

Conditional and Unconditional Categorical Regression Models with Missing Covariates

Glen A. Satten and Raymond J. Carroll *

December 4, 1999

Abstract

We consider methods for analyzing categorical regression models when some covariates (\mathbf{Z}) are completely observed but other covariates (\mathbf{X}) are missing for some subjects. When data on \mathbf{X} is missing at random (i.e., when the probability that \mathbf{X} is observed does not depend on the value of \mathbf{X} itself), we present a likelihood approach for the observed data which allows the same nuisance parameters to be eliminated in a conditional analysis as when data are complete. An example of a matched case-control study is used to demonstrate our approach.

KEY WORDS: Case-Control Study, Endometrial Cancer, Likelihood; Matching, Missing at Random, Missing Data, Two-stage Sample.

Short title: Matched Studies and Missing Data.

*Glen A. Satten, Centers for Disease Control and Prevention, Atlanta, GA 30333. Raymond J. Carroll, Department of Statistics and Department of Biostatistics & Epidemiology, Texas A&M University, College Station TX 77843-3143, and Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia PA 19104.

1 Introduction

A common problem in categorical data analysis is to determine the effect of explanatory variables \mathbf{V} on a binary outcome \mathbf{D} of interest. In addition, the study design may call for a design in which a conditional analysis is used to eliminate nuisance parameters, such as in a matched analysis. However, some components of \mathbf{V} may not be measured for all study subjects. For example, in a case-control study of endometrial cancer conducted among residents of the Leisure World retirement community, a binary variable denoting obesity was missing in approximately 16 percent of respondents. Hence, a missing-data approach is required for the analysis of this variable. Because the study design matched each case with four controls, the missing-data approach taken must remain valid for highly stratified data.

Data on a subset of covariates may be missing for a variety of reasons. For example, in epidemiologic studies using convenience samples, the medical records used to determine covariate values may not be complete for all participants. In some studies, data on a subset of covariates may be missing by design. This is the case in studies that use two-stage sampling, gathering simple-to-measure covariates on all study participants and then only gathering information on complex or expensive-to-measure covariates on a subset of study participants. In either case, a methodology which allows for different rates of missingness as a function of the observed covariates is highly desirable, especially as ‘complete case’ analyses (analyses in which only data from participants with complete information is used) are known to yield biased results when data are not missing completely at random (Little and Rubin, 1987). In addition, a methodology that allows for elimination of nuisance parameters through conditioning is also essential for analyses of highly stratified data such as matched case-control studies.

Satten and Kupper (1993a, 1993b) developed an approach to categorical regression analyses in which covariate information was missing for some people, and in which surrogate variables were to be used in place of the effect of missing covariate information. This methodology required the non-differential errors approximation, but nuisance parameters could be removed in conditional analyses such as for matched sets. The purpose of this paper is to show that, if surrogate variables are not used, the Satten and Kupper approach is exact, and provides a likelihood-based approach to categorical missing data problems in which some covariates are missing for some study participants. As described in Section 5, our work can be described as a generalization of that of Paik and Sacco (2000).

In Section 2, we state the results of Satten and Kupper, adapted for the missing-data case.

In Section 3 we consider maximum likelihood estimation for unconditional and conditional logistic regression, including matched sets. In Section 4 we present some analyses of the Los Angeles endometrial cancer data described above. Section 5 contains concluding remarks.

2 Model

For the i th study participant, Let D_i be a random variable corresponding to a binary outcome (typically disease for epidemiologic studies), let \mathbf{Z}_i be a column vector of covariates that is observed for all study participants, and let \mathbf{X}_i be a column vector of covariates that is only observed for a subset of study participants. Because of the special role of stratification in conditional logistic regression, we define \mathbf{W}_i to be a vector of always-observed covariables that will be used to define strata. Finally, let S_i denote the stratum that the i th participant is in (S_i may itself be a component of \mathbf{W}_i). Let $\mathbf{V}(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i)$ be the vector of covariates that we wish to include in our model; \mathbf{V} may include arbitrary combinations of the components of \mathbf{X}_i , \mathbf{W}_i and \mathbf{Z}_i , although terms involving components of \mathbf{W}_i alone should be excluded for conditional logistic regression as they are not identifiable. In most cases involving conditional analysis, \mathbf{V} will not be a function of \mathbf{W} because stratification is done to allow the effect of \mathbf{W} to be ignored. Let Δ_i be a binary random variable indicating whether \mathbf{X}_i is missing, so that $\Delta_i = 1$ if \mathbf{X}_i is observed and 0 otherwise. Let

