

ON ROBUST ESTIMATION IN CASE–CONTROL STUDIES WITH ERRORS IN COVARIATES

C. Y. Wang & R. J. Carroll
Department of Statistics
Texas A&M University
College Station, TX 77843–3143

March 28, 1996

Abstract

We discuss robust estimates based on leverage downweighting of the coefficients in logistic regression case-control studies when the true predictor X is an unobserved continuous random variable, and instead a surrogate W is observed with nondifferential measurement error. The asymptotic theory of the estimates is derived, and consistent standard error estimates obtained. An illustrative example shows that the adverse effects of outliers upon standard methods can be somewhat ameliorated by our techniques.

Some Key Words: Asymptotics; Case–control studies; Logistic regression; Measurement error; Robust estimates.

1 INTRODUCTION

Prospective studies and retrospective (case-control) studies are two important tools for the study of factors related to diseases. Somewhat simplistically, the difference between the two is a matter of conditioning. One version of a prospective study is a random sample from a source population. In a retrospective study one obtains groups of individuals with and without the disease (usually not by random sampling), and then one ascertains the values of the covariates. A classical case-control sample contains no information about the marginal probabilities $\Pr(D = 1)$.

We consider robust estimation in case-control studies where some of the covariates are subject to nondifferential measurement error. Let D be disease status, where $D = 0$ refers to controls and $D = 1$ to cases. There are n_1 cases and n_0 controls, and $n = n_0 + n_1$. Let X be the continuous true exposure variable and denote other covariates not subject to error by Z . By the nature of the design of a case-control study, the distribution of the covariates (X, Z) is conditioned on case or control status D . We assume that X is a continuous random variable.

The problem of interest here is the case that one cannot easily ascertain X , but instead one can observe a value W , assumed to be a surrogate, i.e., W and D are independent given (X, Z) . The fact that W is a surrogate is equivalent to the notion of nondifferential measurement error, i.e., the distribution of W given (X, Z, D) depends only on (X, Z) but not otherwise on D .

For most purposes in this paper (except §5), the observed data consist of two types. In the primary study, we observe (Z, W, D) , while in addition there is a smaller validation study for which we observe X as well. The data are gathered by first observing (Z, W, D) for all study participants, and next X is measured in the validation data. Let the dimensions of Z , X and W be $k_0 \times 1$, $k_1 \times 1$ and $k_2 \times 1$ respectively.

A standard analysis is to start with a prospective model for the source population, and then use the case-control data to estimate the parameters of that model. We base our analysis on a prospective logistic regression model

$$\Pr(D = 1|Z, X) = H(\beta'_0 + \beta_1^t Z + \beta_2^t X), \tag{1}$$

where $H(\cdot)$ is the logistic distribution function. Prentice and Pyke (1979) show that if X were observable then the retrospective density of (Z, X) given $D = d$ in the validation data is

$$f_{Z,X|D}(z, x|d) = (n/n_d)q(z, x)H^d(\beta_0 + \beta_1^t z + \beta_2^t x)\{1 - H(\beta_0 + \beta_1^t z + \beta_2^t x)\}^{1-d}, \tag{2}$$

where if π_d is the unidentifiable probability that $D = d$ in the source population, then $\beta_0 = \beta'_0 + \log\{(n_0\pi_1)/(n_1\pi_0)\}$ is a new intercept and $q(z, x)$ is the marginal density of (Z, X) in the population induced by the case-control sampling scheme. Because we cannot estimate (π_0, π_1) using a case-control design, the underlying intercept β'_0 in the source population is not identified.

Prentice & Pyke showed that ordinary logistic regression applied to (X, Z, D) data yields consistent estimates of (β_1, β_2) with asymptotically correct standard errors, even though the retrospective sampling scheme is ignored. Wang & Carroll (1993) showed that standard robust logistic regression estimates appropriate for prospective designs are also consistent with asymptotically correct standard errors when applied to case-control studies.

In §2, we review one such class of robust estimates, based on downweighting observations on the basis of their leverage and, in the logistic model, on a misclassification model.

For the measurement error problem, Rosner, et al. (1989), Carroll, et al. (1993) and Satten & Kupper (1993) have proposed different methods for estimating (β_1, β_2) . These methods are nonrobust, in the usual sense that leverage points and outliers can have large effects on the estimates of (β_1, β_2) and their standard errors. We illustrate this phenomenon in an example described in §6.

Our paper considers robust estimation when X is continuous. Carroll, et al. (1993) assume that $E(X|Z, W, D)$ is linear in (Z, W, D) . They propose to use an estimate of $m(Z, W) = E(X|Z, W)$ as an estimated value for X in the primary data. Their algorithm factors naturally into a linear regression of X on (Z, W, D) , followed by a logistic regression. One could robustify each step using the linear results of Simpson, et al. (1992) and the logistic results of Carroll & Pederson (1993). We take a slightly more general approach. Using calculations of Satten & Kupper (1993) for the case that X is normally distributed, we show that the estimating equations for maximum likelihood are of a particularly convenient linear form. This immediately suggests the development of robust estimation methods with bounded influence functions, these being based on separately downweighting leverage and residuals. In §3, the estimation methods are defined and asymptotic distribution theory is stated in §4. In §5, we extend the results to the case that X is not measured. An illustrative example is considered in §6. Some extensions are outlined in §7.

2 LEVERAGE DOWNWEIGHTING

The robust estimates we use are based on downweighting observations on the basis of their outlyingness in factor space. In linear regression of a variable Y on X , leverage based estimates downweight

observations separately on the basis of the distance of the Y 's to the line and the distances of the X 's to their center, see Simpson, et al. (1992). In logistic regression, the estimates take the form of weighted logistic regression estimates with weights depending on the X 's and the regression parameter but not otherwise on disease status D , see Carroll & Pederson (1993). Generally, for some vector-valued function $\hat{\xi}$ of the X -sample, the weights for a given value of the regression parameter $\mathcal{B} = (\beta_0, \beta_1, \beta_2)$ are of the form $\phi(X, \mathcal{B}, \hat{\xi})$.

For example, in logistic regression of D on (Z, X) in a case-control study, the leverage-based estimates of $\mathcal{B} = (\beta_0, \beta_1, \beta_2)$ solve the equation

$$0 = \sum_{i=1}^n \phi(X_i, \mathcal{B}, \hat{\xi}) (1, Z_i^t, X_i^t)^t \left\{ D_i - H(\beta_0 + \beta_1^t Z_i + \beta_2^t X_i) \right\}. \quad (3)$$

3 NONROBUST METHODS

If (X, W) are continuous, Rosner, et al. (1989), Carroll & Stefanski (1990), Gleaser (1990) and Pierce, et al. (1992) note that for rare diseases with moderate effect (β_2 not too large), a good approximation to the *prospective* model (1) is

$$\Pr(D = 1|Z, W) \simeq H\{\beta_0'' + \beta_1^t Z + \beta_2^t m(Z, W)\}. \quad (4)$$

In (4), the intercept β_0'' for the regression depending on (Z, W) is deliberately allowed to be different from the intercept β_0' in (1) for the regression depending on (Z, X) .

