BIOMETRICS 57, 53–61
March 2001

Semiparametric Maximum Likelihood for Measurement Error Model Regression

Daniel W. Schafer

Department of Statistics, Oregon State University, 44 Kidder Hall, Corvallis, Oregon 97331-4606, U.S.A.
email: schafer@stat.orst.edu

SUMMARY. This paper presents an EM algorithm for semiparametric likelihood analysis of linear, generalized linear, and nonlinear regression models with measurement errors in explanatory variables. A structural model is used in which probability distributions are specified for (a) the response and (b) the measurement error. A distribution is also assumed for the true explanatory variable but is left unspecified and is estimated by nonparametric maximum likelihood. For various types of extra information about the measurement error distribution, the proposed algorithm makes use of available routines that would be appropriate for likelihood analysis of (a) and (b) if the true x were available. Simulations suggest that the semiparametric maximum likelihood estimator retains a high degree of efficiency relative to the structural maximum likelihood estimator based on correct distributional assumptions and can outperform maximum likelihood based on an incorrect distributional assumption. The approach is illustrated on three examples with a variety of structures and types of extra information about the measurement error distribution.

KEY WORDS: ECM algorithm; EM algorithm; Errors in variables; Functional model; Generalized linear models; Linear regression; Nonlinear regression; Nonparametric maximum likelihood; Semiparametric mixture model; Structural model.

1. Introduction

The problem of estimating the parameters in the regression of a response variable y on an explanatory variable x from observations on (y, w), where w is a measurement of x, is a special case of what has historically been called the errors-in-variables problem. In 1948, Neyman and Scott used this as one of several examples to demonstrate the inconsistency of maximum likelihood estimators of structural parameters in the presence of infinitely many nuisance parameters, the nuisance parameters in this case being the unknown x's. To bypass this obstacle, many techniques for measurement error model regression have been based on the structural model in which x's are assumed to be random variables with some specified probability distribution. As a reaction to the Neyman and Scott paper, however, Kiefer and Wolfowitz (1956) proposed a different solution, in which the nuisance parameters are assumed to be random variables but from an unspecified distribution. They proved consistency of the resulting maximum likelihood estimator of the semiparametric model but did not provide a computational approach.

The idea laid dormant until the computational clarification for semiparametric mixture models by Laird (1978), who used the term "nonparametric maximum likelihood" for estimation of the unspecified mixture distribution (by a step function with an estimated number of steps and estimated step heights). There has recently been keen attention to this approach for a variety of statistical problems (see Lindsay and Lesperance (1995) for a review) but surprisingly little attention to Kiefer and Wolfowitz's (1956) suggestion for measurement error model regression. Several authors have commented on its potential (e.g., Clayton, 1991; Lindsay and Lesperance, 1995), and Roeder, Carroll, and Lindsay (1996) proposed its application to the particular problem of case–control studies with imprecise exposures, with the suggestion of more general applicability.

This paper elaborates on such a generalization and provides a convenient computational form for data analysis. The work was motivated partly by the success and computational ease of semiparametric maximum likelihood for the related problem of generalized linear models with mixed effects (see Aitkin (1999) and references therein). The traditional reliance on a normally distributed random effect for likelihood analysis of that problem parallels the traditional reliance on normally distributed explanatory variables for likelihood analysis of the errors-in-variables problem. The semiparametric approach circumvents the need for any distributional assumptions about the respective nuisance parameters.

Parametric analysis in measurement error model regression refers to techniques based on the likelihood for the structural model, in which full distributional assumptions are made for (a) the response, (b) the measurement error, and (c) the true, unknown values of the imprecisely measured explanatory variables. Parametric estimators can be substantially more efficient than those based on weaker assumptions, as has been shown by Küchenhoff and Carroll (1997), Zhao and Lee (1996), Higdon and Schafer (1999), Carroll, Freedman, and

Pee (1997), and Carroll, Roeder, and Wasserman (1999); but such estimators are often avoided because of concern about the practical specification and possible misspecification of a distribution for (c) and the unavailability of general programs. These concerns have been addressed somewhat by the advancement of techniques that use rich parametric models (like mixtures of normals) for (c) and that make use of standard routines for likelihood analysis in the absence of measurement error (Carroll et al., 1999; Schafer, unpublished manuscript). The semiparametric approach offers an alternative that is practically attractive since it avoids the need to consider, explore, and fit that distribution altogether. Simulations in a few settings suggest that the semiparametric maximum likelihood estimator is very similar to the maximum likelihood estimator based on rich distributional assumptions.

A general computational approach is based here on the ECM algorithm (Meng and Rubin, 1993) and makes use of existing routines for maximum likelihood estimation of (a) and (b) that would be appropriate if x were available. A key feature is that data analytic steps concerning the regression of interest may be conducted in much the same way as they are in the absence of measurement errors.

Section 2 develops the method for a single explanatory variable measured with error. A recipe is first given for the case that the parameters in the measurement error distribution are known and then is extended to include various other types of extra information. Section 3 extends the solution to include additional covariates. Section 4 demonstrates the approach on an example with two explanatory variables measured with error.

2. Single x and No Additional Covariates

2.1 *Known Measurement Error Distribution*

Let y_i, x_i, and w_i represent the response, the true explanatory variable, and its measurement, respectively, for the ith observation in a sample of size n. For now, suppose that x_i is one dimensional and there are no additional explanatory variables. Let $f(y \mid x; \theta_1)$ represent the response distribution, the mean of which is the regression of interest, and $f(w \mid x; \theta_2)$ the measurement error distribution. It will be supposed that $f(y \mid x, w; \theta_1)$ does not depend on w and that observations indexed by different i's are independent of one another. It will also be supposed that x is a random variable with density function $f(x)$, which is unspecified.

Although the parameter $\theta = (\theta_1, \theta_2)$ may be identifiable for many models, it is unrealistic to expect a useful analysis without extra information (see Fuller, 1987; Carroll, Rupert, and Stefanski, 1995). The approach is developed in this section with the parameter of the measurement error distribution, θ_2, taken to be known, and is modified in Section 2.2 to include various other types of extra information (such as replicate measurements on a subset of cases).