$$\theta(\mathbf{x}, \mathbf{z}, \mathbf{w}) = \frac{\Pr[D = 1 | \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}, \mathbf{W} = \mathbf{w}]}{\Pr[D = 0 | \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}, \mathbf{W} = \mathbf{w}]}, \quad (1)$$

$$\psi(\mathbf{x}, \mathbf{z}, \mathbf{w}; \mathbf{x}', \mathbf{z}', \mathbf{w}') = \theta(\mathbf{x}, \mathbf{z}, \mathbf{w}) / \theta(\mathbf{x}', \mathbf{z}', \mathbf{w}') \quad (2)$$

be odds and odds ratio of disease, respectively, given full information on covariates \mathbf{X} , \mathbf{Z} and \mathbf{W} . The goal of an analysis is to estimate parameters in a model for the odds or odds ratio of D given complete data \mathbf{X} , \mathbf{Z} and \mathbf{W} . If \mathbf{X} , \mathbf{Z} and \mathbf{W} were measured for all study participants, a logistic model for (1) or a conditional logistic model for (2) could be developed using standard methods.

While \mathbf{X} is not measured for all study participants, \mathbf{Z} and \mathbf{W} are, so that we may develop an unconditional regression model using the odds

$$\tilde{\theta}(\mathbf{z}, \mathbf{w}) = \Pr[D = 1 | \mathbf{Z} = \mathbf{z}, \mathbf{W} = \mathbf{w}] / \Pr[D = 0 | \mathbf{Z} = \mathbf{z}, \mathbf{W} = \mathbf{w}], \quad (3)$$

or a conditional regression model using the odds ratio $\tilde{\theta}(\mathbf{z}, \mathbf{w}) / \tilde{\theta}(\mathbf{z}', \mathbf{w}')$. Make the definitions

$$\pi(\mathbf{x} | \mathbf{z}, \mathbf{w}) = \Pr[\mathbf{X} = \mathbf{x} | D = 0, \mathbf{Z} = \mathbf{z}, \mathbf{W} = \mathbf{w}];$$

$$\rho(\mathbf{x} | \mathbf{z}, \mathbf{w}) = \Pr[\mathbf{X} = \mathbf{x} | D = 1, \mathbf{Z} = \mathbf{z}, \mathbf{W} = \mathbf{w}].$$

The first result of Satten and Kupper, restated for the missing-data situation, is that

$$\tilde{\theta}(\mathbf{z}, \mathbf{w}) = \sum_{\mathbf{x}} \theta(\mathbf{x}, \mathbf{z}, \mathbf{w}) \pi(\mathbf{x}|\mathbf{z}, \mathbf{w}), \quad (4)$$

where the sum in (4) is over all possible values of \mathbf{X} . The second result from Satten and Kupper is that

$$\rho(\mathbf{x}|\mathbf{z}, \mathbf{w}) = \frac{\pi(\mathbf{x}|\mathbf{z}, \mathbf{w})\theta(\mathbf{x}, \mathbf{z}, \mathbf{w})}{\sum_{\mathbf{x}} \pi(\mathbf{x}|\mathbf{z}, \mathbf{w})\theta(\mathbf{x}, \mathbf{z}, \mathbf{w})}. \quad (5)$$

If \mathbf{X} is continuous, the sums should be replaced by integrals and $\pi(\mathbf{x}|\mathbf{z}, \mathbf{w})$ and $\rho(\mathbf{x}|\mathbf{z}, \mathbf{w})$ should be considered probability density functions. Finally, these results are easily generalized to outcome measures with more than one category.

