A Semiparametric Mixture Approach to Case-Control Studies With Errors in Covariates

Kathryn Roeder, Raymond J. Carroll, and Bruce G. Lindsay

Methods are devised for estimating the parameters of a prospective logistic model in a case-control study with dichotomous response $D$ that depends on a covariate $X$. For a portion of the sample, both the gold standard $X$ and a surrogate covariate $W$ are available; however, for the greater portion of the data, only the surrogate covariate $W$ is available. By using a mixture model, the relationship between the true covariate and the response can be modeled appropriately for both types of data. The likelihood depends on the marginal distribution of $X$ and the measurement error density $(W | X, D)$. The latter is modeled parametrically based on the validation sample. The marginal distribution of the true covariate is modeled using a nonparametric mixture distribution. In this way we can improve the efficiency and reduce the bias of the parameter estimates. The results also apply when there is no validation data provided the error distribution is known or estimated from an independent data source. Many of the results also apply to the easier case of prospective sampling.

1. INTRODUCTION

In this article we examine logistic case-control studies with errors in covariates, using a semiparametric mixture model approach. Although we use case-control studies to illustrate our points, many of the results and methods apply more generally.

To study the relationship between disease status and exposure level to a suspected disease causing agent $(X)$, epidemiologists often use retrospective sampling in which the diseased population (the "cases," with $D = 1$) and the disease free population (the "controls," with $D = 0$) are sampled separately to determine their levels of exposure $X$. Viewed from the perspective of the joint population of $(X, D)$, we are taking observations from the conditional distributions of $X|D$.

Suppose that the probability of disease in the source population can be described by the prospective logistic model, $Pr(D = 1|X = x) = \kappa(x) = \frac{1 + \exp(-\beta_0 - \beta_1x)}{1}$, and the marginal density of $X$ in the population is $g(x)$, which will be modeled as an unknown density. Then the population is described by parameters $(\beta_0, \beta_1, g)$. It is well known that standard logistic regression, performed as if $D$ were the dependent variable and the covariates $X$ were fixed, leads to the maximum likelihood (ML) estimate of $\beta_1$ for retrospective sampling (Prentice and Pyke 1979). Unless the prevalence of disease in the source population $(\phi)$ is known, $\beta_0$ must be viewed as a nuisance parameter. In Section 2 we shed new light on this well-known result by demonstrating that the retrospective model is identifiable up to a specific equivalence class. If $\phi$ is unknown, then only $\beta_1$ is identifiable. The remaining parameters, $\beta_0$ and $g$, are linked by $\phi$.

Epidemiologists often use a validation design, in which some of the data are "complete" in that the covariates are measured both directly (without error) and indirectly (with error), whereas for the remainder of the data, the covariates are measured only indirectly. The latter sample is called the "reduced" or "incomplete" sample. Let $W$ and $X$ denote the covariates measured with and without error. Carroll, Gail, and Lubin (1993) extended the use of the prospective logistic model to account for this errors-in-variables design. Other papers in the area include the work of Armstrong, Whitemore, and Howe (1989), Buonaccorsi (1990), and Satten and Kupper (1993).

A motivating example that we analyze in Section 8 concerns the effect of low-density lipoprotein (LDL) cholesterol $(X)$ on the probability of heart disease $(D)$. Consider a design in which the case-control sample is split into a group of size $n_C$ and another group of size $n_R$, independent of LDL level. Suppose that total cholesterol $(W)$ and LDL are measured for $n_C$ individuals, of which $n_{CL}$ are cases and $n_{CL}$ are controls. Because LDL is expensive to measure relative to total cholesterol, only total cholesterol is measured in the remaining $n_R$ individuals. The complete and reduced data sets are $\{X_i, W_i, D_i, i = 1, \ldots, n_C\}$ and $\{W_j, D_j, j = 1, \ldots, n_R\}$.

Following Carroll et al. (1993), we assume a parametric conditional distribution for $W$, given $X$ and $D$, denoted by $f_W(W | X, D)|\omega, \phi|\alpha$. We choose a parametric model for the measurement error distribution for two reasons: (a) there is opportunity to test this assumption through the validation data, and (b) in many problems, other studies will also have assessed this assumption, and the error distribution may be transportable from study-to-study. On the other hand, $G$, the distribution of $X$ corresponding to density $g$, is left unspecified, largely because this distribution depends on the source population.

A natural way to model errors-in-variables data is with a semiparametric mixture model (Kiefer and Wolfowitz 1956). Let $h_j(\xi; \alpha, \beta)$ denote the conditional likelihood of $(W_j, D_j | X = \xi)$. For the reduced data, it is immediately obvious that the joint likelihood of $(W_j, D_j)$ can be written in the form of a mixture model, $L_j(\alpha, \beta, G)$.
\[ \int h_i(\xi; \alpha, \beta) \, dG(\xi) \] 

The joint likelihood of the complete data \((X_i, W_i, D_i)\) can also be written in the form of a mixture model, 

\[ L_0(\alpha, \beta, G) = \int h_i(\xi; \alpha, \beta) I(x_i = \xi) \, dG(\xi), \]

with the use of an indicator function.

But the terms in the conditional (retrospective) likelihood are not of the form of a mixture model. In Section 3 we show that if \( \theta \) maximizes the joint likelihood, then it lies in an equivalence class of parameters that also maximize the retrospective likelihood. Hence maximizing the joint likelihood is equivalent to maximizing the retrospective likelihood. Thus a semiparametric mixture model approach can be taken to obtain ML estimates after all. Specifically, the retrospective ML estimate of the parameter of interest, \( \beta_1 \), can be obtained by maximizing the joint likelihood over the parametric \((\alpha, \beta)\) and nonparametric \((G)\) components of the likelihood. A confidence interval for \( \beta_1 \) is obtained from the profile likelihood. Assuming the likelihood satisfies the regularity conditions specified by Kiefer and Wolfowitz (1956), the profile ML estimator, \( \hat{\beta}_1 \), is consistent. In certain models with conditional structure, the estimates of the parametric component are known to be asymptotically efficient (Lindsay, Clogg, and Grego 1991; van der Vaart, in press). Although many simulation studies have shown high efficiency in other models (e.g., Lesperance 1989), a general theory of efficiency has not yet been developed.

Because \( X \) was not measured in the reduced sample, the problem can be viewed as a missing-data problem. In Section 4 we present an EM algorithm to obtain the ML estimate of \( G \) and also note gradient-based algorithms appropriate for semiparametric mixture models.

