The Annals of Statistics
1995, Vol. 23, No. 1, 161–181

DIMENSION REDUCTION IN A SEMIPARAMETRIC REGRESSION MODEL WITH ERRORS IN COVARIATES[1]

BY R. J. CARROLL, R. K. KNICKERBOCKER AND C. Y. WANG

Texas A & M University, Lilly Research Laboratories and Fred Hutchinson Cancer Research Center

We consider a semiparametric estimation method for general regression models when some of the predictors are measured with error. The technique relies on a kernel regression of the "true" covariate on all the observed covariates and surrogates. This requires a nonparametric regression in as many dimensions as there are covariates and surrogates. The usual theory copes with such higher-dimensional problems by using higher-order kernels, but this is unrealistic for most problems. We show that the usual theory is essentially as good as one can do with this technique. Instead of regression with higher-order kernels, we propose the use of dimension reduction techniques. We assume that the "true" covariate depends only on a linear combination of the observed covariates and surrogates. If this linear combination were known, we could apply the one-dimensional versions of the semiparametric problem, for which standard kernels are applicable. We show that if one can estimate the linear directions at the root-n rate, then asymptotically the resulting estimator of the parameters in the main regression model behaves as if the linear combination were known. Simulations lend some credence to the asymptotic results.

1. Introduction.

1.1. *Logistic regression.* We consider semiparametric estimation when the true predictors are measured with error, discussing in detail the logistic regression model in which a binary response Y is related to a scalar predictor X via logistic regression:

$$\text{(1)} \qquad \Pr(Y = 1|X) = H(\beta_{00} + \beta_{01}X); \qquad H(v) = \{1 + \exp(-v)\}^{-1}.$$

As in all measurement error models [Fuller (1987)], the problem is that X is difficult or expensive to observe, but instead one can observe a proxy W for X. As is typical in the nonlinear measurement error model literature, we will assume that W is a *surrogate* for X, that is, Y and W are independent given X. Intuitively, this means that if X could be observed, W would provide no additional information about Y. Under the assumption that W is a surrogate, the conditional distribution of Y given W is the binary regression model

$$\text{(2)} \qquad \Pr(Y = 1|W) = E\{H(\beta_{00} + \beta_{01}X)|W\}.$$

Received August 1992; revised May 1994.
[1]Our research was supported by a grant from the National Cancer Institute (CA-57030).
AMS 1991 *subject classifications*. Primary 62E20; secondary 62M05
Key words and phrases. Dimension reduction, errors-in-variables, kernel regression, logistic regression, semiparametric models.

Parametric inference can be obtained via a model for the distribution of X given W. We are interested in the case that such a distribution is unknown.

The assumption that W is a surrogate might appear to be a strong limiting factor, but this is in fact far from the case. The most common measurement error model is the classical additive error model $W = X + U$, where the measurement error U is a mean-zero random variable independent of Y and X. In this model, W is a surrogate. The classical additive error model occurs throughout Fuller (1987), as well as in many other applications [Rosner, Willett and Spiegelman (1989), Rosner, Spiegelman and Willett (1990) and Carroll and Stefanski (1994)]. Surrogates occur far more generally; for example, W is a surrogate whenever it follows the model $W = \mathscr{F}(X, U)$, where U is independent of (Y, X) and $\mathscr{F}(\cdot)$ is an arbitrary function. This includes standard multiplicative models.

The available data are described as follows. In a sample of size n, (Y_i, W_i) is observed for $i = 1, \ldots, n$. In a random subset of the data, we set $\Delta_i = 1$ and also observe X_i with probability $\pi = \Pr(\Delta_i = 1) = \Pr(\Delta_i = 1 | Y_i, W_i, X_i)$; otherwise $\Delta_i = 0$ and X_i is not observed. The random subset with X_i observed is called a *validation study*. Under this sampling scheme, Carroll and Wand (1991) employ a pseudolikelihood estimation technique. The regression function (2) as a function of (W, β_0, β_1) is estimated via kernel regression in the validation data, by regressing $H(\beta_0 + \beta_1 X)$ on W. This yields an estimated binary regression model and hence an estimated or pseudolikelihood for the primary data. The likelihood for the validation data and the primary data pseudolikelihood are then jointly maximized to yield estimates of (β_{00}, β_{01}), which are asymptotically normally; Pepe and Fleming (1991) derived this method independently.

1.2. *The curse of dimensionality and the new method.* The semiparametric method described above is subject to the curse of dimensionality. Suppose that W is of dimension d and let the order of the kernel be κ, with $\kappa = 2$ being the usual nonnegative second-order kernel. Then, in order to achieve asymptotic normality at the rate $n^{1/2}$, Carroll and Wand (1991) require that $nh^{2d} \to \infty$ and $nh^{2\kappa} \to 0$. Clearly, if $d = 2$, these conditions exclude the use of a second-order kernel. Larger values of d require progressively higher-order kernels. Carroll and Wand (1991) call this "hardly practical."

In Section 2, we sketch an argument in linear regression which shows why the conditions of Carroll and Wand are almost necessary. Our approach to the problem is to exploit the possibility that the distribution of X given W depends only on lower-dimensional linear combinations of W, in particular a single linear combination.

The standard parametric solution to this dilemma is to assume that X given W is normally distributed with mean $W^T\gamma_0$ and variance $\sigma^2_{X|W}$ [see Carroll, Spiegelman, Lan, Bailey and Abbott (1984), Rosner, Willett and Spiegelman (1989) and Crouch and Spiegelman (1990)]. The nonparametric generalization is to assume merely that the distribution of X given W

depends, in an unspecified way, only on $W^T\gamma_0$ for some γ_0 with $\|\gamma_0\| = 1$. If γ_0 were known, then one would run the various algorithms on the surrogate $W^T\gamma_0$, and since the dimension of this surrogate is 1, standard second-order kernels could be used. In practice, γ_0 is unknown, but there exist methods for estimating at the rate $n^{1/2}$, such as average derivative estimation [Härdle and Stoker (1989)], projection pursuit regression [Friedman and Stuetzle (1981) and Hall (1989)] and sliced inverse regression [Li (1991) and Duan and Li (1991)].

Given any $n^{1/2}$-consistent estimate $\hat{\gamma}$ of γ_0, the obvious algorithm is to employ the Carroll–Wand methodology using $W^T\hat{\gamma}$ as the estimated surrogate. In this paper, we show that the resulting estimates $(\hat{\beta}_{00}, \hat{\beta}_{01})$ have the same limit distribution as if γ_0 were known. In other words, one can use one's favorite dimension reduction device, without any asymptotic effect on the resulting parameter estimates.

This paper is organized as follows. In Section 2, we sketch the result for linear regression which shows the necessity of using higher-order kernels for surrogate dimensions greater than 1. In Section 3, we describe the algorithm in detail for the case of logistic regression, stating our result in Section 4. In Section 5, we describe some numerical experience we have had with the method, which indicates that the lack of any asymptotic effect due to estimating the directions can hold for fairly small sample sizes.

In Section 6 we describe extensions to our results which include general likelihood problems, quasilikelihood and variance function models including generalized linear models [Carroll and Ruppert (1988), Chapters 2 and 3] and semiparametric corrections for attenuation.

2. Bandwidth rates. Remember that κ is the order of the kernel, and d is the dimension of W. The results in Carroll and Wand (1991) assume that the bandwidths satisfy $nh^{2d} \to \infty$ and that $nh^{2\kappa} \to 0$. In this section, we will indicate why these rates are about as good as one can do with the methodology. The calculations are easiest in the linear regression model $Y = \beta_{00} + \beta_{01}X + \varepsilon$, where ε is a mean-zero random variable independent of (X, W). Let $m(W) = E(X|W)$. In this case, the regression of Y on W is $\beta_{00} + \beta_{01}m(W)$, so that the correction for attenuation technique of Sepanski, Knickerbocker and Carroll (1994) is first to construct an estimate of $m(W)$ in the validation data and then perform a linear regression of Y on $\hat{m}(W)$ in the primary data. We restrict calculation to a subset of the primary data interior to the support of W, and we accomplish this by defining a function $\phi(\cdot)$ which is supported on such a set. Define

$$A_n = n^{-1}\sum_{i=1}^{n}(1-\Delta_i)\phi(W_i)\begin{pmatrix}1\\ \hat{m}(W_i)\end{pmatrix}\begin{pmatrix}1\\ \hat{m}(W_i)\end{pmatrix}^T,$$

$$B_n = n^{-1}\sum_{i=1}^{n}(1-\Delta_i)\phi(W_i)\begin{pmatrix}1\\ \hat{m}(W_i)\end{pmatrix}Y_i.$$

