# APPROXIMATE QUASILIKELIHOOD ESTIMATION IN MEASUREMENT ERROR MODELS

Raymond J. Carroll
Department of Statistics
Texas A&M University
College Station, TX 77843

Leonard A. Stefanski
Department of Statistics
North Carolina State University
Raleigh, NC 27695

## ABSTRACT

We consider quasilikelihood estimation with estimated parameters in the variance function when some of the predictors are measured with error. We review and extend four approaches to estimation in this problem, all of them based on small measurement error approximations. A taxonomy of the data sets likely to be available in measurement error studies is developed. An asymptotic theory based on this taxonomy is obtained and includes measurement error and Berkson error models as special cases.

# 1. INTRODUCTION AND PRELIMINARIES

## 1.1 QUASILIKELIHOOD MEASUREMENT ERROR MODELS

The general quasilikelihood/variance function model for a scalar response $y$ given a $p$-variate predictor $x$ is

$$E(y \mid x) = f_m(x, \beta); \quad Var(y \mid x) = \sigma^2 f_v(x, \beta, \theta) \tag{1.1}$$

where $\sigma^2$ is a scalar parameter and $\beta$ and $\theta$ are column-vector parameters designated collectively as $\Theta = (\beta^t, \theta^t, \sigma^2)^t$. Model (1.1) includes the usual quasilikelihood models and generalized linear models as well as models in which the variance is an unknown function of the mean or predictors. For additional motivation and background on these models see Carroll and Ruppert (1988) and McCullagh and Nelder (1989).

We consider fitting model (1.1) when a $q$-variate proxy $w$ ($q \geq p$) is observed in place of the random predictor $x$ in some subset of the available data. We assume that $w$ is a *surrogate* for $x$, i.e., the conditional distribution of $y$ given $(x, w)$ equals the conditional distribution of $y$ given $x$.

When the conditional distribution of $x$ given $w$ is specified parametrically, it is possible in principle to calculate the conditional mean and variance of $y$ given $w$ and to estimate unknown parameters using standard quasilikelihood/variance function techniques. This approach frequently requires numerical multiple integration and is computationally unattractive. In addition, experience indicates that this approach can be sensitive to specification of the conditional distribution of $x$ given $w$ (Schafer, 1987). The methods developed in this paper depend only on the first two moments of the error given $w$, which is more in the spirit of quasilikelihood/variance function techniques (Carroll and Ruppert, 1988, Section 2.5). In Section 3 we describe a very general approximate model for the first two conditional moments of $y$ given $w$ when the relationship between $x$ and $w$ is specified either conditionally on $x$ or conditionally on $w$. Thus we incorporate both measurement-error and Berkson-error models.

Measurement error models are subsumed under the general model

$$w = c(x, \eta) + \delta u; \quad E(u \mid x) = 0; \quad Cov(u \mid x) = M(x, \eta, \gamma), \tag{1.2}$$

where $c(\cdot, \cdot)$ and $M(\cdot, \cdot, \cdot)$ are known and $\Lambda = (\eta^t, \gamma^t, \delta^2)^t$ is a column vector of parameters. In certain models some components of $\Lambda$ may be known. This includes the classical measurement model (Fuller, 1987), $c(x, \eta) = x$, $M(x, \eta, \gamma) = \Omega$; as well as linear models, $c(x, \eta) = \eta_0 + \eta_1 x$, $M(x, \eta, \gamma) = \Omega$ where $\eta$ contains all the unique elements of $\eta_0$ and $\eta_1$. Models in which some predictors are measured without error are handled by allowing $M$ to have less than full rank. Of course we can

benefit from the full generality of (1.2) only when sufficient validation/reliability data are available for estimating the unknown components of $\Lambda$, see Section 2.

In Section 3 we impose one significant restriction on (1.2). We require that the equation $t = c(s, \eta)$ can be solved for $s$ as a smooth function of $(t, \eta)$ in a neighborhood of the true parameter $\eta$, that is $s = c^-(t, \eta)$ where $c^-$ deonotes an inverse to $c$. This is always possible when $w$ and $x$ are scalars and $c$ is smooth and strictly monotone in its first argument. The requirement is more stringent when the dimension of $x$ exceeds 1. For example, in the multivariate regression version of (1.2), $w = \eta_0 + \eta_1 x + \delta u$ where $\eta_1$ is a $q \times p$ matrix, it requires that $\text{rank}(\eta_1) = p$ and in this case $x = \eta_1^-(w - \eta_0 - \delta u)$ where $\eta_1^-$ is a generalized inverse of $\eta_1$ satisfying $\eta_1^-\eta_1 = I_{p\times p}$.

Berkson-error models are subsumed under the general model

$$x = c^*(w, \eta) + \delta u; \quad E(u \mid w) = 0; \quad Cov(u \mid w) = M^*(w, \eta, \gamma), \tag{1.3}$$

where $c^*(\cdot, \cdot)$ and $M^*(\cdot, \cdot, \cdot)$ are known and $\Lambda = (\eta^t, \gamma^t, \delta^2)^t$ is a column vector of parameters. In certain models some components of $\Lambda$ may be known. The classical Berkson-error model has $c^*(w, \eta) = w$ and $M^*(w, \eta, \gamma) = \Omega$. When some components of $u$ are equal to 0, $M^*$ has less than full rank. As in (1.2), (1.3) includes most regression models used for data analysis.

New classes of estimates have been developed recently based on small measurement error approximations. Specifically it is assumed that $\delta$ in (1.2)-(1.3) is small. Four such approaches are: (i) correct the naive estimators obtained by fitting (1.1) with $x$ replaced by $g(w, \eta, 0)$ where $g$ is given in (3.1), (Stefanski and Carroll, 1985; Stefanski, 1985; Amemiya and Fuller, 1988); (ii) approximate the quasilikelihood/variance function estimates (Whittemore and Keller, 1988); (iii) approximate the quasilikelihood/variance function model and then estimate the parameters in this model (Rudemo *et al.*, 1989; Carroll, 1989; Fuller, 1987, Section 3.3; Armstrong, 1985); and (iv) replace $x$ by an estimate of $E(x \mid w)$ and perform a standard analysis (Rosner et al., 1989). These estimates are computable, widely applicable and, in our experience, work well in applications.

In this paper we develop a unified asymptotic theory for small measurement error estimates in models (1.1)-(1.3) which encompasses and generalizes the previously cited literature. We extend previous work on this problem both with respect to the classes of the models relating $y$ to $x$ and linking $x$ and $w$ and with respect assumptions on $\Lambda$. Our theory includes the construction of estimates and valid standard errors when the nuisance parameters $\Lambda$ are known as well as when some or all of the components of $\Lambda$ are unknown, provided sufficient data are available for estimating these unknowns.

Section 1.2 explains some notations and definitions used in the paper. In Section 2, we identify

the main types of data that arise in the analysis of measurement error models. In Section 3, we define the general model and discuss several examples. Section 4 studies estimation based on approximating the quasilikelihood/variance function model and Section 5 discusses approximating the estimators of this model. Two examples are discussed in Section 6.

### 1.2 NOTATIONS AND DEFINITIONS

In this paper we will use the notation for derivatives of vector- and matrix-valued functions as described in Fuller (1987, Appendix A.4). To these results we add notations for $\partial A/\partial\theta$ and $\partial A/\partial\theta^t$ where $A$ is a $p \times q$ matrix and $\theta$ is an $r \times 1$ vector. The former is an $rp \times q$ matrix containing the $r \times 1$ vectors $\partial a_{ij}(\theta)/\partial\theta$ in a $p \times q$ block structure. The latter is a $p \times qr$ matrix containing the $1 \times r$ vectors $\partial a_{ij}(\theta)/\partial\theta^t$ in a $p \times q$ block structure.

We extend the definition of the trace function to $ds \times s$, $(d, s = 1, 2, \ldots)$ matricies as follows: define $\text{tr}(A_{ds\times s}) = \{\text{trace}(A_1), \ldots, \text{trace}(A_d)\}^t$, where $A_j$ is the square matrix containing rows $(j-1)s+1$ through $js$ of $A$. Although we use tr more extensively, it is particularly convenient in the handling of quadratic approximations to vector-valued functions of random vectors. For example, if $f$ is $p \times 1$, $z$ is $q \times 1$, $f_z(z) = (\partial/\partial z^t)f(z)$ and $f_{zz}(z) = (\partial/\partial z)f_z(z)$, then the quadratic Maclaurin series approximation to $f(z)$, $f_Q(z)$, can be represented as $f_Q(z) = f(0) + f_z(0)z + (1/2)\text{tr}\{f_{zz}(0)zz^t\}$ and its expectation as $E\{f_Q(z)\} = f(0) + f_z(0)E(z) + (1/2)\text{tr}\{f_{zz}(0)E(zz^t)\}$. The symbolic simplicity of the latter two expressions is due to the fact that tr is a linear operator.

Finally, we use $\dim(v)$ to denote the dimension of a row- or column-vector $v$.

## 2. TYPES OF DATA

In this paper we assume that the data available for estimating the parameters of (1.1) and (1.2) or (1.3) are composed of one or more of five types: *primary* data containing $n_1$ observations $(y_i, w_i)$; *internal validation* data containing $n_2$ observations $(y_i, x_i, w_i)$; *internal reliability* data containing $n_3$ observations $(y_i, w_{i1}, \ldots, w_{ik_i})$ where for fixed $i$, $w_{ij}$, $j = 1, \ldots, k_i$ are independent and identically distributed; *external validation* data containing $n_4$ observations $(x_i, w_i)$; and *external reliability* data containing $n_5$ observations $(w_{i1}, \ldots, w_{ik_i})$ where for fixed $i$, $w_{ij}$, $j = 1, \ldots, k_i$ are independent and identically distributed. Observations within and between data types are stochastically independent.

When necessary for clarity we attach an additional preceeding subscript to an observation to indicate its type. For example, the pairs $(x_{2i}, w_{2i})$ and $(x_{4i}, w_{4i})$ are from internal and external validation data respectively. This notation permits identification of unobservables without ambiguity; for example, $x_{1i}$ and $u_{2i}$.

All five types of data are relevant to model (1.2); however, when model (1.3) is assumed, only

primary and validation data are relevant.