Applying the same logic that led to (2), there is an intercept β_{00} such that the retrospective likelihood of (Z, W) given D is approximately proportional to

$$f_{Z,W|D}(z, w|d) \sim H^d\{\beta_{00} + \beta_1^t z + \beta_2^t m(z, w)\} \left[1 - H\{\beta_{00} + \beta_1^t z + \beta_2^t m(z, w)\} \right]^{1-d}. \quad (5)$$

Our main results are for the case that selection into the validation sample, i.e., observing X , depends only on case or control status and not otherwise on (Z, W) , but see §7 for a description of some alternatives. Under this sampling framework, we can use the validation data in an intuitively pleasing fashion, as well as not be restricted only to a likelihood analysis. We will condition on the number in the primary and validation data sets. There will be n_{P1} cases and n_{P0} controls in the primary data, and n_{V1} cases and n_{V0} controls in the validation data; $n_P = n_{P0} + n_{P1}$ and $n_V = n_{V0} + n_{V1}$. In what follows, we will let $(X_{Vid}, Z_{Vid}, W_{Vid})$ and (Z_{Pid}, W_{Pid}) denote observations for the i^{th} individual having $D = d$ in the validation and primary data sets, respectively.

The algorithm of Carroll, et al. (1993) assumes that the distribution of X given (Z, W, D) is linear with an intercept depending on D but constant variance:

$$X|(Z, W, D) = \alpha_{00} + \alpha_{01}D + \alpha_1 Z + \alpha_2 W + \epsilon, \quad (6)$$

where ϵ is has conditional mean zero and covariance matrix Σ_ϵ , and $(\alpha_{00}, \alpha_{01}, \alpha_1, \alpha_2)$ are matrices of dimension $(k_1 \times 1)$, $(k_1 \times 1)$, $(k_1 \times k_0)$ and $(k_1 \times k_2)$, respectively. Satten & Kupper's results can be used to show that if (6) holds among the controls with normally distributed error and variance σ^2 , then it also holds among the cases, with $\alpha_{01} = \sigma^2 \beta_2$.

It is possible to robustify a likelihood approach. However, the results would depend heavily on the assumption of normality, and we instead have chosen to take an approach which yields approximately consistent estimates in a variety of circumstances.

In most applications of case-control studies the disease is necessarily rare, so that regressions in the source population are similar to regressions among the controls. Hence, in particular, $m(Z, W) \approx E(X|Z, W, D = 0)$. Using (6) and absorbing α_{00} into the intercept in (5), we find that the likelihood of (Z, W) given D is approximately proportional to

$$f_{Z,W|D}(z, w|d) \sim H^d(\cdot) \{1 - H(\cdot)\}^{1-d}; \quad (7)$$

where $H(\cdot) = H\{\beta_{0*} + \beta_1^t z + \beta_2^t (\alpha_1 Z + \alpha_2 W)\}$, and $\beta_{0*} = \beta_{00} + \beta_2^t \alpha_{00}$ is a new intercept. Define $\mathcal{T} = \{1, Z^t, (\alpha_1 Z + \alpha_2 W)^t\}^t$. The likelihood is $f_{X|Z,W,D}(\cdot) f_{Z,W|D}(\cdot)$. Thus, from (7), if (α_1, α_2) were known, an approximately unbiased estimating equation for $(\beta_{0*}, \beta_1, \beta_2)$ is

$$\begin{aligned} 0 &= \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \mathcal{T}_{Vid} \left[d - H \left\{ \beta_{0*} + \beta_1^t Z_{Vid} + \beta_2^t (\alpha_1 Z_{Vid} + \alpha_2 W_{Vid}) \right\} \right] \\ &+ \sum_{d=0}^1 \sum_{i=1}^{n_{Pd}} \mathcal{T}_{Pid} \left[d - H \left\{ \beta_{0*} + \beta_1^t Z_{Pid} + \beta_2^t (\alpha_1 Z_{Pid} + \alpha_2 W_{Pid}) \right\} \right]. \end{aligned} \quad (8)$$

We will exploit (8) in §7.

However, we are mainly interested in the estimation of β_1 and β_2 since they measure the effects of covariates Z and X . In the validation data we actually observe X , so it is intuitively appealing to use the X -data in the first term in (8). We will use a different intercept for each of the two terms, because the likelihood (2) and the approximation (5) have potentially different intercepts. If we define $\mathcal{T}_{Vid} = (1, 0, Z_{Vid}^t, X_{Vid}^t)^t$ and $\mathcal{T}_{Pid} = \{0, 1, Z_{Pid}^t, (\alpha_1 Z_{Pid} + \alpha_2 W_{Pid})^t\}^t$, this leads to the estimating equation

$$0 = \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \mathcal{T}_{Vid} \left\{ d - H \left(\beta_0 + \beta_1^t Z_{Vid} + \beta_2^t X_{Vid} \right) \right\} \quad (9)$$

$$+ \sum_{d=0}^1 \sum_{i=1}^{n_{Pd}} \mathcal{T}_{Pid} \left[d - H \left\{ \beta_{0*} + \beta_1^t Z_{Pid} + \beta_2^t (\alpha_1 Z_{Pid} + \alpha_2 W_{Pid}) \right\} \right].$$

We now turn to constructing estimating equations for (α_1, α_2) . In the validation data, ordinary least squares can be used. Define $S_{id} = (1, d, Z_{Vid}^t, W_{Vid}^t)^t$, $e_{Vidj} =$ the j^{th} component of e_{Vid} , where $e_{Vid} = X_{Vid} - (\alpha_{00} + \alpha_{01}d + \alpha_1 Z_{Vid} + \alpha_2 W_{Vid})$. Define $\mathcal{A} = (\alpha_{00}, \alpha_{01}, \alpha_1, \alpha_2)$, then the estimating equation for \mathcal{A} in the validation data is

$$0 = \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} (S_{id} \otimes e_{Vid}), \quad (10)$$

where \otimes denotes a Kronecker product. Solving (9)–(10) simultaneously yields the estimate of Carroll, et al. (1993). There is, however, information about \mathcal{A} in the primary data, the derivative of the score function being

$$\sum_{d=0}^1 \sum_{i=1}^{n_{Pd}} A_{Pid} \left[d - H \left\{ \beta_{0*} + \beta_1^t Z_{Pid} + \beta_2^t (\alpha_1 Z_{Pid} + \alpha_2 W_{Pid}) \right\} \right], \quad (11)$$

where

$$A_{Pid} = \left[0, 0, \{\text{vec}(Z_{Pid}\beta_2^t)\}^t, \{\text{vec}(W_{Pid}\beta_2^t)\}^t \right]^t. \quad (12)$$

The appeal of solving (9)–(10) is that no special software is necessary. The linear regression parameters (α_1, α_2) are estimated by least squares. The logistic regression parameters (β_1, β_2) are estimated via logistic regression with a dummy variable indicating whether one has validation or primary data, and predictors X or $(\alpha_1 Z + \alpha_2 W)$ in the validation and primary data, respectively.