The likelihood function is the product of the joint densities for y_i and w_i, which can be expressed in terms of the regression model of interest as

$$f(y_i, w_i; \theta) = \int f(y_i \mid x; \theta_1) f(w_i \mid x; \theta_2) f(x) dx. \qquad (2.1)$$

Notice that this is a mixture model with mixing distribution $f(x)$, so Laird's (1978) technique for simultaneous maximum likelihood estimation of θ and nonparametric maximum likelihood estimation of $f(x)$ can be applied.

To help visualize the solution, it is useful to first consider full likelihood analysis when $f(x)$ is specified and when (2.1) is approximated by K-node quadrature, i.e.,

$$f(y_i, w_i; \theta) \approx \sum_{k=1}^{K} \pi_k f\left(y \mid x_k^*; \theta_1\right) f\left(w_i \mid x_k^*; \theta_2\right), \qquad (2.2)$$

where π_k is $\alpha_k f(x_k^*)$ and the α_k's and x_k^*'s are known quadrature masses and nodes. Laird's approach is equivalent to finding maximum likelihood estimators by treating (2.2) as the actual density with $\pi = (\pi_1, \ldots, \pi_K)$ and $x^* = (x_1^*, \ldots, x_K^*)$ as additional parameters to be estimated. Semiparametric maximum likelihood can therefore be thought of as maximum likelihood using a quadrature approximation to the integral in (2.1) but with quadrature nodes x_k^* and scaled masses π_k estimated from the data. The number of support points K is also unknown but may be chosen as one less than the smallest integer for which there are redundant x^*'s or nonpositive π_k's. Typically, K is substantially less than n.

Although faster routines for nonparametric maximum likelihood estimation (e.g., Lesperance and Kabfleisch, 1992) can presumably be adapted to this problem, the ECM algorithm (Meng and Rubin, 1993) offers simplicity and transparency. For the types of problems considered here, the convergence is often rapid or at least is not so slow as to prohibit its use for full data analysis, including the construction of profile likelihoods.

The well-known EM algorithm (Dempster, Laird, and Rubin, 1977) for the finite mixture model (2.2) for a given K (e.g., Redner and Walker, 1984; Liu and Sun, 1997) is based on the introduction of a fictitious multinomial random variable for each i to indicate the appropriate mixture component for that case and the treatment of this introduced variable as missing data. The algorithm requires the updating of estimates of $\psi = (\theta, \pi, x^*)$ as those values that maximize the expected value of the complete data log likelihood given the observed data and with parameters in the expectation replaced by their current estimates, i.e., at iteration $t+1$, $\psi^{(t+1)}$ is found by maximizing

$$Q\left(\psi \mid \psi^{(t)}\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} p_{ik}^{(t)} \log\left\{\pi_k f\left(y_i \mid x_k^*; \theta_1\right) f\left(w_i \mid x_k^*; \theta_2\right)\right\}, \qquad (2.3)$$

where

$$p_{ik}^{(t)} = \pi_k^{(t)} g_{ik}^{(t)} \Big/ \sum_k \pi_k^{(t)} g_{ik}^{(t)}$$

and

$$g_{ik}^{(t)} = f\left(y_i \,\middle|\, x_k^{*(t)}; \theta_1^{(t)}\right) f\left(w_i \,\middle|\, x_k^{*(t)}; \theta_2^{(t)}\right). \qquad (2.4)$$

The term $p_{ik}^{(t)}$ represents the estimated probability that x_i comes from the kth mixture component, given y_i and w_i. The parameter θ_2 has a superscript (t) attached to it in (2.4) for generality in what follows, but its value is taken to be known in this section.

With the ECM Algorithm, the M-step in (2.3) is replaced by the following sequence of constrained maximization (CM) steps:

CM step 1: Find $\theta_1^{(t+1)}$ to maximize

$$Q_1\left(\theta_1 \,\middle|\, \psi^{(t)}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}^{(t)} \log f\left(y_i \,\middle|\, x_k^{*(t)}; \theta_1\right).$$

CM step 2: For $k = 1, \ldots, K$, find $x_k^{*(t+1)}$ to maximize

$$Q_x\left(x_k^* \,\middle|\, \psi^{(t)}\right) = \sum_{i=1}^{n} p_{ik}^{(t)} \left\{\log f\left(y_i \,\middle|\, x_k^*; \theta_1^{(t)}\right) + \log f\left(w_i \,\middle|\, x_k^*; \theta_2^{(t)}\right)\right\}.$$

CM step 3: Compute $\pi_k^{(t+1)}$ as $\Sigma_{i=1}^{n} p_{ik}^{(t)}/n$.

The major benefit is that CM step 1 involves the maximization of a log-likelihood function of the form that would be appropriate for estimating θ_1 if x were available but is applied to an augmented data matrix with nK rows, where, for $i = 1, \ldots, n$ and $k = 1, \ldots, K$, row $n(k-1)+i$ is (y_i, x_k^*) with associated weight $p_{ik}^{(t)}$. Therefore, existing routines for linear, generalized linear, or nonlinear regression can be used for this step. CM step 2 requires Newton–Raphson updating of the x_k^*'s. A reviewer has pointed out that the mixture gradient (Lindsay and Roeder, 1992) is a computationally inexpensive way to check that one has not found a local maximum.

2.2 *Extra Information*

This section details the additional CM step for updating θ_2 when it is unknown and when certain types of extra information are available. CM steps 1–3 remain essentially unchanged except for minor notational changes to incorporate various data structures, as described below. The reader may wish to skip these technical details and proceed to Section 2.3.