The primary data generally do not identify all of the components of $\Theta$ or the unknown components of $\Lambda$. External validation/reliability data generally identify the unknown components of $\Lambda$ and allow for identification of $\Theta$ when combined with primary data. Internal validation/reliability data serve the same purpose as external data with regard to identification but are preferred over external data for obvious reasons. Note that internal validation data can be used to test the conditional independence assumption of Section 1.

Typically the data required for a measurment error analysis has $n_1 > 0$ and one of $n_2 \ldots n_5 > 0$, depending on model (1.2) or (1.3). For example, Tosteson *et al.* (1989) have $n_2 = n_3 = n_5 = 0$, but $n_4 > 0$, see Section 6. However, Rudemo *et al.* (1989) have $n_2 = n_3 = n_4 = n_5 = 0$ but they assume that $x = w + \delta u$ and they exploit the nonlinearity in their model to identify $\delta^2$.

As noted previously, we assume that reliability data are collected only when the measurement error model (1.2) is assumed. In this case if $w_{ir} = c(x_i, \eta) + \delta u_{ir}$, $E(u_{ir} \mid x_i) = 0$ and $Cov(u_{ir} \mid x_i) = M(x_i, \eta, \gamma)$ for $r = 1, \ldots, k_i$, then $\bar{w}_{i\cdot} = c(x_i, \eta) + k_i^{-1/2}\delta u_{i*}$, $E(u_{i*} \mid x_i) = 0$ and $Cov(u_{i*} \mid x_i) = M(x_i, \eta, \gamma)$. Thus $(\bar{w}_{i\cdot}, x_i)$ satisfy (1.2) upon replacing $\delta$ with $k_i^{-1/2}\delta$. This fact is exploited later in the paper.

When $n_1$ and $n_2 > 0$ all parameters are identifiable even when $n_3 = n_4 = n_5 = 0$. Typically $n_1 \gg n_2$ because of additional costs associated with obtaining $x$. Thus, we have a small data set from which consistent estimates of $\Theta$ can be found, and a larger data set from which approximately consistent estimates are available using the methods of Sections 4 and 5. In general, the effect of pooling the two types of data is to obtain a less variable and less biased estimate than would be obtained if only the primary data were available. The estimates are also typically much less variable than if the primary data were ignored, but the bias incurred means that a variance-bias tradeoff determines whether the primary data should be used. Of course, the primary data are useful in testing for no association; see Tosteson and Tsiatis (1988) and Stefanski and Carroll (1989).

## 3. BASIC MODEL AND EXAMPLES

In this section, we determine three general approximate models for $y$ given $w$ when the error in predicting $x$ from $w$ is small. Several specific cases of the model are discussed for illustration and clarification of the general results.

Assume in addition to (1.1) that as $\delta \to 0$, there are known functions $(g, h_2, h_3)$, parameters $\Lambda = (\eta, \gamma, \delta)$ and a mean zero random variable $u$, such that

$$x = g(w, \eta, \delta u); \quad E(\delta u \mid w) = \delta^2 h_2(w, \eta, \gamma) + \mathcal{O}_P(\delta^3);$$

$$Cov(\delta u \mid w) = \delta^2 h_3(w,\eta,\gamma) + \mathcal{O}_P(\delta^3). \tag{3.1}$$

Define

$$f_{mx}(x,\beta) = (\partial/\partial x^t) f_m(x,\beta); \quad f_{mxx}(x,\beta) = (\partial/\partial x) f_{mx}(x,\beta),$$

letting subscripts denote derivatives. Define $f_{vx}$ and $f_{vxx}$ similarly. Also define $g_s(w,\eta,s) = (\partial/\partial s^t)g(w,\eta,s)$ and $g_{ss}(w,\eta,s) = (\partial/\partial s)g_s(w,\eta,s)$ and

$$\begin{aligned} H_m(w,\eta,\gamma,\beta) = {} & f_{mx}\{g(w,\eta,0),\beta\}g_s(w,\eta,0)h_2(w,\eta,\gamma) \\ & + (1/2)\mathrm{tr}\Big[f_{mxx}\{g(w,\eta,0),\beta\}g_s(w,\eta,0)h_3(w,\eta,\gamma)g_s(w,\eta,0)^t\Big] \\ & + (1/2)f_{mx}\{g(w,\eta,0),\beta\}\mathrm{tr}\Big\{g_{ss}(w,\eta,0)h_3(w,\eta,\gamma)\Big\}. \end{aligned}$$

Define $H_v(w,\eta,\gamma,\beta,\theta,\sigma^2)$ similarly except that $f_v$ replaces $f_m$. In addition, define

$$S(w,\eta,\gamma,\beta) = f_{mx}\{g(w,\eta,0),\beta\}g_s(w,\eta,0)h_3(w,\eta,\gamma)g_s(w,\eta,0)^t f_{mx}\{g(w,\eta,0),\beta\}^t.$$

Assumption (3.1), the relationships $E(y^p \mid w) = E\{E(y^p \mid x) \mid w\}$, $p = 0, 1$ and Taylor series expansions of $f_m\{g(w,\eta,\delta u),\beta\}$ and $f_v\{g(w,\eta,\delta u),\beta,\theta\}$ around $\delta = 0$ are used to show that

$$E(y \mid w) \approx U_{mA,1}(w,\beta,\eta,\gamma,\delta^2), \tag{3.2a}$$

$$Var(y \mid w) \approx U_{vA,1}(w,\beta,\theta,\sigma^2,\eta,\gamma,\delta^2), \tag{3.2b}$$

where

$$U_{mA,1}(w,\beta,\eta,\gamma,\delta^2) = f_m\{g(w,\eta,0),\beta\} + \delta^2 H_m(w,\eta,\gamma,\beta);$$

$$U_{vA,1}(w,\beta,\theta,\sigma^2,\eta,\gamma,\delta^2) = \sigma^2\Big[f_v\{g(w,\eta,0),\beta,\theta\} + \delta^2 H_v(w,\eta,\gamma,\beta,\theta,\sigma^2)\Big] + \delta^2 S(w,\eta,\gamma,\beta).$$

The error in the approximation in (3.2) is of order $\mathcal{O}_P(\delta^3)$. The subscript $A$ in (3.2a) and (3.2b) indicates the approximate nature of model. The model specified in (3.2) is an extension of some approximate models studied by Armstrong (1985) and Fuller (1987, Section 3.3) to allow for more general error structures in both the models relating $y$ to $x$ and $x$ with $w$.

A simpler and sometimes appropriate approximate model can be obtained by first modelling $x$ as a funtion of $w$ and then substituting the conditional expectation of $x$ given $w$ for $x$ in a standard analysis of (1.1). This approach is most natural in the generalized Berkson-error model in which case $E(x \mid w)$ is, in our notation, $g(w,\eta,0)$, but it can also be employed when (1.2) is known to hold. In the latter case an approximation to $E(x \mid w)$ would be employed,

$$\tilde{E}(x \mid w) = g(w,\eta,0) + \delta^2 g_s(w,\eta,0)h_2(w,\eta,\gamma) + (\delta^2/2)\mathrm{tr}\{g_{ss}(w,\eta,0)h_3(w,\eta,\gamma)\}.$$

Note that the error in this approximation is of order $\mathcal{O}_P(\delta^3)$ and that $\tilde{E}(x \mid w)$ reduces to $E(x \mid w)$ under (1.3). Thus we use it in the following assuming that either (1.2) or (1.3) holds. Substituting $\tilde{E}(x \mid w)$ for $x$ in (1.1) results in a second approximate model

$$U_{mA,2}(w,\beta,\eta,\gamma,\delta^2) = f_m\{\tilde{E}(x \mid w),\beta\}; \tag{3.3a}$$

$$U_{vA,2}(w,\beta,\theta,\sigma^2,\eta,\gamma,\delta^2) = \sigma^2 f_v\{\tilde{E}(x \mid w),\beta,\theta\}. \tag{3.3b}$$

Taylor series expansions of $U_{mA,2}$ and $U_{vA,2}$ show that to terms of order $\mathcal{O}_P(\delta^2)$,

$$U_{mA,1} - U_{mA,2} = (\delta^2/2)\mathrm{tr}\big[f_{mxx}\{g(w,\eta,0),\beta\}g_s(w,\eta,0)h_3(w,\eta,\gamma)g_s(w,\eta,0)^t\big]; \tag{3.4a}$$

$$\begin{aligned} U_{vA,1} - U_{vA,2} = \sigma^2(\delta^2/2)\mathrm{tr}\big[f_{vxx}\{g(w,\eta,0),\beta\}g_s(w,\eta,0)h_3(w,\eta,\gamma)g_s(w,\eta,0)^t\big] \\ +\delta^2 S(w,\eta,\gamma,\beta). \end{aligned} \tag{3.4b}$$

It follows that the strategy of replacing $x$ by $\tilde{E}(x \mid w)$ can be justified whenever the right hand sides of (3.4) are negligible. Consider for example simple logistic regression, $\Pr(y = 1 \mid x) = F(\beta_0 + \beta_1 x)$, where $F(t) = \{1 + \exp(t)\}^{-1}$, under the Berkson error model (1.3). In this case

$$U_{mA,1} - U_{mA,2} \approx (\beta_1^2/2)Var(x \mid w)F^{(1)}\{\beta_0 + \beta_1 g(w,\eta,0)\};$$

$$U_{vA,1} - U_{vA,2} \approx (\beta_1^2/2)Var(x \mid w)\Big(F^{(2)}\{\beta_0 + \beta_1 g(w,\eta,0)\} + 2\big[F^{(1)}\{\beta_0 + \beta_1 g(w,\eta,0)\}\big]^2\Big),$$

where $F^{(k)}$ is the $k^{th}$ derivative of $F$. Thus the two approximations are essentially equal whenever $\beta_1^2 Var(x \mid w)$ is negligible. This occurs near the null model $\beta_1 = 0$, a situation not uncommon in epidemiologic research (Rosner et al., 1989). Similar justification for (3.3) can be found for other generalized linear models.

