J. R. Statist. Soc. B (1991)
53, *No.* 3, *pp.* 573–585

Semiparametric Estimation in Logistic Measurement Error Models

By R. J. CARROLL† and M. P. WAND

Texas A&M University, College Station, USA

[Received July 1989. Final revision February 1990]

SUMMARY

We describe semiparametric estimation and inference in a logistic regression model with measurement error in the predictors. The particular measurement error model consists of a primary data set in which only the response Y and a fallible surrogate W of the true predictor X are observed, plus a smaller validation data set for which (Y, X, W) are observed. Except for the underlying assumption of a logistic model in the true predictor, no parametric distributional assumption is made about the true predictor or its surrogate. We develop a semiparametric parameter estimate of the logistic regression parameter which is asymptotically normally distributed and computationally feasible. The estimate relies on kernel regression techniques. For scalar predictors, by a detailed analysis of the mean-squared error of the parameter estimate, we obtain a representation for an optimal bandwidth.

Keywords: BANDWIDTH SELECTION; DENSITY ESTIMATION; ERRORS IN VARIABLES; GENERALIZED LINEAR MODELS; KERNEL REGRESSION; LOGISTIC REGRESSION; MAXIMUM LIKELIHOOD; MEASUREMENT ERRORS MODELS; NONPARAMETRIC REGRESSION; PROBIT REGRESSION

1. INTRODUCTION

1.1. *Motivation and Literature*

This paper describes semiparametric estimation and inference in a logistic regression model with measurement error in the predictors. The primary concern is the univariate predictor case, although our main asymptotic normality result can be extended to higher dimensions. We first describe the example which motivated this work, and then the general model.

In the Nurses Health Study described by Rosner *et al.* (1989), the relationship between breast cancer (Y, a binary variable) and long-term dietary saturated fat (X) was examined prospectively. The primary data set consisted of a cohort of 89538 women, but instead of observing X, a surrogate W was observed, namely a self-administered quantitative food frequency questionnaire. To understand the relationship between X and W, 173 nurses became part of a validation study, in which Y, W and X were observed. X was not observed exactly but diet was measured sufficiently often in the validation data set, for one week at four different points in the year, that we may assume that X is known. The question to be addressed is how to use the validation data to obtain good estimates of logistic regression parameters.

†*Address for correspondence*: Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA.

 0035-9246/91/53573 $2.00

The logistic model is $\mathrm{pr}(Y=1\,|\,X=x)=F(\beta_0+\beta_1 x)$, where $F(v)=\{1+\exp(-v)\}^{-1}$. We assume that Y and W are independent given X. If $f_{X|W}$ is the conditional density of X given W, then in the primary study the observed data follow the model

$$\mathrm{pr}(Y=1\,|\,W=w)=\int F(\beta_0+\beta_1 x)\, f_{X|W}(x\,|\,w)\,\mathrm{d}x. \tag{1.1}$$

It is well known that a logistic regression of Y on W leads to inconsistent estimates of β_0 and β_1 (Stefanski and Carroll, 1985). In the Nurses Health Study, the measurement error is large and the asymptotic bias considerable; see Rosner *et al.* (1989).

The topic of binary regression when predictors are measured with error has been the subject of several recent papers. In this literature, Y has been unobserved in the validation data although extensions to our case are straightforward. The methods can be categorized as

(a) fully parametric,
(b) efficient semiparametric and
(c) approximate corrections for attenuation.

Carroll *et al.* (1984) and Schafer (1987) parameterize $f_{X|W}$ with a nuisance vector ξ. They compute a pseudo-maximum-likelihood estimate, using the following algorithm:

(a) estimate ξ from the validation data;
(b) pretend that in model (1.1) ξ is known and equal to its estimated value, thus yielding a pseudolikelihood function;
(c) estimate $\beta=(\beta_0,\beta_1)^{\mathrm{T}}$ by maximizing the pseudolikelihood over the primary data.

This method in the logistic case is difficult to compute, performs poorly in moderate sample sizes and is non-robust to misspecification of the conditional density $f_{X|W}$.

Stefanski and Carroll (1987) take a semiparametric approach. Specifically, W given X is assumed to be normally distributed with constant variance and mean linear in X; the mean and variance parameters are denoted by ξ. The marginal density of X, f_X, is assumed unknown. A sufficiency argument gives rise to a set of estimating equations depending on ξ and β. Again, ξ is estimated from the validation study and β is then estimated from the primary data assuming that ξ is known. This method makes fewer assumptions than the previous method, and works well for even moderate sample sizes, but the robustness against non-normal measurement error has not been explored.

A third approach is based on small measurement error asymptotics; see Stefanski and Carroll (1985), Stefanski (1985), Whittemore and Keller (1988) and Rosner *et al.* (1989). These methods pretend that the difference between X and W is 'small', which gives rise to estimates of β which partially correct for measurement error. There are again nuisance parameters ξ, which are estimated from the validation study. These methods work very well in practice, even though they are only partial corrections for attenuation. An asymptotic theory is given in Carroll and Stefanski (1990).