2.2.1 *Internal replication.* Suppose that, instead of w_i, we observe $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{ir_i})'$, meaning there are r_i measurements of x_i, with r_i being greater than one for at least some i. Then CM steps 1–3 are unchanged except that w_i must be replaced by $\boldsymbol{w}_i$ in CM step 2 and in (2.4). The CM step for updating θ_2 is as follows:

CM step 4: Find $\theta_2^{(t+1)}$ to maximize

$$Q_2\left(\theta_2 \,\middle|\, \psi^{(t)}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}^{(t)} \log f\left(\boldsymbol{w}_i \,\middle|\, x_k^{*(t)}; \theta_2\right).$$

If the measurements are correlated (as in Wang, Carroll, and Liang, 1996) and have a multivariate normal distribution with known $r_i \times r_i$ correlation matrix Λ_i and unknown variance θ_2, then CM step 4 leads to the following:

CM step 4a:

$$\theta_2^{(t+1)} = \sum_{i=1}^{n}\sum_{k=1}^{K} \frac{p_{ik}^{(t)}}{n} \left\{\left(\boldsymbol{w}_i - x_k^{*(t)} e_i\right)' \Lambda_i^{-1} \left(\boldsymbol{w}_i - x_k^{*(t)} e_i\right)\right\},$$

where e_i is an r_i-vector of ones.

If the replicate measurements are independent and normally distributed, then the expression in Section 2.2.1 applies, with Λ_i an $r_i \times r_i$ identity matrix. For other distributional choices, if the replicate measurements are independent given x, then CM step 4 becomes the following:

CM Step 4b: Find $\theta_2^{(t+1)}$ to maximize

$$Q_2\left(\theta_2 \,\middle|\, \psi^{(t)}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}^{(t)} \sum_{j=1}^{r_i} \log f\left(w_{ij} \,\middle|\, x_k^{*(t)}; \theta_2\right).$$

This has the form of a weighted log likelihood of the form appropriate for estimating θ_2 if x were known and can therefore make use of existing routines applied to augmented data. Section 3.2 provides an example of this type, in which the measurement error distribution is taken to be a mixture of two normals.

2.2.2 *External estimate of measurement error variance based on ν degrees of freedom.* Suppose that $f(w \mid x; \theta_2)$ is a normal density with mean x and variance θ_2 and that $\tilde{\theta}_2$ is an estimate, independent of the primary data set, such that $\nu\tilde{\theta}_2/\theta_2$ has a chi-squared distribution on ν d.f. Then CM step 4 is the following formula for $\theta_2^{(t+1)}$:

$$\theta_2^{(t+1)} = \left\{\sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}^{(t)} \left(w_i - x_k^{*(t)}\right)^2 + \nu\tilde{\theta}_2\right\}/(n+\nu).$$

Such an external estimate is sometimes available, but this form may also be useful for incorporating a subjective guess of θ_2 and including uncertainty about the guess by treating it as an estimate based on some small degrees of freedom.

2.2.3 *Internal validation.* Suppose that exact values of x are available on a subset of observations so that (y_i, x_i, w_i) are observed for $i = 1, \ldots, n_1$ while only (y_i, w_i) are observed for $i = n_1 + 1, \ldots, n_1 + n_2$. Then, as shown by Roeder et al. (1996), the density for the first n_1 observations can be written as a mixture by using the indicator function $\int f(y_i \mid x; \theta_1) f(w_i \mid x; \theta_2) I(x = x_i) f(x) dx$, where $I(x = x_i)$ is one if $x = x_i$ and zero otherwise. Then (2.3) becomes

$$Q\left(\psi \,\middle|\, \psi^{(t)}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}^{(t)} \log\left[\pi_k f\left(y \mid x_k^*; \theta_1\right) f\left(w_i \mid x_k^*; \theta_2\right) \times \left\{I_{(x_k^* = x_i)}\right\}^{q_i}\right],$$

where q_i is one for those cases with exact values of x and zero otherwise. Therefore, all of the available x_i's (for $i = 1, \ldots, n_1$) must be included in the set of support points, $x^* = (x_1, \ldots, x_{n_1}, x_{n_1+1}^*, \ldots, x_{n_1+K}^*)$ and $\pi = (\pi_1, \ldots, \pi_{n_1+K})$, with the last K of the x_k's and all of the π_k's unknown. If the number of observations with exact values of x is large, then the computations may become unwieldy. Roeder et al. (1996) suggested using a fixed grid of support points in that case and treating them as known.

2.3 *Tests and Confidence Intervals*

If (2.2) is treated as the true density with K known, the log likelihood is the sum of the logs of (2.2), which are calculated at each iteration as the denominator of $p_{ik}^{(t)}$ in (2.4), so the maximized log likelihood is available as a by-product of the calculations for parameter estimates. This is true for all the forms of extra information presented above except when there is an external estimate of measurement error variance, in which case the external log likelihood must be added. Hypothesis testing can be carried out with likelihood ratio tests in the usual way. There is no theoretical justification for this when K is unknown, but it has been shown to work well in other situations (Lindsay and Lesperance, 1995; Aitkin, 1999).

Following Aitkin (1999), approximate standard errors are obtained without additional programming by dividing the absolute value of the parameter estimate by the square root of the likelihood ratio test statistic for the hypothesis that the parameter is zero (i.e., assuming the likelihood ratio statistic is approximately equal to the Wald statistic). These should be accurate, at least, for large sample sizes. In any case, more reliable confidence intervals can be obtained by inverting the likelihood ratio test.

2.4 *Example*

Schafer (unpublished manuscript) reported recession velocities and measured distances from earth for 211 galaxies. To estimate Hubble's constant and departures from a linear Hubble Flow (regression of velocity on distance), interest is in the normal quadratic regression of velocity on true distance, forced through the origin. The degree of imprecision in the measured distances (based on the Tully–Fisher method) is unknown, but a second measurement of distance (based on cepheid variables) with known measurement error variance is available on 10 of the galaxies.

In this problem, y is recession velocity, x is log of true distance from earth, w_{T} is the log of the Tully–Fisher distance measurement, and w_{C} is the log of the cepheid distance measurement. It is supposed here that the distribution of y given x is normal with a regression function that is quadratic in distance, $\beta_1 \exp(x_i) + \beta_2 \exp(2x_i)$, with constant variance. The measurement error distributions on the log scale, $f(w_{\mathrm{T}i} \mid x_i)$ and $f(w_{\mathrm{C}i} \mid x_i)$, are taken to be normal, the first with constant, unknown standard deviation σ_{T} and the second with known standard deviation $\sigma_{\mathrm{C}i}$ depending on i, as described in Schafer (unpublished manuscript).