The strategy of replacing $x$ by $\tilde{E}(x \mid w)$ has the advantage of always producing a range-preserving model whereas (3.2) does not. However, we saw above that it is not always appropriate. We now describe a third approximate model that is range preserving and agrees with (3.2) to $\mathcal{O}_P(\delta^2)$. Let $a_m = a_m(w,\eta,\beta) = f_{mx}\{g(w,\eta,0),\beta\}^t / \|f_{mx}\{g(w,\eta,0),\beta\}\|^2$ and define $a_v = a_v(w,\eta,\beta,\theta)$ analogously. Define

$$U_{mA,3} = f_m\{g(w,\eta,0) + \delta^2 a_m H_m(w,\eta,\gamma,\beta),\beta\}; \tag{3.5a}$$

$$U_{vA,3} = \sigma^2 f_v\Big[g(w,\eta,0) + \delta^2 a_v\{H_v(w,\eta,\gamma,\beta,\theta,\sigma^2) + \sigma^{-2}S(w,\eta,\gamma,\beta)\},\beta,\theta\Big]. \tag{3.5b}$$

Taylor series expansions of $U_{mA,3}$ and $U_{vA,3}$ show that they agree with (3.2) to $\mathcal{O}_P(\delta^2)$.

In Sections 4 and 5 we study estimation for the approximate model $(U_{mA},\ U_{vA})$ where $(U_{mA},\ U_{vA})$ can be any one of $(U_{mA,i},\ U_{vA,i})$ $i = 1,2,3$, given by (3.2), (3.3) or (3.5) respectively. We now show by examples the flexibility and generality of our modelling framework.

**Example 3.1 (General linear Berkson error model):** Consider the model (1.3) with $c^*(w,\eta) = \eta_0 + \eta_1 w$. In the notation of (3.1), $\eta$ is a vector containing the unique elements of $\eta_0$ and $\eta_1$, $g(w,\eta,\delta u) = \eta_0 + \eta_1 w + \delta u$, $h_2 = g_{ss} = 0$, $g_s = I_{\dim(u)}$, $h_3(w,\eta,\gamma) = M^*(w,\eta,\gamma)$ and

$$S(w,\eta,\gamma,\beta) = f_{mx}(\eta_0 + \eta_1 w,\beta)M^*(w,\eta,\gamma)f_{mx}(\eta_0 + \eta_1 w,\beta)^t;$$

$$H_m(w,\eta,\gamma,\beta) = (1/2)\mathrm{tr}\Big\{ f_{mxx}(\eta_0 + \eta_1 w,\beta)M^*(w,\eta,\gamma)\Big\};$$

with a similar expression for $H_v$.

**Example 3.2:** Consider a homoscedastic linear regression Berkson error model $f_m(x,\beta) = x^t\beta$, $f_v(x,\beta,\theta) = 1$, $x = \eta_0 + \eta_1 w + \delta u$ and $Cov(u \mid w) = M^*(w,\eta,\gamma)$. It follows that $E(y \mid w) = \beta^t(\eta_0 + \eta_1 w)$ and $Var(y \mid w) = \sigma^2 + \delta^2\beta^t\eta_1 M^*(w,\eta,\gamma)\eta_1^t\beta$. If $M^*(w,\eta,\gamma)$ is constant, then the observed data will have constant variance and we can estimate $\beta$ only if $\eta_0$, $\eta_1$ are known, as in the classical Berkson case, or estimated from additional data. In the Berkson case, if only one variable, say the last, $x^{(p)}$, with proxy $w^{(p)}$, is measured with error, then the model is sufficiently identified to check for heteroscedastic measurement error. For example, if $e = (0,\ldots,0,1)^t$, then when $M^*(w,\eta,\gamma) = ee^t\exp(\gamma w^{(p)})$, we have $Var(y \mid w) = \sigma^2 + \delta_*^2 exp(\gamma w^{(p)})$. Graphical and formal devices for checking whether $\gamma = 0$ here and estimating $\gamma$ are given in Carroll and Ruppert (1988, Section 2.7, Chapters 3, 6).

**Example 3.3:** In assay models (Davidian *et al.*, 1988; Rudemo *et al.*, 1989), $f_m(x,\beta)$ is usually nonlinear and $f_v(x,\beta,\theta) = f_m(x,\beta)^\theta$. Here $x$ is the univariate log concentration, with zero concentration measured without error and handled separately. In the linear Berkson error version of model (1.3), $\eta_0 = 0$, $\eta_1 = 1$. In practice, both $\sigma$ and $\delta$ are fairly small (Davidian et al., 1988). We have $h_2 = g_{ss} = 0$, $g_s = 1$, and to order $\mathcal{O}_p(\sigma^2 + \delta^2)$, as $\sigma \to 0$ and $\delta \to 0$, equations (3.2) yield

$$E(y \mid w) \approx f_m(w,\beta) + (\delta^2/2)f_{mxx}(w,\beta)h_3(w,\gamma); \tag{3.6a}$$

$$Var(y \mid w) \approx \sigma^2 f_v(w,\beta,\theta) + \delta^2 f_{mx}^2(w,\beta)h_3(w,\gamma). \tag{3.6b}$$

Model (3.6) is typically identifiable. Whether measurement error has constant variance can be assessed by positing forms for $h_3$, e.g., $h_3(w,\gamma) = exp(\gamma w)$, and then using standard techniques for variance analysis. Model (3.6) may not be range preserving and in such cases we suggest using the approximate model given in (3.5). In this example, the variance function is range preserving.

**Example 3.4 (General linear measurement error model):** Consider the linear measurement error model version of (1.2). It follows that for a generalized inverse $\eta^-$ satisfying $\eta^-\eta = I_{p\times p}$, $g(w,\eta,\delta u) = \eta_1^-(w - \eta_0 - \delta u)$, $g_s = -\eta_1^-$, $g_{ss} = 0$. It is easily shown that $E(u \mid \eta_0 + \eta_1 x) = 0$ and that $Cov(u \mid \eta_0 + \eta_1 x) = E\{M(x,\eta,\gamma) \mid \eta_0 + \eta_1 x\} = M_*(x,\eta,\gamma)$,where $M_*(x,\eta,\gamma) = M\{\eta_1^-(x - \eta_0),\eta,\gamma\}$. Let $\kappa_w$ be the marginal density of $w$ with gradient $\kappa_w^{(1)}$. An appeal to Lemma A.1 in the appendix shows that

$$h_2(w,\eta,\gamma) = -\left[\mathrm{tr}\left\{\frac{\partial M_*(w,\eta,\gamma)}{\partial w}\right\} + M_*(w,\eta,\gamma)\frac{\kappa_w^{(1)}(w)}{\kappa_w(w)}\right]; \tag{3.7a}$$

$$h_3(w,\eta,\gamma) = M_*(w,\eta,\gamma). \tag{3.7b}$$

This model includes the possibility that $w$ is a biased measurement for $x$ but can be calibrated with estimated parameters $\eta_0$, $\eta_1$. If $w$ is unbiased so that $\eta_0 = g_{ss} = 0$, $\eta_1 = I_{\dim(x)} = -g_s$, we have the classical measurement error model. For identifiability, one of the diagonal elements of $M$ corresponding to a predictor measured with error must have value 1.0.

In some instances, exact forms for $h_2$ and $h_3$ can be computed. For example, suppose that $u$ and $x$ are independent and normally distributed, the latter with mean $\mu_x$ and covariance $M_x$, and that $E(w \mid x) = x$ and $Cov(w \mid x) = M$. Here $M$ has first diagonal element equal to 1.0. In the notation of (3.1), $g(w,\eta,\delta u) = w - \delta u$, $\gamma = (\mu_x, M_x, M)$, $\Lambda = (I, \gamma^t, \delta^2)^t$ and

$$h_2(w,\Lambda) = M(M_x + \delta^2 M)^{-1}(w - \mu_x); \quad h_3(w,\Lambda) = M\{I - (M_x + \delta^2 M)^{-1}\delta^2 M\}.$$

Often sample sizes are large enough that when the dimension of $w$ is small, the location score $\kappa_w^{(1)}/\kappa_w$ in (3.7a) can be estimated nonparametrically. Related work on hypothesis testing (Stefanski and Carroll, 1989) shows that estimating $\kappa_w^{(1)}/\kappa_w$ is feasible and advantageous when $\dim(w) = 1$. Alternatively, a flexible parametric density could be fit to $\{w_i\}$ thereby providing an estimator of $\kappa_w^{(1)}/\kappa_w$. Note that $\kappa_w^{(1)}/\kappa_w$ is linear if and only if $\kappa_w$ is normal.

**Example 3.3 continued:** Consider the mean and variance functions for Example 3.3 but for a measurement error model instead of a Berkson error model, with $x$ scalar so that $M = 1$. Assuming normality and letting $\sigma \to 0$, $\delta \to 0$, to terms of $\mathcal{O}_P(\sigma^2 + \delta^2)$ we obtain

$$\begin{aligned} Var(y \mid w) &\approx \sigma^2 f_v(w,\beta,\theta) + \delta^2 f_{mx}^2(w,\beta)\sigma_x^2/(\delta^2 + \sigma_x^2); \\ E(y \mid w) &\approx f_m(w,\beta) + \left\{(1/2)\sigma_x^2 f_{mxx}(w,\beta) - f_{mx}(w,\beta)(w - \mu_x)\right\}\delta^2/(\delta^2 + \sigma_x^2) \\ &\approx f_m\left[w + \left(\frac{\delta^2}{\delta^2 + \sigma_x^2}\right)\left\{\frac{f_{mxx}(w,\beta)\sigma_x^2}{2f_{mx}(w,\beta)} - w + \mu_x\right\},\beta\right]. \end{aligned}$$

The two approximations for the mean are from (3.2a) and (3.5a) respectively, the latter appropriate when the true mean is positive. Note that in this case a simpler model can be obtained by replacing $(\delta^2 + \sigma_x^2)$ with $\sigma_x^2$ in both the mean and variance function without affecting the order of the approximation.