Then the semiparametric correction for attenuation estimate of $\mathscr{B}_0 = (\beta_{00}, \beta_{01})^T$ is $\hat{\mathscr{B}} = A_n^{-1}B_n$. Define $\xi_i = Y_i - \beta_{00} - \beta_{01}m(W_i)$, which given W_i has mean zero. Simple algebra shows that

$$
\begin{aligned}
A_n n^{1/2}\left(\hat{\mathscr{B}} - \mathscr{B}_0\right) &= n^{-1/2}\sum_{i=1}^{n}(1-\Delta_i)\phi(W_i)\begin{pmatrix}1\\ m(W_i)\end{pmatrix}\xi_i \\
&\quad + n^{-1/2}\sum_{i=1}^{n}(1-\Delta_i)\phi(W_i)\begin{pmatrix}0\\ \hat{m}(W_i) - m(W_i)\end{pmatrix}\xi_i \\
&\quad - n^{-1/2}\sum_{i=1}^{n}(1-\Delta_i)\beta_{01}\phi(W_i)\begin{pmatrix}1\\ m(W_i)\end{pmatrix}\{\hat{m}(W_i) - m(W_i)\} \\
&\quad - n^{-1/2}\sum_{i=1}^{n}(1-\Delta_i)\beta_{01}\phi(W_i)\begin{pmatrix}0\\ \{\hat{m}(W_i) - m(W_i)\}^2\end{pmatrix}.
\end{aligned} \tag{3}
$$

The rate of convergence for a kernel regression estimate is of the order $\mathscr{O}_p\{h^{2\kappa} + (nh^d)^{-1}\}$. The first term on the right-hand side of (3) is clearly $\mathscr{O}_p(1)$. Because $E(\xi_i|W_i) = 0$, the second term can easily be shown to be of order $\mathscr{O}_p\{h^{\kappa} + (nh^d)^{-1/2}\}$. Long, detailed and tedious calculations can be used to show that the third term is of the order $\mathscr{O}_p(1) + \mathscr{O}_p\{nh^{2\kappa} + (nh^d)^{-1}\}^{1/2}$. The fourth term, however, is essentially the average mean squared error of a kernel regression estimate times $n^{1/2}$, and hence it is of order

$$n^{1/2}\mathscr{O}_p\left\{h^{2\kappa} + (nh^d)^{-1}\right\},$$

so for it to converge in probability to zero, we require that $nh^{2d} \to \infty$ and $nh^{4\kappa} \to 0$. Combining the results, we see that it is required that $nh^{2d} \to \infty$ (from the fourth term) and $nh^{2\kappa} \to 0$ (from the third term).

3. Details of the method. Let $\mathscr{B}_0 = (\beta_{00}, \beta_{01})^T$ and write arbitrary $\mathscr{B} = (\beta_0, \beta_1)^T$. We observe (Y_i, W_i) for $i = 1, \ldots, n$, and on a random subset of the data set have $\Delta_i = 1$ and observe X_i as well. We assume that measurements of X occur for a nonvanishing fraction of the data, so that if $\pi = \Pr(\Delta = 1)$, then $0 < \pi < 1$.

We assume the existence of a $n^{1/2}$-consistent estimate $\hat{\mathscr{B}}_0$ of $\mathscr{B}_0$, for example, coming from the validation data, which are those observations with $\Delta_i = 1$. While the validation data provide one estimate of $\mathscr{B}_0$, such data usually will form only a small subset of all the available observations, and we wish to use the remaining data with $\Delta_i = 0$ to improve the estimate of $\mathscr{B}_0$.

We will also assume that there is a surrogate W for X and a vector γ_0 such that

(4) the distribution of X given W depends only on $W^T\gamma_0$.

In other words X depends on W only through $W^T\gamma_0$. Without loss of generality we assume that $\|\gamma_0\| = 1$.

Define $G(w^T\gamma, \mathscr{B}, \gamma) = E\{H(\beta_0 + \beta_1 X)|W^T\gamma = w^T\gamma\}$. Also define

$$\dot{G}(w^T\gamma, \mathscr{B}, \gamma) = G(w^T\gamma, \mathscr{B}, \gamma)\{1 - G(w^T\gamma, \mathscr{B}, \gamma)\},$$

$$\dot{H}(v) = H(v)\{1 - H(v)\}.$$

If γ_0 and $G(\cdot)$ were known, by (1) and (2) a one-step likelihood-based scoring estimate of $\mathscr{B}$ is

$$\hat{\mathscr{B}} = \hat{\mathscr{B}}_0 + B_{2n}^{-1}(\hat{\mathscr{B}}_0)B_{1n}(\hat{\mathscr{B}}_0),$$

where

$$B_{1n}(\mathscr{B}) = n^{-1}\sum_{i=1}^{n}\Delta_i\begin{pmatrix}1\\X_i\end{pmatrix}\{Y_i - H(\beta_0 + \beta_1 X_i)\}$$
$$+ n^{-1}\sum_{i=1}^{n}(1-\Delta_i)G_\beta(W_i^T\gamma_0, \mathscr{B}, \gamma_0)\frac{\{Y_i - G(W_i^T\gamma_0, \mathscr{B}, \gamma_0)\}}{\dot{G}(W_i^T\gamma_0, \mathscr{B}, \gamma_0)},$$

$$B_{2n}(\mathscr{B}) = n^{-1}\sum_{i=1}^{n}\Delta_i\begin{pmatrix}1\\X_i\end{pmatrix}\begin{pmatrix}1\\X_i\end{pmatrix}^T\dot{H}(\beta_0 + \beta_1 X_i)$$
$$+ n^{-1}\sum_{i=1}^{n}(1-\Delta_i)\frac{G_\beta(W_i^T\gamma_0, \mathscr{B}, \gamma_0)G_\beta^T(W_i^T\gamma_0, \mathscr{B}, \gamma_0)}{\dot{G}(W_i^T\gamma_0, \mathscr{B}, \gamma_0)},$$

with subscripts denoting derivatives. However, we generally do not know γ_0 or G, so as in Carroll and Wand (1991) we will estimate $G(W^T\gamma_0, \mathscr{B}, \gamma_0)$ with the nonparametric regression of $H(\beta_0 + \beta_1 X)$ on $W^T\hat{\gamma}$ in a fixed compact set $\mathscr{C}$ interior to the support of $W^T\hat{\gamma}$. This restriction to the set $\mathscr{C}$ decreases the efficiency, but increases the robustness of the estimator. We estimate G with the Nadaraya–Watson estimator

$$G_n(w^T\gamma, \mathscr{B}, \gamma) = \frac{C_n(w^T\gamma, \mathscr{B}, \gamma)}{D_n(w^T\gamma, \gamma)} = \frac{\sum_{i=1}^n\Delta_i H(\beta_0 + \beta_1 X_i)K_h\{\gamma^T(W_i - w)\}}{\sum_{i=1}^n\Delta_i K_h\{\gamma^T(W_i - w)\}},$$

where $K(\cdot)$ is a symmetric density function with bounded support and $K_h(\cdot) = h^{-1}K(\cdot/h)$. Replacing $G(W^T\gamma, \mathscr{B}, \gamma)$ by $G_n(w^T\gamma, \mathscr{B}, \gamma)$, $G_\beta(w^T\gamma, \mathscr{B}, \gamma)$ by $G_{n\beta}(w^T\gamma, \mathscr{B}, \gamma)$ and γ_0 by its estimator $\hat{\gamma}$, we propose the following estimator of $\mathscr{B}$:

$$\hat{\mathscr{B}}(\hat{\gamma}) = \hat{\mathscr{B}}_0 + B_{4n}^{-1}(\hat{\mathscr{B}}_0, \hat{\gamma})B_{3n}(\hat{\mathscr{B}}_0, \hat{\gamma}), \tag{5}$$

where

$$B_{3n}(\mathscr{B}, \gamma) = n^{-1}\sum_{i=1}^{n}\Delta_i\begin{pmatrix}1\\X_i\end{pmatrix}\{Y_i - H(\beta_0 + \beta_1 X_i)\}$$
$$+n^{-1}\sum_{i=1}^{n}(1-\Delta_i)G_{n\beta}(W_i^T\gamma, \mathscr{B}, \gamma)\frac{\{Y_i - G_n(W_i^T\gamma, \mathscr{B}, \gamma)\}}{\dot{G}_n(W_i^T\gamma, \mathscr{B}, \gamma)} \tag{6}$$
$$\times \phi(W_i^T\gamma)$$

and

$$B_{4n}(\mathscr{B},\gamma) = n^{-1}\sum_{i=1}^{n}\Delta_i\begin{pmatrix}1\\X_i\end{pmatrix}\begin{pmatrix}1\\X_i\end{pmatrix}^T \dot{H}(\beta_0+\beta_1 X_i)$$

$$+n^{-1}\sum_{i=1}^{n}(1-\Delta_i) \tag{7}$$

$$\times\frac{G_{n\beta}(W_i^T\gamma,\mathscr{B},\gamma)G_{n\beta}^T(W_i^T\gamma,\mathscr{B},\gamma)}{\dot{G}_n(W_i^T,\mathscr{B},\gamma)}\phi(W_i^t\gamma).$$

4. Statement of main result. For technical purposes, instead of working directly with $n^{1/2}$-consistent estimate of $\mathscr{B}_0$ and γ, we work with *discretized* versions of them, as follows. Let c be an arbitrary (small) constant, and let $\mathscr{F}$ be the set $\{0, \pm c/n^{1/2}, \pm 2c/n^{1/2}, \ldots\}$.