Equation (5) has important implications. The functions $\theta(\cdot)$ and hence $\psi(\cdot)$ are functions only of the disease model, and $\pi(\cdot)$ is a modeling assumption among the controls. Equation (5) shows that $\rho(\cdot)$ is completely specified by the disease model and the missing data model among the controls. There is thus no flexibility in specifying $\rho(\cdot)$ once $\theta(\cdot)$ and $\pi(\cdot)$ are specified.

3 Maximum Likelihood Estimation with Missing Covariates

The results of Section 2 can be used to construct a likelihood for parameters of interest, using only the observed data. A strength of this approach is that it allows the conditional analyses commonly used in epidemiologic studies to be extended to problems in which some covariates are missing.

We assume a logistic model in which $\theta(\mathbf{x}, \mathbf{z}, \mathbf{w}) = \exp\{\beta_0(\mathbf{w}) + \mathbf{V}^T(\mathbf{x}, \mathbf{z}, \mathbf{w})\boldsymbol{\beta}\}$. In unconditional analyses, the goal is maximum likelihood estimation of the $\beta_0(\mathbf{w})$'s as well as $\boldsymbol{\beta}$. In conditional analyses, usually we assume a separate intercept for each stratum, in which case we take $\beta_0(\mathbf{w}) = \beta_{0j}$ for the j th stratum (formally, take S to be one of the components of \mathbf{W} and let $\beta_0(\mathbf{w})$ depend only on this component). In conditional analyses, the goal is estimation of $\boldsymbol{\beta}$ while eliminating the nuisance parameters β_{0j} . In the full data case, conditional logistic regression eliminates β_{0j} by conditioning on the number of persons with disease in each stratum; the resulting conditional likelihood is independent of β_{0j} (Breslow and Day, 1980). We show here that a similar approach is possible with missing data.

We write the likelihood of the full data as

$$\Pr(D, \mathbf{X}, \Delta|\mathbf{Z}, \mathbf{W}) = \Pr(D|\mathbf{Z}, \mathbf{W})\Pr(\Delta|D, \mathbf{Z}, \mathbf{W})\Pr(\mathbf{X}|D, \mathbf{Z}, \Delta, \mathbf{W}), \quad (6)$$

and assume that $\Pr(\mathbf{X}|D, \mathbf{Z}, \Delta, \mathbf{W}) = \Pr(\mathbf{X}|D, \mathbf{Z}, \mathbf{W})$, i.e. that the data are missing at random in the sense of Little and Rubin (1987). We assume that the missing data probability $\Pr(\Delta|D, \mathbf{Z}, \mathbf{W})$

does not depend on $\beta_0(\mathbf{w})$ or β , although the score equations we derive below will have mean zero even if this is not the case. For the observed data, the unconditional likelihood, excluding terms involving $\Pr(\Delta|D, \mathbf{Z}, \mathbf{W})$, can be written as

$$\begin{aligned} L^U &= \prod_{i=1}^N \tilde{\theta}(\mathbf{Z}_i, \mathbf{W}_i)^{d_i} \{1 + \tilde{\theta}(\mathbf{Z}_i, \mathbf{W}_i)\}^{-1} \pi(\mathbf{X}_i|\mathbf{Z}_i, \mathbf{W}_i)^{\delta_i(1-d_i)} \rho(\mathbf{X}_i|\mathbf{Z}_i, \mathbf{W}_i)^{\delta_i d_i} \\ &= \prod_{i=1}^n \frac{\theta(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i)^{\delta_i d_i} \left\{ \sum_{\mathbf{x}} \theta(\mathbf{x}, \mathbf{Z}_i, \mathbf{W}_i) \pi(\mathbf{x}|\mathbf{Z}_i, \mathbf{W}_i) \right\}^{d_i(1-\delta_i)} \pi(\mathbf{X}_i|\mathbf{Z}_i, \mathbf{W}_i)^{\delta_i}}{1 + \sum_{\mathbf{x}} \theta(\mathbf{x}, \mathbf{Z}_i, \mathbf{W}_i) \pi(\mathbf{x}|\mathbf{Z}_i, \mathbf{W}_i)}. \end{aligned}$$