In this paper, we consider a fourth approach, namely using nonparametric kernel

regression methods in the validation data to estimate the probability function (1.1). A semiparametric estimate of β results from this method, with an estimated bandwidth and an asymptotically normal limit distribution. The next subsection outlines this approach. Independently, Pepe and Fleming (1991) consider a problem similar to ours with W a discrete random variable.

1.2. *Description of Method*

We assume throughout the paper that the joint, marginal and conditional densities of X and W have two bounded and continuous derivatives.

Define

$$B(x, y, \beta) = F(\beta_0 + \beta_1 x)^y \{1 - F(\beta_0 + \beta_1 x)\}^{1-y}. \tag{1.2}$$

The likelihood function for $Y = y$ given $W = w$ is given by

$$L(y, w, \beta) = E\{B(X, y, \beta) \mid W = w\}. \tag{1.3}$$

Equation (1.3) describes a nonparametric regression problem: regressing $B(X, y, \beta)$ on W. No numerical integration is required to obtain an estimate of the likelihood function. We propose to estimate equation (1.3) by a kernel regression.

In what follows, n_1 is the size of the validation data set, n_2 the size of the primary data set and $n_2/n_1 = \lambda$. Derivatives with respect to β are denoted by subscripts.

Let K be a symmetric density function, and let h be a bandwidth or window width. Define

$$\hat{f}_W(w) = (n_1 h)^{-1} \sum_1^{n_1} K\{(w - W_i)/h\},$$

$$D_n(y, w, \beta) = (n_1 h)^{-1} \sum_1^{n_1} B(X_i, y, \beta) K\{(w - W_i)/h\},$$

$$C_n(y, w, \beta) = (n_1 h)^{-1} \sum_1^{n_1} B_\beta(X_i, y, \beta) K\{(w - W_i)/h\}.$$

The estimated likelihood function for $Y = y$ given $W = w$ is $L_n(y, w, \beta) = D_n(y, w, \beta)/\hat{f}_W(w)$, while the estimated likelihood score is $H_n(y, w, \beta) = C_n(y, w, \beta)/D_n(y, w, \beta)$. Let $l(y, x, \beta) = (1, x)^T \{y - F(\beta_0 + \beta_1 x)\}$ be the likelihood score for $(Y, X) = (y, x)$.

In principle, there is information about β in both validation and primary data sets. The validation data contribute terms to the likelihood based on (Y, X), while the primary data are values of (Y, W). We use both types of information in our estimate.

Recall that $n = n_1 + n_2$. We propose that we estimate β by $\hat{\beta}$, the solution to

$$n^{-1/2} \sum_{j=1}^{n_1} l(Y_j, X_j, \beta) + n^{-1/2} \sum_{i=n_1+1}^{n} H_n(Y_i, W_i, \beta) = 0. \tag{1.4}$$

If $f_{X|W}$ were known, then we would replace H_n by H in equation (1.4) to obtain the likelihood equations. We shall quantify how much would be gained by making this replacement, in effect considering the cost due to estimating H by nonparametric regression.

In solving equation (1.4) we can use a scoring method with the starting value of β taken to be $\hat{\beta}^{(0)}$, the maximum likelihood estimate based on the validation data. An alternative estimator can be found for β by performing just one iteration of Newton's method with $\hat{\beta}^{(0)}$ as the starting value. Let $\hat{\beta}^{(1)}$ be this estimator. It may be shown that $\hat{\beta}$ and $\hat{\beta}^{(1)}$ have the same limit distribution.

An interesting feature of this problem is that when $\beta_1 = 0$, for $i \geqslant n_1 + 1$, it can be shown that the estimated score is unbiased, i.e. $EH_n(Y_i, W_i, \beta) = 0$. Thus, the classic bias–variance trade-off that we associate with nonparametric regression disappears when $\beta_1 = 0$.

1.3. *Arrangement of Paper*

In the next section, we summarize the asymptotic behaviour of $\hat{\beta}$ and $\hat{\beta}^{(1)}$ when $h \to 0$, where h is deterministic. In Section 3, we indicate methods for selecting h from the validation data, one of which is easy to implement. Section 4 describes the results of a simulation study.

2. ASYMPTOTICS FOR DETERMINISTIC BANDWIDTHS

2.1. *Introduction*

A practical problem occurs with this method as a result of edge effects. The estimated likelihood $H_n(y, w, \beta)$ will be unreliable near the boundaries of the validation data, where there are few observations and the weighted averaging of kernel regression becomes asymmetric. In addition, the primary data set, being larger, is expected to have observations W_i outside the range of the primary data. Used blindly, this would mean extrapolating the kernel fit $H_n(y, w, \beta)$ outside the range of the data used in its construction. Such extrapolation is dangerous, and robustness considerations dictate that it be avoided.

One method for overcoming this is to evaluate $H_n(y, w, \beta)$ only for those w in a fixed set interior to the support of W. Such a restriction is similar in spirit to the so-called Mallows method of robust regression, which downweights observations on the basis of leverage.

What follows are formal calculations, rather than detailed proofs. These calculations can be justified if the summations for $i \geqslant n_1 + 1$ in equation (1.4) are taken over those W_i in a fixed set interior to the support of W.

