et al. 1984; Schafer 1987 has received the most attention. The usefulness of these methods is limited by the fact that they may not be robust for alternative measurement error structures. Whittemore and Keller (1988) and Stefanski and Carroll (1985) have proposed bias adjustments to the naive estimator (which ignores the measurement error), and these are appropriate under models more general than the additive normal structure. However, the adjustments are justified only when the size of the measurement error is small. In contrast, the estimated-likelihood method makes no restriction on the size, or the form, of the measurement error. Imputation methods for dealing with missing covariate data (Little and Rubin 1987) also generally require specification of a parametric model for the joint distribution of the covariates. A semiparametric version of multiple imputation, termed HOT DECK, was described by Rubin (1987). In HOT DECK, imputed values of X are drawn with replacement from the empirical distribution of X given (Y, Z) generated by the validation sample. The usual multiple inputation variance is, however, inappropriate for the resultant semiparametric estimator (Rubin 1987, p. 122). Without an expression for the variance of the HOT DECK estimator, inferential techniques are therefore limited.

Much work remains to be done. Some further evaluation of the small sample properties of the estimated-likelihood method is warranted. Furthermore, extensions to continuous mismeasured or surrogate covariates will be needed for many practical applications. Carroll and Wand (1989) provide some results for this problem. Design considerations such as the size of the validation sample fraction and procurement of validation samples, which are not simple random samples from the total, will also require further study.

APPENDIX

Asymptotic Results

Assume (a) the validation sample V is a random sample of size n^V from the total sample of n and that

$$\lim_{n\to\infty}\frac{n^V}{n} = \rho^V > 0,$$

and (b) the usual regularity conditions for maximum likelihood estimation hold (Cox and Hinkley 1974, Ch. 9) for both $P_\beta(Y \mid X, Z)$ and $P_\beta(Y \mid Z)$. Also, with probability one, the first and second partial derivatives of $P_\beta(Y \mid X, Z)$ are bounded uniformly in a neighborhood of β, and $P_\beta(Y \mid X, Z) \geq C$ for some positive constant C.

Theorem 1. (Consistency) In a neighborhood of the true parameter β, the solution $\hat{\beta}$ to the estimated score equation (a) is consistent and (b) maximizes the estimated likelihood $\hat{L}(\beta)$, with probability converging to 1 as $n \to \infty$.

Proof. It can be shown that in a neighborhood of the true parameter β_0,

$$\frac{1}{n}\left[\frac{d^2\log\hat{L}(\beta)}{d\beta^2} - \frac{d^2\log L(\beta)}{d\beta^2}\right] \xrightarrow{P} 0$$

uniformly as $n \to \infty$. Therefore, uniformly in β,

$$\frac{1}{n}\frac{d^2\log\hat{L}(\beta)}{d\beta^2} \xrightarrow{P} I(\beta) \equiv E\left[\frac{-d^2\log(\beta)}{d\beta^2}\right],$$

the information matrix which is positive definitive in a neighborhood of β_0 by regularity assumptions. This implies that log $\hat{L}(\beta)$ is concave with probability converging to 1 as $n \to \infty$ near

β_0, proving the second assertion of the theorem. Using the same techniques as in Andersen and Gill (1982, App. II), it can be shown that $(1/n) \log L(\beta)$, and in turn, $(1/n) \log \hat{L}(\beta)$, converges to a concave function $L(\beta)$ with local maximum at β_0. Therefore, in a neighborhood of β_0, the value that maximizes $(1/n) \log \hat{L}(\beta)$ must be close to the value that maximizes L(β), namely β_0, with probability converging to 1 as $n \rightarrow \infty$.