For this model, the score equations for the $\beta_0(\mathbf{w})$ and β are

$$\begin{aligned} \sum_i I(\mathbf{W}_i = \mathbf{w}) \{d_i - \hat{d}_u(\mathbf{Z}_i, \mathbf{W}_i)\} &= 0; \\ \sum_i \left\{ \mathbf{V}(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i) d_i \delta_i + \hat{\mathbf{V}}(\mathbf{Z}_i, \mathbf{W}_i) d_i (1 - \delta_i) - \hat{\mathbf{V}}(\mathbf{Z}_i, \mathbf{W}_i) \hat{d}_u(\mathbf{Z}_i, \mathbf{W}_i) \right\} &= 0, \end{aligned}$$

where $\hat{\mathbf{V}}(\mathbf{Z}_i, \mathbf{W}_i)$ is the column vector with j th component $\sum_{\mathbf{x}} V(\mathbf{x}, \mathbf{Z}_i, \mathbf{W}_i) \rho(\mathbf{x}|\mathbf{Z}_i, \mathbf{W}_i)$ and $\hat{d}_u(\mathbf{Z}_i, \mathbf{W}_i) = \tilde{\theta}(\mathbf{Z}_i, \mathbf{W}_i) / \{1 + \tilde{\theta}(\mathbf{Z}_i, \mathbf{W}_i)\} = \Pr(D = 1 | \mathbf{Z}_i, \mathbf{W}_i)$.

As with the full data case, we construct a conditional likelihood by dividing the unconditional likelihood by the probability that n_{1j} persons with disease are observed in the j th stratum conditional on observed values of \mathbf{Z} and \mathbf{W} and $n_j = \sum_i I(s_i = j)$, for each j . This is easily accomplished because of the factorization (6) that we have used. The resulting conditional likelihood L^C is given by

$$L^C = \left\{ \prod_{j=1}^J L_j^C \right\} \prod_{i=1}^n \pi(\mathbf{X}_i|\mathbf{Z}_i, \mathbf{W}_i)^{\delta_i(1-d_i)} \rho(\mathbf{X}_i|\mathbf{Z}_i, \mathbf{W}_i)^{\delta_i d_i}, \text{ where} \quad (7)$$

$$L_j^C = \frac{\prod_{\{i|S_i=j\}} \tilde{\theta}(\mathbf{Z}_i, \mathbf{W}_i)^{d_i}}{\sum_{\mathbf{d}'_j} \prod_{\{i|S_i=j\}} \tilde{\theta}(\mathbf{Z}_i, \mathbf{W}_i)^{d'_{ji}}} \quad (8)$$

and where \mathbf{d}'_j is a vector of length n_j that has n_{1j} elements equal to one and $n_{0j} = n_j - n_{1j}$ elements equal to zero.

Because $\tilde{\theta}(\mathbf{z}, \mathbf{w})$ is linear in $\theta(\mathbf{x}, \mathbf{z}, \mathbf{w})$, each factor of $\tilde{\theta}(\mathbf{z}, \mathbf{w})$ is proportional to $\exp(\beta_{0j})$. Since there are an equal number of terms $\tilde{\theta}(\mathbf{z}, \mathbf{w})$ in the numerator and denominator of (8), we see L_j is not a function of the nuisance parameters β_{0j} . A similar argument using (5) shows that $\rho(\mathbf{x}|\mathbf{z}, \mathbf{w})$ is also not a function of β_{0j} . Hence, the overall conditional likelihood is not a function of the β_{0j} 's as claimed. An alternate way to see this is to recognize that the L_j 's and $\rho(\mathbf{x}|\mathbf{z}, \mathbf{w})$ can be expressed entirely in terms of odds ratios $\psi(\mathbf{x}, \mathbf{z}, \mathbf{w}; \mathbf{x}', \mathbf{z}', \mathbf{w}')$ and $\pi(\mathbf{x}|\mathbf{z}, \mathbf{w})$. The score equations for β in the conditional model are identical to those in the unconditional model, except that $\hat{d}_u(z)$

is replaced by the conditional probability $\hat{d}_c(z) = \Pr[D = 1 | \mathbf{Z} = \mathbf{z}, \mathbf{W} = \mathbf{w}, n_{1j}, n_j]$ calculated using the extended hypergeometric distribution (8).