By definition, an estimator $\hat{\theta}$ is a discretized version of an estimator $\hat{\theta}_*$ if each component of $\hat{\theta}$ takes on that value in $\mathscr{F}$ closest to the corresponding component of $\hat{\theta}_*$.

Note that our use of the term "discretize" is completely different from the idea of binning in nonparametric regression. Our meaning is that all the components of $\hat{\mathscr{B}}_0$ and $\hat{\gamma}$ are constrained to take values in $\mathscr{F}$.

The use of discretization is a technical tool which leads to great simplification of proofs, because it enables use of the following trick due to Le Cam. Let $\hat{\theta}_n$ be a discretized $n^{1/2}$-consistent estimate of a parameter θ_0, and consider a random variable $A_n(\theta)$. To show that $A_n(\hat{\theta}_n) - A_n(\theta_0) = o_p(1)$, it suffices to show that $A_n(\theta_n) - A_n(\theta_0) = o_p(1)$, where $\theta_n = \theta_0 + t_n/n^{1/2}$ is a *deterministic* sequence with $t_n \to t_0$, where t_0 is a finite constant. We will discretize both $\hat{\gamma}$ and the starting value $\hat{\mathscr{B}}_0$.

THEOREM. *Under the conditions stated in the Appendix,*

$$n^{1/2}\{\hat{\mathscr{B}}(\hat{\gamma}) - \hat{\mathscr{B}}(\gamma_0)\} = o_p(1).$$

The implication of this result is that one can estimate $\mathscr{B}_0$ asymptotically just as well as if γ_0 were known. The proof of the theorem is in the Appendix.

Applying the main result of Carroll and Wand (1991), we see that $n^{1/2}\{\hat{\mathscr{B}}(\gamma_0) - \mathscr{B}_0\}$ is asymptotically normally distributed. The asymptotic covariance of $n^{1/2}\{\hat{\mathscr{B}}(\gamma_0) - \mathscr{B}_0\}$ has three terms: (i) the Fisher information for $\mathscr{B}$ from the validation data; (ii) the Fisher information from the primary data set if $f_{X|W^T\gamma_0}$ were known; and (iii) the cost for not knowing $f_{X|W^T\gamma_0}$. None of these terms involves the choice of $K(\cdot)$ or h. Thus, it follows that $n^{1/2}\{\hat{\mathscr{B}}(\hat{\gamma}) - \mathscr{B}_0\}$ will be asymptotically normal with the same variance as when γ_0 is known.

5. Simulations. A small simulation study was undertaken to compare the estimates of the regression parameters using the Carroll–Wand proce-

dure using both $W^T\gamma_0$ and $W^T\hat{\gamma}$ as the surrogate. The main point of the simulation is to investigate the main result, namely, that there is little effect to the dimension reduction proposed in this paper.

The logistic regression model used was $\Pr(Y = 1|X) = H(-1 + 0.693X)$, with $H(v) = \{1 + \exp(-v)\}^{-1}$, the logistic distribution function. The surrogates W were generated as five-dimensional standard normal random variables, while X given W was normally distributed with mean $W^T\gamma$ and variance δ^2. We let δ take on the values $(0.25, 0.5, 1.0)$, these representing instances of small, moderate and very large measurement error. We let $\gamma = (0.894, 0.447, 0, 0, 0)^T$, and we estimate γ using sliced inverse regression with 10 observations per slice. The sample sizes generated were 150 and 600, and in each case exactly $\pi = \frac{1}{3}$ of the observations were validation data in which X was observed. This is slightly different from selecting items into the validation study randomly with probability $\pi = \frac{1}{3}$, but the main theoretical result that there is no asymptotic cost to dimension reduction holds in this case as well. We used the ad hoc bandwidth selection procedure described by Carroll and Wand and let the bandwidth be $h = \hat{\sigma}(n/3)^{-1/3}$, where $\hat{\sigma}$ is the sample standard deviation from the validation data (of size $n/3$) for $W^T\gamma_0$ and $W^T\hat{\gamma}$ for the two respective estimators. For the two estimators, the set $\mathscr{C}$ was from h plus the minimum to h minus the maximum value of $\gamma_0^t W$ and $\hat{\gamma}^t W$. We simulated 1000 data sets and report the mean, standard deviation, mean squared error, median absolute error and the 95th percentile of the absolute error for each of the estimates of the slope. The results are tabulated in Tables 1–3.

The estimators in our simulation were based on fully iterating (5), starting from the (undiscretized) validation data estimate. The sliced inverse regression estimate we used assumed that the distribution of X is described by a one-dimensional linear combination of W.

The simulations indicate that the two estimators are very close both in terms of mean square error and in the percentiles of the absolute errors even for the smallest sample sizes. For example, consider the case $\delta = 1.0$ and $n = 150$. In Figure 1 we plot kernel density estimates of the estimated slopes when γ is known or estimated. While there is some right skewness, the two plots are similar.

TABLE 1

1000 *simulated estimates of slope using the logistic model with* $\delta = 0.25$ *and* $\pi = \frac{1}{3}$

Estimator	Mean	Std. Dev.	MSE	MAE	95% AE
$n = 150$					
γ_0 known	0.7498	0.2483	0.0649	0.1643	0.5267
γ_0 estimated	0.7427	0.2490	0.0644	0.1580	0.5017
$n = 600$					
γ_0 known	0.7122	0.1113	0.0127	0.0736	0.2248
γ_0 estimated	0.7112	0.1115	0.0127	0.0729	0.2243

TABLE 2
1000 *simulated estimates of slope using the logistic model with* $\delta = 0.5$ *and* $\pi = \frac{1}{3}$

Estimator	Mean	Std. Dev.	MSE	MAE	95% AE
$n = 150$					
γ_0 known	0.7567	0.2573	0.0702	0.1549	0.5519
γ_0 estimated	0.7450	0.2620	0.0713	0.1616	0.5433
$n = 600$					
γ_0 known	0.7137	0.1134	0.0133	0.0754	0.2306
γ_0 estimated	0.7106	0.1128	0.0130	0.0743	0.2242

We have simulated other models such as X given W distributed normally with mean $(W^T\gamma)^2$ and variance δ^2. The results for this model are similar to those reported above for the larger sample sizes, namely, that there is little cost due to dimension reduction. Other bandwidths such as $C\hat{\sigma}(n/3)^{-1/3}$ for $C = (0.5, 0.75, 1.5, 2)$ produced similar results, with $h = \hat{\sigma}(n/3)^{-1/3}$ generally performing better than the other bandwidths in terms of mean square error.

The results indicate that for large enough sample sizes, there is little effect due to dimension reduction. We have not addressed directly the question of whether dimension reduction itself leads to an improvement over doing brute-force multidimensional kernel regression. Our only evidence on this point is indirect. We attempted to make such a comparison in a Monte Carlo study, but ran into numerical difficulties. The brute force method was numerically unstable in the sense that there were convergence difficulties with the algorithm. Even when convergence occurred, the computation took many times longer than the dimension reduction method. Finally, we have no idea how one would select a multidimensional bandwidth in this context.