Theorem 2. (Asymptotic Normality)

$$\frac{1}{\sqrt{n}} \frac{d \log \hat{L}(\beta)}{d\beta} \xrightarrow{d} N\left(0, I(\beta) + \frac{(1-\rho^V)^2}{\rho^V} \Sigma\right)$$

where

$$\begin{aligned} I(\beta) &= \rho^V E\left[-\frac{d^2 \log P_\beta(Y \mid X, Z)}{d\beta^2}\right] \\ &\quad + (1-\rho^V) E\left[-\frac{d^2 \log P_\beta(Y \mid Z)}{d\beta^2}\right] \\ &= E\left[-\frac{1}{n}\frac{d^2 \log L(\beta)}{d\beta^2}\right] \end{aligned}$$

and

$$\Sigma = \text{var}\{E[d \log P_\beta(Y \mid Z)/d\beta | X, S]\}.$$

Proof.

$$\begin{aligned} \frac{1}{\sqrt{n}} \frac{d \log \hat{L}(\beta)}{d\beta} &= \frac{1}{\sqrt{n}} \sum_{j \in \bar{V}} \frac{d\hat{P}_\beta(Y_j \mid Z_j)/d\beta}{\hat{P}_\beta(Y_j \mid Z_j)} \\ &\quad + \frac{1}{\sqrt{n}} \sum_{i \in V} \frac{dP_\beta(Y_i \mid X_i, Z_i)/d\beta}{P_\beta(Y_i \mid X_i, Z_i)}. \end{aligned}$$

Let a derivative of P with respect to β be denoted by D, so that

$$D_\beta(Y \mid X, Z) = \frac{dP_\beta(Y \mid X, Z)}{d\beta}, D_\beta(Y \mid Z) = \frac{dP_\beta(Y \mid Z)}{d\beta},$$

$$\text{and} \quad \hat{D}_\beta(Y \mid Z) = \frac{d\hat{P}_\beta(Y \mid Z)}{d\beta}.$$

The score function can be rewritten as

$$\begin{aligned} \frac{1}{\sqrt{n}} \frac{d \log \hat{L}(\beta)}{d\beta} &= \frac{1}{\sqrt{n}} \sum_{j \in \bar{V}} \frac{D_\beta(Y_j \mid Z_j)}{P_\beta(Y_j \mid Z_j)} \\ &\quad + \frac{1}{\sqrt{n}} \sum_{i \in V} \frac{D_\beta(Y_i \mid X_i, Z_i)}{P_\beta(Y_i \mid X_i, Z_i)} \\ &\quad + \frac{1}{\sqrt{n}} \sum_{j \in \bar{V}} \left\{\frac{\hat{D}_\beta(Y_j \mid Z_j)}{\hat{P}_\beta(Y_j \mid Z_j)} - \frac{D_\beta(Y_j \mid Z_j)}{P_\beta(Y_j \mid Z_j)}\right\}. \end{aligned} \tag{A.1}$$

The sum of the first two terms is the score function if $P(X \mid Z)$ were known.

Suppose $n^V(s)$ and $n^{\bar{V}}(s)$ are, respectively, the number of validation and nonvalidation sample members with $S = s$ and $(\mathbf{Y}^{\bar{V}}, \mathbf{Z}^{\bar{V}}, \mathbf{S}^V)$ denotes all outcome and covariate information in the nonvalidation sample and all surrogate covariate information in the validation sample. Then, the third term in (A.1) is

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{j \in \bar{V}} \left\{\frac{\hat{D}_\beta(Y_j \mid Z_j)}{P_\beta(Y_j \mid Z_j)} - \frac{D_\beta(Y_j \mid Z_j)}{[P_\beta(Y_j \mid Z_j)]^2} \hat{P}_\beta(Y_j \mid Z_j)\right\} + O_P\left(\frac{1}{\sqrt{n}}\right) \\ &\overset{a}{=} \frac{1}{\sqrt{n}} \sum_{i \in V} \sum_{j \in \bar{V}} W_{ij} \frac{I(S_j = S_i)}{n^V(S_j)} \\ &= \frac{1}{\sqrt{n}} \sum_{i \in V} \frac{n^{\bar{V}}(S_i)}{n^V(S_i)} \sum_{j \in \bar{V}} W_{ij} I(S_j = S_i)/n^{\bar{V}}(S_i) \\ &\overset{a}{=} \frac{1}{\sqrt{n}} \frac{1-\rho^V}{\rho^V} \sum_{i \in V} \overline{W}_{X_i, S_i}(\mathbf{Y}^{\bar{V}}, \mathbf{Z}^{\bar{V}}, \mathbf{S}^V), \end{aligned}$$