Score equations for the parameters in $\pi(\mathbf{x}|\mathbf{z}, \mathbf{w})$ can be derived once a model for π is chosen. A particularly attractive case is when \mathbf{X} , \mathbf{Z} and \mathbf{W} take only finitely many values. In this case the model $\pi(\mathbf{x}|\mathbf{z}, \mathbf{w}) = \exp(\gamma_{xzw}) / \sum_{x'} \exp(\gamma_{x'zw})$ yields the score equations

$$\sum_i I(\mathbf{Z}_i = \mathbf{z}, \mathbf{W}_i = \mathbf{w}) \left[\{d_i - \hat{d}_{u,c}(\mathbf{Z}_i, \mathbf{W}_i)\} \{\rho(\mathbf{x}|\mathbf{z}, \mathbf{w}) - \pi(\mathbf{x}|\mathbf{z}, \mathbf{w})\} - \delta_i d_i \{I(\mathbf{X}_i = \mathbf{x}) - \rho(\mathbf{x}|\mathbf{z}, \mathbf{w})\} - \delta_i (1 - d_i) \{I(\mathbf{X}_i = \mathbf{x}) - \pi(\mathbf{x}|\mathbf{z}, \mathbf{w})\} \right] = 0.$$

Estimates of the variance-covariance matrix of the β and γ parameters can be obtained by inverting the observed information matrix. In our analyses, the observed information matrix was calculated numerically as the Jacobian of the score equations.

Our methods require a relative risk model $\theta(\mathbf{x}, \mathbf{z}, \mathbf{w})$ and a model $\pi(\mathbf{x}|\mathbf{z}, \mathbf{w})$ for the missing covariate among the controls. Because this determines the model $\rho(\mathbf{x}|\mathbf{z}, \mathbf{w})$ for the missing covariate among the cases, and because $\rho(\mathbf{x}|\mathbf{z}, \mathbf{w})$ depends on parameters β in the relative risk model through equation (5), when all study participants have both \mathbf{X} and (\mathbf{Z}, \mathbf{W}) available, our analyses do not reduce to those obtained using standard logistic or conditional logistic regression. This is discussed further in the context of an example in Section 4.

Note that in standard logistic regression, it is possible to develop a likelihood-based approach to this missing data problem that does reduce to standard logistic regression when no data is missing. This can be accomplished by using a different factorization of the likelihood than (6) in which the model for the missing \mathbf{X} given (\mathbf{Z}, \mathbf{W}) applies to the entire population. However, when this approach is taken in matched studies, the resulting conditional likelihood no longer eliminates the nuisance parameters $\beta_0(\mathbf{W})$ when data are missing.

4 Example: The L. A. Endometrial Cancer Case Control Study

The Los Angeles Endometrial Cancer case control study is a 4:1 matched study of the effect of various risk factors on endometrial cancer, conducted among residents of the Leisure World retirement community (Mack et al. 1976). These data were used by Breslow and Day (1980) in their book *Statistical Methods in Cancer Research*, a standard reference for this field. Even in this ‘textbook’ example, one variable (obesity) is missing in 6 of 63 (9.5%) of cases and 45 of 252 (17.9%) of controls. Although this difference is not statistically significant ($p = 0.13$ using Fisher’s exact test), the magnitude of the difference suggests a method which allows for different rates of

missing data between cases and controls should be used in analyzing these data.

In this section, we present several analyses of these data to illustrate a number of aspects of our approach. To begin, we consider several analyses using the binary variables GALL (history of gall bladder disease) and EST (history of use of estrogen therapy). These two variables were selected as they are the only significant predictors of endometrial cancer in these data. Both GALL and EST are available for all study participants; in some of our analyses, we artificially removed some data on EST values for some study participants. A third variable, OB (a binary variable indicating obesity) is missing in approximately 16% of study participants. A ‘complete case’ analysis indicates OB is not a significant predictor of endometrial cancer; we apply our approach to add OB to GALL and EST to determine if this finding holds up in a more complete analysis. In all analyses, the conditional likelihood (7)-(8) was used along with model (10) for $\pi(\mathbf{x}|\mathbf{z}, \mathbf{w})$. As Breslow and Day (1980) do not provide the values of the matching variables we take \mathbf{w}_i to have only one component corresponding to S_i and assume $\pi(\mathbf{x}|\mathbf{z}, \mathbf{w})$ is independent of \mathbf{w} .