**Example 3.5:** The error in using (3.7) is of order $\mathcal{O}_p(\delta^3)$ when $x$ and $u$ are normally distributed. However, (3.7) suggests flexible models which can cope with nonnormal and/or heteroscedastic error. Consider for example, a logistic regression study such as described by Jones *et al.* (1987). In this study, the outcome is incidence of breast cancer, and the predictor $x$ is average daily dietary saturated fat intake. We have analyzed a subset of these data consisting of a cohort of 2888 women under the age of 50. In this group there were 37 cases of breast cancer. Here $w$ is derived from 24-hour diet recall questionnaire and is very variable. It is reasonable in this case to suppose that $w$ is unbiased for $x$. Thus, we set $\eta_0 = 0$ and $\eta_1 = 1$. If measurement error variance is loglinear in $x$, and if we replace the score $\kappa_w^{(1)}(w)/\kappa_w(w)$ by a linear function of $w$, then (3.7) and a little algebra yield

$$E(w \mid x) = x; \quad E(\delta u \mid w) = \delta^2 exp(\alpha_3 w)(\alpha_1 + \alpha_2 w) = \delta^2 h_2(w, \Lambda); \tag{3.8a}$$

$$Var(\delta u \mid w) = \delta^2 exp(\alpha_3 w) = \delta^2 h_3(w, \Lambda). \tag{3.8b}$$

Equations (3.8) identify $h_2$, $h_3$. For binary regression models, $\Pr(y = 1 \mid x) = F(\beta_0 + \beta_1 x)$, (3.2) yields an approximate model which is not range preserving and we suggest using (3.5). In this case (3.5a) gives the approximate mean

$$\Pr(y = 1 \mid w) \approx F\left[\beta_0 + \beta_1\{w - \delta^2 h_2(w, \Lambda)\} + (\delta^2/2)\beta_1^2 F_{21}(\beta_0 + \beta_1 w) h_3(w, \Lambda)\right], \tag{3.9}.$$

with $F_{21}(v) = F_2(v)/F_1(v)$, where $F_1$ and $F_2$ are the first two derivative of $F$. Note that for binary regression models, the approximations in (3.5) satisfy $U_{vA,3} = U_{mA,3}(1 - U_{mA,3})$.

## 4. ESTIMATORS BASED ON APPROXIMATE QUASILIKELIHOOD

Estimation of the parameters in the approximate quasilikelihood/variance function models (3.2), (3.3) and (3.5) is complicated by the fact that data of different types are frequently available. We now describe a general method of obtaining estimators with asymptotically-valid standard errors from data composed of combinations of observations of the types described in Section 2. Our approach to estimation is based on the principles set forth in Davidian and Carroll (1987) and Carroll and Ruppert (1988) but it is tailored to the specifics of the problem at hand.

Presentation of the general theory is facilitated by a judicious partitioning of the parameters in $\Lambda = (\eta^t, \gamma^t, \delta^2)^t$ which we now describe. Recall that $(U_{mA},\ U_{vA})$ is used to represent the approximations in either (3.2), (3.3) or (3.5).

Depending on the particular model assumed in (1.2) or (1.3) some components of $\Lambda$ may be known. Designate the unknown components by $\Lambda_U$. It is implicitly assumed that the components of $\Lambda_U$ are identified by the available data. Let $\Lambda_P^*$ designate those components of $\Lambda_U$ that are identified by the primary data. Partition $\Lambda_P^*$ into those components appearing in the mean function $U_{mA}$, $\Lambda_{Pm}^*$, and those not appearing in the mean function, $\Lambda_{Pv}^*$. Let $\Lambda_{Pm}$ and $\Lambda_{Pv}$ be subsets of $\Lambda_{Pm}^*$ and $\Lambda_{Pv}^*$ respectively. The parameters in $\Lambda_U$ that are capable of being estimated with the primary data are contained in $\Lambda_P^*$ whereas $\Lambda_P = (\Lambda_{Pm}^t, \Lambda_{Pv}^t)^t$ contains the parameters that we choose to estimate using the primary data. The two sets need not be equal and will sometimes differ for reasons of convenience and/or model robustness. Frequently either $\Lambda_{Pm}$ or $\Lambda_{Pv}$ will be empty. Finally, let $\Lambda_{VR}$ contain all of the parameters in $\Lambda_U$ not contained in $\Lambda_{Pm}$ or $\Lambda_{Pv}$. Thus depending on context we can write either $\Lambda_U = (\Lambda_{Pm}^t, \Lambda_{Pv}^t, \Lambda_{VR}^t)^t$ or $\Lambda = (\eta^t, \gamma^t, \delta^2)^t$ or $\Lambda = (\Lambda_U, \Lambda_U^c)$, where the last equality denotes set equivalence. It is also convenient to write $\hat{\Lambda} = (\hat{\eta}^t, \hat{\gamma}^t, \hat{\delta}^2)^t$ even when some components of $\Lambda$ are known.

Most of the information for estimating the components of $\Lambda_U$ is contained in the validation and/or reliability data depending on the particular model under study. We assume that unbiased score equations, $\Upsilon_{R,\Lambda_{Pm}}, \Upsilon_{R,\Lambda_{Pv}}, \Upsilon_{R,\Lambda_{VR}}, \Upsilon_{V,\Lambda_{Pm}}, \Upsilon_{V,\Lambda_{Pv}}$ and $\Upsilon_{V,\Lambda_{VR}}$ are available for obtaining consistent M-estimators of the components of $\Lambda_U$ from the available reliability and validation data. For example, in Example 3.1, $\eta_0$ and $\eta_1$ would often be estimated by the usual normal equations for linear regression. The scores with first subscript $R$ are functions of $\Lambda$ and the replicates of $w$ in the reliability data. The scores with first subscript $V$ are functions of $\Lambda$ and the pairs $(x, w)$ from the validation data.

We propose to estimate $\Theta$ and $\Lambda_U$ with M-estimators $\hat{\Theta}$ and $\hat{\Lambda}_U$ solving equations of the form

$$0 = n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} F_{ji}(\Theta, \Lambda_U) \tag{4.1}$$

where $j$ indexes the five types of observations described in Section 2. We now describe the functional form of $F_{ji}$ for each of the five types of data described in Section 2.

Write $U_{mA}$ and $U_{vA}$ for the right hand sides of (3.2), (3.3) and (3.5), and define $r_A = y - U_{mA}$. Let $\Psi_{P,\beta}, \Psi_{P,\theta}, \Psi_{P,\sigma^2}, \Psi_{P,\Lambda_{Pm}}, \Psi_{P,\Lambda_{Pv}}$ and $\Psi_{P,\Lambda_{VR}}$ be functions of $(y, w, \Theta, \Lambda)$ defined via

componentwise matching in the following equations:

$$(\Psi^t_{P,\beta}, \Psi^t_{P,\Lambda_{Pm}}) = \left(\frac{y - U_{mA}}{U_{vA}}\right)\frac{\partial U_{mA}}{\partial(\beta^t, \Lambda^t_{Pm})};$$

$$(\Psi^t_{P,\theta}, \Psi_{P,\sigma^2}, \Psi^t_{P,\Lambda_{Pv}}) = \left(\frac{r_A^2 - U_{vA}}{U_{vA}}\right)\frac{\partial \log U_{vA}}{\partial(\theta^t, \sigma^2, \Lambda^t_{Pv})};$$

$$\Psi_{P,\Lambda_{VR}} = 0_{\dim(\Lambda_{VR})\times 1}.$$

With these definitions, $F_{1i}(\Theta, \Lambda_U) = (\Psi^t_{P,\beta}, \Psi^t_{P,\theta}, \Psi_{P,\sigma^2}, \Psi^t_{P,\Lambda_{Pm}}, \Psi^t_{P,\Lambda_{Pv}}, \Psi^t_{P,\Lambda_{VR}})^t$ evaluated at $(y_{1i}, w_{1i}, \Theta, \Lambda)$.

Write $f_m$ and $f_v$ for the mean and variance function in (1.1), and define $r = y - f_m$. Let $\Psi_{IV,\beta}, \Psi_{IV,\theta}, \Psi_{IV,\sigma^2}, \Psi_{IV,\Lambda_{Pm}}, \Psi_{IV,\Lambda_{Pv}}$ and $\Psi_{IV,\Lambda_{VR}}$ be functions of $(y, x, w, \Theta, \Lambda)$ defined via componentwise matching in the following equations:

$$\Psi_{IV,\beta} = \left(\frac{y - f_m}{f_v}\right)\frac{\partial f_m}{\partial \beta};$$

$$(\Psi^t_{P,\theta}, \Psi_{P,\sigma^2}) = \left(\frac{r^2 - f_v}{f_v}\right)\frac{\partial \log f_v}{\partial(\theta^t, \sigma^2)};$$

$$(\Psi^t_{IV,\Lambda_{Pm}}, \Psi^t_{IV,\Lambda_{Pv}}, \Psi^t_{IV,\Lambda_{VR}}) = (\Upsilon^t_{V,\Lambda_{Pm}}, \Upsilon^t_{V,\Lambda_{Pv}}, \Upsilon^t_{V,\Lambda_{VR}}).$$

With these definitions, $F_{2i}(\Theta, \Lambda_U) = (\Psi^t_{IV,\beta}, \Psi^t_{IV,\theta}, \Psi_{IV,\sigma^2}, \Psi^t_{IV,\Lambda_{Pm}}, \Psi^t_{IV,\Lambda_{Pv}}, \Psi^t_{IV,\Lambda_{VR}})^t$ evaluated at $(y_{2i}, x_{2i}, w_{2i}, \Theta, \Lambda)$.

For internal reliability data we summarize the observation $(y_i, w_{i1}, \ldots, w_{ik_i})$ as $(y_i, \bar{w}_i, k_i)$; see Section 2. Define $U_{mA*}$ and $U_{vA*}$ as functions of $(\bar{w}, k, \Theta, \Lambda)$ via

$$U_{mA*} = U_{mA}(\bar{w}, \beta, \eta, \gamma, \delta^2/k); \quad U_{vA*} = U_{vA}(\bar{w}, \beta, \theta, \sigma^2, \eta, \gamma, \delta^2/k);$$

and define $r_{A*} = y - U_{mA*}$. Using established conventions, define

$$\Psi_{IR,\beta} = \left(\frac{y - U_{mA*}}{U_{vA*}}\right)\frac{\partial U_{mA*}}{\partial \beta};$$

$$(\Psi^t_{IR,\theta}, \Psi_{IR,\sigma^2}) = \left(\frac{r_{A*}^2 - U_{vA*}}{U_{vA*}}\right)\frac{\partial \log U_{vA*}}{\partial(\theta^t, \sigma^2)};$$

$$(\Psi^t_{IR,\Lambda_{Pm}}, \Psi^t_{IR,\Lambda_{Pv}}, \Psi^t_{IR,\Lambda_{VR}}) = (\Upsilon^t_{R,\Lambda_{Pm}}, \Upsilon^t_{R,\Lambda_{Pv}}, \Upsilon^t_{R,\Lambda_{VR}}).$$

With these definitions, $F_{3i}(\Theta, \Lambda_U) = (\Psi^t_{IR,\beta}, \Psi^t_{IR,\theta}, \Psi_{IR,\sigma^2}, \Psi^t_{IR,\Lambda_{Pm}}, \Psi^t_{IR,\Lambda_{Pv}}, \Psi^t_{IR,\Lambda_{VR}})^t$ evaluated at $(y_{3i}, \bar{w}_{3i\cdot}, k_i, \Theta, \Lambda)$.