6. Extensions. We have deliberately phrased this problem in the context of logistic regression, because it is one of the most important nonlinear measurement error models and also has some of the simplest notation. Our purpose in working with this special case is that it makes the theoretical

TABLE 3
1000 *simulated estimates of slope using the logistic model with* $\delta = 1.0$ *and* $\pi = \frac{1}{3}$

Estimator	Mean	Std. Dev.	MSE	MAE	95% AE
$n = 150$					
γ_0 known	0.7340	0.2655	0.0721	0.1573	0.5266
γ_0 estimated	0.7027	0.2541	0.0646	0.1563	0.4883
$n = 600$					
γ_0 known	0.7000	0.1161	0.0135	0.0779	0.2214
γ_0 estimated	0.6915	0.1148	0.0131	0.0749	0.2269

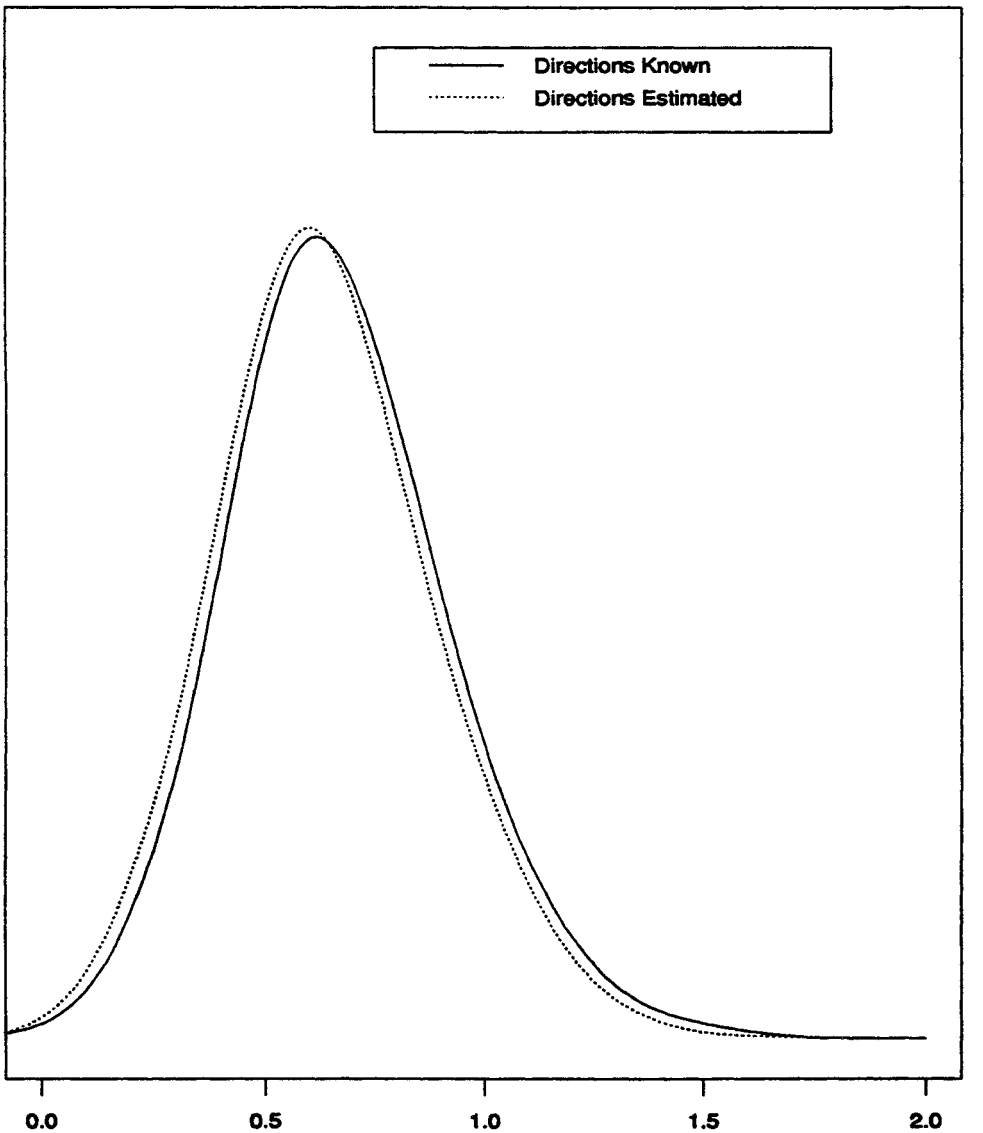

FIG. 1. *Kernel density estimates for known and estimated directions when* $\delta = 1.0$, $n = 150$ *and* $\pi = \frac{1}{3}$.

calculations and the basic idea of dimension reduction transparent. However, the methods we have described can be greatly generalized.

For instance, the results hold (under regularity conditions) not just for logistic regression but for any likelihood problem. In the general likelihood case, if $l(Y|X, \mathscr{B})$ is the underlying conditional likelihood for $\mathscr{B} = (\beta_0, \beta_1)^t$, then the conditional likelihood for an observed data pair $(Y, W) = (y, w)$ is

$E\{l(y|X, \mathscr{B})|W = w\}$. This likelihood can be estimated by kernel regression techniques, and the result maximized to obtain an estimate of $\mathscr{B}$. The resulting limit distribution requires only a notational change from the logistic model. If W is discrete, a similar technique has been proposed by Pepe and Fleming (1991).

The results are not restricted to likelihood problems, but also apply to general quasilikelihood and variance function models. If the conditional mean and variance of Y given X are $f(X, \mathscr{B})$ and $g^2(X, \mathscr{B}, \theta)$, say, then the conditional mean and variance of Y given W can be estimated using formulae similar to (2). For example, the conditional mean of Y given W is $E\{f(X, \mathscr{B})|W\}$. Sepanski and Carroll (1993) estimate such regressions nonparametrically and then apply quasilikelihood and variance function estimating equations for $(\mathscr{B}, \theta)$. They run into the same curse of dimensionality problems that concern us, and the same methods we have proposed apply here as well, that is, dimension reduction can alleviate the curse of dimensionality.

In generalized linear models especially, it is well known that a remarkably accurate approximation to the likelihood of Y given W can be achieved by replacing X where it is not observed by $E(X|W)$ [see Rosner, Willett and Spiegelman (1989), Rosner, Spiegelman and Willett (1990), Carroll and Stefanski (1990), Gleser (1990) and Pierce, Stram, Vaeth and Schafer (1992), among others]. For example, this replacement strategy, closely related to the "correction for attenuation" in linear regression, would suggest that in the logistic regression model (1), $\text{pr}(Y = 1|W) = H\{\beta_{00} + \beta_{01}E(X|W)\}$. This is not exactly true, but it very nearly is in many practical problems. If we are willing to pretend this approximation is exact, then one strategy is to estimate the function $E(X|W)$ via nonparametric regression using that part of the data for which (X, W) is observed. This method is trivial to compute: a single nonparametric regression to estimate $E(X|W)$, followed by generalized linear model program. We actually prefer the resulting estimators to the Carroll–Wand method because of this ease of computation, as well as the good performance of the method in simulation studies not reported here. One can show that the curse of dimensionality described in Section 1.2 holds here as well. Under regularity conditions, it again may be shown that our results concerning dimension reduction still apply.

If we are willing to treat the replacement model as exact, there is no need to observe X at all. Instead, in many problems one can observe a variable $X_* = X + U$, where U is uncorrelated with (Y, W). For instance, as described by Carroll and Stefanski (1994), in the Framingham Heart Study W would be observed blood pressure at baseline, while X_* would be observed blood pressure four years earlier. It follows that $E(X_*|W) = E(X|W)$ and the results of the previous paragraph apply when one regresses X_* on W. The use of such "replication" data greatly expands the possible applications of our results.

In the simulations (Section 5), we did not discuss the gains to be made by our estimators over using the validation data alone (i.e., $\hat{\mathscr{B}}_0$), but they are

considerable. In these and many other simulations we have done with $\pi = \frac{1}{3}$ of the data being validation, the Carroll–Wand estimator is typically at least 50% more efficient than using validation data only, while the semiparametric replacement algorithms are typically at least twice as efficient as using only the validation data.

In principle, it is possible to extend the results to the case that there are two independent data sets, a primary one in which only (Y, W) is observed $(\Delta = 0)$ and an independent *external* data set in which (X, W) is observed $(\Delta = 1)$. Use of such external data requires that the distribution of (X, W) be the same as in the primary data with $\Delta = 0$. The algorithm (5) changes here by deleting the first terms in (6) and (7). While it is clear that our dimension reduction result holds in this case, the technical problem with pushing the theoretical results through lies in constructing a $n^{1/2}$-consistent preliminary estimate of $\mathscr{B}_0$. The outline of what to do is standard. Consistent estimation of $\mathscr{B}_0$ is possible because the external data allow consistent estimation of the distribution of X given W. Once consistency is proved, $n^{1/2}$-consistency then needs to be checked.

We have not considered here the important case that there are some covariates Z measured without error, so that the logistic regression model (1) has mean $H(\beta_{00} + \beta_{01}X + \beta_{02}^T Z)$. In this problem, the expectation (2) must be conditioned on (Z, W), while replacement algorithms require estimating $E(X|Z, W)$. Hence the previously published methods almost automatically lead to higher-order kernels. If in this problem we assume that X given (Z, W) depends only on $W^T\gamma_0 + Z^T\gamma_1$, and if (γ_0, γ_1) can be estimated at the rate $n^{1/2}$, then it can be shown that there is a (usually small) asymptotic effect due to estimating (γ_0, γ_1), but second-order kernels may be employed.