where

$$W_{ij} = \frac{D_\beta(Y_j \mid X_i, Z_j)}{P_\beta(Y_j \mid Z_j)} - \frac{D_\beta(Y_j \mid Z_j)}{(P_\beta(Y_j \mid Z_j))^2} P_\beta(Y_j \mid X_i, Z_j),$$

$$\overline{W}_{X_i, S_i} = \sum_{j \in \bar{V}} W_{ij} \frac{I(S_j = S_i)}{n^{\bar{V}}(S_i)},$$

and where $\overset{a}{=}$ denotes asymptotic equivalence, in the sense that the difference converges to zero in probability.

By the strong law of large numbers, almost surely,

$$\begin{aligned} \overline{W}_{x,s}(\mathbf{Y}^{\bar{V}}, \mathbf{Z}^{\bar{V}}, \mathbf{S}^V) &\rightarrow E\left[\frac{D_\beta(Y \mid x, Z)}{P_\beta(Y \mid Z)} \right. \\ &\qquad \left. - \frac{D_\beta(Y \mid Z)}{(P_\beta(Y \mid Z))^2} P_\beta(Y \mid x, Z) \,\middle|\, S = s\right] \\ &= E\left[\int \frac{D_\beta(Y \mid x, Z) P_\beta(Y \mid Z) dY}{P_\beta(Y \mid Z)} \right. \\ &\qquad \left. - \int \frac{D_\beta(Y \mid Z) P_\beta(Y \mid x, Z) P_\beta(Y \mid Z) dY}{[P_\beta(Y \mid Z)]^2} \,\middle|\, S = s\right] \\ &= -E\left[\frac{D_\beta(Y \mid Z)}{P_\beta(Y \mid Z)} \,\middle|\, X = x, \quad S = s\right]. \end{aligned}$$

Moreover, conditional on $(\mathbf{Y}^{\bar{V}}, \mathbf{Z}^{\bar{V}}, \mathbf{S}^V)$, the $\{\overline{W}_{X_i, S_i}, i \in V\}$ are independent, with variance–covariance matrix converging to $\Sigma(\beta)$ with probability 1, and are uncorrelated with the score contributions of the validation sample members. Asymptotic normality follows using the Lyapounov central limit theorem.

[*Received April 1989. Revised June 1990.*]

REFERENCES

Andersen, P. K., and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics*, 10, 1100–1120.

Anderson, G. A. (1989), "Mismodelling Covariates in Cox Regression," unpublished Ph.D. dissertation, University of Washington, Dept. of Biostatistics.

Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Gailey, K. T., and Abbott, R. D. (1984), "On Errors in Variables for Binary Regression Models," *Biometrika* 71, 19–25.

Carroll, R. J., and Wand, M. P. (in press), "Semiparametric Estimation in Logistic Measurement Error Models," *Journal of the Royal Statistical Society*, Ser. B.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, New York: Chapman & Hall.

Gail, M. H., Wieand, S., and Piantadosi, S. (1984), "Biased Estimates of Treatment Effect in Randomized Experiments With Non-Linear Regressions and Omitted Covariates," *Biometrika*, 71, 431–444.

GAUSS System Version 2.0 (1988), Kent, Washington: Aptech Systems, Inc.

Little, R. J. A., and Rubin, D. B. (1987), *Analysis of Missing Data*, New York: John Wiley.

Pepe, M. S., Self, S. G., and Prentice, R. L. (1989), "Further Results on Covariate Measurement Errors in Cohort Studies With Time to Response Data," *Statistics in Medicine*, 8, 1167–1178.

Rubin, D. B. (1987), *Multiple Imputation for Non-Response in Surveys*, New York: John Wiley.

Schafer, D. W. (1987), "Covariate Measurement Error in Generalized Linear Models," *Biometrika* 74, 385–391.

Stefanski, L. A., and Carroll, R. J. (1985), "Covariate Measurement Error in Logistic Regression," *The Annals of Statistics*, 13, 1335–1351.

Whittemore, A. S., and Keller, J. B. (1988), "Approximations for Regression With Covariate Measurement Error," *Journal of the American Statistical Association*, 83, 1057–1066.