In this paper, our numerical computations will be based on the simplicity inherent in this method. While we have not encountered them, there may be situations where the extra information in the primary data for estimating (α_1, α_2) will be worth the additional computing complexity. Hence, we will allow for the use of the extra information in the unbiased estimating equation (11).

The combined estimating equation is

$$0 = \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \begin{pmatrix} \mathcal{T}_{Vid} \\ 0 \end{pmatrix} \{d - H(\beta_0 + \beta_1^t Z_{Vid} + \beta_2^t X_{Vid})\} + \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \begin{pmatrix} 0 \\ S_{id} \otimes e_{Vid} \end{pmatrix} \\ + \sum_{d=0}^1 \sum_{i=1}^{n_{Pd}} \begin{pmatrix} \mathcal{T}_{Pid} \\ A_{Pid} \end{pmatrix} \left[d - H \left\{ \beta_{0*} + \beta_1^t Z_{Pid} + \beta_2^t (\alpha_1 Z_{Pid} + \alpha_2 W_{Pid}) \right\} \right], \quad (13)$$

where A_{Pid} is set equal to zero if one wants to use the simple linear–logistic method, and equal to (12) if one wants to attempt a more complex but possibly more efficient analysis.

It is obvious that the estimates which solve (13) are not robust. In the next section, we will use leverage downweighting to construct new estimates with bounded influence functions.

4 ROBUST ESTIMATION

We will construct a bounded influence estimate for $\Theta = \{\beta_0, \beta_{0*}, \beta_1^t, \beta_2^t, \text{vec}(\mathcal{A})^t\}^t$ by leverage downweighting of each of the components of (13). Let $\mathcal{B} = (\beta_0, \beta_{0*}, \beta_1^t, \beta_2^t)^t$. As in Carroll & Pederson (1993), the weights for the first component are $\phi_1(\mathcal{T}_{Vid}, \mathcal{B}, \hat{\xi}_V)$, while that for the third component is $\phi_3(\mathcal{T}_{Pid}, A_{Pid}, \mathcal{B}, \hat{\xi}_P)$. Following Simpson, et al. (1992), we will denote the linear regression weights for the second component by $\phi_2(S_{id}, \mathcal{A}, \hat{\xi}_W)$.

With robust linear regression estimates as defined by Simpson, et al., we downweight residuals using a fixed odd function $\psi(\cdot)$, which are scaled by an estimate of variability for each component of the residuals. Define $\hat{\Sigma} = (\hat{\sigma}_1, \dots, \hat{\sigma}_{k_1})$, where each component $\hat{\sigma}_j$ ($j = 1, \dots, k_1$) denotes the variability of the j^{th} component of the residuals, and we assume $\hat{\sigma}_i$ converges to σ_j . These estimates of scale should be fairly robust, e.g., the median absolute deviation of the residuals from their median.

The resulting estimating equation is

$$\begin{aligned} 0 = & \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_1(\mathcal{T}_{Vid}, \mathcal{B}, \hat{\xi}_V) \begin{pmatrix} \mathcal{T}_{Vid} \\ 0 \end{pmatrix} \{d - H(\beta_0 + \beta_1^t Z_{Vid} + \beta_2^t X_{Vid})\} \\ & + \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_2(S_{id}, \mathcal{A}, \hat{\xi}_W) \hat{\mathcal{F}}_{Vid} + \sum_{d=0}^1 \sum_{i=1}^{n_{Pd}} \left(\phi_3(\mathcal{T}_{Pid}, A_{Pid}, \mathcal{B}, \hat{\xi}_P) \begin{pmatrix} \mathcal{T}_{Pid} \\ A_{Pid} \end{pmatrix} \right. \\ & \left. \times \left[d - H \left\{ \beta_{0*} + \beta_1^t Z_{Pid} + \beta_2^t (\alpha_1 Z_{Pid} + \alpha_2 W_{Pid}) \right\} \right] \right), \end{aligned} \quad (14)$$

where

$$\hat{\mathcal{F}}_{Vid} = \begin{pmatrix} 0 \\ S_{id} \otimes \left\{ \hat{\sigma}_1 \psi\left(\frac{e_{Vid1}}{\hat{\sigma}_1}\right), \dots, \hat{\sigma}_{k_1} \psi\left(\frac{e_{Vidk_1}}{\hat{\sigma}_{k_1}}\right) \right\}^t \end{pmatrix}.$$

Starting values for estimating Θ can be obtained as follows. First, we estimate \mathcal{A} by applying to the validation data any of the robust linear regression techniques described by Simpson, et al. (1992). Then we estimate $(\beta_0, \beta_{0*}, \beta_1, \beta_2)$ by a dummy variable logistic regression using any of the robust logistic regression techniques described by Carroll & Pederson (1993). We use a scoring algorithm to update the initial estimates and hence solve (14). Let $\hat{\xi} = (\hat{\xi}_V, \hat{\xi}_W, \hat{\xi}_P)$ and let $h_n(\Theta, \hat{\xi}, \hat{\Sigma})$ be the right hand side of (14). Define $H^{(1)} = H(1 - H)$. Then a scoring-type algorithm updates the current $\hat{\Theta}$ via the formula:

$$\tilde{\Theta} = \hat{\Theta} + G_n^{-1}(\hat{\Theta}, \hat{\xi}, \hat{\Sigma}) h_n(\hat{\Theta}, \hat{\xi}, \hat{\Sigma}), \text{ where}$$