In Section 5 we partially extend the results to models for which the probability of disease depends on additional covariates. In Section 6 we develop an extension to a study design without direct validation data. Suppose that no complete data are available in the sample, but that an independent study has been conducted in which both \( X \) and \( W \) have been measured. Although none of the other measurement error models can handle this situation, the semiparametric mixture method can analyze data of this form. This model requires that the measurement error be nondifferential (i.e., the distribution of \( W \mid X, D \) does not depend on \( D \)).

When validation data are available, a natural competitor to the mixture method is the pseudolikelihood method proposed by Carroll et al. (1993). They estimated the marginal distribution of \( X \) using a weighted average of the empirical distributions of \( X \mid D = d \) obtained from the complete data. This estimate is plugged into the likelihood, from which the maximum pseudolikelihood estimates of the remaining parameters is obtained. By modeling the relationship between \( W \) and \( D \), using a rough-and-ready estimate of the unobserved distribution of \( X \), the information about \( \beta_1 \) contained in the reduced data can be partially recovered. But because the distribution of \( W \) depends on \( X \), there is some additional information in the reduced sample about the distribution of \( X \). Consequently, jointly maximizing the full likelihood should yield more information about \( \beta_1 \) than is available when only the complete data are used to estimate the distribution of \( X \).

In Section 7 we present a simulation experiment that evaluates the performance of the mixture method for various sample sizes and amounts of measurement error. The mixture method always performs as well or better than the pseudolikelihood method. The amount of improvement depends on the sample size and the amount of error in the measurements. In Section 8 we analyze an epidemiological data set and present profile likelihoods.

2. IDENTIFIABILITY

Throughout this section we treat \( X \) as a discrete random variable for ease of exposition; however, all of the results extend directly to the continuous case. Assume that the probability of disease in the source population for a given level of exposure can be modeled by the prospective logistic model \( K(x) \). Then it is readily determined that

\[
\begin{align*}
    f_D(d) &= \left\{ \sum_x K(x) \phi(x) \right\}^d \left\{ \sum_x \tilde{K}(x) \phi(x) \right\}^{1-d}, \\
    \Pr(X = x \mid D = d) &= \frac{K(x)^d \tilde{K}(x)^{1-d} \phi(x)}{\left( \sum_x K(x) \phi(x) \right)^d \left( \sum_x \tilde{K}(x) \phi(x) \right)^{1-d}}.
\end{align*}
\]

For the \( p \)-dimensional parameter \( \beta_1 \) to be identifiable from this distribution, it will be assumed that \( g(x_i) > 0 \) for some set \( x_0, x_1, \ldots, x_p \) of affinely independent vectors. The following lemma provides useful information for the consideration of identifiability issues and, later, for maximizing the likelihood.

**Lemma 1.** Suppose that \((\beta_0, \beta_1, g)\) and \((\beta_0^*, \beta_1^*, g^*)\) are two logistic regression models satisfying

\[ \Pr(D = 1; \beta_0, \beta_1, g) = \phi \]

and

\[ \Pr(D = 1; \beta_0^*, \beta_1^*, g^*) = \phi^*. \]

Then

\[ \Pr(X = x \mid D = d; \beta_0, \beta_1, g) = \Pr(X = x \mid D = d; \beta_0^*, \beta_1^*, g^*) \]

if and only if

\begin{enumerate}
  \item \( \beta_1 = \beta_1^* \);
  \item \( \beta_0 = \beta_0 + \log(\phi / \phi^*) \); and
  \item \( g^*(x) = \frac{[1 + \exp(\beta_0^* + \beta_1^* x)]/[1 + \exp(\beta_0 + \beta_1 x)]}{\sum_x [1 + \exp(\beta_0^* + \beta_1^* x)]/[1 + \exp(\beta_0 + \beta_1 x)]} g(x), \)
\end{enumerate}

where \( \phi = 1 - \phi \) and \( \phi^* = 1 - \phi^* \).

**Proof.** See the Appendix.

Thus we see that from the retrospective sample, only the parameter \( \beta_1 \) is fully identifiable. The marginal distribution
of \( X \) can be determined only up to an equivalence class of functions. But if the true population probability of disease, \( \phi \), is otherwise known, then the foregoing formulas show that \( \beta_0 \) and \( g \) are identifiable.

3. THE ERRORS-IN-VARIABLES MODEL

We assume that the true covariates and the error-prone measurements are available in a validation study consisting of a random sample of \( n_{C1} \) cases and \( n_{C0} \) controls. Thus the complete data consist of \( n_{C} = n_{C0} + n_{C1} \) observations, \{ \( X_i, W_i, D_i \), \( i = 1, \ldots, n_C \} \). In addition, we have \( n_{R} = n_{R0} + n_{R1} \) incomplete or reduced observations, \{ \( W_j, D_j \), \( j = 1, \ldots, n_R \} \), obtained from a random sample of \( n_{R1} \) cases and \( n_{R0} \) controls. The number of cases overall is \( n_1 = n_{C1} + n_{R1} \), and the number of controls is \( n_0 = n_{C0} + n_{R0} \). It is assumed that the data are missing at random (Little and Rubin 1987).

Because we are assuming that the division of the cases and controls into the reduced and complete subsamples is done conditionally on \( D \) but independently of the values of \( X \), the conditional distribution of \( X = x \) given \( D = d \) and the subsample information is still just the conditional distribution of \( X = x \) given \( D = d \) that was described by (2) in the previous section: \( \Pr(X = x | D = d) = f_{X,D}(x, d) / f_D(d) \), where \( f_{X,D}(x, d) = \kappa(x)^d \kappa(x)^{1-d} g(x) \). The marginal distribution of \( D \), given in (1), can be reexpressed in the form of a mixture,

\[
f_D(d) = \left\{ \int \kappa(x) \, dG(x) \right\}^d \left\{ \int \kappa(x) \, dG(x) \right\}^{1-d},
\]

where \( G \) may be discrete or continuous. Recall from Lemma 1 that there is an equivalence class of choices for \( \{ \beta_0, g \} \) for which the same retrospective model is obtained. Now \( \phi = f_D(1) \) identifies a particular pair. If it were necessary to evaluate \( f_D(1) \) for each subsample (reduced and complete) separately, then it would be necessary to incorporate a differential shift for each \( \beta_0 \) and to estimate a different \( g \) for each subsample. The following argument shows why this is not necessary.