Robins, Hsieh and Newey (1994) generalize the Carroll–Wand and Pepe–Fleming techniques by computing an optimal semiparametric score function for likelihood problems. Their methods do not apply directly to quasilikelihood models and corrections for attenuation, although their nonoptimal estimating equations can be extended to the former. For likelihood problems, when W is multidimensional the optimal score becomes difficult to compute. The use of dimension reduction should improve their method by increasing large-sample efficiency as well as by making computation far easier. We are currently studying ways to implement dimension reduction ideas in this context, as well as whether there is any asymptotic effect to estimating the direction of the reduced variable.

APPENDIX

Proof of the Theorem.

A.1. *Preliminaries and assumptions.* All results are proved for the case that W is a bivariate random variable, the general case being only notationally more complex. As described in Section 4, both $\hat{\gamma}$ and $\hat{\mathscr{B}}_0$ are discretized

$n^{1/2}$-consistent estimators. Remember that $\Delta_i = 1$ means that X_i is observed, and that this occurs with probability π independent of (Y_i, W_i, X_i).

Define $\mathscr{B}_n = \mathscr{B}_0 + s_n n^{-1/2}$, and $\gamma_n = \gamma_0 + t_n n^{-1/2}$, where $(s_n, t_n) \to (s_0, t_0)$ for fixed, finite s_0, t_0. Also define $f(\cdot)$ as the joint density of $W = (W_1, W_2)^T$.

We will use the notation outlined in Section 3, with the following additions:

$$\pi = \mathrm{pr}(\Delta_i = 1) = \mathrm{pr}(\Delta_i = 1 | Y_i X_i W_i), \qquad 0 < \pi < 1;$$

$$D(a, \gamma) \text{ is the density of } W^T\gamma \text{ at } a;$$

$$C(a, \mathscr{B}, \gamma) = D(a, \gamma) E\{H(\beta_0 + \beta_1 X) | W^T\gamma = a\};$$

$$P(a, x) = \frac{\{D(a, \gamma_0) H(\beta_{00} + \beta_{01} x) - C(a, \mathscr{B}_0, \gamma_0)\}}{D^2(a, \gamma_0)};$$

$$R(a) = \frac{G_\beta(a, \mathscr{B}_0, \gamma_0)}{\dot{G}(a, \mathscr{B}_0, \gamma_0)} \quad \text{and} \quad Q(a, x) = R(a) P(a, x);$$

$$M(a, b) = E\{Q(a, X) | W^T\gamma = b\}, \qquad M_2(a) = \left.\frac{\partial M(b, a)}{\partial b}\right|_{b=a},$$

$$M_3(a, b, c) = E\{Q(a, X) Q^T(b, X) | (W^T\gamma_0 = c)\}.$$

Also define

$$D_n(a, \gamma) = \frac{\sum_{i=1}^n \Delta_i K\{(W_i^T\gamma - a)/h\}}{\sum_{i=1}^n \Delta_i},$$

$$C_n(a, \mathscr{B}, \gamma) = \frac{\sum_{i=1}^n \Delta_i H(\beta_0 + \beta_1 X_i) K\{(W_i^T\gamma - a)/h\}}{\sum_{i=1}^n \Delta_i}.$$

Make the following assumptions:

ASSUMPTION 1. In (a, γ), $D(a, \gamma)$ is thrice continuously and bounded differentiable and bounded away from zero in a neighborhood of γ_0 and on an open set in a containing $\mathscr{E}$, the support of $\phi(\cdot)$.

ASSUMPTION 2. $nh^{2+\varepsilon} \to \infty$ and $nh^{4-\varepsilon} \to 0$ for some $\varepsilon > 0$.

ASSUMPTION 3. $n^{1/2}(\hat{\gamma} - \gamma_0) = \mathscr{O}_p(1)$.

ASSUMPTION 4. The function $K(\cdot)$ is a thrice continuously differentiable symmetric density function with bounded support.

ASSUMPTION 5. $M(a, b)$ and its first two partial derivatives are continuous and uniformly bounded.

ASSUMPTION 6. X has finite fourth moment.

ASSUMPTION 7. $f(\cdot,\cdot)$ and its first two partial derivatives are uniformly bounded.

ASSUMPTION 8. $M_3(a, b, c)$ has an uniformly bounded continuous derivative in (a, b) for each fixed c.

ASSUMPTION 9. Uniformly on $a \in \mathscr{C}$, there exists $\varepsilon > 0$ such that

$$|D_n(a, \gamma_n) - D(a, \gamma_n)| = \mathscr{O}_p\left\{h^{2-\varepsilon} + (nh^{1+\varepsilon})^{-1/2}\right\},$$

$$|D_n(a, \gamma_0) - D(a, \gamma_0)| = \mathscr{O}_p\left\{h^{2-\varepsilon} + (nh^{1+\varepsilon})^{-1/2}\right\},$$

$$|C_n(a, \mathscr{B}_0, \gamma_n) - C(a, \mathscr{B}_0, \gamma_n)| = \mathscr{O}_p\left\{h^{2-\varepsilon} + (nh^{1+\varepsilon})^{-1/2}\right\}$$

and

$$|C_n(a, \mathscr{B}_0, \gamma_0) - C(a, \mathscr{B}_0, \gamma_0)| = \mathscr{O}_p\left\{h^{2-\varepsilon} + (nh^{1+\varepsilon})^{-1/2}\right\}.$$

Additionally, uniformly for $a \in \mathscr{C}$ and $b \in \mathscr{C}$, such that $(a, b) = (w^T\gamma_n, w^T\gamma_0)$ for some w, we have that there exists $\varepsilon > 0$ such that

$$|D_n(a, \gamma_n) - D(b, \gamma_0)| = \mathscr{O}_p\left\{h^{2-\varepsilon} + (nh^{1+\varepsilon})^{-1/2}\right\}$$

and

$$|C_n(a, \mathscr{B}_0, \gamma_n) - C(b, \mathscr{B}_0, \gamma_0)| = \mathscr{O}_p\left\{h^{2-\varepsilon} + (nh^{1+\varepsilon})^{-1/2}\right\}.$$

We note that by adapting the results of Mack and Silverman (1982) or Marron and Härdle (1986), a set of sufficient conditions that imply Assumption 9 can be found.

ASSUMPTION 10. $\int af(a, b)\, da < \infty$ for every b.

ASSUMPTION 11. $\phi(\cdot)$ has bounded support and two bounded derivatives.

A.2. *The proof.*

THEOREM. *Under the model outlined in Section* 3 *and Assumptions* 1–11, *we have*

$$n^{1/2}\left\{\hat{\mathscr{B}}(\hat{\gamma}) - \hat{\mathscr{B}}(\gamma_0)\right\} \to_P 0,$$

where $\hat{\mathscr{B}}(\gamma)$ *is defined in* (5), (6), *and* (7).

PROOF. Due to the discretization of $\hat{\gamma}$ and $\hat{\mathscr{B}}_0$, it suffices to show that

$$n^{1/2}\left\{\hat{\mathscr{B}}(\gamma_n) - \hat{\mathscr{B}}(\gamma_0)\right\} = o_p(1),$$

for starting values $\mathscr{B}_n = \mathscr{B}_0 + s_n n^{1/2}$ and $\gamma_n = \gamma_0 + t_n n^{-1/2}$.

We prove the result in two steps:

(i) $$T_n = n^{-3/2}\{\pi(1-\pi)\}^{-1} \times \sum_{i=1}^{n}\sum_{j=1}^{n}(1-\Delta_i)\Delta_j\Big\{Q\big(W_i^T\gamma_0, X_j\big)\phi\big(W_i^T\gamma_n\big) \times\big[K_h\{\gamma_n^t(W_j - W_i)\} - K_h\{\gamma_0^T(W_j - W_i)\}\big]\Big\} \to_P 0;$$

(ii) $T_n \to_P 0$ implies that $n^{1/2}\{\hat{\mathscr{B}}(\gamma_n) - \hat{\mathscr{B}}(\gamma_0)\} \to_P 0$.