$$\begin{aligned}
G_n(\Theta, \xi, \Sigma) &= \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_1(\mathcal{T}_{Vid}, \mathcal{B}, \xi_V) \begin{pmatrix} \mathcal{T}_{Vid} \\ 0 \end{pmatrix} \begin{pmatrix} \mathcal{T}_{Vid} \\ 0 \end{pmatrix}^t H^{(1)}(\beta_0 + \beta_1^t Z_{Vid} + \beta_2^t X_{Vid}) \\
&+ \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_2(S_{id}, \mathcal{A}, \xi_W) \begin{pmatrix} 0 \\ S_{id} S_{id}^t \otimes \left\{ \psi^{(1)}\left(\frac{e_{Vid1}}{\sigma_1}\right), \dots, \psi^{(1)}\left(\frac{e_{Vidk_1}}{\sigma_{k_1}}\right) \right\} \\ 0 \end{pmatrix} \\
&+ \sum_{d=0}^1 \sum_{i=1}^{n_{Pd}} \phi_3(\mathcal{T}_{Pid}, A_{Pid}, \mathcal{B}, \xi_P) \begin{pmatrix} \mathcal{T}_{Pid} \\ A_{Pid} \end{pmatrix} \begin{pmatrix} \mathcal{T}_{Pid} \\ A_{Pid} \end{pmatrix}^t \\
&\quad \times H^{(1)} \left\{ \beta_{0*} + \beta_1^t Z_{Pid} + \beta_2^t (\alpha_1 Z_{Pid} + \alpha_2 W_{Pid}) \right\}.
\end{aligned}$$

Let the total sample size be $n = n_V + n_P$, and $(n_\lambda/n) \rightarrow \lambda, 0 < \lambda < 1$. Let G be the asymptotic limit of $G_n(\Theta, \xi)/n$ as $n \rightarrow \infty$, which is assumed necessarily to exist.

THEOREM 1: Under the conditions stated in the appendix,

$$n^{1/2}(\tilde{\Theta} - \Theta) \Rightarrow \text{Normal}(0, G^{-1}LG^{-t}),$$

where $L = L_1 + L_2 + L_3 + L_4 + L_4^t$, with the terms defined as follows:

$$L_1 = (n_V/n) \left\{ L_{12} - \sum_{d=0}^1 \frac{n_V}{n_{Vd}} L_{11} L_{11}^t \right\}, \quad (15)$$

$$L_{12} = \int \phi_1^2(\mathcal{T}_V, \mathcal{B}, \xi_V) \begin{pmatrix} \mathcal{T}_V \\ 0 \end{pmatrix} \begin{pmatrix} \mathcal{T}_V \\ 0 \end{pmatrix}^t H^{(1)}(\beta_0 + \beta_1^t z + \beta_2^t x) dQ_V(z, x),$$

$$L_{11} = \int \phi_1(\mathcal{T}_V, \mathcal{B}, \xi_V) \begin{pmatrix} \mathcal{T}_V \\ 0 \end{pmatrix} H^{(1)}(\beta_0 + \beta_1^t z + \beta_2^t x) dQ_V(z, x), \quad \mathcal{T}_V = (1, 0, z^t, x^t)^t,$$

$$L_2 = \sum_{d=0}^1 (n_{Vd}/n) E[\phi_2^2(S_{1d}, \mathcal{A}, \xi_W) \hat{\mathcal{F}}_{V1d} \hat{\mathcal{F}}_{V1d}^t | D = d],$$

$$L_3 = (n_P/n) \left\{ L_{32} - \sum_{d=0}^1 \frac{n_P}{n_{Pd}} L_{31} L_{31}^t \right\},$$

$$L_{32} = \int \phi_3^2(\mathcal{T}_P, A_P, \mathcal{B}, \xi_P) \begin{pmatrix} \mathcal{T}_P \\ A_P \end{pmatrix} \begin{pmatrix} \mathcal{T}_P \\ A_P \end{pmatrix}^t H^{(1)} \{ \beta_{0*} + (\beta_1^t + \beta_2^t \alpha_1) z + \beta_2^t \alpha_2 w \} dQ_P(z, w),$$

$$\mathcal{T}_P = \left\{ 0, 1, z^t, (\alpha_1 z + \alpha_2 w)^t \right\}^t, \quad A_P = \left[0, 0, \{ \text{vec}(z \beta_2^t) \}^t, \{ \text{vec}(w \beta_2^t) \}^t \right]^t,$$

$$L_{31} = \int \phi_3(\mathcal{T}_P, \mathcal{B}, \xi_P) \begin{pmatrix} \mathcal{T}_P \\ A_P \end{pmatrix} H^{(1)} \{ \beta_{0*} + (\beta_1^t + \beta_2^t \alpha_1) z + \beta_2^t \alpha_2 w \} dQ_P(z, w),$$

$$L_4 = \sum_{d=0}^1 (n_{Vd}/n) E \left[\phi_1(\mathcal{T}_{V1d}, \mathcal{B}, \xi_V) \begin{pmatrix} \mathcal{T}_{V1d} \\ 0 \end{pmatrix} \{ d - H(\cdot) \} \phi_2(S_{1d}, \mathcal{A}, \xi_W) \hat{\mathcal{F}}_{V1d}^t | D = d \right],$$

where in L_4 , $H(\cdot) = H(\beta_0 + \beta_1^t Z_{V1d} + \beta_2^t X_{V1d})$.

A consistent estimate of G is given by $\hat{G} = G_n(\hat{\Theta}, \hat{\xi}, \hat{\Sigma})/n$. An estimate of L is given in the appendix. This estimate has the properties that if there is no primary data set and hence the X 's are fully observable, then the parameter estimates are the same as those of Wang & Carroll

(1993) and the covariance estimate is the same as that obtained from a weighted logistic regression analysis. The same is true if the α 's are known.

5 NO AVAILABLE VALIDATION

In many instances, it is impossible to measure X on even a subset of the data. Instead, in a small subset of the data we observe an unbiased measure $T = X + U$, where $E(U|Z, W, D) = 0$ and hence $E(T|Z, W, D) = E(X|Z, W, D)$; an example in a prospective study is given by Rosner, et al. (1989). This suggests that one perform a linear regression of T on (Z, W) in the validation data, and then pool all the data and perform a logistic regression of Y on Z and $(\alpha_1 Z + \alpha_2 W)$. Necessarily, the weight function for this logistic regression has the same form in the validation and primary data sets, and we enforce this by using a common ξ in defining the leverage weights. If we define $\mathcal{T}_{Vid} = \{1, Z_{Vid}^t, (\alpha_1 Z_{Vid} + \alpha_2 W_{Vid})^t\}^t$ and $A_{Vid} = [0, 0, \{\text{vec}(Z_{Vid}\beta_2^t)\}^t, \{\text{vec}(W_{Vid}\beta_2^t)\}^t]^t$, this algorithm leads to the estimating equation

$$\begin{aligned} 0 = & \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \left(\phi_3(\mathcal{T}_{Vid}, A_{Vid}, \mathcal{B}, \hat{\xi}) \begin{pmatrix} \mathcal{T}_{Vid} \\ 0 \end{pmatrix} \right. \\ & \left. \times \left[d - H \left\{ \beta_{0*} + \beta_1^t Z_{Vid} + \beta_2^t (\alpha_1 Z_{Vid} + \alpha_2 W_{Vid}) \right\} \right] \right) \\ & + \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_2(S_{id}, \mathcal{A}, \hat{\xi}_W) \hat{\mathcal{F}}_{Vid} \\ & + \sum_{d=0}^1 \sum_{i=1}^{n_{Pd}} \left(\phi_3(\mathcal{T}_{Pid}, A_{Pid}, \mathcal{B}, \hat{\xi}) \begin{pmatrix} \mathcal{T}_{Pid} \\ 0 \end{pmatrix} \right. \\ & \left. \times \left[d - H \left\{ \beta_{0*} + \beta_1^t Z_{Pid} + \beta_2^t (\alpha_1 Z_{Pid} + \alpha_2 W_{Pid}) \right\} \right] \right), \end{aligned}$$

where the residuals are from the regression of T on (Z, W) .