In the reduced sample, the likelihood contribution from an observation \( \{ W, D \} \) is \( f_{W,D}(w, d) = f_{W,D}(w, d) / f_D(d) \), where

\[
f_{W,D}(w, d) = \int h(x) \, dG(x),
\]

and

\[
h(x) = \kappa(x)^d \kappa(x)^{1-d} f_{W,X,D}(w, x, d).
\]

In the complete subsample, the likelihood contribution from an observation \( \{ W, X, D \} \) is \( f_{W,X,D}(w, x, d) = f_{W,X,D}(w, x, d) / f_D(d) \), where

\[
f_{W,X,D}(w, x, d) = \int h(x) I \{ x = x \} \, dG(x).
\]

Note the device of using an indicator function so that both (5) and (7) are written as integrals \( dG(x) \).

Putting these together, we get that the likelihood to maximize is of the form

\[
L_C(\alpha, \beta, G) = \frac{L_1(\alpha, \beta, G)}{L_M(\beta, G)},
\]

where

\[
L_1(\alpha, \beta, G) = \prod_{j=1}^{n_R} f_{W,D}(w_j, d_j) \prod_{i=1}^{n_C} f_{W,X,D}(w_i, x_i, d_i)
\]

and

\[
L_M(\beta, G) = \{ f_D(1) \}^{m_1} \{ f_D(0) \}^{m_0}.
\]

In the foregoing, \( C, J \), and \( M \) stand for conditional, joint, and marginal. Note that as functions of \( G \), both \( L_1 \) and \( L_M \) have the product-of-integrals form \( \prod_i f_i(\xi) \, dG(\xi) \), for some nonnegative functions \( f_i(\xi) \). It follows that the conditional likelihood \( L_{C|}(\alpha, \beta, G) \) is a ratio of two products of integrals. Consequently, we cannot apply the theory of non-parametric mixture models to the retrospective likelihood. Here now we run into some good fortune, as we can simply estimate the parameters from \( L_1(\alpha, \beta, G) \), which does have mixture form. The parameter estimates that we so obtain will be in an equivalence class of estimators that maximize \( L_{C|}(\alpha, \beta, G) \), being that member of that equivalence class that fits the observed fractions of cases and controls. Moreover, the profile likelihood for \( \beta_1 \) from \( L_1 \) is the same as from \( L_C \).

We state the result in a general manner, as it has applications in other problems (see, e.g., Lindsay et al. 1991). Suppose that for a two-parameter model \( (\theta, \phi) \), we have a likelihood decomposition of the form

\[
L_C(\theta, \phi) = \frac{L_1(\theta, \phi)}{L_M(\theta, \phi)},
\]

where \( L_M(\theta, \phi) \) is of multinomial form:

\[
L_M(\theta, \phi) = \prod_{k} p_k(\theta, \phi)^{m_k},
\]

where \( m = m_1 + \cdots + m_K \) and \( p_k(\theta, \phi) = p_k(\theta, \phi) \). In this setting we have the following simple theorem.

**Theorem 2.** Let \( \Omega \) be the set of all \( (\theta, \phi) \) in the parameter space such that there exists \( \phi^* \) depending on \( (\theta, \phi) \) satisfying (a) \( L_C(\theta, \phi) = L_C(\theta, \phi^*) \); and (b) \( p_k(\theta, \phi^*) = m_k / m \). Suppose that in the set of \( (\theta, \phi) \) maximizing \( L_C(\theta, \phi) \), there exists \( (\theta_C, \phi_C) \) in \( \Omega \). Then the following hold:

1. \( (\tilde{\theta}_C, \tilde{\phi}_C) \) maximizes \( L_D(\theta, \phi) \).
2. If \( (\tilde{\theta}_1, \tilde{\phi}_1) \) is any maximizer of \( L_D(\theta, \phi) \), then it satisfies (b) and also maximizes \( L_C(\theta, \phi) \).

Moreover, if the entire parameter space is in \( \Omega \), then \( L_D \) and \( L_C \) generate the same profile likelihoods for \( \theta \).

**Proof.** See the Appendix.

**Corollary 3.** If \( (\alpha, \beta, G) \) maximizes \( L_1(\alpha, \beta, G) \), then \( (\tilde{\alpha}, \tilde{\beta}, \tilde{G}) \) also maximizes \( L_C(\alpha, \beta, G) \) and satisfies the equation \( \Pr(D = 1; \tilde{\alpha}, \tilde{\beta}, \tilde{G}) = n_1 / n \). The profile likelihoods for \( \beta_1 \) from \( L_C \) and \( L_1 \) are identical.
Proof. See the Appendix.

Besides extending the result of Prentice and Pyke (1979) to the situation with errors in covariates, this manner of proof considerably clarifies the kind of structural features necessary for an equivalence between a retrospective and prospective likelihood analysis. One needs sufficiently rich structure to fit perfectly the marginal distribution of $D$ without diminishing the ability to fit the conditional distributions of the covariates given $D$. In particular, it is important to note that if we had modeled the marginal distribution $G$ parametrically, then exact equivalence between prospective and retrospective inferences would not necessarily follow.

4. ALGORITHMS

In this section we present two algorithms for maximizing the likelihood with respect to the mixing distributions for a fixed value of $(\alpha, \beta)$. Algorithms for estimating $(\alpha, \beta)$ given $G$ have been discussed by Carroll et al. (1992). To obtain the joint ML estimates for $(\alpha, \beta)$ and $G$, we alternate between the two estimation problems. Throughout the discussion of estimates for $G$, we suppress the dependence of the likelihood on the parametric component of the model $(\alpha, \beta)$.

4.1 Geometric Results About Mixture Models

We first summarize a number of results about nonparametric mixture estimators due to Lindsay (1983). For a fixed value of $(\alpha, \beta)$, the problem reduces to one of maximizing a concave functional, $l(G) = \sum_{i=1}^{n} \log L_i(G) + \sum_{j=1}^{m} \log L_j(G)$, over a convex set. The likelihood vector, evaluated at the ML estimate $(L_1(G), \ldots, L_i(G), \ldots, L_m(G), \ldots, L_n(G))$, is unique. Furthermore, the ML estimate $G$ is known to be a discrete distribution that has a fixed upper bound, $K$, on the number of support points: $K$ equals the number of distinct terms in the likelihood vector. Because the maximization problem has these convenient properties, it is not difficult to construct an algorithm that walks up the likelihood surface and converges to the ML estimate regardless of the starting value of $G$.