2.2. *Main Result*

Make the following definitions:

$$C(y, w, \beta) = f_W(w) E\{B_\beta(X, y, \beta) \mid W = w\};$$

$$D(y, w, \beta) = f_W(w) E\{B(X, y, \beta) \mid W = w\};$$

$$H(y, w, \beta) = C(y, w, \beta)/D(y, w, \beta);$$

$$M_\beta(\beta) = (1 + \lambda)^{-1} E\{l_\beta(Y, X, \beta) + \lambda H_\beta(Y, W, \beta)\}.$$

Also, let $\hat{\beta}^*$ stand for either $\hat{\beta}$ or $\hat{\beta}^{(1)}$. Since $H_n \to H$, by a Taylor series argument

$$n^{1/2}(\hat{\beta}^* - \beta) \approx -\{G_1(n, \beta) + G_2(n, \beta)\}^{-1} n^{1/2}\{G_3(n, \beta) + G_4(n, \beta)\}, \quad (2.1)$$

for random variables $G_i(n, \beta)$, $i = 1, \ldots, 4$, specified in equation (A.2) of Appendix A. If we assume that $h^4 n \to 0$ and $nh^2 \to \infty$, then calculations outlined in Appendix A indicate that $n^{1/2}(\hat{\beta}^* - \beta)$ is asymptotically normally distributed with mean zero and covariance matrix given by

$$n \operatorname{cov}(\hat{\beta}^* - \beta) \to M_\beta(\beta)^{-1} \Gamma(\beta) M_\beta(\beta)^{-1}, \tag{2.2}$$

where

$$\Gamma(\beta) = (1+\lambda)^{-1}[E\{l(Y, X, \beta)\, l(Y, X, \beta)^T\} + \lambda E\{H(Y, W, \beta)\, H(Y, W, \beta)^T\} + \lambda^2 \zeta(\beta)] \tag{2.3}$$

$$\zeta(\beta) = \sum_{y=0}^{1} \sum_{z=0}^{1} E\{L(z, W, \beta)\, L(y, W, \beta)\, Q(X, z, W, \beta)\, Q(X, y, W, \beta)^T f_W^2(W)\}$$

$$Q(x, y, w, \beta) = \frac{B_\beta(x, y, \beta)\, D(y, w, \beta) - B(x, y, \beta)\, C(y, w, \beta)}{D^2(y, w, \beta)} \tag{2.4}$$

We indicate in Section 5 and Appendix A that this result can be extended to arbitrary dimensions for X and W.

Remark 1. Each term in equation (2.3) has a distinct source. The first is the contribution of the validation data set: the Fisher information for β from validation. The second term is the Fisher information from the primary data set if $f_{X|W}$ were known. The third term represents the cost due to not knowing $f_{X|W}$. For the Nurses Health Study, if we assume that (X, W) are jointly normally distributed, then from information in Rosner *et al.* (1989) we conclude that X and W have standard deviations 4.6 and 5.9 respectively, and that, given W, X has a mean linear in W with slope 0.47 and standard deviation 3.7. Rosner *et al.* (1989) conclude for this example that the slope estimate obtained by regressing Y on W approximates $0.5\beta_1$, not β_1 itself; this is the effect of the measurement error. We also assume that X and W have the same means, which we take to be zero by centring. If we choose $\beta_0 = -5.0$ and, following Rosner *et al.* (1989), $\beta_1 = -0.018$, then the contribution due to estimating H is less than 1% of the total standard error of $\hat{\beta}_1$. If, however, $\beta_1 = -0.3$, then nearly 70% of the standard error is due to estimating H. This latter value of β_1 is used merely as an illustration, as it is much larger than would be expected in this study. □

Remark 2. In most applications, the primary data set is large relative to the validation data, i.e. λ is large. In the Nurses Health Study, $\lambda = 517.6$. In such cases, there is little information about β in the validation data set, and there will be little difference between solving equation (1.4) and solving

$$n^{-1/2} \sum_{i=n_1+1}^{n} H_n(Y_i, W_i, \beta) = 0. \tag{2.5}$$

The changes in the asymptotics when solving equation (2.5) is that $M_\beta(\beta) = E\{H_\beta(Y, W, \beta)\}$, $\Gamma(\beta) = E\{H(Y, W, \beta)\, H(Y, W, \beta)^T\} + \lambda \zeta(\beta)$ and $n = n_2$ in expression (2.2). □

Remark 3. When λ is extremely large, the main component in covariance (2.2) is $\zeta(\beta)$, which comes from the uncertainty in nonparametric estimation in the validation data. This gives one motivation to make the validation data set sufficiently

large that the randomness incurred by the nonparametric regression does not dominate. □

Remark 4. Equations (2.2) and (2.3) give some insight into the design of a study, in particular the choice of size of the validation study. If we assume that (X, W) are jointly normally distributed, then we can compute covariance (2.2) for various values of (β, λ) to obtain a sense of an acceptable λ. For the Nurses Health Study, following the assumptions of remark 1, with $\beta_0 = -5$ and $\beta_1 = -0.018$, we estimate that the effect of using a validation sample size $n_1 = 3750$ instead of the actual size $n_1 = 173$ is to decrease the standard error for estimating β_1 by less than 5%. The standard error can be reduced by approximately 40% by observing X for all study participants. However, if $\beta_1 = -0.3$, these figures are 73% and 88% respectively. □