The limiting distribution of the resulting estimates takes the same form as that of Theorem 1. Formulae for estimates of G and L are given in the appendix.

6 AN EXAMPLE

As an illustrative example, we use data described by Satten & Kupper (1993). These data were extracted from the Lipids Research Clinics (LRC) prospective study concerning blood cholesterol levels and risk of heart disease, and we used the nonsmokers. Let D be coronary heart disease

history, X the value of elevated low density lipoprotein cholesterol level (LDL), Z participant's age (range 60–70), and W total cholesterol (TC).

In this particular data set, X is available for all study participants, and the study is prospective, so our intent is to illustrate what might happen to a case–control analysis if there were outliers and leverage points. In Figure 1, we plot X against W for all study participants. Note that there are two unusual points, one a clear outlier to the main trend in the data, the other a large leverage point. We will call the former #1, and the latter #2.

To simulate a case–control study, we randomly divided the data into a primary data set of size $n_P = 232$ and a validation data set of size $n_V = 26$. The validation data consists of 13 cases and 13 controls. In this scheme, the outlying observation #1 was allocated to the validation data, while the leverage point #2 was allocated to the primary data.

Since X is continuous, in this example leverage weights were constructed basically along the lines of Simpson, et al. (1992), as follows. Let $X_{Vid*} = (Z_{Vid}^t, X_{Vid}^t)^t$, and let $\xi_V = (\mu_V, M_V)$ be the center and covariance matrix of X_{Vid*} . We define

$$\phi_1(\mathcal{T}_{Vid}, \mathcal{B}, \xi_V) = u_{1b} \left[\left\{ (X_{Vid*} - \mu_V)^t M_V^{-1} (X_{Vid*} - \mu_V) / 2 \right\}^{1/2} \right],$$

where u_{1b} is defined in §9.5 and $b = 8$. Let $W_{Vid*} = (Z_{Vid}^t, W_{Vid}^t)^t$, and let $\xi_W = (\mu_W, M_W)$ be the center and covariance matrix of W_{Vid*} . We define

$$\phi_2(S_{id}, \mathcal{A}, \xi_W) = \min \left[1, \left\{ c / \left\{ (W_{Vid*} - \mu_W)^t M_W^{-1} (W_{Vid*} - \mu_W) \right\} \right\} \right],$$

where $c = 2$. Let $W_{Pid*} = (Z_{Pid}^t, (\alpha_1 Z_{Pid} + \alpha_2 W_{Pid})^t)^t$, and let $\xi_P = (\mu_P, M_P)$ be the center and covariance matrix of W_{Pid*} , then we define

$$\phi_3(\mathcal{T}_{Pid}, A_{Pid}, \mathcal{B}, \xi_P) = u_{1b} \left[\left\{ (W_{Pid*} - \mu_P)^t M_P^{-1} (W_{Pid*} - \mu_P) / 2 \right\}^{1/2} \right].$$

We estimated σ iteratively by the median absolute residual times 1.48. For $\psi(\cdot)$, we used the Hampel function with cutoff points (1.5, 3.0, 8.0), see Carroll & Welsh (1989).

We applied the nonrobust technique with estimating equation (13), setting $A_{Pid} = 0$ for computational convenience, and compared it to the more robust technique with estimating equation (14). In Table 1 we list the effects of deleting the unusual points. Fairly clearly, the robust estimates are far less sensitive to case deletion, the reason being that the two unusual cases are given nearly no weight in the analysis.

7 DIFFERENT VALIDATION SAMPLING SCHEMES

In some circumstances, selection into the validation study may depend on (Z, W, D) and not just on D , see Zhao & Lipsitz (1992) for a review. Under such sampling schemes, for continuous exposures the estimation methods we have proposed will no longer be consistent.

Breslow & Cain (1988) and Zhao & Lipsitz (1992) describe nonrobust estimates using only validation data. These methods are weighted logistic regression, and as such are easily robustified.

With X continuous, solving (8) and (10) simultaneously will also result in a consistent estimate. Leverage downweighting modifications are clear. Asymptotic distribution theory can be obtained via the same techniques used in the appendix.

8 CONCLUSIONS

Our focus has been on showing that standard ideas from robust estimation lead to useful and computable methods in case-control studies with errors in predictors. The methods rely on defining approximately unbiased estimating equations for the parameters of interest. We extended the methods of Carroll, et al. (1993) to allow more effective use of validation data. In §5, we showed how to extend previous techniques to allow for the possibility that X cannot be measured, and that instead an unbiased version of it is available for part of the study. Such methods are particularly important in diet studies, where it is effectively impossible to record diet for a long period of time. Finally, for a continuous exposure, in §7 we briefly considered the possibility that selection into the validation study depends on more than just case-control status.

It is possible to improve efficiency somewhat in the continuous case by weighting the primary and validation data separately. For example, in (9) the validation data might be given a higher weight than the primary data, because the former produce direct information about the parameters of interest. Obtaining optimal weights is easy enough in principle, although doing so will result in some computational complications. We plan to explore this issue in a future paper.

ACKNOWLEDGEMENT

Our research was supported by a grant from the National Cancer Institute. We thank Ed Davis for providing the data used in the example, and Glen Satten and Larry Kupper for providing a preprint of their paper.