The gradient method described later can be applied directly to obtain the ML estimate of $G$. But the EM algorithm can be conveniently applied only if the problem is simplified somewhat. We presuppose that $G$ has support on a fixed grid, $\xi = (\xi_1, \ldots, \xi_k)$. The problem of finding the ML estimate on this grid inherits all of the convenient properties of the full maximization problem. In particular, algorithms can be constructed that are guaranteed to converge regardless of the starting value.

4.2 EM Algorithms

Because the reduced data are missing $X$, the problem is suited to the EM algorithm, which estimates missing data and then maximizes the likelihood, given these estimates. The missing data can be thought of as a membership indicator variable for the $M$ possible values of the covariate $\xi = (\xi_1, \ldots, \xi_M)$. For convenience, we assume throughout that $\xi$ is known, although it is possible to implement a more general version of the EM algorithm to estimate the best grid for $G$. This assumption is justified shortly. The group membership for the complete data is obvious because $X$ is observed; in other words, for the $i$th observation, the posterior probability of group membership is one for $\xi = X_i$. Consequently, the set of support points, $\xi$, must include all distinct observed $X_i$'s. The group membership for reduced data can be estimated in the usual way (see, e.g., Titterington, Smith, and Makov 1985). Let $L_j(G^{(m)}) = \sum_i g^{(m)}(\xi_i) h_j(\xi_i)$. The EM algorithm, at the $(m+1)$-st step, puts mass $g^{(m+1)}(\xi_i) = \{A_{O_k} + A_{R_k}\}/(n_G + n_R)$, at $\xi_i$, where

$$A_{R_k} = \sum_{j=1}^{n_R} \frac{g^{(m)}(\xi_i) h_j(\xi_i)}{L_j(G^{(m)})}$$

and

$$A_{O_k} = \sum_{i=1}^{n_O} I(x_i = \xi_k).$$

For a fixed $(\alpha, \beta)$, $\lim_{m \to \infty} G^{(m)}$ maximizes the likelihood, provided that the support points are known.

Remark. Although the EM algorithm is typically used to estimate both the location of the support $\xi$ and the mass associated with each point of support, we chose to select a fixed grid of support points and then estimate the probability associated with each point of support. Selection of $\xi$ can be based on the observed $X_i$'s, $W_i$'s, and the distributional relationship between $X$ and $W$. We found it sufficient to include each distinct $X$ from the complete data set as well as a grid of points separated by at most $1/5\sigma_{W|X}$ if $X$ is observed. We found that $\sigma_{W|X}$ may depend on $X$. The range of the grid can be determined by estimating $E[X|W]$ for the minimum and maximum value of $W$ observed in the experiment; call the minimum predicted value $w_1$ and the maximum predicted value $w_2$. Define the grid on the interval $[w_1 - 2\sigma_{W|X} \times w_2 + 2\sigma_{W|X} \times w_2]$. Provided that a sufficiently dense grid is used, the solution will not differ measurably from the exact ML estimate. Time required to perform a single iteration of the algorithm increases directly with the number of grid points. We chose to use a fixed grid in the interest of computational feasibility; the EM algorithm is notoriously slow, especially when both the location and the weights must be estimated.

4.3 Gradient Methods

Because the optimization problem reduces to one of maximizing a concave function over a convex set, gradient-based algorithms are ideally suited to this maximization problem. The gradient of the log-likelihood at $G$ toward $\delta(\xi)$, a point mass at $\xi$, is

$$D(G, \xi) = \sum_i \left[ I(x_i = \xi) - 1 \right] + \sum_j \left[ h_j(\xi) \frac{I(G)}{L_j(G)} - 1 \right].$$

The algorithms are based on the following characterization due to Lindsay (1983). The mixing distribution $G$ maximizes the likelihood if and only if
a. \( D(\hat{G}, \xi) \leq 0 \) for all \( \xi \); and 
b. \( D(\hat{G}, \xi) = 0 \) in the support of \( \hat{G} \).

This result is the basis of the gradient methods. We present a simple but somewhat inefficient algorithm known as the VDM (Federov 1972; Wynn 1970). At the \( n \)th step, let \( \mathcal{G}^{(n)} \) be the current estimator. At each step, find \( \xi^{(n)} \) to maximize \( D(\mathcal{G}^{(n)}; \xi) \), and then find \( \varepsilon^{(n)} \) to maximize the likelihood evaluated at \( (1 - \varepsilon)\mathcal{G}^{(n)} + \varepsilon \delta(\xi^{(n)}) \); set \( \mathcal{G}^{(n+1)} = (1 - \varepsilon^{(n)})\mathcal{G}^{(n)} + \varepsilon \delta(\xi^{(n)}) \). Iterate until condition (a) is virtually satisfied. Other, more efficient versions of this sort of algorithm have been developed (see Boehning 1985, Lespeance and Kalbfleisch 1992, and Wu 1978a,b).

4.4 Combining Algorithms for the Parametric and Nonparametric Components of the Model

Although the likelihood is concave as a function of \( G \), there are no guarantees that it will even be unimodal when viewed as a function of \( (\alpha, \beta) \). Therefore, it is essential that good starting values be obtained for the parametric component of the model. We used the following algorithm with good success.

1. Estimate \( \beta \) from the complete observations \( (X, D) \) using a standard logistic regression algorithm. Call this \( \beta_C \).
2. Estimate \( \alpha \) from the complete observations \( (X, W, D) \). Call this \( \alpha_C \).
3. From the complete data, estimate \( G \) independently from \( (\alpha, \beta) \) using a weighted average of the empirical distribution of \( X \) for cases and controls. Call it \( G_{PL} \); this is the estimator used by Carroll et al. (1993).
4. Set \( G = G_{PL} \), and estimate \( (\alpha, \beta) \), using an algorithm such as the modified Newton–Raphson to maximize the likelihood. The complete-data estimators, \( (\alpha_C, \beta_C) \), are natural choices for a starting value. The resulting estimators are the partial likelihood estimators of Carroll et al. (1993). Call these \( (\alpha_{PL}, \beta_{PL}) \).
5. Fix the parametric components at \( (\alpha_{PL}, \beta_{PL}) \). Using either the gradient algorithm or the EM algorithm, maximize the likelihood over \( G \). If a gradient algorithm is used, then any starting value will work; a natural choice is \( G_{PL} \). If the EM algorithm is used, then any starting value will suffice, provided that positive mass is associated with each grid point. We used a starting value equal to \( G_{PL} + 1 \times \text{Uniform}(\xi) \), where the discrete uniform had equal mass on the grid points \( \xi \).
6. Maximize the likelihood over \( (\alpha, \beta) \), fixing \( G \) at the maximum obtained in the previous iteration.
7. Maximize the likelihood over \( G \), fixing \( (\alpha, \beta) \) at the maximum obtained in the previous iteration.
8. Repeat Steps 6 and 7 until convergence.