The structure of extra information is nonstandard in this problem but can be incorporated in a straightforward way. Since 10 observations contain both kinds of measurements, the formula for the weights, $p_{ik}^{(t)} = \pi_k^{(t)} g_{ik}^{(t)} / \Sigma_k \pi_k^{(t)} g_{ik}^{(t)}$, must use

$$g_{ik}^{(t)} = f\left(y \,\middle|\, x_k^{*(t)}; \theta_1^{(t)}\right) f\left(w_{\mathrm{T}i} \,\middle|\, x_k^{*(t)}; \theta_2^{(t)}\right) \times f\left(w_{\mathrm{C}i} \,\middle|\, x_k^{*(t)}\right)^{v_i} f\left(x_k^{*(t)}; \theta_3^{(t)}\right),$$

where v_i is one for those observations with both measurements and is zero otherwise. Then CM steps 2 and 4 become the following:

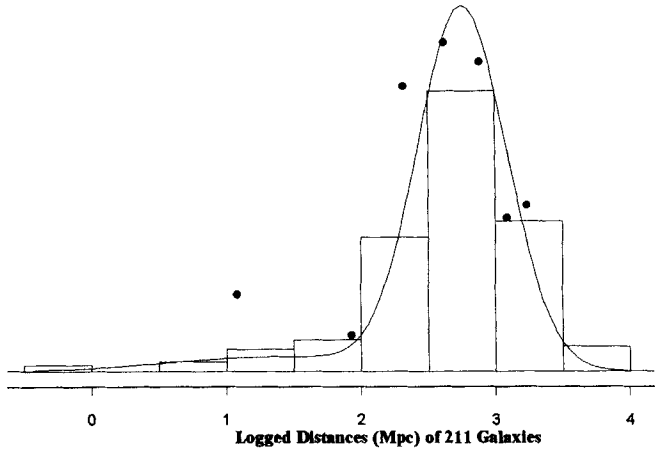

Figure 1. Histogram of measured distances for 211 galaxies, with fitted mixture of two normals and scaled masses from the nonparametric maximum likelihood estimates of the density of actual log distance.

CM step 2: Find $x_k^{*(t+1)}$ to maximize

$$Q_x\left(x_k^* \,\middle|\, \psi^{(t)}\right) = \sum_{i=1}^{n} p_{ik}^{(t)} \left\{ \log f\left(y_i \,\middle|\, x_k^*; \theta_1^{(t)}\right) + \log f\left(w_{\mathrm{T}i} \,\middle|\, x_k^*; \theta_2^{(t)}\right) + v_i \log f\left(w_{\mathrm{C}i} \mid x_k^*\right) \right\}.$$

CM step 4: Compute

$$\sigma_{\mathrm{T}}^{2(t+1)} = \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{ p_{ik}^{(t)} \left(w_{\mathrm{T}i} - x_k^{*(t)}\right)^2 \right\} / n.$$

The values of the maximized log likelihood associated with $K = 6$, 7, and 8 support points are -1729.40, -1724.64, and -1724.64. With $K = 8$, there are only seven distinct values of x^*, so an appropriate choice for K is seven. The resulting support points, $x^* = (1.1, 1.9, 2.3, 2.6, 2.9, 3.1, 3.2)$ with masses $\pi = (.06, .03, .21, .24, .23, .11, .12)$, are shown in Figure 1 along with a histogram of the "measured" distances w_{T} and, for reference, the fitted density for x when it is assumed to be a mixture of two normals.

The estimated regression coefficients for the linear and quadratic terms are 42.6 (with SE = 7.1) and 2.46 (SE = .41). In Figure 2, the fit is compared to (A) the naive least squares fit (101.3 × distance − 0.83 × distance^2), (B) the maximum likelihood fit based on a simplifying assumption of normality for log distance (65.9 × distance + .99 × distance^2), and (C) the maximum likelihood fit based on the mixture of two normals for x (47.3 × distance + 2.33 × distance^2). The standard errors for the latter are 8.4 and .44. It is interesting to note that the naive fit indicates significant downward curvature, which is consistent with an accelerating universe, while adjustment for measurement errors shows the opposite, a decelerating universe.

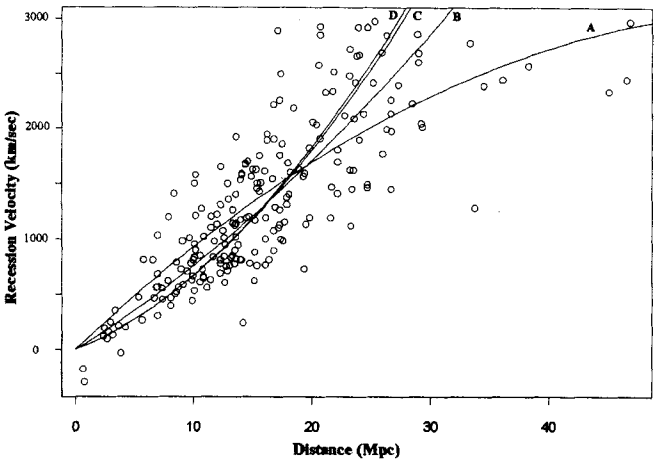

Figure 2. Velocity versus measured distance for 211 galaxies, with four fitted curves for quadratic regression through the origin. **A.** Naive. **B.** Maximum likelihood assuming normal distribution for log distance. **C.** Maximum likelihood assuming a mixture of normals for log distance. **D.** Semiparametric maximum likelihood.

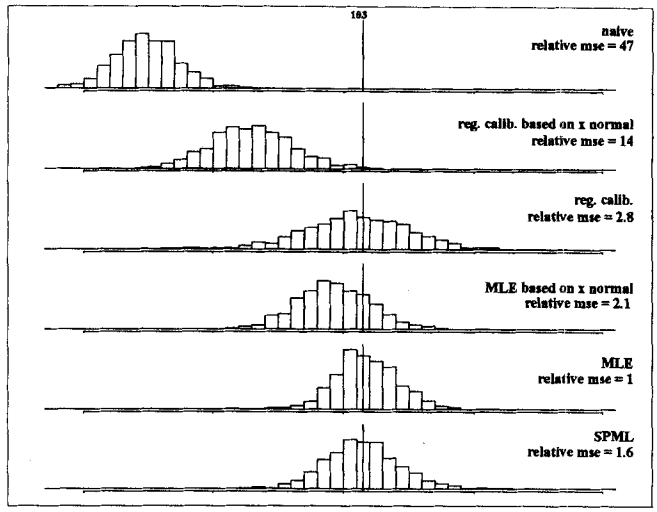

Figure 3. Monte Carlo distributions of six slope estimators for normal linear regression, with true slope 103. The distribution of x, the log of the explanatory variable, is a mixture of two normals; the distribution of the measurement w, given the true x, is normal.