Finally, since external data generally provide no information on $\Theta$,

$$F_{4i}(\Theta, \Lambda_U) = (0_{1\times\dim(\Theta)}, \Upsilon^t_{V,\Lambda_{Pm}}, \Upsilon^t_{V,\Lambda_{Pv}}, \Upsilon^t_{V,\Lambda_{VR}})^t;$$

$$F_{5i}(\Theta, \Lambda_U) = (0_{1\times\dim(\Theta)}, \Upsilon^t_{R,\Lambda_{Pm}}, \Upsilon^t_{R,\Lambda_{Pv}}, \Upsilon^t_{R,\Lambda_{VR}})^t,$$

evaluated at $(x_{4i}, w_{4i})$ and $(w_{5i1}, \ldots, w_{5ik_i})$ respectively.

The scores we defined for estimating $\theta, \sigma^2$ and $\Lambda_{Pv}$ use squared residuals. Other scores based on functions of absolute residuals can also be employed; see Davidian and Carroll (1987).

Let $n = n_1 + \cdots + n_5$ and $p_{j,n} = n_j/n$. The asymptotic distribution of $(\hat{\Theta}, \hat{\Lambda}_U)$ is given for the case that $n \to \infty$ and $p_{j,n} \to p_j \geq 0$.

The M-estimators $(\hat{\Theta}, \hat{\Lambda}_U)$ converge in probability to $(\Theta_*, \Lambda_{U*})$ satisfying

$$0 = \lim_{n\to\infty} n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} E\Big\{F_{ji}(\Theta_*, \Lambda_{U*})\Big\}.$$

Let $G_{ji}(\Theta, \Lambda_U)$ be the matrix of partial derivatives of $F_{ji}$ with respect to $(\Theta^t, \Lambda_U^t)$, and define

$$A = \lim_{n\to\infty} n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} E\Big\{G_{ji}(\Theta_*, \Lambda_{U*})\Big\};$$

$$B = \lim_{n\to\infty} n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} E\Big\{F_{ji}(\Theta_*, \Lambda_{U*})F_{ji}(\Theta_*, \Lambda_{U*})^t\Big\};$$

Moment estimators of these matrices are given by

$$\hat{A} = n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} G_{ji}(\hat{\Theta}, \hat{\Lambda}_U); \quad \hat{B} = n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} F_{ji}(\hat{\Theta}, \hat{\Lambda}_U)F_{ji}(\hat{\Theta}, \hat{\Lambda}_U)^t.$$

Standard asymptotic results imply that under sufficient regularity conditions,

$$n^{1/2}\Big\{(\hat{\Theta} - \Theta_*)^t, \ (\hat{\Lambda}_U - \Lambda_{U*})^t\Big\}^t \xrightarrow{D} \mathcal{N}\{0, \ A^{-1}B(A^t)^{-1}\};$$

$$\hat{A}^{-1}\hat{B}(\hat{A}^t)^{-1} \xrightarrow{P} A^{-1}B(A^t)^{-1}. \tag{4.2}$$

**Example 4.1 (Externally validated studies):** Suppose that $\Lambda_U = \Lambda$, $\Lambda_{Pm}$ and $\Lambda_{Pv}$ are both empty, and that $n_2 = n_3 = n_5 = 0$. Thus estimation of $\Lambda$ depends entirely on an external validation data set. In this case the only nonzero $F_{ji}$, are

$$F_{1i}(\Theta, \Lambda_U) = (\Psi^t_{P,\beta}, \Psi^t_{P,\theta}, \Psi_{P,\sigma^2}, 0_{1\times\dim(\Lambda_U)})^t;$$

$$F_{4i}(\Theta, \Lambda_U) = (0_{1\times\dim(\Theta)}, \Upsilon^t_{V,\Lambda_{Pm}}, \Upsilon^t_{V,\Lambda_{Pv}}, \Upsilon^t_{V,\Lambda_{VR}})^t.$$

Let $\Psi$ and $\Upsilon$ denote the non-identically zero components of $F_{11}$ and $F_{41}$ respectively. The matricies in the asymptotic covariance matrix of (4.2) for this case have the forms:

$$A = \begin{pmatrix} p_1 A_{\Psi,\Theta} & p_1 A_{\Psi,\Lambda} \\ 0 & p_4 A_{\Upsilon,\Lambda} \end{pmatrix}; \qquad B = \begin{pmatrix} p_1 B_\Psi & 0 \\ 0 & p_4 B_\Upsilon \end{pmatrix};$$

where for example, $A_{\Psi,\Theta} = E(\partial\Psi/\partial\Theta^t)$ and $B_\Psi = E(\Psi\Psi^t)$. In this case the asymptotic covariance matrix of $n^{1/2}(\hat{\Theta} - \Theta_*)$ has the form

$$p_1^{-1} A_{\Psi,\Theta}^{-1} B_\Psi (A_{\Psi,\Theta}^t)^{-1} + p_4^{-1}\Delta, \tag{4.3}$$

where $\Delta$ is a nonnegative matrix. The matrix $\Delta$ depends on the submatricies of $A$ and $B$ in a simple but not very informative way. The term $p_4^{-1}\Delta$ is the contribution to the asymptotic covariance matrix of $n^{1/2}(\hat{\Theta} - \Theta_*)$ due to estimating $\Lambda$.

If costs, $c_1$, $c_2$ can be assigned to obtaining observations $(y, w)$, $(x, w)$ respectively, then (4.3) could be used for design purposes.

## 5. ALTERNATIVE ESTIMATORS

Denote the right hand side of (4.1) by $L^*(\Theta, \eta, \gamma, \delta^2)$. Frequently $\Lambda_{Pm}$ and $\Lambda_{Pv}$ are both empty, and then $L^* = (L^t, L_*^t)^t$ where $L_*$ does not depend on $\Theta$. It follows that $\hat{\Lambda}_U$ is found by solving $L_*(\Lambda_U) = 0$ and $\hat{\Theta}$ is found by solving $L(\Theta, \hat{\eta}, \hat{\gamma}, \hat{\delta}^2) = 0$. In this case there are two additional approaches to estimation that deserve attention. Below we adapt the methods proposed by Whittemore and Keller (1988) and Stefanski (1985) to our estimation problem.

In the following we drop the distiction between $\hat{\Lambda}_U$ and $\hat{\Lambda}$ keeping in mind that the latter is $(\hat{\Lambda}_U, \Lambda_U^c)$. Note that

$$L(\Theta, \Lambda) = n^{-1} \sum_{j=1}^{3} \sum_{i=1}^{n_j} F_{ji}^\dagger(\Theta, \Lambda),$$

where $F_{ji}^\dagger(\Theta, \Lambda)$ denotes the first $\dim(\Theta)$ rows of $F_{ji}(\Theta, \Lambda)$, $(j = 1, 2, 3,\ i = 1, \ldots, n_j)$.

### 5.1 APPROXIMATING THE QUASILIKELIHOOD ESTIMATORS

Let $\hat{\Theta}(\tau)$ be a function of $\tau$ defined implicitly by the equation $L(\hat{\Theta}(\tau), \hat{\eta}, \hat{\gamma}, \tau\hat{\delta}^2) = 0$. Note that $\hat{\Theta}(1) = \hat{\Theta}$ and $\hat{\Theta}(0)$ is the so-called naive estimator obtained by fitting model (1.1) to the pairs $(y_{ji}, \hat{x}_{Cji})_{i=1}^{n_j}, j = 1, 2, 3$, using standard estimation techniques where $(y_{ji}, \hat{x}_{Cji}) = (y_{ji}, \tilde{x}_{ji}(\hat{\eta}, 0))$ and

$$(y_{ji}, \tilde{x}_{ji}(\eta, \delta)) = \begin{cases} (y_{1i}, g(w_{1i}, \eta, \delta u_{1i})), & \text{if } j = 1; \\ (y_{2i}, x_{2i}), & \text{if } j = 2; \\ (y_{3i}, g(\bar{w}_{3i\cdot}, \eta, \delta \bar{u}_{3i\cdot})), & \text{if } j = 3. \end{cases} \tag{5.1}$$

In fact the naive estimator defined above satisfies

$$0 = n^{-1} \sum_{j=1}^{3} \sum_{i=1}^{n_j} \Psi_T(y_{ji}, \hat{x}_{Cji}, \hat{\Theta}(0))$$

where $\Psi_T$ is the appropriate score for model (1.1) in the absence of measurement error.

With this notation we now derive the estimator proposed by Whittemore and Keller (1988). Taylor series expansions of $L(\hat{\Theta}(\tau), \hat{\eta}, \hat{\gamma}, \tau\hat{\delta}^2)$ and $\hat{\Theta}(\tau)$ lead to the approximations

$$L_\Theta(\hat{\Theta}(0), \hat{\eta}, \hat{\gamma}, 0)\hat{\Theta}_\tau(0) + \hat{\delta}^2 L_{\delta^2}(\hat{\Theta}(0), \hat{\eta}, \hat{\gamma}, 0) \approx 0$$

$$\hat{\Theta}_\tau(0) \approx \hat{\Theta}(1) - \hat{\Theta}(0);$$

where for example $L_\Theta = \partial L/\partial \Theta^t$ and $\hat{\Theta}_\tau = \partial\hat{\Theta}/\partial\tau$. Thus, $\hat{\Theta} \approx \hat{\Theta}_{c,1}$ where,

$$\hat{\Theta}_{c,1} = \hat{\Theta}(0) - \hat{\delta}^2\{L_\Theta(\hat{\Theta}(0), \hat{\eta}, \hat{\gamma}, 0)\}^{-1} L_{\delta^2}(\hat{\Theta}(0), \hat{\eta}, \hat{\gamma}, 0).$$

The utility of $\hat{\Theta}_{c,1}$ lies in its computability. It is an explicit function of $\hat{\Theta}(0)$ which in turn often can be obtained using standard statistical software. The difference, $\hat{\Theta} - \hat{\Theta}_{c,1}$ is $\mathcal{O}(\hat{\delta}^4)$ ***almost surely*** when $L$ is a well-behaved function of $\Theta$ and $\delta^2$.