Remark 5. If we model $f_{X|W}$ parametrically, a result similar to equations (2.2) and (2.3) occurs: an extra component in covariance due to estimating parameters via the validation study. See Section 4 for details. □

Remark 6. The covariance matrix (2.2) can be estimated by replacing $M_\beta(\beta)$, $\Gamma(\beta)$ and $\zeta(\beta)$ by their method-of-moments estimators. □

3. BANDWIDTH SELECTION

The proof of asymptotic normality with covariance (2.2) is sketched in Appendix A, under the condition that $nh^4 + (nh^2)^{-1} \to 0$. Unfortunately, this tells us nothing about selection of the bandwidth h. We might take the view that h should be varied over a wide range to see whether the estimates and inference are sensitive to h. We have sympathy with this data analytic viewpoint, but there is also value in letting h be determined by the data. In this section, we discuss automatic bandwidth selection. As a first step we derive a higher order expansion of the covariance of an asymptotically equivalent form of $n^{1/2}(\hat{\beta} - \beta)$.

Define

$$a_1 = \frac{\lambda}{2(1+\lambda)} \int z^2 K(z)\, dz \sum_{y=0}^{1} E\{L(y, W, \beta)\, Q(X, y, W, \beta)\, f_W(W)\, f_2(X, W)\} \tag{3.1}$$

$$a_2 = \lambda \int K^2(z)\, dz \sum_{y=0}^{1} E\{L(y, W, \beta)\, Q(X, y, W, \beta)\, f_W(W)\, B(X, y, \beta)/D(y, W, \beta)\} \tag{3.2}$$

$$f_2(x, w) = \{(\partial^2/\partial w^2) f_{X,W}(x, w)\}/f_{X,W}(x, w) \tag{3.3}$$

$$a_3(n, h) = M_\beta(\beta)^{-1}\{(nh^4)^{1/2} a_1 - (nh^2)^{-1/2} a_2\}.$$

In Appendix B, we outline a result showing that, for some random variable Z_n^* and matrix A, $n^{1/2}(\hat{\beta}^* - \beta) = Z_n^* + o_p\{(nh^4)^{1/2} + (nh^2)^{-1/2}\}$, where

$$E\, Z_n^* (Z_n^*)^T = M_\beta(\beta)^{-1} \Gamma(\beta) M_\beta(\beta)^{-1} + a_3(n, h)\, a_3(n, h)^T + (nh)^{-1} A. \tag{3.4}$$

Equation (3.4) shows that the bandwidth does not affect the covariance except to

smaller order terms. However, our second-order expansion does suggest a plug-in method for bandwidth selection; see below.

At the end of Section 1.2, we discussed the fact that $EH_n(Y, W, \beta) = 0$ when $\beta_1 = 0$. We can show in this case that $a_1 = a_2 = a_3(n, h) = (0, 0)^T$. Hence, equation (3.4) leads to choosing $h = \infty$, which suggests one reason why reliable automatic bandwidth selection based on plugging into equation (3.4) will be difficult, in general.

Remark 7. In terms of rates of convergence for estimating the linear combination $\gamma^T\beta$, the 'optimal' h minimizes $\{\gamma^T a_3(n, h)\}^2$. Except when $\beta_1 = 0$, this h is of the order $n^{-1/3}$, much smaller than the usual order of $n^{-1/5}$ common in nonparametric regression. In fact, the usual rate is prohibited in our calculations, if we are to estimate β at the rate $n^{1/2}$. □

The matrix A in equation (3.4) is complicated. For example, the contribution to A from the first-order linear expansion of H_n about H is $M_\beta(\beta)^{-1} J_2 M_\beta(\beta)^{-1}$, where

$$J_2 = \lambda \int K^2(z)\,dz \sum_{y=0}^{1} E\{L(y, W, \beta)\, Q(X, y, W, \beta)\, Q^T(X, y, W, \beta)\, f_W(W)\}. \tag{3.5}$$

The higher order terms are even more complex, perhaps too much so to be useful in plug-in bandwidth selection.

In the simulations to follow, we used a simple *ad hoc* bandwidth selection method. We took $h = \hat{\sigma}_W n^{-1/3}$, where $\hat{\sigma}_W$ is the estimated standard deviation of W in the validation data set; a robust scale could be used as well. This method, while *ad hoc*, does have the correct rate of convergence and is easily programmed. Unlike plug-in rules based on equation (3.4), it has the virtue of being stable even when $\beta_1 \approx 0$.

4. PARAMETRIC PROBLEMS

In parametric problems, the form of the limit distribution of $\hat{\beta}$ is the same as equation (2.2). Writing the densities as $f_{X|W}(x|w, \eta)$ and $f_W(w|\eta)$, the likelihood of an observed $(Y, X, W) = (y, x, w)$ in the validation data is $B(x, y, \beta) f_{X|W}(x|w, \eta) f_W(w|\eta)$. If

$$L(y, w, \beta, \eta) = \int B(x, y, \beta) f_{X|W}(x|w, \eta)\,dx,$$

then $L_*(y, w, \beta, \eta) = f_W(w|\eta) L(y, w, \beta, \eta)$ is the likelihood of an observed $(Y, W) = (y, w)$ in the primary data set. The full maximum likelihood estimate of (β, η) follows standard lines.