2.5 *A Simulation Study*

Samples of size 211 were simulated according to the model of the previous section, with x (which portrays log distance) generated from the mixture of two normal distributions displayed in Figure 1. Specifically, with probability .08, x comes from a normal distribution with mean 1.34 and variance .62, and with probability .92, it comes from a normal distribution with mean 2.76 and variance .11. The response distribution was simulated as normal with mean $-300 + 103\exp(x)$ and variance 26,000 to match the best fitting straight line in the regression of observed velocity on distance from the actual data set. The first measurements, w_1, were taken to be normal with mean x and variance .06. For 10 of the 211 cases, a second measurement was generated from a normal distribution with mean x and with variance equal to the known variances of the 10 cepheid measurements in the data set (0.0056, 0.0075, 0.011, 0.0069, 0.0079, 0.010, 0.023, 0.0036, 0.025, 0.016).

For each simulated sample, six estimators of slope were computed: (1) the naive estimator that ignores measurement error, (2) a regression calibration estimator (in which the missing x is replaced by an estimate of $\mathrm{E}(x \mid w)$; see Carroll et al., 1995) based on an (incorrect) assumption of normality for x, (3) a regression calibration estimator based on a (correct) assumption of a mixture of two normals for x, (4) a maximum likelihood estimator based on the (incorrect) assumption of normality for x, (5) a maximum likelihood estimator based on a (correct) assumption of a mixture of two normals for x, and (6) the semiparametric maximum likelihood estimator (SPML).

The Monte Carlo mean square error (MSE) for the maximum likelihood estimator is 24.0. The relative MSEs of the other estimators are listed in Figure 3 along with histograms of the Monte Carlo distributions. Those based on the incorrect normality assumption tend to be biased. The semiparametric maximum likelihood estimator has a mean square error that is 60% larger than that of the maximum likelihood estimator but nevertheless performs better than the maximum likelihood estimator based on the misspecified model for $f(x)$ and better than the regression calibration estimator based on the correct distributional assumptions.

A second simulation duplicated conditions of the first except using functional model sampling. The observed log distances in the data example, scaled to have the same variance as x in the previous simulation, were used as x. These fixed values of x were used for all 1000 simulated samples; there was no repeated sampling of x's from some distribution. The MSE for the maximum likelihood estimator in this case is 44. The MSE for the SPML estimator is 60% of this. The Monte Carlo distributions are displayed in Figure 4.

2.6 *Simulations for a Binary Logistic Regression Model*

Figure 5 shows histograms of Monte Carlo distributions for six estimators in a binary logistic, structural regression model. Samples of size 500 were generated with x from a mixture of two normal distributions with mixing probabilities .75 and .25, means 50 and 80, and variances 100 and 300. The logit of the binomial probability was $-5 + .1x$, and the distribution of w given x was normal with mean x and variance 80. The reliability of the measurements (the squared correlation of the measurements and the true values of the explanatory variable) with these choices is about .8. For this simulation, the measurement error variance was taken to be known.

The SPML estimator has similar operating characteristics in this setting to the maximum likelihood estimator based on the correct distributional assumptions. The maximum likelihood estimator based on the incorrect normality assumption for x has a smaller mean squared error but with some appreciable bias.

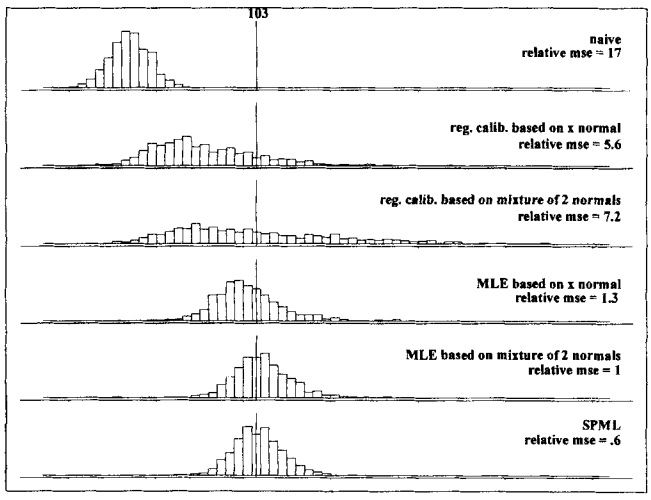

Figure 4. Monte Carlo distributions over 1000 samples of six slope estimators from a normal linear regression with slope 103 in a functional setting: x-values were taken to be the observed distances from Section 2.4, scaled to have variance (and other conditions) matching that of Figure 3.

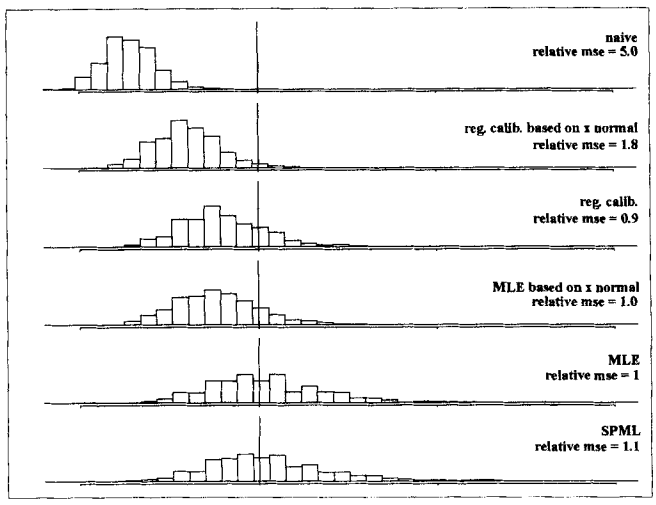

Figure 5. Monte Carlo distributions over 1000 samples of six slope estimators from a binary logistic regression with slope 0.1. The structural model is simulated with x from a mixture of two normals and with additive measurement errors; the reliability of the measurements is .8.