Because the approach presented here does not reduce to the usual conditional logistic regression analysis when no data is missing, it is interesting to compare the results obtained using complete data, while interchanging which variables are considered the X and Z variables. We conducted two analyses using the binary variables GALL and EST. The results are summarized in Table 1, and show that the differences between point estimates for the two analyses are small compared to the standard errors of the point estimates. Neither analysis agrees exactly with the usual conditional maximum likelihood analysis, but again differences are small compared to the statistical uncertainty in the point estimates.

To assess the effect of missing data in a situation where we have an idea of the true value with full data, we removed some of the information on EST values and reanalyzed the data using the conditional likelihood approach described in Section 3. Among cases, participants with EST= 1 were removed with probability 0.20 while those with EST= 0 were all kept. Among controls, persons with GALL= 1 were removed with probability 0.10; among controls with GALL= 0, persons with EST= 0 were removed with probability 0.30 and those with EST= 1 were removed with probability 0.40. We generated a data set using these probabilities, which resulted in removal of EST values from 97 (31%) of study participants. We then fit a saturated model using our conditional likelihood approach; the results of this analysis are shown in column 4 of Table 1. We also fit a ‘complete case’ analysis using standard conditional maximum likelihood; these results are shown in column 5 of Table 1. As expected, the performance of our new conditional likelihood

estimators is superior to the standard complete case cmle in both bias and variance.

Because some study participants have missing OB values, the effect of obesity on the occurrence of endometrial cancer can only be established by standard methodology using a ‘complete case’ analysis. Using the methods of Section 3, we can fit a model which includes GALL and EST as components of \mathbf{Z} and OB as the sole component of \mathbf{X} , including main effects and all second-order interactions. The results of this analysis, along with the complete case analysis using standard conditional logistic regression, are shown in Table 2. In this case, the results of the complete case analysis are supported.

5 Discussion

In many situations where logistic regression or conditional logistic regression are to be used to determine the effect of explanatory variables on a binary outcome, some of the explanatory variables are only available for a subset of study participants. We have demonstrated that a modification of the approach of Satten and Kupper (1993a) applies in this case, allowing likelihood-based inference for this type of data. This approach also has the advantage that a conditional likelihood can be easily constructed that allows elimination of the same nuisance parameters through conditioning as are eliminated in the complete data case. As an example, we applied our approach to a case-control study in which each case was matched to four controls.

Besides the basic logistic model relating disease to covariates, the only thing that needs to be specified is a density or mass function $\pi(\mathbf{x}|\mathbf{z}, \mathbf{w})$ for the missing covariates given the observed covariates *among the controls*. Our work, see equation (5), shows that the two together completely specify the density/mass function $\rho(\mathbf{x}|\mathbf{z}, \mathbf{w})$ of the missing covariates given the observed covariates among the cases. Typically, the latter can have a complex form, but if $\pi(\mathbf{x}|\mathbf{z}, \mathbf{w})$ is the density/mass function of a univariate member of the class of canonical exponential families of distributions, then $\rho(\mathbf{x}|\mathbf{z}, \mathbf{w})$ is from the same canonical exponential family, differing only by an offset (Wang, et al., 1997; Paik and Sacco, 2000).

In conditional logistic regression, strata will be sparsely occupied by design. Hence, it is not possible to model the effect of stratum in $\pi(\mathbf{x}|\mathbf{z}, \mathbf{w})$ without additional information. Fortunately, strata are usually defined by matching variables \mathbf{w} which can be included in the model for π . In this way, data from strata or matched sets with similar values of \mathbf{w} combine to estimate $\pi(\mathbf{x}|\mathbf{z}, \mathbf{w})$. For example, if matching was done on 5-year age group, and if there were enough sets in each age category, a separate model for $\pi(\mathbf{x}|\mathbf{z}, \text{age group})$ could be constructed in each age group. Note

that without matching variables \mathbf{w} , the sparseness would make it impossible to model $\pi(\mathbf{x}|\mathbf{z}, S)$ in each stratum.