Because the steps involved in estimating \( G \) for a fixed value of \( (\alpha, \beta) \) are guaranteed to converge, all of the difficulties encountered in this algorithm are shared by other parametric ML estimation problems. For example, if the Newton–Raphson algorithm is not modified, then it can overstep the maximum and fail to converge. All of the usual warnings appropriate to maximizing a function apply. By using multiple starting values and checking the likelihood at each step of the parametric estimation scheme, one can ensure convergence eventually. There is no need to monitor the algorithms that estimate the nonparametric component of the model, as they are guaranteed to walk up the likelihood surface.

5. ADDITIONAL COVARIATES MEASURED WITHOUT ERROR

Frequently, additional covariates measured without error will be available. For instance, indicator variables such as sex and smoking status may have been recorded. In this section we extend our methodology to incorporate this extra information. Let \( Z \) denote an arbitrary set of covariates measured without error. Model the probability of disease, given all the covariates, as a prospective logistic regression model, where

\[
\Pr(D = 1 | X = \xi, Z = z) = \{1 + \exp(-\beta_0 - \beta_1'\xi - \beta_2'z)\}^{-1} \equiv \mathcal{K}(\xi, z).
\]

 Mimicking the development of the likelihood given in (5)–(7), let

\[
h_j(\xi) = \mathcal{K}_j(\xi, z_j) \mathcal{K}(\xi, z_j)^{-1} \mathcal{K}_j(\xi, z_j)^{-1} \mathcal{K}_j(\xi, z_j)^{-1} f_{W|X, Z}(w_j | \xi, z_j, d_j).
\]

Let \( dG_{X,Z}(x, z) = dG_X(x) \times G_{X|Z}(z|x) \equiv dG_X(x) \times F_{Z|X}(z|x) \) represent the joint distribution of \( (X, Z) \) in the case-control population. The contributions to the likelihood from the reduced and complete observations are then

\[
L_1(\alpha, \beta, G) = \int_X h_1(\xi; \alpha, \beta) \, dG_{X,Z}(\xi, z)
\]

and

\[
L_2(\alpha, \beta, G) = \int_X h_2(\xi; \alpha, \beta) I(z = \xi) \, dG_{X,Z}(\xi, z).
\]

We now arrive at a curious situation where ML procedures for estimating the joint distribution \( G_{X,Z} \) will break down. To simplify the issues, consider the case in which we observe variables \( (W, Z) \), where \( W \) is \( X \) observed with error but \( Z \) is observed directly. Suppose that the distribution of \( W \) given \( X = x \) is symmetric and unimodal about \( x \). If the observed \( Z_i \) are all distinct, then by writing \( G_{X,Z} = G_Z G_{X|Z} \), one can show that the ML estimator for the joint distribution \( G_{X,Z} \) is \( n^{-1} \sum_i \delta(w_i, z_i) \), where \( \delta \) is the point mass distribution. This follows because \( w_i \) is the value of \( x \) maximizing the error density \( f_{W|X}(w | x) \). Thus the ML estimator converges to \( G_{W,Z} \), not \( G_{X,Z} \). Curiously, regardless of whether the variables are both observed without error or both observed with error, ML still produces appropriate estimates. The problem seems to be that the sharpness of the \( Z \) observations prevents the pooling together of information over \( i \) that is necessary to obtain good estimates of conditional distribution of \( X \) given \( Z \). For example, if \( Z \) is discrete with a finite number of values, then this problem does not arise, as then the conditional distribution for each \( Z = z \) can be consistently estimated and will simply be the mixture estimator of \( G_X \) obtained from the set of \( W_i \)'s that have \( Z = z \).
With this in mind, we must consider alternative strategies for this case. One strategy would be to attempt to model the Z given X distribution. If \( f_{Z|X} \) were known, then for the reduced observations, (11) becomes

\[
L_3 = \int_X h_3(\xi; \alpha, \beta) f_{Z|X}(z|x) dG_X(x),
\]

and the problem is of the same form as that solved previously. Furthermore, if \( f_{Z|X} \) is known to follow some parametric form depending on \( \eta \) and \( W\{X, D, Z|\) \ depends on \( \alpha \), then the results of the previous sections apply with parametric component \( (\eta, \alpha, \beta) \). The drawback of this approach is that if the form of \( Z|X \) is unknown, then the results will depend on parametric modeling assumptions. This problem is especially difficult if \( Z \) is multivariate with discrete and continuous components.

Another possibility that retains the nonparametric flavor of our approach and that has the same mathematical structure would be to model \( G_{X, Z} \) as a discrete mixture of continuous densities, such as \( \sum_{\delta} N(\mu_\delta, h_\delta) \), where \( h \) would have to be chosen so as to balance the criteria of providing a flexible family of distributions (small) and providing sufficient smoothing to alleviate the foregoing problem (large). This approach has been studied in a slightly different context by Magder and Zeger (in press). Further research is necessary to resolve these issues.

6. NO COMPLETE DATA OBSERVED

Consider a situation in which only \( W, D \) have been measured (no validation data). If the measurement error distribution, \( f_{W|X, D} \), is known, then the likelihood is of the form \( L_3(\beta, G) = \prod_i f_{W, D}(w_i, d_i) \), where \( f_{W, D} \) is given by (5). Notice that this is equivalent to (8) for \( \alpha_C = 0 \) and \( \alpha \) known. Provided that the model satisfies the identifiability constraints of Kiefer and Wolfowitz (1956), \( \beta \) and \( G \) can be consistently estimated by the ML estimates even though no complete data are available. In fact this is the usual form of a semiparametric mixture model (see, e.g., Butler and Louis 1992; Lindsay 1995).