3. Inclusion of Additional Explanatory Variables, Free of Error

3.1 *Modification to Include the Regression of x on z*

Suppose interest is in the regression of y on x and z, where z is a vector of additional explanatory variables, which are assumed to be free of measurement error. Assuming that the measurement error distribution does not depend on z, the conditional density of y_i and w_i given z_i may be written as

$$f(y_i, w_i \mid z_i; \theta) = \int f(y_i \mid x, z_i; \theta_1) f(w_i \mid x; \theta_2) f(x \mid z_i) dx.$$

Some choice must be made for incorporating the dependence of $f(x \mid z_i)$ on z_i without making a complete parametric specification. One simple choice is to take $x_i = \gamma' z_i + e_i$, where γ is a vector of regression coefficients and with the e_i's independent and identically distributed according to some unspecified density $f(e)$. Then (2.2) becomes

$$f(y_i, w_i; \theta) = \sum_{k=1}^{K} \pi_k f\left(y_i \mid x_{ik}^*, z_i; \theta_1\right) f\left(w_i \mid x_{ik}^*; \theta_2\right)$$

with $x_{ik}^* = \gamma' z_i + e_k$, where the e_k's are the support points to be estimated. The only modification to the program of Section 2 then is that CM step 2 has two parts as follows:

CM step 2a: For $k = 1, \ldots, K$, find $e_k^{(t+1)}$ to maximize

$$Q_e\left(e_k \,\middle|\, \psi^{(t)}\right) = \sum_{i=1}^{n} p_{ik}^{(t)} \left\{ \log f\left(y_i \,\middle|\, x = \gamma^{(t)'} z_i + e_k, z_i; \theta_1^{(t)}\right) + \log f\left(w_i \,\middle|\, x = \gamma^{(t)'} z_i + e_k; \theta_2^{(t)}\right) \right\}.$$

CM step 2b: Find $\gamma^{(t+1)}$ to maximize

$$Q_\gamma\left(\gamma \,\middle|\, \psi^{(t)}\right) = \sum_{i=1}^{n} p_{ik}^{(t)} \left\{ \log f\left(y_i \,\middle|\, x = \gamma' z_i + e_k^{(t)}, z_i; \theta_1^{(t)}\right) + \log f\left(w_i \,\middle|\, x = \gamma' z_i + e_k^{(t)}; \theta_2^{(t)}\right) \right\}.$$

Since the derivatives of x_{ik}^* with respect to e_k and γ are straightforward, the modification to the algorithm for using CM steps 2a and 2b is minor. The approaches for incorporating extra information and for carrying out tests and computing standard errors are the same as in Section 2. One choice for starting values is equally spaced e_k's between $m - 2s$ and $m + 2s$, where m is the intercept in the fitted regression of w on z and s is an estimate of SD $(x \mid w, z)$ based on simplifying normality assumption for $x \mid z$ and $w \mid x$.

3.2 *Logistic Regression Example*

Clayton (1991) reported data from a study in which the ratio of polyunsaturated to saturated fat intake was measured for 336 males by a 1-week, full-weighed dietary survey. The survey was repeated for a subset of 76 subjects approximately 6 months after their initial measurement. By the end of the study, 46 of the 336 men had died of ischaemic heart disease. Let y represent the binary response for heart disease mortality, x the log of the true (long-term mean) value of the polyunsaturated to saturated fat ratio, w_1 the log of the first measurement, w_2 the log of the second measurement if available, and z the person's age in years at entry into the study and suppose the observations are ordered so that the first 76 are those with two measurements. Schafer (unpublished manuscript) fit the logistic regression of y on x and z with $\text{logit}(\pi) = \beta_0 + \beta_1 e^x + \beta_2 e^{2x} + \beta_3 z$, assuming $x \mid z$ to be normal and $w \mid x$ to be a mixture of two normals. The

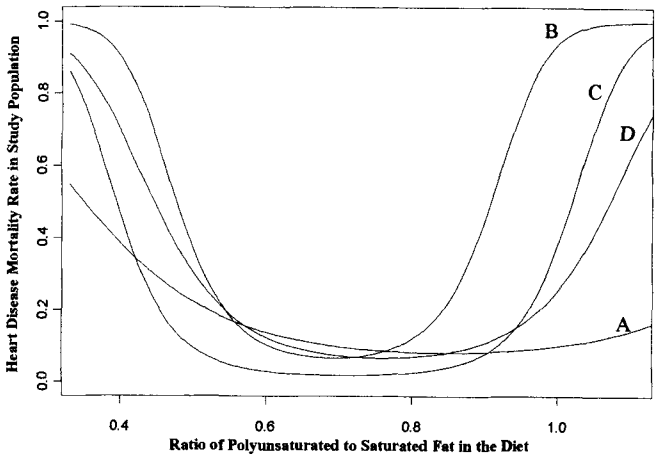

Figure 6. Fitted models for heart disease mortality as a function of the ratio of polyunsaturated to saturated fat in the diet for a 50-year-old man in the study population. **A.** Naive. **B.** Maximum likelihood assuming $f(x \mid z)$ and $f(w \mid x)$ normal. **C.** Maximum likelihood assuming $f(x \mid z)$ normal and $f(w \mid x)$ a mixture of two normals. **D.** semiparametric maximum likelihood assuming $f(w \mid x)$ a mixture of two normals.