Write $\hat{\Theta}_{c,1} = \hat{\Theta}_{c,1}(U_{mA}, U_{vA})$ to emphasize the dependence of $\hat{\Theta}_{c,1}$ on $U_{mA}$ and $U_{vA}$ and let $U_m$ and $U_v$ denote the left hand sides of (3.2a) and (3.2b) respectively. Then since the approximations in (3.2) and (3.5) are of order $\mathcal{O}_P(\delta^3)$ it follows that $\hat{\Theta}_{c,1}(U_{mA}, U_{vA}) = \hat{\Theta}_{c,1}(U_m, U_v)$ in these two cases. Thus $\hat{\Theta}_{c,1}$ is the estimator proposed by Whittemore and Keller (1988) whenever the approximations in (3.2) or (3.5) are employed.

Note that $\hat{\Theta}_{c,1}$ can be written in the form

$$\hat{\Theta}_{c,1} = \hat{\Theta}(0) + (\hat{\delta}^2/2)Q_n(\hat{\Theta}(0), \hat{\Lambda})^{-1} H_n(\hat{\Theta}(0), \hat{\Lambda})$$

where

$$Q_n(\Theta, \Lambda) = n^{-1}\sum_{j=1}^{3}\sum_{i=1}^{n_j} q_{ij}(\Theta, \Lambda); \quad H_n(\Theta, \Lambda) = n^{-1}\sum_{j=1}^{3}\sum_{i=1}^{n_j} h_{ij}(\Theta, \Lambda);$$

$$h_{ij}(\Theta, \Lambda) = -2\left\{\frac{\partial}{\partial\delta^2}F^{\dagger}_{ji}(\Theta, \Lambda)\right\}_{\delta^2=0}; \quad q_{ij}(\Theta, \Lambda) = \Psi_{T\Theta}\left(y_{ji}, \tilde{x}_{ji}(\eta, 0), \Theta\right);$$

and $\Psi_{T\Theta} = (\partial/\partial\Theta^t)\Psi_T$.

## 5.2 CORRECTING THE NAIVE ESTIMATORS

Assume temporarily that for internal reliability data $k_i = k$ for $i = 1, \ldots, n_3$ and suppose that $\hat{\eta}$ and $\hat{\gamma}$ are consistent for $\eta$ and $\gamma$. Then the naive estimator, $\hat{\Theta}(0)$, converges in probability to $\Theta_N$ satisfying

$$0 = \sum_{j=1}^{3} p_j \Psi_T(y_{j1}, x_{Cj1}, \Theta_N),$$

where $(y_{j1}, x_{Cj1}) = (y_{j1}, \tilde{x}_{j1}(\eta, 0))$, $j = 1, 2, 3$. Let $(y_{j1}, x_{j1}) = (y_{ji}, \tilde{x}_{ji}(\eta, \delta))$, $j = 1, 2, 3$. In the appendix we show that under both (1.2) and (1.3) there exist functions $d_{1j}(y, x, \eta, \gamma)$ and $d_{2j}(y, x, \eta, \gamma)$ such that

$$E(x_{Cj1} - x_{j1} \mid y_{j1}, x_{j1}) = \delta^2 d_{1j}(y_{j1}, x_{j1}, \eta, \gamma) + \mathcal{O}_P(\delta^3);$$

$$Cov(x_{Cj1} - x_{j1} \mid y_{j1}, x_{j1}) = \delta^2 d_{2j}(y_{j1}, x_{j1}, \eta, \gamma) + \mathcal{O}_P(\delta^3).$$

Note that $d_{12}$ and $d_{22}$ are identically zero under both (1.2) and (1.1). Also $d_{13}$ and $d_{23}$ are defined only under (1.2) and in this case $d_{13} = d_{11}/k$ and $d_{23} = d_{21}/k$.

An adaptation of the Taylor series argument in Stefanski (1985) shows that $\Theta_N = \Theta - (\delta^2/2)Q^{-1}H + \mathcal{O}(\delta^3)$ where

$$Q = E\left\{\sum_{j=1}^{3} p_j \Psi_{T\Theta}\left(y_{j1}, x_{Cj1}, \Theta\right)\right\};$$

$$H = E\left(\sum_{j=1}^{3} p_j \left[2\Psi_{Tx}\left(y_{j1}, x_{j1}, \Theta\right) d_{1j}\left(y_{j1}, x_{j1}, \eta, \gamma\right) + \mathrm{tr}\left\{\Psi_{Txx}\left(y_{j1}, x_{j1}, \Theta\right) d_{2j}\left(y_{j1}, x_{j1}, \eta, \gamma\right)\right\}\right]\right).$$

Let $\kappa_{i1} = 1$, $\kappa_{i2} = 0$, and $\kappa_{i3} = k_i^{-1}$ and define

$$h_{ij}(\Theta, \Lambda) = \kappa_{ij}\Bigg(2\Psi_{Tx}\{y_{ji}, \tilde{x}_{ji}(\eta, 0), \Theta\} d_{11}\{y_{ji}, \tilde{x}_{ji}(\eta, 0), \eta, \gamma\}$$
$$+ \mathrm{tr}\Big[\Psi_{Txx}\{y_{ji}, \tilde{x}_{ji}(\eta, 0), \Theta\} d_{21}\{y_{ji}, \tilde{x}_{ji}(\eta, 0), \Theta\}\Big]\Bigg);$$

$$q_{ij}(\Theta, \Lambda) = \Psi_{T\Theta}\{y_{ji}, \tilde{x}_{ji}(\eta, 0), \Theta\};$$

$$Q_n(\Theta, \Lambda) = n^{-1} \sum_{j=1}^{3} \sum_{i=1}^{n_j} q_{ij}(\Theta, \Lambda); \quad H_n(\Theta, \Lambda) = n^{-1} \sum_{j=1}^{3} \sum_{i=1}^{n_j} h_{ij}(\Theta, \Lambda).$$

Then $\hat{\Theta}_{c,2}$ defined as

$$\hat{\Theta}_{c,2} = \hat{\Theta}(0) + (\hat{\delta}^2/2) Q_n(\hat{\Theta}, \hat{\Lambda})^{-1} H_n(\hat{\Theta}, \hat{\Lambda})$$

is the estimator proposed in Stefanski (1985) adapted to to our model.

### 5.3 ASYMPTOTIC DISTRIBUTIONS

We now derive the asymptotic joint distribution of $\hat{\Theta}(0)$, $\hat{\Lambda}$ and $\hat{\Theta}_c$ where $\hat{\Theta}_c = \hat{\Theta}_{c,1}$ or $\hat{\Theta}_{c,2}$. This is accomplished by representing $(\hat{\Theta}(0)^t, \hat{\Lambda}^t, \hat{\Theta}_c^t)^t$ as an M-estimator and appealing to standard asymptotic theory.

Let

$$\Psi_{Tji}(\Theta_1, \Lambda) = \begin{cases} \Psi_T(y_{ji}, \tilde{x}_{ji}(\eta, 0), \Theta_1), & j = 1, 2, 3; \\ 0_{\dim(\Theta)\times 1} & j = 4, 5; \end{cases}$$

$$\Upsilon_{ji}(\Lambda) = \begin{cases} 0_{\dim(\Lambda)\times 1}, & j = 1; \\ \Psi_{IV,\Lambda}(x_{ji}, w_{ji}, \Lambda), & j = 2, 4; \\ \Psi_{IR,\Lambda}(w_{ji1}, \ldots, w_{jik_i}, \Lambda), & j = 3, 5; \end{cases}$$

$$\Delta_{ji}(\Theta_1, \Lambda, \Theta_2) = \begin{cases} q_{ij}(\Theta_1, \Lambda)(\Theta_2 - \Theta_1) - (\delta^2/2)h_{ij}(\Theta_1, \Lambda), & j = 1, 2, 3; \\ 0_{\dim(\Theta)\times 1} & j = 4, 5. \end{cases}$$

Define $C_{ji}(\Theta_1, \Lambda, \Theta_2) = (\Psi_{Tji}^t, \Upsilon_{ji}^t, \Delta_{ji}^t)^t$. Then, $\hat{\Theta}(0)$, $\hat{\Lambda}$ and $\hat{\Theta}_c$ satisfy

$$0 = \sum_{j=1}^{5} \sum_{i=1}^{n_j} C_{ji}(\hat{\Theta}(0), \hat{\Lambda}, \hat{\Theta}_c).$$

The M-estimators $\hat{\Theta}(0)$, $\hat{\Lambda}$ and $\hat{\Theta}_c$ converge in probability to $\Theta_*$, $\Lambda_*$ and $\Theta_{c*}$ respectively where

$$0 = \lim_{n\to\infty} n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} E\left\{C_{ji}(\Theta_*, \Lambda_*, \Theta_{c*})\right\}.$$

Let $D_{ji}(\Theta_1, \Lambda, \Theta_2)$ be the matrix of partial derivatives of $C_{ji}$ with respect to $(\Theta_1^t, \Lambda^t, \Theta_2^t)$, and define

$$A = \lim_{n\to\infty} n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} E\left\{D_{ji}(\Theta_*, \Lambda_*, \Theta_{c*})\right\};$$

$$B = \lim_{n\to\infty} n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} E\left\{C_{ji}(\Theta_*, \Lambda_*, \Theta_{c*}) C_{ji}(\Theta_*, \Lambda_*, \Theta_{c*})^t\right\}.$$

Moment estimators of these matrices are given by

$$\hat{A} = n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} D_{ji}(\hat{\Theta}(0), \hat{\Lambda}, \hat{\Theta}_c); \quad \hat{B} = n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} C_{ji}(\hat{\Theta}(0), \hat{\Lambda}, \hat{\Theta}_c) C_{ji}(\hat{\Theta}(0), \hat{\Lambda}, \hat{\Theta}_c)^t.$$

Standard asymptotic results imply that under sufficient regularity conditions,

$$n^{1/2} \left\{(\hat{\Theta}(0) - \Theta_*)^t, \ (\hat{\Lambda} - \Lambda_*)^t, \ (\hat{\Theta}_c - \Theta_{c*})^t\right\}^t \xrightarrow{D} \mathcal{N}\left\{0, \ A^{-1}B(A^t)^{-1}\right\};$$

$$\hat{A}^{-1}\hat{B}(\hat{A}^t)^{-1} \xrightarrow{P} A^{-1}B(A^t)^{-1}.$$

## 6. EXAMPLES

We now present two examples, one with a Berkson error structure and the second with a measurement error structure. We do not intend to compare various methods of modelling/estimation.