We will show that $T_n = o_p(1)$ by showing its mean and covariance converge to 0. Note that

$$\begin{aligned} M(a,a) &= E\{R(a)P(a,X)|W^T = a\} \\ &= R(a)\frac{D(a,\gamma_0)E\{H(\beta_{00}+\beta_{01}X)|W^T\gamma_0 = a\} - C(a,\mathscr{B}_0,\gamma_0)}{D^2(a,\gamma_0)} \qquad (8) \\ &= 0. \end{aligned}$$

Define $\gamma_0 = (\gamma_{01}, \gamma_{02})^T$, $\gamma_n = (\gamma_{n1}, \gamma_{n2})^T$, $W_1 = (W_{11}, W_{12})^T$ and $W_2 = (W_{21}, W_{22})^T$. Conditioning on $\mathscr{G} = \{W_i^T\gamma_0, W_i^T\gamma_n, W_j^T\gamma_0, W_j^T\gamma_n\}$, $i, j = 1, \ldots, n$, it follows that

$$\begin{aligned} E(T_n) = n^{1/2}\int & M\big(W_1^T\gamma_0, W_2^T\gamma_0\big) \\ & \times\big[K_h\{\gamma_n^T(W_2 - W_1)\} - K_h\{\gamma_0^T(W_2 - W_1)\}\big] \qquad (9) \\ & \times\phi\big(W_1^T\gamma_n\big)f(W_1)f(W_2)\,dW_1\,dW_2. \end{aligned}$$

Next make the substitutions $a_1 = W_{11}$, $a_2 = \gamma_0^T W_1$, $b_1 = W_{21}$ and $b_2 = \gamma_0^T W_2$, and note that (9) simplifies to

$$\begin{aligned} & \frac{n^{1/2}}{\gamma_{02}^2}\int M(a_2,b_2) \\ & \times\left[K_h\left\{\frac{\gamma_{n2}}{\gamma_{02}}(a_2-b_2) + u_n n^{-1/2}(a_1-b_1)\right\} - K_h(a_2-b_2)\right] \qquad (10) \\ & \times\phi\left(\frac{\gamma_{n2}}{\gamma_{02}}a_2 + u_n a_1 n^{-1/2}\right)f\left(a_1, \frac{a_2-\gamma_{01}a_1}{\gamma_{02}}\right) \\ & \times f\left(b_1, \frac{b_2-\gamma_{01}b_1}{\gamma_{02}}\right)da_1\,da_2\,db_1\,db_2, \end{aligned}$$

where $u_n = (\gamma_{n1} - \gamma_{n2}\gamma_{01}/\gamma_{02})n^{1/2}$. Note that $u_n = \mathcal{O}(1)$. Next make the substitution $a_2 = b_2 + zh$ and note that (10) is

$$
\begin{aligned}
&\frac{n^{1/2}}{\gamma_{02}^2}\int\left[K\left\{z\frac{\gamma_{n2}}{\gamma_{02}} - \frac{u_n(a_1 - b_1)}{n^{1/2}h}\right\} - K(z)\right] \\
&\quad\times\phi\left\{\frac{\gamma_{n2}}{\gamma_{02}}(b_2 + zh) + u_n a_1 n^{-1/2}\right\} \\
&\quad\times M(b_2 + zh, b_2) f\left(a_1, \frac{b_2 + zh - \gamma_{01}a_1}{\gamma_{02}}\right) \\
&\quad\times f\left(b_1, \frac{b_2 - \gamma_{01}b_1}{\gamma_{02}}\right) dz\, da_1\, db_1\, db_2.
\end{aligned} \tag{11}
$$

Now expand $M(b_2 + zh, b_2)$ in a Taylor series about $b_2 + zh = b_2$ and note that (11) is equivalent to

$$
\begin{aligned}
&\frac{n^{1/2}h}{\gamma_{02}^2}\int\left[K\left\{z\frac{\gamma_{n2}}{\gamma_{02}} - \frac{u_n(a_1 - b_1)}{n^{1/2}h}\right\} - K(z)\right] \\
&\quad\times\phi\left\{\frac{\gamma_{n2}}{\gamma_{02}}(b_2 + zh) + u_n a_1 n^{-1/2}\right\} \\
&\quad\times zM_2(b_2) f\left(a_1, \frac{b_2 + zh - \gamma_{01}a_1}{\gamma_{02}}\right) \\
&\quad\times f\left(b_1, \frac{b_2 - \gamma_{01}b_1}{\gamma_{02}}\right) dz\, da_1\, db_1\, db_2 + o(1),
\end{aligned} \tag{12}
$$

using the fact that $M(a_1, a_1) = 0$ and Assumptions 2, 4, 5, 7 and 11. Next expand $\phi(\cdot)$ and $f(\cdot)$ in Taylor's series expansions and note that (12) simplifies to

$$
\begin{aligned}
&\frac{n^{1/2}h}{\gamma_{02}^2}\int zM_2(b_2) \\
&\quad\times\left[K\left\{z\frac{\gamma_{n2}}{\gamma_{02}} - \frac{u_n(a_1 - b_1)}{n^{1/2}h}\right\} - K(z)\right]\phi(b_2) \\
&\quad\times f\left(a_1, \frac{b_2 - \gamma_{01}a_1}{\gamma_{02}}\right) f\left(b_1, \frac{b_2 - \gamma_{01}b_1}{\gamma_{02}}\right) dz\, da_1\, db_1\, db_2 + o(1),
\end{aligned} \tag{13}
$$

again using Assumptions 2, 4, 5, 7 and 11. For the symmetric density function $K(\cdot)$, $\int zK(z-b)\,dz = b$ and $\int zK(z)\,dz = 0$. Thus (13) is

$$\frac{u_n}{\gamma_{n2}^2}\int M_2(b_2)\phi(b_2)(a_1-b_1)f\left(a_1, \frac{b_2-\gamma_{01}a_1}{\gamma_{02}}\right) \times f\left(b_1, \frac{b_2-\gamma_{01}b_1}{\gamma_{02}}\right) da_1\,db_1\,db_2 + o(1) = o(1).$$

Thus $E(T_n) \to 0$ as was to be shown.

Next we show $\mathrm{Var}(T_n) = o(1)$. First note that

$$\begin{aligned} E\left(T_nT_n^T\right) \approx n^{-3}\sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}(1-\Delta_i)(1-\Delta_k)\Delta_j\Delta_l \\ \times E\Big(Q\left(W_i^T\gamma_0, X_j\right)Q^T\left(W_k^T\gamma_0, X_l\right) \\ \times\phi\left(W_i^T\gamma_n\right)\left[K_h\left\{\gamma_n^T(W_j-W_i)\right\}\right. \\ \left.-K_h\left\{\gamma_0^T(W_j-W_i)\right\}\right] \\ \times\phi\left(W_k^T\gamma_n\right)\left[K_h\left\{\gamma_n^T(W_l-W_k)\right\}\right. \\ \left.-K_h\left\{\gamma_0^T(W_l-W_k)\right\}\right]\Big). \end{aligned} \tag{14}$$

To show $\mathrm{Var}(T_n) \to 0$, first note that the terms where $i \neq k$ and $j \neq l$ are negated asymptotically by the term $E(T_n)E(T_n^T)$. Hence it suffices to study the terms where $(i = k, j = l)$, $(i = k, j \neq l)$ and $(i \neq k, j = l)$, which we will denote T_{1n}, T_{2n} and T_{3n}, respectively.

Let "$\sim$" denote proportionality. As before, condition on $\mathscr{G}$ and note that

$$\begin{aligned} E(T_{1n}) &\sim \left(n\gamma_{02}^2\right)^{-1}\int M_3(a_2,a_2,b_2)\phi^2\left(\frac{\gamma_{n2}}{\gamma_{02}}a_2 + u_na_1n^{-1/2}\right) \\ &\quad\times\left[K_h\left\{\frac{\gamma_{n2}}{\gamma_{02}}(a_2-b_2) + u_nn^{-1/2}(a_1-b_1)\right\} - K_h(a_2-b_2)\right]^2 \\ &\quad\times f\left(a_1, \frac{a_2-\gamma_{01}a_1}{\gamma_{02}}\right)f\left(b_1, \frac{b_2-\gamma_{01}b_1}{\gamma_{02}}\right)da_1\,da_2\,db_1\,db_2 \\ &= (nh)^{-1}\gamma_{02}^{-2}\int M_3(a_2,a_2,a_2+zh)\phi^2\left(\frac{\gamma_{n2}}{\gamma_{02}}a_2 + u_na_1n^{-1/2}\right) \\ &\quad\times\left[K\left\{z\frac{\gamma_{n2}}{\gamma_{02}} + \frac{u_n(a_1-b_1)}{n^{1/2}h}\right\} - K(z)\right]^2 \\ &\quad\times f\left(a_1, \frac{a_2-\gamma_{01}a_1}{\gamma_{02}}\right)f\left(b_1, \frac{a_2+zh-\gamma_{01}b_1}{\gamma_{02}}\right)dz\,da_1\,da_2\,db_1 \\ &= o(1). \end{aligned}$$