REFERENCES

- Breslow, N. E. & Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75, 11–20.
- Carroll, R.J., Gail, M.H. & Lubin, J.H. (1993). Case-control studies with errors in covariates. *Journal of the American Statistical Association*, 88, 177–191.
- Carroll, R.J. & Pederson, S. (1993). Further remarks on robustness in the logistic regression model. *Journal of the Royal Statistical Society, Series B*, to appear.
- Carroll, R. J. & Stefanski, L. A. (1990). Approximate quaslikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85, 652–663.
- Carroll, R. J. & Welsh, A. H. (1989). A note on asymmetry and robustness in linear regression. *The American Statistician*, 42, 285–287.
- 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, editors. American Mathematics Society, Providence.
- Prentice, R. L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403–411.
- Pierce, D. A., Stram, D. O., Vaeth, M., Schafer, D. (1992). The errors in variables problem: considerations provided by radiation dose-response analyses of the A-bomb survivor data. *Journal of the American Statistical Association*, 87, 351–359.
- Rosner, B., Willett, W. C. & 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.
- Satten, G. A. & Kupper, L. L. (1993). Inferences about exposure-disease association using probability of exposure information. *Journal of the American Statistical Association*, to appear.
- Simpson, D.G., Ruppert, D. & Carroll, R.J. (1992). On one-step GM-estimates and stability of inferences in linear regression. *Journal of the American Statistical Association*, 87, 439–450.
- Wang, C.Y. & Carroll, R.J. (1993). Robust estimation in case-control studies. *Biometrika*, to appear.
- Zhao, L. P. & Lipsitz, S. (1992). Designs and analysis of two-stage studies. *Statistics in Medicine*, 11, 769–782.

9 APPENDIX

9.1 Technical Conditions for Theorem 1

- (A1) The initial estimates $\hat{\Theta}$ and $\hat{\xi}$ converge to their limits at the rate $o_p(n^{-1/4})$.
- (A2) The parameter space of (Θ, ξ) is open and convex.
- (A3) The weight functions ϕ_1, ϕ_2 and ϕ_3 are bounded and continuous.

(A4) ϕ_1, ϕ_2 and ϕ_3 are smooth enough such that the second derivatives with respect to parameters are bounded in absolute value by a statistic which has finite expectation at the true parameters.

(A5) Z and W have bounded third moments.

(B1) The score function ψ is bounded and continuous.

(B2) ψ has derivative $\psi^{(1)}$ such that (i) $\|\psi^{(1)}\|_{sup} < \infty$ and (ii) $\|(\cdot)\psi^{(1)}(\cdot)\|_{sup} < \infty$, where $\|\cdot\|_{sup}$ is the supremum norm.

(B3) $\psi^{(1)}$ has derivative $\psi^{(2)}$ such that $\|\psi^{(2)}\|_{sup} + \|(\cdot)\psi^{(2)}(\cdot)\|_{sup} + \|(\cdot)^2\psi^{(2)}(\cdot)\|_{sup} < \infty$.

(B4) $E[\psi(\epsilon v)|D] = 0$ and $E[\epsilon v\psi^{(1)}(\epsilon v)|D] = 0$ for any $v > 0$.

(B5) $E\|(Z, W)\|^4\phi_2^2 = O(1)$ and $E\|(Z, W)\|^3\phi_2 = O(1)$.

The condition (B4) is essentially equivalent to symmetry of the errors ϵ . Under this condition, the asymptotic distribution theory and covariance estimates which we obtain are appropriate even when the ϵ 's are heteroscedastic.

If the ϵ 's are homoscedastic and independent of (D, Z, W) , it can be shown that symmetry is not required. Following Carroll & Welsh (1989), it can be shown that the distribution theory of the intercept terms is affected by asymmetry because the method of estimating $(\sigma_1, \dots, \sigma_{k_1})$ comes into play. However, asymmetry does not change the distribution theory of the other parameters.

9.2 Sketch Proof of Theorem 1

Under the conditions (A1) – (B4), it is easy to prove that

$$n^{1/2}(\tilde{\Theta} - \Theta) = G^{-1}\{n^{-1/2}h_n(\Theta, \xi, \Sigma)\} + o_p(1).$$

Write $h_n(\Theta, \xi, \Sigma)$ as $h_{n1}(\theta, \xi) + h_{n2}(\Theta, \xi, \Sigma) + h_{n3}(\Theta, \xi)$. Then

$$\begin{aligned} \text{cov}\{h_n(\Theta, \xi, \Sigma)\} &= \text{cov}\{h_{n1}(\Theta, \xi)\} + \text{cov}\{h_{n2}(\Theta, \xi, \Sigma)\} + \text{cov}\{h_{n3}(\Theta, \xi)\} \\ &\quad + \text{cov}\{h_{n1}(\Theta, \xi), h_{n2}(\Theta, \xi, \Sigma)\} + \text{cov}\{h_{n1}(\Theta, \xi), h_{n3}(\Theta, \xi)\}^t. \end{aligned}$$

By easy calculations, we have

$$\begin{aligned} n^{-1}\text{cov}\{h_{n1}(\Theta, \xi)\} &= (n_V/n)\text{cov}\{n_V^{-1/2}h_{n1}(\Theta, \xi)\} = L_1, \\ n^{-1}\text{cov}\{h_{n2}(\Theta, \xi, \Sigma)\} &= L_2, \\ n^{-1}\text{cov}\{h_{n3}(\Theta, \xi)\} &= (n_P/n)\text{cov}\{n_P^{-1/2}h_{n3}(\Theta, \xi)\} = L_3, \\ n^{-1}\text{cov}\{h_{n1}(\Theta, \xi), h_{n2}(\Theta, \xi, \Sigma)\} &= L_4. \end{aligned}$$

9.3 Covariance Estimates with Validation

Define $H_{P_i}^{(1)}(\cdot) = H^{(1)}\{\hat{\beta}_{0*} + \hat{\beta}_1^t Z_{Pid} + \hat{\beta}_2^t(\hat{\alpha}_1 Z_{Pid} + \hat{\alpha}_2 W_{Pid})\}$. For estimating L , we use $\hat{L} = \hat{L}_1 + \hat{L}_2 + \hat{L}_3 + \hat{L}_4 + \hat{L}_4^t$, where

$$\hat{L}_1 = (n_V/n)\{\hat{L}_{12} - \sum_{d=0}^1 \frac{n_V}{n_{Vd}} \hat{L}_{11} \hat{L}_{11}^t\}, \quad (16)$$

$$\hat{L}_{12} = n_V^{-1} \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_1^2(\mathcal{T}_{Vid}, \hat{\mathcal{B}}, \hat{\xi}_V) \begin{pmatrix} \mathcal{T}_{Vid} \\ 0 \end{pmatrix} \begin{pmatrix} \mathcal{T}_{Vid} \\ 0 \end{pmatrix}^t H^{(1)}(\hat{\beta}_0 + \hat{\beta}_1^t Z_{Vid} + \hat{\beta}_2^t X_{Vid}),$$

$$\hat{L}_{11} = n_V^{-1} \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_1(\mathcal{T}_{Vid}, \hat{\mathcal{B}}, \hat{\xi}_V) \begin{pmatrix} \mathcal{T}_{Vid} \\ 0 \end{pmatrix} H^{(1)}(\hat{\beta}_0 + \hat{\beta}_1^t Z_{Vid} + \hat{\beta}_2^t X_{Vid}),$$