If the measurement error distribution is unknown, then it clearly cannot be estimated from the reduced sample. But if an independent data set is available in which both \( X \) and \( W \) have been measured, then it is possible to proceed as indicated earlier, provided that the measurement error distribution is nondifferential (i.e., does not depend on \( D \)). The idea is to use these data to obtain an estimate of the distribution of \( W|X \) in the case-control population. The assumption of nondifferential error is necessary, because otherwise \( f_{W|X} \) is a function of the study population. In particular,

\[
f_{W|X}(w|x) = f_{W|X, D}(w|x, D = 0)f_D(0) + f_{W|X, D}(w|x, D = 1)f_D(1),
\]

which depends on the prevalence of diseased individuals in the population unless \( f_{W|X, D} = f_{W|X} \). Consequently, the measurement error distribution is transportable only if it is nondifferential. We conclude that the ML methods presented so far are just as applicable in this situation as they are in the usual validation study, provided that the measurement error is nondifferential.

Contrast this scenario to the situation encountered when using methods that require a model for \( X|W \) (see, e.g., Saten and Kupper 1993). Clearly, by the same argument, the distribution of \( X|W \) is not transportable unless the effect is null; hence such methods cannot be applied in the no-validation situation. The pseudolikelihood method proposed by Carroll et al. (1993) clearly is not applicable either, as it requires an empirical estimate of the distribution of \( X \).

The only other literature that allows for consistent estimation when there is no validation and the error model is estimated independently is the paper by Stefanski and Carroll (1987), who assumed normally distributed measurement error. Our results allow for any error distribution.

To fit the model in this situation, use the following modification of the algorithm given in Section 4.4:

1. Obtain an ad hoc estimate of \( E[X|W = w_i] \), \( i = 1, \ldots, n \) using the known distribution of \( W|X \): Call it \( X_\text{est} \) and refer to \( X, D, W \) as the pseudo-complete data.
2. Estimate \( \beta \) from the pseudo-complete data \( \{X_\text{est}, D, i = 1, \ldots, n\} \) using a standard logistic regression algorithm. Call this \( \beta_{\text{PC}} \).
3. From the pseudo-complete data, estimate \( G \) independently from \( \beta \) using a weighted average of the empirical distribution of \( \hat{X} \) for cases and controls. Call it \( G_{\text{PC}} \).
4. Set \( G = G_{\text{PC}} \), and estimate \( \beta \) using an algorithm such as the modified Newton–Raphson to maximize the likelihood. The pseudo-complete estimator, \( \{\beta_{\text{PC}}\} \), is a natural choice for a starting value. Call the result \( \beta_{\text{PL}} \).
5. Fix the parametric component at \( \beta_{\text{PL}} \). Using either the gradient algorithm or the EM algorithm, maximize the likelihood over \( G \). If a gradient algorithm is used, then any starting value will work; a natural choice is \( G_{\text{PL}} \). If the EM algorithm is used, then any starting value will suffice, provided that positive mass is associated with each grid point. We used a starting value equal to \( 0.9 \times G_{\text{PL}} + 0.1 \times \text{Uniform} \{\xi\} \), where the discrete uniform had equal mass on the grid points \( \xi \) which are chosen as in Section 4.
6. Maximize the likelihood over \( \beta \), fixing \( G \) at the maximum obtained in the previous iteration.
7. Maximize the likelihood over \( G \), fixing \( \beta \) at the maximum obtained in the previous iteration.
8. Repeat Steps 6 and 7 until convergence.

7. A SIMULATION EXPERIMENT

To examine the performance of our estimator, we simulated data with measurement error under two scenarios: with and without a validation study. We chose a lognormal distribution for both \( X \) and \( W|X, D \) because this distribution often arises in practice (see, e.g., Nero, Schwed, and Nazaroff 1986). Distributions and parameter values were chosen to be similar to or identical with those simulated by Carroll et al. (1993). Data were generated prospectively, and then a case-control sample was obtained from this simulated source sample. The prospective logistimodel was
given by $Pr(D = 1|X = x) = K(x)$, with $\beta_0 = -3.09$ and $\beta_1 = .5$. The true covariate, $X$, was generated as a lognormal random variable so that $\log(X)$ had mean $-1/2\sigma_x^2$ and variance $\sigma_x^2$, where $\sigma_x = 1.08$. The surrogate predictor was also lognormal, with $\log(W)$, given $(X, D)$, having mean $\log(X)$ and standard deviation $\sigma$, which varied. We repeated each experiment 50 times.

The general form of the measurement error model that we fit is $\log(W) = \alpha_0 + \alpha_2 \log(X) + \sigma z$, with $z \sim N(0, 1)$; the simulated data fell into this class, with $\alpha_0 = 0$, $\alpha_2 = 1$, and $\sigma$ varying. In addition to the unknown mixing distribution $G$, the model also has five free parameters, $(\alpha_0, \alpha_2, \sigma, \beta_0, \beta_1)$; hence we call this the five-parameter model. In the following subsection we assume that $(\alpha_0, \alpha_2, \sigma)$ are known. This leaves only two parametric unknowns, $(\beta_0, \beta_1)$; hence we call this the two-parameter model.

We used the EM algorithm to estimate the mixing distribution. We chose sample sizes and parameters to provide valid comparisons for realistic sample sizes and to illustrate the limits of the method. A value of $\sigma = 1$ represents a large amount of measurement error: $\sigma_x^2 \approx \sigma^2$. The method failed to converge occasionally with this amount of measurement error. In practice one could fit this model when $\sigma = 1$, provided that a number of starting values could be used to find the ML estimate. Alternatively, $\sigma < .25$ represents a rather small measurement error. In principle, this method will perform well when $\sigma < .25$; however, the EM grid would have to be quite dense to yield an estimate similar to the actual ML estimate for $G$. The computations are prohibitively slow for simulations in this situation. Consequently, we report only the performance of the method for $\sigma = .25, .50, .75$.

### 7.1 Known Measurement Error

In this section we assume that the measurement error distribution is known and examine the performance of the two-parameter model with a small, medium, and large amount of measurement errors ($\sigma = .25, .50, .75$). Samples $\{X_i, W_i, D_i, i = 1, \ldots, n\}$ of size 80 and 240 were generated, each consisting of half cases and half controls. First, using $\{X_i, D_i, i = 1, \ldots, n\}$, we estimated $\beta$ using the logistic model (equivalent to $\sigma = 0$). Next, omitting the gold standard variables $X_i$ and using only $\{W_i, D_i, i = 1, \ldots, n\}$, we estimated $(\beta, G)$ using the mixture model (the MIX method). Results are presented in Table 1. Notice that as the variance of the measurement error distribution increased, the mean squared error (MSE) of the MIX method increased rapidly. In fact, when $\sigma = 1$, the reduced data appeared to have almost no information remaining relative to when $\sigma = 0$ (results not reported). As the sample size increased, the MSE decreased substantially, as would be expected. Of greater interest is the increase in the disparity between the performance with and without measurement error. This is especially apparent for models with larger measurement error.