Let $H(y, w, \beta, \eta) = (\partial/\partial\beta) \log L(y, w, \beta, \eta)$. Let $\hat{\eta}$ be the maximum likelihood estimate of η in the validation data and define

$$\psi(x, w, \eta) = (\partial/\partial\eta) \log\{f_{X|W}(x|w, \eta) f_W(w|\eta)\}.$$

If $\hat{\beta}$ is the pseudo-maximum-likelihood estimate obtained by replacing η by $\hat{\eta}$ and maximizing the pseudolikelihood in β, then $\hat{\beta}$ has the limit distribution (2.2), with the exception that in equation (2.3) we replace $\zeta(\beta)$ by

$$\zeta_{\text{parm}}(\beta) = EH_\eta (E\psi_\eta)^{-1} E\psi\psi^T (E\psi_\eta)^{-1} EH_\eta^T.$$

5. MONTE CARLO STUDY

To understand the performance of the method when applied to data, we undertook a small Monte Carlo study.

The performance of our method was assessed in comparison with some other standard methods. The first was the usual logistic coefficient replacing X by W. Second was the estimate of Rosner *et al.* (1989), which divides the usual estimate by the slope of the regression of X on W in the validation data. The third was a modification of the Stefanski and Carroll (1985) estimate; see Stefanski (1989). We modified this slightly by applying the method not to W but to $W_* = (W - v_0)/v_1$, where v_0 and v_1 are the least squares intercept and slope from the regression of W on X. The final estimate was that of Whittemore and Keller (1988), p. 1060. Their parameters A and Ω were estimated from the validation data by assuming a linear regression of X on W.

The method proposed here started with the usual logistic regression estimates and used five iterations of the scoring method. The functions $\hat{f}_W$, C_n and D_n were assessed on a grid of 41 points covering the range of the validation data, with grid points being the equally spaced percentiles of W, i.e. the minimum, 0.025 percentile, 0.05 percentile, etc. Between grid points, linear interpolation was used. The sums over i in the formulae for these functions were assessed only for those W_i in the primary data which were in the range of the validation data.

The competing estimates are all parametric in nature and are based on specific parametric assumptions. If these assumptions hold, then we expect that these methods will outperform our method. Our simulations were based on the idea of calibrating our method in situations ideal for the parametric methods, i.e. linear additive normal measurement error. We also tested the methods against two moderate model departures, and against a severe model departure.

We took the primary data set to be of size $n_2 = 2000$ with the validation study of size $n_1 = 250$. We took $\beta_0 = -1.10$, -2.20, -3.66 and $\beta_1 = 0.80$. The three choices of β_0 represent cases where the expected numbers of times $Y = 1$ are 500, 200 and 50 respectively.

There were three sampling situations. In the first, the random variables (X, W) were normally distributed according to the model $W = X + U$, where (X, U) are independent with zero means and standard deviations 0.50.

The second sampling situation consisted of $W = XU$, where X is as in the previous example and U is the negative exponential distribution, so that the variance of W is twice that of X, as in the previous case. This is only a moderately heteroscedastic situation. However, a plot of X against W for a single data set, given in Fig. 1, suggests that the semiparametric method might do poorly in this case, because of the rapid change of the regression near $W = 0$.

The third sampling situation is of a moderate model breakdown for the parametric methods. We took X to be uniform on the interval (0, 5), and $W = X^2/15$. The values of β_0 were -2.7, -3.8 and -5.26. We say that this is a moderate model breakdown because the plot of X against W is reasonably linear for much of the range of W.

The fourth model situation represents severe model breakdown. We took W to be uniform on the interval $(-\pi/2, 5\pi/2)$, and $X = \cos W$.

Clearly this last choice can be criticized on the grounds that the parametric

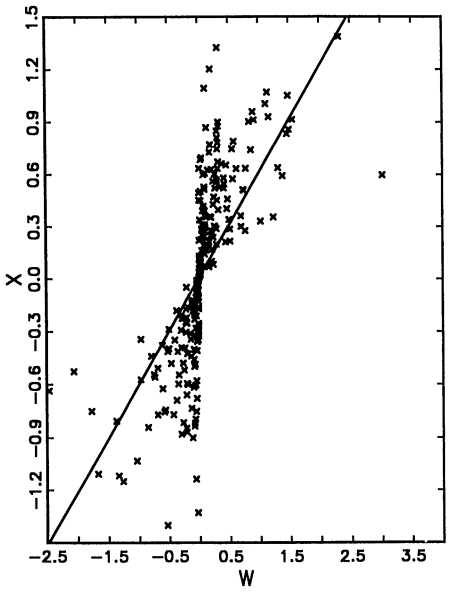

Fig. 1. Plot of X against W for a randomly generated sample of size 250: here, X is normally distributed with mean zero and standard deviation 0.5, while $W = XU$ and U follows a negative exponential distribution with unit mean

methods are inappropriately applied. However, if we only simulated situations where the parametric assumptions were appropriate, then we would find the unsurprising result that parametric is better than semiparametric in practice. What is of interest here is to see whether the semiparametric method does reasonably well in straightforward cases, and better in cases of model breakdown.