Paik and Sacco (2000) have proposed an estimating equation approach to two-stage sampling for pair-matched case-control studies. They make the same parametric assumptions that we do, except that they require that $\pi(\mathbf{x}|\mathbf{z}, \mathbf{w})$ be the density or mass function of a canonical exponential family, while we allow for any distribution. While their approach reduces to the usual full data analysis when all study participants have full data, and thus may be preferable in that sense, it is not clear that their approach is optimally efficient, since they are, in effect, throwing away the information contained in equation (5). The approach given here should enjoy optimal efficiency properties, as it is based on standard likelihood methods. Besides the optimality properties, our work can be described as a second generalization of theirs, because their methods do not appear to allow for interactions between \mathbf{X} and \mathbf{Z} .

Finally, although we have only presented results for binary outcomes, all results in Sections 2-3 are easily extended to polychotomous or longitudinal categorical outcomes, allowing for a likelihood-based treatment of missing data in these situations.

Acknowledgments

Carroll's research was supported by a grant from National Cancer Institute (CA-57030), and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). The authors thank Myounghee Cho Paik for useful conversations and for pointing out the importance of this problem.

References

- Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research, Volume 1 - The Analysis of Case-Control Studies*. International Agency for Research on Cancer: Lyon, France.
- Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analyses with Missing Data*. New York: John Wiley & Sons.
- Mack T. M., Pike, M. C., Henderson, B. E., Pfeiffer, R. I., Gerkins, V. R., Arthur, B. S., and Brown, S. E. (1976). Estrogens and endometrial cancer in a retirement community. *New Eng. J. Med.*, 294, 1262-1267.
- Paik, M. C. and Sacco R. (2000). Matched case-control data analyses with missing covariates. *Applied Statistics*, to appear.
- Satten, G. A. and Kupper, L. (1993a). Inferences about exposure-disease associations using probability-of-exposure information. *Journal of the American Statistical Association*, 88, 200-208.
- Satten, G. A. and Kupper, L. (1993b). Conditional regression analysis of the odds ratio between

two binary variables when one is not measured with certainty: a method for epidemiologic studies. *Biometrics*, 49, 429-440.

Wang, C. Y., Wang, S. & Carroll, R. J. (1997). Estimation in choice-based sampling with measurement error and bootstrap analysis. *Journal of Econometrics*, 77, 65-86.

| | Full Data Analysis | | | Missing Data Analysis | |
|-------------------|---------------------|---------------------|--------------------|-----------------------|--------------------------------|
| | X = EST Z = GALL | X = GALL Z = EST | Standard CMLE | X = EST Z = GALL | Complete Case Standard CMLE |
| Variable | | | | | |
| GALL | 3.021 (0.854) | 2.970 (0.847) | 2.894 (0.883) | 2.876 (0.843) | 2.139 (0.908) |
| EST | 2.715 (0.612) | 2.725 (0.612) | 2.700 (0.612) | 2.615 (0.626) | 2.425 (0.632) |
| GALL \times EST | -2.230 (0.943) | -2.230 (0.943) | -2.0523 (0.995) | -2.163 (0.964) | - 1.759 (1.094) |

Table 1: Analyses of GALL and EST from Los Angeles endometrial cancer data. Columns 1-3 illustrate the asymmetric treatment of X and Z when data on X is available from all study participants. Columns 4-5 compare the results of our proposed conditional maximum likelihood estimators with a ‘complete case’ analysis using standard conditional maximum likelihood estimators, for data in which approximately 30% of EST values have been artificially removed.

| | CMLE | Complete Case |
|-------------------|-------------------|-------------------|
| OB | 1.411 (1.405) | 1.442 (1.385) |
| GALL | 3.251 (1.188) | 3.245 (1.283) |
| EST | 3.465 (1.366) | 3.510 (1.401) |
| OB \times GALL | -0.186 (0.886) | -0.142 (0.915) |
| OB \times EST | -0.884 (1.393) | -1.098 (1.393) |
| GALL \times EST | -2.317 (1.056) | -2.261 (1.168) |

Table 2: Comparison of Proposed CMLE with complete case analysis using standard conditional maximum likelihood.