Instead the examples are used to illustrate the effect of estimating the nuisance parameter $\Lambda$ on selected methods.

**Lung Function (Berkson Model):** Tosteson *et al.* (1989) fit a probit model to data from a study of respiratory health and exposure to nitrogen dioxide. They had access to two external validation data sets in addition to the primary data set. Here $y$ = presence of wheeze, $x$ = log personal exposure and $w$ = ( log bedroom exposure, log kitchen exposure )$^t$.

They fit the model $x = \eta_0 + \eta_1^t w + \delta u$; $E(u|w) = 0$; $Var(u \mid w) = 1$. In our notation, $g_{ss} = h_2 = 0$, $g_s = h_3 = 1$. Tosteson *et al.* assume a probit binary model with $u$ standard normal. Using the smaller of the two validation data sets (the Portage data), we find that $\delta = 0.265$, $\eta_0 = 1.28$ and $\eta_1^t = (0.28, 0.33)$. Let $w_\eta = \eta_0 + \eta_1^t w$. For probit regression (3.9) gives that to order $O_P(\delta^3)$ with no assumptions on the distribution of $u$,

$$\begin{aligned} \Pr(y = 1 \mid w_\eta) &\approx \Phi\Big\{(\beta_0 + \beta_1 w)(1 - \delta^2\beta_1^2/2)\Big\} \\ &\approx \Phi\Big\{(\beta_0 + \beta_1 w_\eta)/(1 + \delta^2\beta_1^2)^{1/2}\Big\}. \end{aligned} \tag{6.1}$$

The second approximation follows easily from the first and is exact when $u$ is normally distributed. The observed probit regression slope estimate is 0.05. This estimate can be related to $\beta_1$ by (6.1), from which it is intuitively obvious that there is essentially no effect due to estimating $\Lambda$. This intuition is confirmed by an application of (4.2).

**Diet and Breast Cancer (Measurement Error Model):** For illustration, we consider a sample of 2888 women under the age of 50, the data being a subset of those used by Jones *et al.* (1987). The response $y$ is an indicator of breast cancer and the predictor $x$ is the logarithm of long term saturated fat in diet. There were 37 cases of breast cancer in the study. Instead of observing $x$ exactly, we observe a 24-hour diet recall measure $w$ of the log saturated fat. Such dietary measures exhibit great variability. In the study, $w$ had mean 2.98 and standard deviation 0.635. A logistic regression fit to the $(y, w)$ data yielded an estimated slope of -0.40, with estimated standard error 0.24 and p-value 0.08. This is clearly a measurement error model, and we use the approximate model $w = x + \delta u$, where $u$ given $x$ has mean 0 and variance 1. We do not have access to validation/reliability data, but it seems reasonable as a first approximation to use the validation results of Willett *et al.* (1985). They do not transform fat intake. They use four seven-day diet record measurements and find that the correlation between any two weekly measurements is approximately 0.55. If we assume normally distributed measurement error in the log scale, then their data suggests that for their study, $\delta = 0.34$. Their seven day diet records differ from our

24-hour recall measurements and should be more precise. It is not clear how to reconcile these figures, but as a reasonable guess to illustrate our methods, we use $\delta = 0.53$, i.e., seven-day diet records are about 2.5 times less variable than 24-hour recall.

Using this value for $\delta$, the method of Section 5.2 yields estimated slope of -0.62 respectively. When we treat $\hat{\delta}$ as fixed, the slope estimate has an estimated standard error of 0.24 and a p-value of 0.01. To assess the effect of estimating $\delta$ under the constraint that we do not have actual external reliability data, we assumed that $u$ was normally distributed, so that with the sample size of 150 in the paper by Willett *et al.* (1985), $\hat{\delta}^2$ has approximate variance $2\delta^2/150 \approx (0.033)^2$. Using the theory of Section 5, we find that the adjusted standard error for the slope is 0.23, an insignificant change.

Because we expected the standard error of the corrected estimate to be larger than that of the naive estimate, we also performed a small bootstrap simulation. We resampled the $(y, w)$ pairs to form bootstrap samples, and repeated the experiment 200 times. The usual analysis had bootstrap mean -0.39 and standard error 0.16, while the corrected analysis had bootstrap mean -0.60 and standard error 0.23. Kernel density estimates using a Gaussian kernel with bandwidth 0.20 are given in Figure 1. The asymptotic theory and the bootstrap analysis are remarkably similar for the corrected estimate, but for the usual analysis the estimated standard error seems to be a bit too high. The latter can be explained as follows. Let $x_i$ be the vector consisting of 1.0 and the observed diet measurement. Note that the usual logistic regression analysis gives an estimated asymptotic covariance matrix of $\left\{\sum_1^n x_i x_i^t F^{(1)}(x_i^t\hat{\beta})\right\}^{-1} = Q_n^{-1}$, see (6.3), while the usual theory of M-estimates yields an estimated covariance of $Q_n^{-1}\left[\sum_1^n x_i x_i^t\{y_i - F(x_i^t\hat{\beta})\}^2\right]^{-1}Q_n^{-1}$. Since, for the most part, $y_i \approx F(x_i^t\hat{\beta}) \approx 0$, one might expect that the ordinary logistic standard errors may be a bit too large.

Because of the imprecision in relating 24-hour recall measures with seven-day diet records, as well as the fact that we have not included other predictors, we wish to emphasize that the preceeding analysis was illustrative only.

## 7. CONCLUSION

The class of quasilikelihood/variance function models (1.1) is very broad and of recognized importance in statistical practice. We have examined some general methods of constructing parameter estimates when the predictors in (1.1) are measured with error. A major part of this paper, Sections 4 and 5, developed a comprehensive asymptotic theory for estimates derived using these methods. The theory provides usable standard error estimates and allows for the presence of

validation and/or replication data.

In Sections 2 and 3, we developed range–preserving models for the observed data based upon (1.1). One such class of models, (3.5), was shown to be correct to order $O(\delta^3)$. A second set of models, which replace $x$ by an estimate of $E(x|w)$, is given by (3.3). We showed that these are correct only to order $O(\delta^2)$, although in (3.4) we note that in many applications the difference between (3.3) and (3.5) will be negligible. Taken together, the range–preserving model classes (3.3) and (3.5) include as special cases most of the suggestions made previously in the literature.

One can thus summarize the paper as having: (i) developed very broad classes of models and estimators; (ii) taken explicit account of the types of data sets, including validation and replication data, that are likely to be available in a measurement error model; and (iii) provided the asymptotic distribution theory for parameter estimates including formulae for obtaining consistent standard errors. We view these combined results as a necessary first step toward addressing the question of which method works best in practice. One advantage of the range–preserving models (3.3) and (3.5) is that, being models for the observed data, they can be checked for fit and compared to one another, as is done in a specific example by Rudemo *et al.* (1989). In our first example, we found that the two models were essentially the same for the likely values of the parameters. In other cases, the two models may be very different. If forced to choose between the two in the absence of an applied context, we would recommend (3.5) over (3.3). The former is based on higher-order expansions in terms of $\delta$, and it is more flexible with respect to modelling, as it can accomodate heterogeneous variability in the relationship between $x$ and $w$. As seen in the discussion of logistic regression just following (3.4), such heterogeneity can be important in model fitting.

## ACKNOWLEDGEMENT

Our research was supported by the National Institutes of Health, the National Science Foundation and a cooperative agreement among Harvard School of Public Health, Societal Institute of Mathematical Sciences and the Environmental Protection Agency. We thank Alice Whittemore, David Ruppert and Donna Spiegelman for making advance copies of their papers available to us.

## REFERENCES

Amemiya, Y. and Fuller, W. A. (1988). Estimation for the nonlinear functional relationship. *Annals of Statistics*, 16, 147-160.

Armstrong, B. (1985). Measurement error in the generalized linear model. *Communications in Statistics*, **B**, 14, 529-544.

Carroll, R. J. (1989). Covariance analysis in generalized linear measurement error models. *Statistics in Medicine*, to appear.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall, London.

Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American*

*Statistical Association*, 82, 1079-1091.

Davidian, M., Carroll, R. J., Smith, W. (1988). Variance functions and the minimum detectable concentration in assays. *Biometrika*, 75, 549-556.

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

Jones, D. Y., Schatzkin, A., Green, S. B., Block, G. Brinton, L. A., Ziegler, R. G., Hoover, R. and Taylor, P. R. (1987). Dietary fat and breast cancer in the National Health and Nutrition Survey I: epidemiologic follow-up study. *Journal of the National Cancer Institute*, 79, 465-471.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* London: Chapman and Hall.

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,* to appear.

Rudemo, M., Ruppert, D. and Streibig, J. C. (1989). Random effect models in nonlinear regression with applications to bioassay. *Biometrics*, 45, 349-362.

Schafer, D. (1987). Covariate measurement error in generalized linear models. *Biometrika*, 74, 385-391.

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

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

Stefanski, L. A. and Carroll, R. J. (1989). Score tests in generalized linear measurement error models. *Journal of the Royal Statistical Society, Series B*, to appear.

Tosteson, T., Stefanski, L. A. and Schafer D.W. (1989). A measurement error model for binary and ordinal regression. *Statistics in Medicine*, to appear.

Tosteson, T. and Tsiatis, A. (1988). The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates. *Biometrika*, 75, 507-514.

Whittemore, A. S. and Keller, J. B. (1988). Approximations for errors in variables regression. *Journal of the American Statistical Association*, 83, 1057-1066.

Willett, W. C., Sampson, L., Stampfer, M. J., Rosner, B., Bain, C., Witschi, J., Hennekens, C. H. and Speizer, F. E. (1985). Reproducibility and validity of a semiquantitative food frequency questionairre. *American Journal of Epidemiology*, 122, 51-65.