There were 100 iterations of the experiment. The results are given in Table 1, where we report median values for $\hat{\beta}_1$, the median absolute error (MAE) and the 90th percentile of the absolute errors. The semiparametric method does remarkably well in the normal model and is the clear winner in the quadratic and cosine models. As expected, its performance for rare events in the heteroscedastic model is relatively poorer compared with the method of Rosner *et al.* (1989), but it is still not unacceptable.

6. GENERALIZATIONS AND DISCUSSION

Most applications of logistic regression involve non-scalar predictors. One important example consists of the case where the predictors are (Z, X). Here Z is observable, X is scalar and unobservable, and W is the scalar surrogate for X. If the distribution of X given W is independent of Z, then our results apply with only the following minor modifications. In equation (1.2), define $B(x, z, y, \beta) = F(\beta_0^T z + \beta_1 x)^y \{1 - F(\beta_0^T z + \beta_1 x)\}^{1-y}$, while in equation (1.3) define $L(y, z, w, \beta) = E\{B(X, z, y, \beta) \mid W = w\}$. These changes lead to redefined functions such as $D_n(y, z, w, \beta)$, $D(y, z, w, \beta)$ and $l(Y, X, Z, \beta)$. Otherwise, the results are unchanged.

TABLE 1
Results of a simulation study†

Estimate	*Results for the following values of β:*								
	$\beta_0 = -1.09$			$\beta_0 = -2.20$			$\beta_0 = -3.66$		
	Median	*MAE*	*90th*	*Median*	*MAE*	*90th*	*Median*	*MAE*	*90th*
Additive normal									
Usual	0.39	0.413	0.496	0.40	0.404	0.532	0.36	0.436	0.680
W-K	0.78	0.100	0.259	0.80	0.135	0.362	0.72	0.314	0.688
R-W-S	0.78	0.096	0.244	0.78	0.126	0.340	0.72	0.312	0.726
S-C	0.77	0.164	0.376	0.82	0.153	0.467	0.78	0.333	0.757
Semi-p	0.79	0.099	0.309	0.80	0.135	0.380	0.72	0.314	0.688
Heteroscedastic normal									
Usual	0.41	0.395	0.508	0.40	0.404	0.524	0.40	0.401	0.685
W-K	1.36	0.563	1.052	1.25	0.453	0.805	1.09	0.421	0.871
R-W-S	0.79	0.137	0.311	0.78	0.152	0.370	0.77	0.255	0.606
S-C	0.55	0.289	0.462	0.53	0.290	0.483	0.48	0.331	0.653
Semi-p	0.88	0.109	0.245	0.90	0.133	0.405	0.91	0.329	0.776
Cosine model									
Usual	0.00	0.800	0.803	0.00	0.800	0.806	0.00	0.800	0.807
W-K	−0.01	0.806	0.856	−0.01	0.812	0.879	0.00	0.801	0.928
R-W-S	−0.29	1.975	8.926	0.31	2.183	14.949	0.30	4.765	20.277
S-C	0.00	0.800	0.803	0.00	0.800	0.806	0.00	0.800	0.808
Semi-p	0.90	0.101	0.214	0.90	0.133	0.307	0.92	0.217	0.513
	$\beta_0 = -2.70$			$\beta_0 = -3.80$			$\beta_0 = -5.26$		
Quadratic model									
Usual	0.67	0.131	0.191	0.63	0.173	0.235	0.64	0.158	0.288
W-K	1.92	1.116	1.615	0.86	0.154	0.423	0.53	0.300	0.536
R-W-S	0.74	0.064	0.126	0.69	0.109	0.177	0.70	0.101	0.238
S-C	0.67	0.127	0.188	0.63	0.170	0.232	0.65	0.155	0.286
Semi-p	0.80	0.035	0.095	0.80	0.045	0.116	0.82	0.091	0.213

†Median is the median of the estimates, MAE is the median absolute error and 90th is the 90th percentile of the absolute error. Usual is the ordinary logistic regression estimate, W-K is the Whittemore and Keller (1988) estimate, R-W-S is the estimate of Rossner *et al.* (1989), S-C is the modified Stefanski and Carroll (1985) method discussed in the text and Semi-p is the semiparametric method. The models are as discussed in the text.

If W is non-scalar or if X given W is not independent of Z, then nonparametric regression must be performed in more than one dimension. We show in Appendix A that if W has dimension d, if all densities have p bounded and continuous derivatives and if K is a pth-order kernel function, then result (2.2) holds as long as $nh^{2p} \to 0$ and $nh^{2d} \to \infty$. The case discussed in Section 2 is $d = 1$ and $p = 2$. Note that the dimension of W need not be the same as the dimension of X. However, such a method is hardly practical if $d = 10$. The problem of suitable methods for higher dimensional surrogates remains open.

ACKNOWLEDGEMENTS

We are grateful to the referees for their helpful comments which led to significant improvements in the presentation of our results. Our research was supported by

the National Institutes of Health and Sonderforschungsbereich 303, University of Bonn.