$$\hat{L}_2 = n^{-1} \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_2^2(S_{id}, \hat{\mathcal{A}}, \hat{\xi}_W) \hat{\mathcal{F}}_{Vid} \hat{\mathcal{F}}_{Vid}^t, \quad \hat{L}_3 = (n_P/n)\{\hat{L}_{32} - \sum_{d=0}^1 \frac{n_P}{n_{Pd}} \hat{L}_{31} \hat{L}_{31}^t\},$$

$$\hat{L}_{32} = n_P^{-1} \sum_{d=0}^1 \sum_{i=1}^{n_{Pd}} \phi_3^2(\mathcal{T}_{Pid}, \hat{\mathcal{A}}_{Pid}, \hat{\mathcal{B}}, \hat{\xi}_P) \begin{pmatrix} \mathcal{T}_{Pid} \\ \hat{\mathcal{A}}_{Pid} \end{pmatrix} \begin{pmatrix} \mathcal{T}_{Pid} \\ \hat{\mathcal{A}}_{Pid} \end{pmatrix}^t H_{P_i}^{(1)}(\cdot);$$

$$\hat{L}_{31} = n_P^{-1} \sum_{d=0}^1 \sum_{i=1}^{n_{Pd}} \phi_3(\mathcal{T}_{Pid}, \hat{\mathcal{A}}_{Pid}, \hat{\mathcal{B}}, \hat{\xi}_P) \begin{pmatrix} \mathcal{T}_{Pid} \\ \hat{\mathcal{A}}_{Pid} \end{pmatrix} H_{P_i}^{(1)}(\cdot);$$

$$\begin{aligned} \hat{L}_4 &= n^{-1} \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_1(\mathcal{T}_{Vid}, \hat{\mathcal{B}}, \hat{\xi}_V) \begin{pmatrix} \mathcal{T}_{Vid} \\ 0 \end{pmatrix} \{d - H(\hat{\beta}_0 + \hat{\beta}_1^t Z_{Vid} + \hat{\beta}_2^t X_{Vid})\} \\ &\quad \times \phi_2(S_{id}, \hat{\mathcal{A}}, \hat{\xi}_W) \hat{\mathcal{F}}_{Vid}^t. \end{aligned}$$

9.4 Covariance Estimates with No Validation

Define $H_{V_i}^{(1)}(\cdot) = H^{(1)}\{\hat{\beta}_{0*} + \hat{\beta}_1^t Z_{Vid} + \hat{\beta}_2^t(\hat{\alpha}_1 Z_{Vid} + \hat{\alpha}_2 W_{Vid})\}$ and let \hat{e}_{Vidj} be the j th component of the estimated residuals. Then the quantities necessary for estimation of the covariance matrix are

$$\begin{aligned} \hat{G} &= n^{-1} \left[\sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_3(\mathcal{T}_{Vid}, \hat{\mathcal{A}}_{Vid}, \hat{\mathcal{B}}, \hat{\xi}) \begin{pmatrix} \mathcal{T}_{Vid} \\ 0 \end{pmatrix} \begin{pmatrix} \mathcal{T}_{Vid} \\ 0 \end{pmatrix}^t H_{V_i}^{(1)}(\cdot) \right. \\ &\quad \left. + \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_2(S_{id}, \hat{\mathcal{A}}, \hat{\xi}_W) \begin{pmatrix} 0 \\ S_{id} S_{id}^t \otimes \left\{ \psi^{(1)}\left(\frac{\hat{e}_{Vid1}}{\hat{\sigma}_1}\right), \dots, \psi^{(1)}\left(\frac{\hat{e}_{Vidk_1}}{\hat{\sigma}_{k_1}}\right) \right\} \right] \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ &\quad \left. + \sum_{d=0}^1 \sum_{i=1}^{n_{Pd}} \phi_3(\mathcal{T}_{Pid}, \hat{\mathcal{A}}_{Pid}, \hat{\mathcal{B}}, \hat{\xi}) \begin{pmatrix} \mathcal{T}_{Pid} \\ \hat{\mathcal{A}}_{Pid} \end{pmatrix} \begin{pmatrix} \mathcal{T}_{Pid} \\ \hat{\mathcal{A}}_{Pid} \end{pmatrix}^t H_{V_i}^{(1)}(\cdot) \right] \end{aligned}$$

$$\hat{L} = \hat{L}_1 + \hat{L}_2 + \hat{L}_3 + \hat{L}_4 + \hat{L}_4^t, \quad \text{where } \hat{L}_1 = \lambda_V \{\hat{L}_{12} - \sum_{d=0}^1 \frac{n_V}{n_{Vd}} \hat{L}_{11} \hat{L}_{11}^t\},$$

$$\hat{L}_{12} = n_V^{-1} \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_3^2(\mathcal{T}_{Vid}, \hat{\mathcal{A}}_{Vid}, \hat{\mathcal{B}}, \hat{\xi}) \begin{pmatrix} \mathcal{T}_{Vid} \\ \hat{\mathcal{A}}_{Vid} \end{pmatrix} \begin{pmatrix} \mathcal{T}_{Vid} \\ \hat{\mathcal{A}}_{Vid} \end{pmatrix}^t H_{V_i}^{(1)}(\cdot);$$