## APPENDIX

The following lemma is used in Example 3.4 and in the derivations of the functions $d_{1j}$ and $d_{2j}$, $(j = 1, 2, 3)$ defined in Section 5.2.

**LEMMA A.1.** Suppose that $v_1$, $v_2$ and $v_3$ are random vectors such that $v_1 = v_2 + \delta v_3$ where $\delta > 0$ is a constant scalar. If $v_1$ and $v_2$ have joint density $f_{v_1 v_2}(v_1, v_2)$, $E(v_3 \mid v_2) = 0$, $Cov(v_3 \mid v_2) = M(v_2)$ and $E(v_3 \mid v_1)$ and $Cov(v_3 \mid v_1)$ are three-times differentiable functions of $\delta$ almost

surely, then

$$E(v_3 \mid v_1) = -\delta \left[ \mathrm{tr} \left\{ \frac{\partial M(v_1)}{\partial v_1} \right\} + M(v_2) \frac{\partial \log\{f_{v_2}(v_1)\}}{\partial v_1} \right] + \mathcal{O}_P(\delta^2)$$
$$= -\delta \left[ \mathrm{tr} \left\{ \frac{\partial M(v_1)}{\partial v_1} \right\} + M(v_2) \frac{\partial \log\{f_{v_1}(v_1)\}}{\partial v_1} \right] + \mathcal{O}_P(\delta^2);$$

$$Cov(v_3 \mid v_1) = M(v_1) + \mathcal{O}_P(\delta).$$

**PROOF.** Under the assumptions of the lemma the assertion about $Cov(v_3 \mid v_1)$ is obvious and the second expression for $E(v_3 \mid v_1)$ follows easily from the first. Thus, it only remains to identify the terms in the Maclaurin series expansion of $E(v_3 \mid v_1)$. Note that

$$E(v_3 \mid v_1) = \frac{\int v_3 f_{v_3|v_2}(v_3 \mid t - \delta v_3) f_{v_2}(t - \delta v_3) dv_3}{\int f_{v_2 v_3}(t - \delta v_3, v_3) dv_3}. \tag{A.1}$$

The denominator in (A.1) equals $f_{v_2}(v_1) + \mathcal{O}_P(\delta)$. After a Taylor series expansion, the numerator in (A.1) is shown to be

$$-\delta \left[ \int v_3 (\partial/\partial v_1) f_{v_3|v_2}(v_3 \mid v_1) v_2 dv_3 + \int f_{v_3|v_2}(v_3 \mid v_1) v_3 v_3^t (\partial/\partial v_1) f_{v_2}(v_1) dv_3 \right] + \mathcal{O}_P(\delta^2)$$

$$= -\delta \left[ \mathrm{tr} \left\{ \frac{\partial M(v_1)}{\partial v_1} \right\} f_{v_2}(v_1) + M(v_1) \frac{\partial f_{v_2}(v_1)}{\partial v_1} \right] + \mathcal{O}_P(\delta^2),$$

completing the proof. ••

We now derive the functions $d_{1j}$ and $d_{2j}$, $(j = 1, 2, 3)$ defined in Section 5.2. Note that $x_{Cj1}$ is function of $w_{j1}$ and $x_{j1}$, $(j = 1, 2, 3)$ and thus the conditional independence assumption of Section 1.1 implies that $d_{1j}$ and $d_{2j}$ are functions of $x_{j1}$ almost surely for $j = 1, 2, 3$.

Consider first the generalized measurement error model (1.2). Note that $g\{c(x_{11}, \eta), \eta, 0\} = x_{11}$ almost surely. A Taylor series expansion shows that

$$x_{C11} = g(w_{11}, \eta, 0) = x_{11} + \delta g_w\{c(x_{11}, \eta), \eta, 0\} u + (\delta^2/2) \,\mathrm{tr}\left[ g_{ww}\{c(x_{11}, \eta), \eta, 0\} u_{11} u_{11}^t \right] + \mathcal{O}_P(\delta^3).$$

Thus under (1.2),

$$d_{11}(y_{j1}, x_{j1}, \eta, \gamma) = \mathrm{tr}\Big[ g_{ww}\{c(x_1 1, \eta), \eta, 0\} M(x_{11}, \eta, \gamma) \Big];$$

$$d_{21}(y_{j1}, x_{j1}, \eta, \gamma) = g_w\{c(x_{11}, \eta), \eta, 0\} M(x_{11}, \eta, \gamma) g_w^t\{c(x_{11}, \eta), \eta, 0\}.$$

Both $d_{12}$ and $d_{22}$ are identically 0 and $d_{13} = d_{11}/k$ and $d_{23} = d_{21}/k$ where $k$ is the number of replicates.

Now consider the generalized Berkson error model (1.3). Let $f_{c^*}(c^*)$ be the density of $c^*(w,\eta)$, assumed to exist. Conditioning first on $\{c^*(w,\eta), w\}$ and then on $\{c^*(w,\eta)\}$ shows that

$$E\{u \mid c^*(w,\eta)\} = 0, \qquad Cov\{u \mid c^*(w,\eta)\} = M^{**}(w,\eta,\gamma),$$

where $M^{**}(w,\eta,\gamma) = E\{M^*(w,\eta,\gamma) \mid c^*(w,\eta)\}$. Frequently $M^*$ is a function of $c^*$ and then $M^{**} = M^*$. However, this is not always the case.

It follows directly from Lemma A.1 that

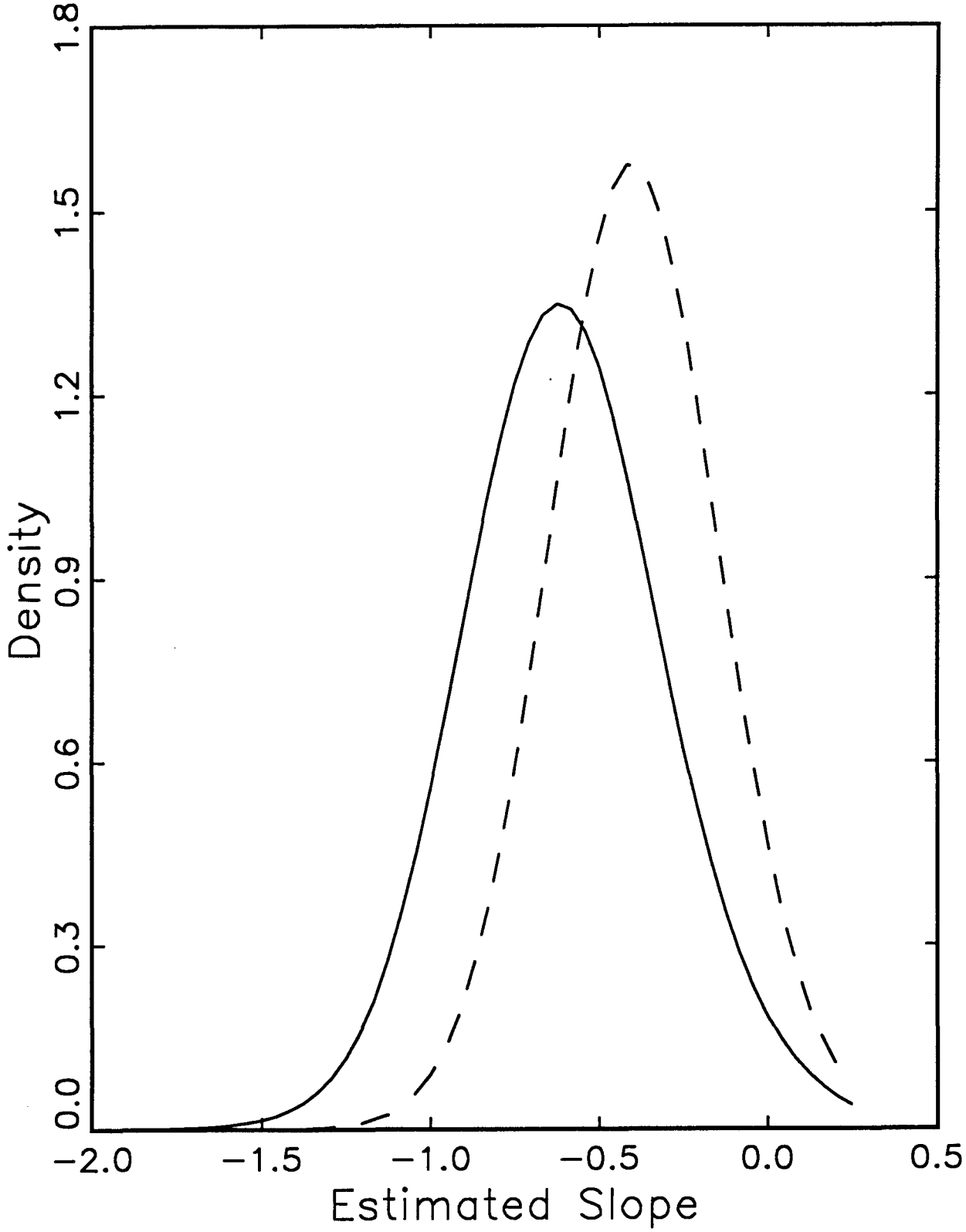
$$d_{11}(y_{j1}, x_{j1}, \eta, \gamma) = \mathrm{tr}\left\{\frac{\partial M^{**}(x_{11},\eta,\gamma)}{\partial x_{11}}\right\} + M^{**}(x_{11},\eta,\gamma)\frac{\partial \log\{f_{c^*}(x_{11},\eta)\}}{\partial x_{11}};$$

$$d_{21}(y_{j1}, x_{j1}, \eta, \gamma) = M^{**}(x_{11},\eta,\gamma).$$

Both $d_{12}$ and $d_{22}$ are identically 0 and $d_{13}$ and $d_{23}$ are not defined for the generalized Berkson error model.

## FIGURE 1

The results of the bootstrap study in Section 6 are displayed as kernel density estimates. The solid line represents the values obtained for the usual logistic regression estimate, while the dashed line represents the values for the corrected estimate.


1.8
1.5
1.2
0.9
0.6
0.3
0.0
Density
−2.0
−1.5
−1.0
−0.5
0.0
0.5
Estimated Slope