For sample sizes smaller than 80, it is unlikely that the semiparametric model provides any advantages over a model that assumes a parametric form for the unobserved covariate. For smaller sample sizes, the data simply do not provide sufficient information from which to estimate the distribution of this unobservable accurately in a nonparametric setting.

Next, we simulate data from a validation study. To see the effect of increasing amount of complete data, first $\frac{1}{2}$ and then $\frac{1}{4}$ of the data were considered as complete. The performance of the MIX method is compared with the pseudolikelihood (PL) method proposed by Carroll et al. (1993) in Table 2. We found that if a substantial proportion of the data are complete ($\frac{1}{4}$ or more), then the PL and MIX methods perform similarly. When only $\frac{1}{4}$ of the data are complete, the pattern of results is markedly different. The MSE of both methods increased dramatically; however, the MIX method performed much better than the PL method for this scenario. The difference in performance was greatest when the variance of the measurement error was greatest.

For the most part, we found that the MIX method either provided the same estimate as the PL method or provided one that was better. When the number of reduced observations was not large relative to the number of complete

| Table 1. Known Measurement Error Model With No Complete Data, $\beta_1 = .5$ |
|------------------|--|--|--|---|---|---|
| $n_{10}$ | $n_{11}$ | $n_{10}$ | $n_{11}$ |
| Mean | .53 | .54 | .57 | .59 |
| MSE | .061 | .078 | .111 | .180 |
| RMSE | 1.00 | 1.28 | 1.82 | 2.95 |
| Mean | .51 | .50 | .53 | .55 |
| MSE | .012 | .022 | .026 | .054 |
| RMSE | 1.03 | 1.83 | 2.16 | 4.50 |

**Note:** Mean and MSE of $\beta_1$ are calculated based on 50 repetitions of the simulation experiment. RMSE is the ratio of the MSE of the PL estimator to the MSE of the MIX estimator.

| Table 2. Known Measurement Error Model With Some Complete Data, $\beta_1 = .5$ |
|------------------|--|--|--|---|---|---|
| $n_{10}$ | $n_{11}$ | $n_{10}$ | $n_{11}$ | $n_{10}$ | $n_{11}$ |
| Mean | .496 | .496 | .520 | .520 | .512 | .505 |
| MSE | .016 | .016 | .021 | .020 | .038 | .032 |
| RMSE | 1.00 | 1.06 | 1.05 | 1.19 | 1.19 | 1.19 |
| Mean | .512 | .501 | .519 | .515 | .611 | .522 |
| MSE | .040 | .017 | .089 | .033 | .143 | .038 |
| RMSE | 2.35 | 2.70 | 3.70 | 3.70 | 3.70 | 3.70 |

**Note:** Mean and MSE of $\beta_1$ are calculated based on 50 repetitions of the simulation experiment. RMSE is the ratio of the MSE of the PL estimator to the MSE of the MIX estimator.
observations, it was often the case that no information beyond that provided by the PL estimate was available in the data about the distribution of \( X \). In these cases the algorithm essentially stopped at Step 4 (Sec. 4.4) and produced the PL estimate. Thus it is apparent from these simulations that the MIX method outperforms the PL method only when there are substantially more reduced observations than complete observations; however, as \( n_R \) grows, the discrepancy between the performance will continue to increase.

When the measurement error distribution is known (or obtained from another study), there is no lower bound on the number of complete cases required. Comparing Table 1 to Table 2, one can see the reduction in the MSE when at least some of the observations are complete (compare simulations with total sample size of 240). As expected, when \( \sigma = .25 \), the improvement is not substantial: .022 versus .016 and .017. When \( \sigma = .75 \), the difference is more notable: .054 versus .032 and .038.

### 7.2 Unknown Measurement Error

In this section we assume that the measurement error distribution is known to be lognormal, but the three parameters of the lognormal model are unknown. This puts us in the framework of the five-parameter model. To compare the MIX method and the PL method, we simulated validation data sets with two different sample sizes and three different measurement errors: \( n_{c0} = n_{c1} = 30, n_{R0} = n_{RL} = 30; n_{c0} = n_{c1} = 30, n_{R0} = n_{RL} = 90; \) and \( \sigma = .25, .50, .75 \).

For the first choice of sample sizes, the MIX method and the PL method performed almost identically, suggesting that for data like those simulated, little extra information can be extracted by maximizing over the mixing distribution. But again, the MIX method outperforms the PL method when the number of complete observations is small relative to the number of reduced observations.

The disparity between the methods increased with the variance in the measurement error distribution. The PL method also tends to perform poorly when the measurement error is small, relative to the number of complete observations (not reported). We believe that PL performs poorly in this setting, because when \( W \) is not close to any \( X \) in the sample, the contribution to the pseudolikelihood is nearly zero and hence the efficiency of the method is reduced.

Comparing the two-parameter and five parameter results (Tables 2 and 3), it is clear that the variance of the estimator increases substantially when the parameters of the measurement error distribution must be estimated.

We discovered that the algorithm was not stable with less than 60 complete observations, when the five-parameter model was fit. Again, we found that the MIX method tended to provide the same estimate as the PL method, or one that was better. Comparing the two-parameter and five parameter results with sample sizes 30, 30, 90, 90, we see that the MIX method outperformed the PL method more often when the measurement error distribution was not known; (for example, for \( \sigma = .75 \), the root mean squared error (RMSE) was 1.40 for the five-parameter model, but only 1.19 for the two-parameter model.)

The measurement error model can be extended to allow for differential error by incorporating two extra parameters: \( \log(\omega_W) = \alpha_0 + \alpha_1 \log(x) + \alpha_2 d + \alpha_3 e \). Carroll et al. (1993) performed extensive simulations, comparing the seven-parameter PL model to competing methods in the validation study setting, and found that incorporating the two extra parameters did not significantly reduce the efficiency of the PL model, and, moreover, it provided a substantial increase in the robustness of the model. Our methods demonstrated a serious bias in the estimates of the \( \beta_i \) when the nondifferential error was ignored. We found that for the sample sizes and parameter values given in Table 3, the seven-parameter MIX method performed equal to or better than the seven-parameter PL method (not reported). In toto, these results suggest that it is worthwhile to use a richer measurement error model unless there is evidence that the measurement error is nondifferential.