Next study the terms for which $(i = k, j \neq l)$:

$$E(T_{2n}) \sim \gamma_{02}^{-3} \int M(a_2, b_2) M^T(a_2, c_2) \phi^2\left(\frac{\gamma_{n2}}{\gamma_{02}} a_2 + u_n a_1 n^{-1/2}\right)$$
$$\times \left[K_h\left\{\frac{\gamma_{n2}}{\gamma_{02}}(a_2 - b_2) + u_n n^{-1/2}(a_1 - b_1)\right\} - K_h(a_2 - b_2)\right]$$
$$\times \left[K_h\left\{\frac{\gamma_{n2}}{\gamma_{02}}(a_2 - c_2) + u_n n^{-1/2}(a_1 - c_1)\right\} - K_h(a_2 - c_2)\right]$$
$$\times f\left(a_1, \frac{a_2 - \gamma_{01} a_1}{\gamma_{02}}\right) f\left(b_1, \frac{b_2 - \gamma_{01} b_1}{\gamma_{02}}\right)$$
$$\times f\left(c_1, \frac{c_2 - \gamma_{01} c_1}{\gamma_{02}}\right) da_1\, da_2\, db_1\, db_2\, dc_1\, dc_2$$

$$(15) \quad = \gamma_{02}^{-3} \int M(a_2, a_2 + z_1 h) M^T(a_2, a_2 + z_2 h) \phi^2\left(\frac{\gamma_{n2}}{\gamma_{02}} a_2 + u_n a_1 n^{-1/2}\right)$$
$$\times \left[K\left\{\frac{\gamma_{n2}}{\gamma_{02}} z_1 + u_n (n^{1/2} h)^{-1}(a_1 - b_1)\right\} - K(z_1)\right]$$
$$\times \left[K\left\{\frac{\gamma_{n2}}{\gamma_{02}} z_2 + (u_n n^{1/2} h)^{-1}(a_1 - c_1)\right\} - K(z_2)\right]$$
$$\times f\left(a_1, \frac{a_2 - \gamma_{01} a_1}{\gamma_{02}}\right) f\left(b_1, \frac{a_2 + z_1 h - \gamma_{01} b_1}{\gamma_{02}}\right)$$
$$\times f\left(c_1, \frac{a_2 + z_2 h - \gamma_{01} c_1}{\gamma_{02}}\right) dz_1\, dz_2\, da_1\, da_2\, db_1\, dc_1$$
$$= o(1)$$

by dominated convergence.

Finally, study the terms for which $(i \neq k, j = l)$:

$$E(T_{3n}) \sim \gamma_{02}^{-3} \int M_3(a_2, c_2, b_2) \phi\left(\frac{\gamma_{n2}}{\gamma_{02}} a_2 + u_n a_1 n^{-1/2}\right) \phi\left(\frac{\gamma_{n2}}{\gamma_{02}} c_2 + u_n c_1 n^{-1/2}\right)$$
$$\times \left[K_h\left\{\frac{\gamma_{n2}}{\gamma_{02}}(b_2 - a_2) + u_n n^{-1/2}(b_1 - a_1)\right\} - K_h(b_2 - a_2)\right]$$
$$\times \left[K_h\left\{\frac{\gamma_{n2}}{\gamma_{02}}(b_2 - c_2) + u_n n^{-1/2}(b_1 - c_1)\right\} - K_h(b_2 - c_2)\right]$$
$$\times f\left(a_1, \frac{a_2 - \gamma_{01} a_1}{\gamma_{02}}\right) f\left(b_1, \frac{b_2 - \gamma_{01} b_1}{\gamma_{02}}\right)$$
$$\times f\left(c_1, \frac{c_2 - \gamma_{01} c_1}{\gamma_{02}}\right) da_1\, da_2\, db_1\, db_2\, dc_1\, dc_2$$

$$
\begin{aligned}
&= \gamma_{02}^{-3} \int M_3(b_2 + z_1 h, b_2 + z_2 h, b_2)\, \phi\left(\frac{\gamma_{n2}}{\gamma_{02}}(b_2 + z_1 h) + u_n a_1 n^{-1/2} \right) \\
&\quad \times \phi\left(\frac{\gamma_{n2}}{\gamma_{02}}(b_2 + z_2 h) + u_n a_1 n^{-1/2} \right) \\
&\quad \times \left[K\left\{ \frac{\gamma_{n2}}{\gamma_{02}} z_1 + u_n (n^{1/2} h)^{-1}(b_1 - a_1) \right\} - K(z_1) \right] \\
&\quad \times \left[K\left\{ \frac{\gamma_{n2}}{\gamma_{02}} z_2 + (u_n n^{1/2} h)^{-1}(c_1 - a_1) \right\} - K(z_2) \right] \\
&\quad \times f\left(a_1, \frac{b_2 + z_1 h - \gamma_{01} a_1}{\gamma_{02}} \right) f\left(b_1, \frac{b_2 - \gamma_{01} b_1}{\gamma_{02}} \right) \\
&\quad \times f\left(c_1, \frac{b_2 + z_2 h - \gamma_{01} c_1}{\gamma_{02}} \right) dz_1\, dz_2\, da_1\, db_1\, db_2\, dc_1 \\
&= o(1)
\end{aligned}
$$

by dominated convergence.

So we have shown that $\mathrm{Var}(T_n) \to 0$, hence $T_n \to_P 0$. We now must show that $T_n \to_P 0$ implies that the theorem holds.

Referring to Sepanski and Carroll (1993), we see that the difficult step is to show that

$$
n^{1/2}\{B_{3n}(\mathscr{B}_n, \gamma_n) - B_{3n}(\mathscr{B}_0, \gamma_0)\} = o_p(1), \tag{16}
$$

where $B_{3n}(\cdot, \cdot)$ is defined in (6). Taking a Taylor series expansion, it follows that

$$
\begin{aligned}
G_n(a, \mathscr{B}_n, \gamma) &= G_n(a, \mathscr{B}_0, \gamma) + (\mathscr{B}_n - \mathscr{B}_0) G_{n\beta}(a, \mathscr{B}_0, \gamma) + \mathscr{O}_p(n^{-1}) \\
&= G_n(a, \mathscr{B}_0, \gamma) + \mathscr{O}_p(n^{-1/2}),
\end{aligned}
$$

since $\mathscr{B}_n - \mathscr{B}_0 = s_n n^{-1/2}$. Similarly, $G_{n\beta}(a, \mathscr{B}_n, \gamma) = G_{n\beta}(a, \mathscr{B}_0, \gamma) + \mathscr{O}_p(n^{-1/2})$ and $\dot{G}_n(a, \mathscr{B}_n, \gamma) = \dot{G}_n(a, \mathscr{B}_0, \gamma) + \mathscr{O}_p(n^{-1/2})$. Thus, (16) holds if

$$
n^{1/2}\{B_{3n}(\mathscr{B}_0, \gamma_n) - B_{3n}(\mathscr{B}_0, \gamma_0)\} = o_p(1). \tag{17}
$$

The terms in $B_{3n}(\mathscr{B}_0, \gamma_n)$ and $B_{3n}(\mathscr{B}_0, \gamma_0)$ from the validation data ($\Delta_i = 1$) are the same, so (17) holds if

$$
\{S_n(\gamma_n) - S_n(\gamma_0)\} = o_p(1), \tag{18}
$$

where

$$
\begin{aligned}
S_n(\gamma) = n^{-1/2} \sum_{i=1}^{n} &(1 - \Delta_i) G_{n\beta}(W_i^T \gamma, \mathscr{B}_0, \gamma) \\
&\times \frac{\{Y_i - G_n(W_i^T \gamma, \mathscr{B}_0, \gamma)\}}{\dot{G}_n(W_i^T \gamma, \mathscr{B}_0, \gamma)} \phi(W_i^T \gamma).
\end{aligned}
$$

Next note that

$$
\begin{aligned}
&S_n(\gamma_0) - n^{-1/2}\sum_{i=1}^{n}(1-\Delta_i)G_{n\beta}\big(W_i^T\gamma_0, \mathscr{B}_0, \gamma_0\big) \\
&\qquad \times \frac{\{Y_i - G_n(W_i^T\gamma_0, \mathscr{B}_0, \gamma_0)\}}{\dot{G}_n(W_i^T\gamma_0, \mathscr{B}_0, \gamma_0)}\phi(W_i^T\gamma_n) \\
&= n^{-1/2}\sum_{i=1}^{n}(1-\Delta_i)G_{n\beta}\big(W_i^T\gamma_0, \mathscr{B}_0, \gamma_0\big) \\
&\qquad \times \frac{\{Y_i - G_n(W_i^T\gamma_0, \mathscr{B}_0, \gamma_0)\}}{\dot{G}_n(W_i^T\gamma_0, \mathscr{B}_0, \gamma_0)}\{\phi(W_i^T\gamma_0) - \phi(W_i^T\gamma_n)\} \\
&= o_p(1).
\end{aligned} \tag{19}
$$