APPENDIX A

This section discusses asymptotic normality of the estimates of β. We shall assume that W is of dimension d, that the kernel K is a pth-order kernel, that $nh^{2d} \to \infty$ and that $nh^{2p} \to 0$. The case discussed in Section 2 is $p = 2$ and $d = 1$.

Since the size of the validation data set is $O(n)$, $n^{1/2}$-consistent estimates $\hat{\beta}^{(0)}$ of β are already available from the validation data. Thus, the analysis of the estimate $\hat{\beta}^{(1)}$ will follow standard lines of argument, as will that of $\hat{\beta}$. Rather than to use considerable space in tedious but standard arguments, we have chosen to take approximation (2.1) as our starting point.

Dropping the dependence on β, define $Q(x, y, w)$ by equation (2.4) and

$$Q_*(x, y, w) = Q(x, y, w)/D(y, w, \beta).$$

Where appropriate, we shall suppress arguments to individual terms, e.g. $H_{n,i}$ for $H_n(Y_i, W_i, \beta)$. In the subsequent calculations, we shall also use the notation

$$\tilde{Q}_{n,i} = (n_1 h^d)^{-1} \sum_{j=1}^{n_1} K\{(W_j - W_i)/h\}\, Q(X_j, Y_i, W_i). \tag{A.1}$$

An analogous definition is ascribed to $\tilde{Q}_{*,n,i}$. Let $\eta_n = nh^{2p} + (nh^{2d})^{-1}$. To order $o_p(\eta_n^{1/2})$, we can show that

$$n^{1/2}(\hat{\beta}^{(1)} - \beta) \approx -\left(n^{-1} \sum_{1}^{n_1} l_{i,\beta} + n^{-1} \sum_{n_1+1}^{n} H_{n,i,\beta}\right)^{-1} n^{1/2}\left(n^{-1} \sum_{1}^{n_1} l_i + n^{-1} \sum_{n_1+1}^{n} H_{n,i}\right). \tag{A.2}$$

Note that $G_i(n, \beta)$, $i = 1, \ldots, 4$, from approximation (2.1) are the four normalized sums in this expression. Except for terms of order $o_p(\eta_n^{1/2})$, we can write $H_{n,i} = H_i + \tilde{Q}_{n,i} - \tilde{Q}_{*,n,i}(D_{n,i} - D_i)$. Making this substitution and then taking derivatives with respect to β, we can show by computing first and second moments that

$$n^{-1} \sum_{1}^{n_1} l_{i,\beta} + n^{-1} \sum_{n_1+1}^{n} H_{n,i,\beta} - M_\beta(\beta) = o_p(\eta_n^{1/2}). \tag{A.3}$$

We shall use equations (A.2) and (A.3) in Appendix B.

Since $H_{n,i} - H_i - \tilde{Q}_{n,i} = o_p(n^{-1/2})$ under our conditions on the bandwidth h, for result (2.2) we need to show that Z_n is asymptotically normal with mean zero and covariance $\Gamma(\beta)$, where

$$Z_n = n^{-1/2}\left\{\sum_{j=1}^{n_1} l_j + \sum_{i=n_1+1}^{n} (H_i + \tilde{Q}_{n,i})\right\}.$$

Let $\mathcal{G}_{n_1}$ denote the observations in the validation data, i.e. $(X_i, W_i, Y_i)_1^{n_1}$. Define $\alpha_n = E(\tilde{Q}_{n,n} | \mathcal{G}_{n_1})$. A standard bias calculation shows that, except for terms of order $o_p(1)$,

$$\alpha_n = (\lambda/n_2) \sum_{j=1}^{n_1} S_j,$$

where

$$S_j = \sum_{y=0}^{1} L(y, W_j, \beta)\, Q(X_j, y, W_j)\, f_W(W_j).$$

It is easy to show by a covariance calculation that

$$n^{-1/2} \sum_{n_1+1}^{n} (\tilde{Q}_{n,i} - \alpha_n) = o_p(1),$$

and hence that

$$Z_n = n^{-1/2} \left\{ \sum_{j=1}^{n_1} (l_j + \lambda S_j) + \sum_{i=n_1+1}^{n} H_i \right\} + o_p(1). \tag{A.4}$$

The right-hand side of equation (A.4) has covariance $\Gamma(\beta)$, and it is clearly asymptotically normally distributed.

APPENDIX B

The purpose of this section is to verify equation (3.4). We are taking the dimension of W and X to be $d = 1$, and the order of the kernel to be $p = 2$; see equation (A.1). On the basis of the introductory comments in Appendix A, specifically equations (A.2) and (A.3), it suffices to compute the mean and covariance of

$$Z_n^* = n^{-1/2} \left[\sum_{1}^{n_1} l_i + \sum_{n_1+1}^{n} \{H_i + \tilde{Q}_{n,i} - \tilde{Q}_{*,n,i}(D_{n,i} - D_i)\} \right] = U_{1n} + U_{2n} + U_{3n} - U_{4n}.$$

Obviously, since they are scores, $E(l_i) = E(H_i) = 0 = E(U_{1n}) = E(U_{2n})$.