### 7.3 Profile Likelihoods

There is currently no standard methodology for computing the variance of a profile ML estimate in a semi-
8. A CHOLESTEROL STUDY

In this example we analyze a data set concerning the risk of coronary heart disease (CHD) as a function of blood cholesterol level. These data were extracted from the Lipids Research Clinics study, which was previously discussed by Satten and Kupper (1993). We use a portion of these data involving men age 60–70 who do not smoke (256 records: 4 outliers were removed). A subject is recorded as having CHD ($D = 1$) if they have had a previous heart attack, an abnormal exercise electrocardiogram, history of angina pectoris, and so forth. The measured covariates are low-density lipoprotein (LDL) cholesterol level and total cholesterol (TC) level. Direct measurement of LDL levels is time-consuming and requires costly special equipment. For this reason, we are interested in whether TC serves as a useful surrogate for LDL. Note that the measurement error of TC is not the source of error of primary interest; rather, the unknown quantity of the other components of TC (triglycerides and high density lipoproteins) lead to the “measurement error.” Henceforth CHD, LDL/100, and TC/100 play the roles of $D, X$, and $W$.

In this data set, both $X$ and $W$ have been recorded for each subject. In the full data set there are 113 cases, of which 47 had LDL levels higher than 160. Among the 143 controls, 43 had elevated LDL levels. Figure 1 presents boxplots of the data. Using $X$ as the predictor, the prospective logistic regression estimate for $\beta_1$ was .656 with a standard error of .336. Contrast this with the attenuated estimate (.540) obtained when measurement error was ignored and $W$ was used as the predictor.

A nondifferential lognormal measurement error model provided a good fit to the data (Fig. 2), with the exception of a slight increase in the variance of $W|X$ for small values of $X$. The differential measurement error model fit significantly better but did not change the parameters enough to have a practical impact on the estimation procedure. Consequently, the measurement error was modeled using a nondifferential error model.

To illustrate the information present in the reduced data, we analyzed a sample of data with and without the reduced observations. From the 113 cases and 143 controls, 32 cases and 40 controls were randomly selected to serve as complete data. The remaining observations were treated as re...
duced observations. Using only the 72 complete observations, we obtained $\hat{\beta}_1 = .943$ with standard error of .62. Figure 3 illustrates the profile likelihood for the non-differential model when both complete and reduced data are used; $\beta_1 = .765$. To compare, notice that a 95% confidence interval using all of the data as complete had length 1.34, a 95% profile interval for the 72 complete observations and 184 reduced observations had length 1.75, and a 95% confidence interval for the 72 complete observations only had length 2.48. We conclude that the precision of the estimate increases substantially when both complete and reduced data are used.

9. CONCLUSIONS

In this article, we have suggested using nonparametric mixture methods to estimate regression parameters when one or more of the regression parameters are measured with error. The theoretical results and implementation of the MIX method have both concentrated on the logistic case-control study, allowing for differential measurement error. We have considered situations where the predictor $X$ is observed in a subset of the study (Secs. 3 and 4) or cannot be observed at all (Sec. 6). Simulations (Sec. 7) and an example (Sec. 8) indicate the feasibility of the methodology.

The use of mixture methods, and the MIX method itself, is not restricted to logistic case-control studies but also applies to any prospective (as opposed to case-control) likelihood problem. In principle, one needs only a likelihood for the distribution of the response $Y$ given the predictor $X$, as well as a parametric error model relating the observed predictor $W$ to $(Y, X)$ (differential error) or relating $W$ to $X$ (non-differential error). When $X$ is partially observed in this context, there are many other competing techniques (see Robins, Hsich, and Newey 1995 and references therein). When $X$ is unobserved, as occurs in the classical measurement error problem, non-differential measurement error is required. Mixture methods apply as discussed in Section 6, as long as there is sufficient information to identify the parameters of the relevant distributions, especially the error distribution of $W$ given $X$. We discussed in Section 6 the case where there is an independent experiment that estimates the distribution of $W$ given $X$. It is possible to extend the results to a second case, where there are replicates of $W$ that are sufficient in themselves to identify the error distribution.

APPENDIX: PROOF OF RESULTS FROM SECTIONS 2 AND 3

Proof of Lemma 1

a. $\beta_1 = \beta_1^*$, because the log odds ratio is proportional to $\beta_1$ for each retrospective model.
b. Follows from Bayes's theorem.
c. When $d = 1$, (3) implies

$$K(x)g(x)/\phi = K^*(x)g^*(x)/\phi^*.$$  

It follows that

$$g^*(x) = g(x) \phi^* [1 + \exp(\beta_1 + \beta_1^* x)]/\phi [1 + \exp(\beta_1 + \beta_1^* x)].$$

For $d = 0$, the same relationship holds. The result follows from noting that $\sum_s g_s(x) = 1$.

Proof of Theorem 2

The key to the proof is that any $(\theta, \phi^*)$ satisfying (b) necessarily maximizes $L_{\theta}(\theta, \phi^*)$, as this term is maximized over all possible multinomial probabilities $p_1, \ldots, p_K$. Thus given any $(\theta_0, \phi)$ maximizing $L_C(\theta, \phi)$, the corresponding $(\theta_0, \phi^*)$ satisfying (a) and (b) simultaneously maximizes both terms in the product $L_C(\theta, \phi) L_M(\theta, \phi)$. If the entire parameter space is in $\Omega$, then for any $\theta_0$ there is a $\phi_0^*$ that maximizes $L_C(\theta_0, \phi_0)$ over $\phi$ and satisfies $p_0((\theta_0, \phi_0^*) = n_0/m$. Consequently, the joint and conditional profile likelihoods for $\theta$ are equal.

Proof of Corollary 3

From Lemma 1, for any fixed set of parameters $(\alpha, \beta, \gamma)$, we can find a set $(\alpha^*, \beta, \gamma^*)$ giving the same conditional distributions for $X$ given $D$ but having the prospectively value of $P(E|D = 1, \alpha^*, \beta, \gamma^*) = n_0/m$. But if it follows that the conditional distributions of $W$ given $D$ are also identical for $(\alpha, \beta, \gamma^*)$ and $(\alpha, \beta, \gamma)$; hence the values of $L_C$ are identical, and so the hypothesis of the theorem is met.

[Received May 1994. Revised July 1995.]

REFERENCES