Hence a sufficient condition for (18) to hold is that

$$
\begin{aligned}
&S_n(\gamma_n) - n^{-1/2}\sum_{i=1}^{n}(1-\Delta_i)G_{n\beta}\big(W_i^T\gamma_0, \mathscr{B}_0, \gamma_0\big) \\
&\qquad \times \frac{\{Y_i - G_n(W_i^T\gamma_0, \mathscr{B}_0, \gamma_0)\}}{\dot{G}_n(W_i^T\gamma_0, \mathscr{B}_0, \gamma_0)}\phi(W_i^T\gamma_n) = o_p(1).
\end{aligned}
$$

Applying Assumption 9 on uniform convergence, it follows that

$$
\begin{aligned}
&\frac{G_{n\beta}(a, \mathscr{B}_0, \gamma_n)}{\dot{G}_n(a, \mathscr{B}_0, \gamma_n)} - R(a) = \mathscr{O}_p(n^{-1/2}) \quad \text{and} \\
&\frac{G_{n\beta}(a, \mathscr{B}_0, \gamma_0)}{\dot{G}_n(a, \mathscr{B}_0, \gamma_0)} - R(a) = \mathscr{O}_p(n^{-1/2}),
\end{aligned}
$$

uniformly over $a \in \mathscr{C}$. Therefore it is sufficient to show that

$$
\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}(1-\Delta_i)R(W_i^T\gamma_0)\phi(W_i^T\gamma_n) \\
&\qquad \times\{G_n(W_i^T\gamma_n, \mathscr{B}_0, \gamma_n) - G_n(W_i^T\gamma_0, \mathscr{B}_0, \gamma_0)\} = o_p(1).
\end{aligned} \tag{20}
$$

Consider the term $G_n(W_i^T\gamma_n, \mathscr{B}_0, \gamma_n) - G_n(W_i^T\gamma_0, \mathscr{B}_0, \gamma_0)$. Uniformly over $w^T\gamma_n \in \mathscr{C}$ and $w^T\gamma_0 \in \mathscr{C}$ we have

$$
\begin{aligned}
&G_n(w^T\gamma_n, \mathscr{B}_0, \gamma_n) - G_n(w^T\gamma_0, \mathscr{B}_0, \gamma_0) \\
&= \frac{C_n(w^T\gamma_n, \mathscr{B}_0, \gamma_n)}{D_n(w^T\gamma_n, \gamma_n)} - \frac{C_n(w^T\gamma_0, \mathscr{B}_0, \gamma_0)}{D_n(w^T\gamma_0, \gamma_0)} \\
&= \frac{D_n(w^T\gamma_0, \gamma_0)C_n(w^T\gamma_n, \mathscr{B}_0, \gamma_n) - C_n(w^T\gamma_0, \mathscr{B}_0, \gamma_0)D_n(w^T\gamma_n, \gamma_n)}{D_n(w^T\gamma_n, \gamma_n)D_n(w^T\gamma_0, \gamma_0)}.
\end{aligned} \tag{21}
$$

Assumption 9 and straightforward algebra show that (21) is

$$
\begin{aligned}
&\big[D(w^t\gamma_0,\gamma_0)\{C_n(w^T\gamma_n,\mathscr{B}_0,\gamma_n)-C_n(w^T\gamma_0,\mathscr{B}_0,\gamma_0)\}\\
&\quad+C(w^T\gamma_0,\mathscr{B}_0,\gamma_0)\{D_n(w^T\gamma_n,\gamma_n)-D_n(w^T\gamma_0,\gamma_0)\}\big]\\
&\quad\times\{D(w^T\gamma_0,\gamma_0)\}^{-2}+o_p(n^{-1/2})\\
&=\frac{\sum_{j=1}^n\Delta_jP(w^T\gamma_0,X_j)\big[K_h\{\gamma_n^T(W_j-w)\}-K_h\{\gamma_0^T(W_j-w)\}\big]}{\sum_{j=1}^n\Delta_j}\\
&\quad+o_p(n^{-1/2}).
\end{aligned}
\tag{22}
$$

Now use the fact that $n^{-1}\sum_{i=1}^n\Delta_i\to\pi$. Thus, we substitute (22) in (20) to get the sufficient condition

$$
\begin{aligned}
&n^{-3/2}\sum_{i=1}^n\sum_{j=1}^n(1-\Delta_i)\Delta_jQ(W_i^T\gamma_0,X_j)\phi(W_i^T\gamma_n)\\
&\qquad\times\big[K_h\{\gamma_n^T(W_j-W_i)\}-K_h\{\gamma_0^T(W_j-W_i)\}\big]\\
&\quad=o_p(1),
\end{aligned}
$$

or, equivalently,

$$
\begin{aligned}
T_n&=n^{-3/2}\{\pi(1-\pi)\}^{-1}\\
&\quad\times\sum_{i=1}^n\sum_{j=1}^n(1-\Delta_i)\Delta_jQ(W_i^T\gamma_0,X_j)\phi(W_i^T\gamma_n)\\
&\quad\times\big[K_h\{\gamma_n^T(W_j-W_i)\}-K_h\{\gamma_0^T(W_j-W_i)\}\big]\\
&=o_p(1),
\end{aligned}
$$

which we have already shown. This completes the proof. □

Acknowledgment. The authors wish to thank the referees for many helpful suggestions.

REFERENCES

CARROLL, R. J. and RUPPERT, D. (1988). *Transformation and Weighting in Regression.* Chapman & Hall, London.

CARROLL, R. J., SPIEGELMAN, C., LAN, K. K. G., BAILEY, K. T. and ABBOTT, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika* **71** 19–26.

CARROLL, R. J. and STEFANSKI, L. A. (1990). Approximate quasilikelihood estimation in models with surrogate predictors. *J. Amer. Statist. Assoc.* **85** 652–663.

CARROLL, R. J. and STEFANSKI, L. A. (1994). Measurement error, instrumental variables and corrections for attenuation with applications to meta-analysis. *Statistics in Medicine* **13** 1265–1282.

CARROLL, R. J. and WAND, M. P. (1991). Semiparametric estimation in logistic measurement error models. *J. Roy. Statist. Soc. Ser. B* **53** 573–585.

CROUCH, E. A. and SPIEGELMAN, D. (1990). The evaluation of integrals $\int_{-\infty}^{\infty}f(t)\exp(-t^2)\,dt$ and their applications to logistic–normal models. *J. Amer. Statist. Assoc.* **85** 464–467.

DUAN, N. and LI, K. C. (1991). Slicing regression: a link-free regression method. *Ann. Statist.* **19** 505–530.

FRIEDMAN, J. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

FULLER, W. A. (1987). *Measurement Error Models*. Wiley, New York.

GLESER, L. J. (1990). Improvement of the naive approach to estimation in nonlinear error-in-variables regression models. In *Statistical Analysis of Measurement Error Models and Application* (P. J. Brown and W. A. Fuller, eds.) 99–114. Amer. Math. Soc., Providence, RI.

HALL, P. (1989). On projection pursuit regression. *Ann. Statist.* **17** 573–588.

HÄRDLE, W. and STOKER, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995.

LI, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86** 337–342.

MACK, Y. and SILVERMAN, B. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Verw. Gebiete* **60** 405–415.

MARRON, J. S. and HÄRDLE, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation. *J. Multivariate Anal.* **20** 91–113.

PEPE, M. S. and FLEMING, T. R. (1991). A general nonparametric method for dealing with errors in missing or surrogate covariate data. *J. Amer. Statist. Assoc.* **86** 108–113.

PIERCE, D. A., STRAM, D. O., VAETH, M. and SCHAFER, D. (1992). The errors in variables problem: considerations provided by radiation dose–response analyses of the A-bomb survivor data. *J. Amer. Statist. Assoc.* **87** 351–359.

ROBINS, J. M., HSIEH, F. and NEWEY, W. (1994). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. Unpublished manuscript.

ROSNER, B., SPIEGELMAN, D. and WILLETT, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology* **132** 734–745.

ROSNER, B., WILLETT, W. C. and SPIEGELMAN, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* **8** 1051–1070.

SEPANSKI, J. H. and CARROLL, R. J. (1993). Semiparametric quasilikelihood and variance function estimation in measurement error models. *J. Econometrics* **58** 226–253.

SEPANSKI, J. H., KNICKERBOCKER, R. K. and CARROLL, R. J. (1994). A semiparametric correction for attenuation. *J. Amer. Statist. Assoc.* **89** 1366–1373.

R. J. CARROLL
DEPARTMENT OF STATISTICS
TEXAS A & M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143

R. K. KNICKERBOCKER
LILLY RESEARCH LABORATORIES
LILLY CORPORATE CENTER
INDIANAPOLIS, INDIANA 46285

C. Y. WANG
DIVISION OF PUBLIC HEALTH SCIENCES
FRED HUTCHINSON CANCER RESEARCH CENTER
SEATTLE, WASHINGTON 98104