Define $f_2(x, w)$ as in equation (3.3). Standard bias calculations show that to order $o(\eta_n^{1/2})$

$$\begin{aligned} E(U_{3n}) &= n^{1/2} h^2 a_1 \\ E(U_{4n}) &= (n^{1/2} h)^{-1} a_2, \end{aligned} \tag{B.1}$$

where a_1 and a_2 are given by equations (3.1) and (3.2) respectively.

The covariance terms take some effort to compute. Of course, $E(U_{1n} U_{1n}^{\mathrm{T}}) = (1 + \lambda)^{-1} E(ll^{\mathrm{T}})$, $E(U_{2n} U_{2n}^{\mathrm{T}}) = \lambda E(HH^{\mathrm{T}})/(1 + \lambda)$ and $E(U_{1n} U_{2n}) = 0$. By direct and fairly easy calculation, to order $o(\eta_n)$, $E(U_{1n} U_{3n}) = E(U_{1n} U_{4n}) = 0$.

Define $c_1(K) = \int K^2(z)\, dz$. It is also relatively easy to show that $E(U_{2n} U_{3n}^{\mathrm{T}}) = O(h^2)$, and that to order $o(\eta_n)$

$$E(U_{2n} U_{4n}^{\mathrm{T}}) = \frac{\lambda}{nh} c_1(K) \sum_{y=0}^{1} E\{L(y, W, \beta)\, H(y, W, \beta)\, Q_*^{\mathrm{T}}(X, y, W)\, B(X, y, W)\, f_W(W)\}. \tag{B.2}$$

A somewhat lengthier calculation yields that

$$\begin{aligned} \operatorname{cov}(U_{3n}) = {} & \lambda^2 \zeta(\beta)/(1 + \lambda) + \lambda (nh)^{-1} c_1(K) \\ & \times \sum_{y=0}^{1} E\{L(y, W, \beta)\, Q(X, y, W)\, Q^{\mathrm{T}}(X, y, W)\, f_W(W)\}. \end{aligned} \tag{B.3}$$

We can show that $E(U_{2n} U_{3n}^{\mathrm{T}}) = O(h^2)$, and that $E(U_{3n} U_{4n}^{\mathrm{T}})$ can be written as

$$E(U_{3n}U_{4n}^{\mathrm{T}}) = \lambda^2(nh)^{-1}c_1(K)\sum_{y,z=0}^{1} E\{L(y, W, \beta)L(z, W, \beta)f_W^2(W) \times B(X, z, \beta)Q(X, y, W)Q_*(X, z, W)^{\mathrm{T}}\} + o(\eta_n). \quad \text{(B.4)}$$

Finally, we note that $E(U_{4n}U_{4n}^{\mathrm{T}})$ is, to order $o(\eta_n)$, the sum of the following two terms:

$$\lambda^2(nh)^{-1}\alpha(K)\sum_{y,z=0}^{1}\int L(z, w, \beta)L(y, w, \beta)Q_*(x_2, z, w)Q_*^{\mathrm{T}}(x_2, y, w)B(x_1, z, \beta) \times B(x_1, y, \beta)f_{X,W}(x_1, w)f_{X,W}(x_2, w)f_W^2(w)\,dx_1\,dx_2\,dw; \quad \text{(B.5)}$$

$$\lambda^2(nh)^{-1}\alpha(K)\sum_{y,z=0}^{1}\int L(z, w, \beta)L(y, w, \beta)Q_*(x_1, z, w)Q_*^{\mathrm{T}}(x_2, y, w)B(x_1, z, \beta) \times B(x_2, y, \beta)f_{X,W}(x_1, w)f_{X,W}(x_2, w)f_W^2(w)\,dx_1\,dx_2\,dw, \quad \text{(B.6)}$$

where

$$\alpha(K) = \int K(z_1)K(z_2)K(z_1+z_3)K(z_2-z_3)\,dz_1\,dz_2\,dz_3.$$

We can collect terms to compute equation (3.4). The matrix J_2 in equation (3.5) comes from the second part of equation (B.3). The terms (B.2), (B.4), (B.5) and (B.6) arise from U_{4n}, which is the error in linearizing H_n about H.

REFERENCES

Carroll, R. J., Spiegelman, C. H., 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 quasi-likelihood estimation in models with surrogate predictors. *J. Am. Statist. Ass.*, **85**, 652–663.

Pepe, M. S. and Fleming, T. R. (1991) A general nonparametric method for dealing with errors in missing or surrogate data. *J. Am. Statist. Ass.*, to be published.

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. *Statist. Med.*, **8**, 1075–1093.

Schafer, D. (1987) Covariate measurement error in generalized linear models. *Biometrika*, **74**, 385–389.

Stefanski, L. A. (1985) The effects of measurement error on parameter estimation. *Biometrika*, **72**, 583–592.

——— (1989) Correcting data for measurement error in generalized linear models. *Communs Statist. A*, **18**, 1715–1734.

Stefanski, L. A. and Carroll, R. J. (1985) Covariate measurement error in logistic regression. *Ann. Statist.*, **13**, 1335–1351.

——— (1987) Conditional scores and optimal scores for generalized linear measurement error models. *Biometrika*, **74**, 703–716.

Whittemore, A. S. and Keller, J. B. (1988) Approximations for errors in variables regression. *J. Am. Statist. Ass.*, **83**, 1057–1066.