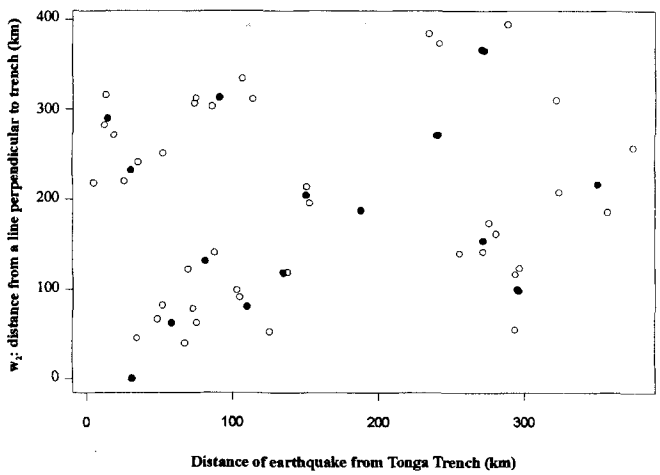

Figure 7. Scatterplot of the measurements w_2 versus w_1 from the earthquake data with 15 estimated support points indicated by solid circles.

same model is used here but with the distribution of $x \mid z$ unspecified.

CM step 4b in Section 2.2.2 for internal replication is used for updating the parameters of the measurement error distribution. A program for weighted maximum likelihood estimation of a mixture of two normals with known means can be used in this step. Since nine is the smallest K for which there are redundant support points, a value of $K = 8$ is appropriate. The resulting support points are (3.89, 3.96, 4.05, 4.15, 4.25, 4.37, 4.46, 4.53) with masses (.04, .03, .18, .18, .28, .20, .05, .04). The resulting fit for the logit, labeled D in Figure 6, is $22.6 - .79e^{x} + .0057e^{2x} + .043z$, with standard errors of coefficients 7.5, .24, .0018, and .012, respectively. For comparison, the maximum likelihood fit assuming normality for $x \mid z$ (labeled C in Figure 6) is $16.3 - .58e^{x} + .0041e^{2x} + .039z$, with similar standard errors. Also shown in Figure 6 are A, the naive fit that ignores measurement error, and B, the fit using maximum likelihood when it is assumed that the measurement error distribution is normal.

4. Regression with Several Imprecisely Measured x's

4.1 *Modifications*

Suppose now that x has p components. Then the CM steps in Section 2.1 are still appropriate except that each of the K support points x_k^* is p dimensional. If there are additional explanatory variables, z, then the modifications in Section 3.1 are also appropriate, but the regression model for x on z is a multivariate one. The recipes for including extra information, in Section 2.2, are for a single x. Since there are many combinations of possible structures for the multivariate distribution of the multiple measurement errors and of the types of extra information about each, no attempt will be made to list them here. For some data problems, this will no doubt be a dreadful mess. For some, however, the semiparametric maximum likelihood computations are straightforward or even, as in the case of the following example, surprisingly easy. In these cases, the proposed approach offers an important alternative to techniques based on an assumption of multivariate normality for x.

4.2 *Multiple Regression Example*

Fuller (1987, Section 3.1) listed measurements of y, the depth of each of 43 earthquakes near the Tonga trench; x_1, the perpendicular distance of the earthquake to the trench; and x_2, the distance of the earthquake from an arbitrary line perpendicular to the Tonga trench. It is thought that the regression of y on x_1 and x_2 is of the form $\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$ and that the available observations of the x's—w_1 and w_2—contain measurement errors with variances roughly 100 km^2. It is reasonable to assume that the measurement errors in the two directions are independent of one another. There is no available information for judging the shape of the measurement error distribution in this data set, but normality seems reasonable and sensitivity to this choice could be explored by trying others.

To determine the number of support points, it is useful to first consider values of K that are squares of integers. Initial values for the support points can be the grid coordinates on the square with corners $(-1, -1)$ and $(1, 1)$, rotated by a square root of an estimate of the variance–covariance matrix of x given w and shifted by an estimate of the mean of x given w (with these conditional moments based on simplifying normality assumptions). After finding the smallest integer for which there are redundant x's, a final K can be found by deleting the redundancies and lumping their masses.

In the earthquake example, K was determined to be 15, with support points (31, 1), (82, 132), (31, 233), (14, 290), (59, 63), (110, 81), (91, 313), (135, 119), (18, 188), (151, 205), (241, 271), (296, 99), (272, 154), (351, 217), (273, 365) and masses .02, .05, .09, .07, .14, .07, .12, .02, .02, .05, .02, .07, .09, .07, .09. The support points are shown in Figure 7 on a plot of w_2 versus w_1.

It is convenient to partition CM step 2 into two separate CM steps, one for each dimension of x. In the first, the first coordinates of the support points, i.e., x^*_{1k}, are updated by maximizing the expected complete data log likelihood with x^*_{2k} and other parameters held fixed at their current estimates. In the second, x^*_{2k} is updated with x^*_{1k} held fixed. In this way, the updating of the support points can be accomplished with a series of steps that are identical to the one used for a single x.

The SPML fit using 15 support points is $-20.4 + .51x_1 + .00124x_1^2 + .071x_2$, with standard errors of coefficients 13.2, .19, .00051, and .041. These results are very nearly the same as those of naive least squares that ignore measurement error. This isn't surprising since the measurement error variances are quite small relative to the total variances of the measurements (about .01 for each). The example is nevertheless useful for demonstration because the bivariate normal distribution is very obviously a poor choice for structural modeling of the x's. Further, the locations of the estimated support points in Figure 7, for this example in which the w's and x's are nearly the same, provide a useful image of how the SPML support points map around clusters in the x-space.

5. Conclusions

In the structural measurement error regression model, the joint distribution of the response variable y and the measurement w has the form of a mixture model, as shown in (2.1), where the mixing distribution is the density of the true, unknown explanatory variable x. If the distribution of x, $f(x)$, is modeled without parametric assumptions, it is a semiparametric mixture model for (y, w). The proposed solution is for the joint maximum likelihood estimation of θ in $f(y, w \mid x; \theta)$ and nonparametric maximum likelihood estimation of $f(x)$.

The obvious benefit for measurement error model regression is the absence of a need to specify a parametric form for the distribution of the true explanatory variables. This avoids the difficult task of modeling an unobservable random variable and, more importantly, alleviates concern over the consequences of misspecification of that part of the model. Simulations indicate that maximum likelihood inference can be misleading if a parametric form for $f(x)$ is misspecified, and further that the semiparametric maximum likelihood estimators retain a high degree of efficiency relative to maximum likelihood based on the correct parametric form for $f(x)$. This is consistent with what has been found in other applications that use the semiparametric mixture model (Lindsay and Lesperance, 1995; Aitkin, 1999).