$$\begin{aligned}
\hat{L}_{11} &= n_V^{-1} \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_3(\mathcal{T}_{Vid}, \hat{A}_{Vid}, \hat{\mathcal{B}}, \hat{\xi}) \left(\begin{matrix} \mathcal{T}_{Vid} \\ \hat{A}_{Vid} \end{matrix} \right) H_{Vi}^{(1)}(\cdot); \\
\hat{L}_2 &= n^{-1} \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_2^2(S_{id}, \hat{\mathcal{A}}, \hat{\xi}_W) \left(S_{id} \otimes \left\{ \hat{\sigma}_1 \psi\left(\frac{\hat{e}_{Vid1}}{\hat{\sigma}_1}\right), \dots, \hat{\sigma}_{k_1} \psi\left(\frac{\hat{e}_{Vidk_1}}{\hat{\sigma}_{k_1}}\right) \right\}^0 \right) \\
&\quad \times \left(S_{id} \otimes \left\{ \hat{\sigma}_1 \psi\left(\frac{\hat{e}_{Vid1}}{\hat{\sigma}_1}\right), \dots, \hat{\sigma}_{k_1} \psi\left(\frac{\hat{e}_{Vidk_1}}{\hat{\sigma}_{k_1}}\right) \right\}^t \right), \\
\hat{L}_3 &= \lambda_P \left\{ \hat{L}_{32} - \sum_{d=0}^1 \frac{n_P}{n_{Pd}} \hat{L}_{31} \hat{L}_{31}^t \right\}, \\
\hat{L}_{32} &= n_P^{-1} \sum_{d=0}^1 \sum_{i=1}^{n_{Pd}} \phi_3^2(\mathcal{T}_{Pid}, \hat{A}_{Pid}, \hat{\mathcal{B}}, \hat{\xi}) \left(\begin{matrix} \mathcal{T}_{Pid} \\ \hat{A}_{Pid} \end{matrix} \right) \left(\begin{matrix} \mathcal{T}_{Pid} \\ \hat{A}_{Pid} \end{matrix} \right)^t H_{Pi}^{(1)}(\cdot); \\
\hat{L}_{31} &= n_P^{-1} \sum_{d=0}^1 \sum_{i=1}^{n_{Pd}} \phi_3(\mathcal{T}_{Pid}, \hat{A}_{Pid}, \hat{\mathcal{B}}, \hat{\xi}) \left(\begin{matrix} \mathcal{T}_{Pid} \\ \hat{A}_{Pid} \end{matrix} \right) H_{Pi}^{(1)}(\cdot); \\
\hat{L}_4 &= n^{-1} \sum_{d=0}^1 \sum_{i=1}^{n_{Vd}} \phi_3(\mathcal{T}_{Vid}, \hat{A}_{Vid}, \hat{\mathcal{B}}, \hat{\xi}) \left(\begin{matrix} \mathcal{T}_{Vid} \\ \hat{A}_{Vid} \end{matrix} \right) \{d - H(\hat{\beta}_0 + \hat{\beta}_1^t Z_{Vid} + \hat{\beta}_2^t X_{Vid})\} \\
&\quad \times \phi_2(S_{id}, \hat{\mathcal{A}}, \hat{\xi}_W) \hat{\mathcal{F}}_{Vid}^t.
\end{aligned}$$

9.5 Weights for Logistic Regression

Carroll & Pederson (1993) consider the following method. Let $p = \dim(Z, X)$. Let $(\hat{\mu}, \hat{M})$ be a “robust” estimate of the center and covariance matrix of the (Z, X) ’s. Let ψ_{1b} be any odd function, and define $\psi_{2b}(v) = \psi_{1b}^2(v)/\xi$, where $\xi = E\psi_{1b}^2(\|Z_{p-1}\|)$ and Z_{p-1} is a $(p-1)$ -dimensional normal random variable with zero mean and identity covariance. Define $u_{ib}(v) = \psi_{ib}(v)/v$. Defining $X_{i*} = (Z_i^t, X_i^t)^t$, the estimates $\hat{\xi} = (\hat{\mu}, \hat{M})$ which we use are the solutions to:

$$\begin{aligned}
n^{-1} \sum_{i=1}^n u_{1b} \left[\left\{ (X_{i*} - \mu)^t M^{-1} (X_{i*} - \mu) \right\}^{1/2} \right] (X_{i*} - \mu) &= 0; \\
n^{-1} \sum_{i=1}^n u_{2b} \left\{ (X_{i*} - \mu)^t M^{-1} (X_{i*} - \mu) \right\} (X_{i*} - \mu) (X_{i*} - \mu)^t &= M.
\end{aligned}$$

One possible choice is the trisquared redescending function $\psi_{1b}(v) = v \{1 - (v/b)^2\}^3 I(|v| \leq b)$. The leverage-based weights are $\phi(Z, X, \xi) = u_{1b} \left[\left\{ (X_{i*} - \mu)^t M^{-1} (X_{i*} - \mu) / (p-1) \right\}^{1/2} \right]$. Note that these weights can redescend to zero, so that points which are extremely outlying in the design space receive zero weight. Different estimates of center and location can be used, for example those with higher breakdown as discussed in Simpson, et al. (1992).

Points Deleted	Nonrobust Estimate	Nonrobust Std. err.	Robust Estimate	Robust Std. Err.
None	1.6965	.8912	.3785	.2227
#1	.4383	.2403	.3481	.2283
#2	1.5628	.8185	.3615	.2315
#1, #2	.3738	.2573	.3283	.2373

Table 1: *The effects of deleting observations on estimates of the effect of LDL, when treated as a continuous variable.*

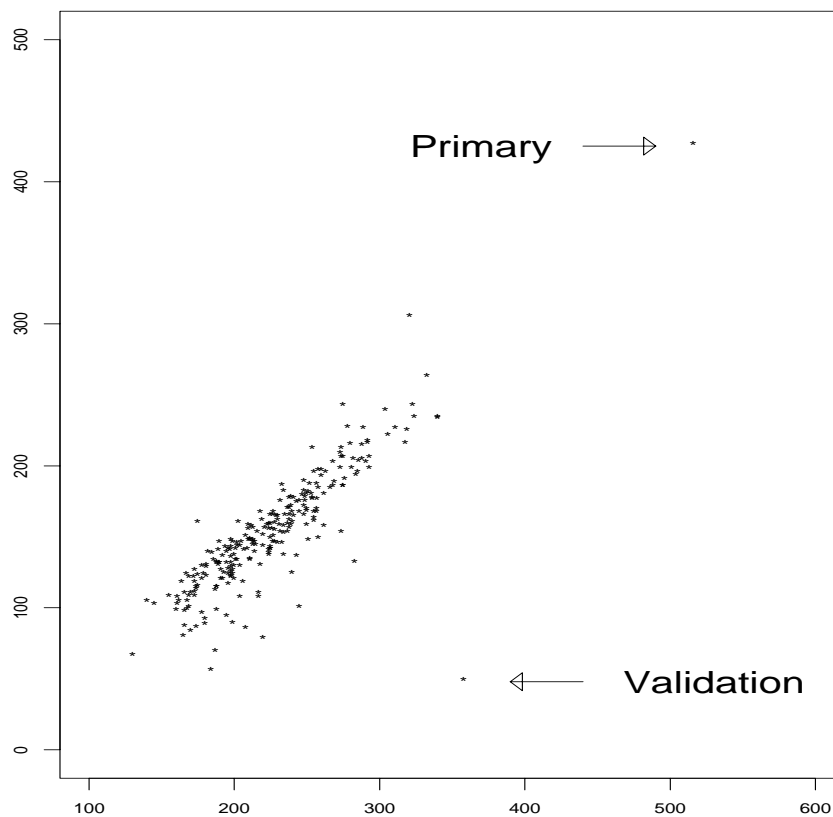


Figure 1: *Plot of $X = LDL$ against $W = total\ cholesterol$ in the example. In the examples, the leverage point was assigned to the primary data, while the outlier was assigned to the validation data.*