The only theoretical result associated with the semiparametric maximum likelihood inference, so far, is consistency, as shown by Kiefer and Wolfowitz (1956). For now, the collective evidence from a variety of types of applications indicates that the method seems to provide good estimators and that inferences based on treating the observed likelihood (for the determined value of K) as an actual likelihood seem to be accurate. There would be considerable interest, naturally, in the development of asymptotic results regarding efficiency and likelihood ratio tests.

Using a relative convergence criterion of .001 for all parameter estimates, the ECM algorithm converged in 84, 237, and 39 iterations for the examples in Sections 2.4, 3.2, and 4.2, respectively. Although improvements in computational efficiency are undoubtedly possible, the approach here is particularly attractive in its use of existing routines for estimating the regression parameters. This means that, whatever the model for $f(y \mid x; \theta_1)$, if a routine is available for its maximum likelihood estimation in the absence of measurement errors and if the routine can accommodate weights, then that routine may be used in CM step 1. This includes the relatively easy analysis for linear and generalized linear models with polynomial and interaction terms and nonlinear regression models using existing nonlinear routines. In all cases, likelihood ratio testing may be performed via the fitting of hierarchical models.

CM step 2 requires the Newton–Raphson updating of the support points at each iteration of the procedure. Although this is easily accomplished, it requires different calculations for different specifications of the model for $f(y \mid x; \theta_1)$ and thus a retooling of the program associated with model changes. This may be avoidable with direct optimization and clever programming.

In Section 2.5, the behavior of the semiparametric maximum likelihood estimator was investigated with simulations in a functional model (fixed x) setting. The distinction between the structural and functional models is subtle and will not be discussed here. One point, though, is that the semiparametric maximum likelihood approach can accommodate rough structural models since it does not rely on smooth choices for $f(x)$. This is evident in Figure 7, where the support points map the clusters in the x-space. Another point is that the step of estimating $f(x)$ nonparametrically in the structural model is the same as estimating the nuisance parameters $x_1, x_2, \ldots, x_n$ in the functional model by an empirical Bayes estimate.

Acknowledgement

The author wishes to thank Ari Verbyla for suggesting this research problem.

Résumé

Ce papier présente un algorithme EM pour l'analyse semiparamétrique de la vraisemblance des modèles linéaire, linéaire généralisé et de régression non-linéaire avec erreur de mesure pour les variables explicatives. On utilise le modèle structurel, dans lequel on spécifie des distributions de probabilité pour (a) la réponse et (b) l'erreur de mesure. On suppose aussi une distribution de probabilité non spécifiée pour la vraie variable explicative, et qui est alors estimée par un maximum de vraisemblance non-paramétrique. L'algorithme proposé utilise des routines disponibles qui seraient appropriées pour l'analyse de la vraisemblance de (a) et (b) si le vrai x était connu, et dans différents cas d'information additionnelle sur la distribution de l'erreur de mesure. Nos simulations suggèrent que l'estimateur semiparamétrique du maximum de vraisemblance conserve un haut degré d'efficacité par rapport à l'estimateur structurel du maximum de vraisemblance établi à partir des bonnes distributions, et peut avoir de meilleures performances que l'estimateur du maximum de vraisemblance basé sur une hypothèse distributionnelle incorrecte. Cette approche est illustrée par trois exemples avec diverses structures et types d'information additionnelle concernant la distribution de l'erreur de mesure.

References

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55,** 117–128.

Carroll, R. J., Rupert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models.* New York: Chapman & Hall.

Carroll, R. J., Freedman, L. S., and Pee, D. (1997). Design aspects of calibration studies in nutrition, with analysis of missing data in linear measurement error models. *Biometrics* **53,** 1440–1451.

Carroll, R. J., Roeder, K., and Wasserman, L. (1999). Flexible parametric measurement error models. *Biometrics* **53,** 44–54.

Clayton, D. G. (1991). Models for the analysis of cohort and case–control studies with inaccurately measured exposures. In *Statistical Models for Longitudinal Studies of Health,* J. H. Dwyer, M. Feinleib, P. Lipsert, et al. (eds), 301–331. New York: Oxford University Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

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

Higdon, R. and Schafer, D. W. (1999). Maximum likelihood computations for regression with measurement error. *Statistical Computing and Data Analysis,* in press.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Annals of Mathematical Statistics* **27,** 886–906.

Küchenhoff, H. and Carroll, R. J. (1997). Biases in segmented regression with errors in predictors. *Statistics in Medicine* **16,** 169–188.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73,** 805–811.

Lesperance, M. L. and Kabfleisch, J. D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *Journal of the American Statistical Association* **97,** 120–126.

Lindsay, B. G. and Lesperance, M. L. (1995). A review of semiparametric mixture models. *Journal of Statistical Planning and Inference* **47,** 29–39.

Lindsay, B. G. and Roeder, K. (1992). Residual diagnostics for mixture models. *Journal of the American Statistical Association* **87,** 785–794.

Liu, C. and Sun, D. X. (1997). Acceleration of the EM algorithm for mixture models using ECME. *American Statistical Association Proceedings of the Statistical Computing Section* 109–114.

Meng, X. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80,** 267–278.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16,** 1–32.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood, and the EM Algorithm. *SIAM Review* **26,** 195–202.

Roeder, K., Carroll, R. J., and Lindsay, B. G. (1996). A semiparametric mixture approach to case–control studies with errors in covariables. *Journal of the American Statistical Association* **91,** 722–732.

Wang, N., Carroll, R. J., and Liang, K. Y. (1996). Quasi-likelihood and variance functions in measurement error models with replicates. *Biometrics* **52,** 401–411.

Zhao, Y. and Lee, A. H. (1996). A simulation study of estimators for generalized linear measurement error models. *Journal of Statistical Computation and Simulation* **54,** 55–74.

Received January 2000. *Revised May* 2000.
Accepted June 2000.