Iturria, Stephen J.; Carroll, Raymond J.; Firth, David

Working Paper

Polynomial regression and estimation function in the presence of multiplication measurement error, with application to nutrition

SFB 373 Discussion Paper, No. 1997,10

Provided in Cooperation with:
Collaborative Research Center 373: Quantification and Simulation of Economic Processes, Humboldt University Berlin

Suggested Citation: Iturria, Stephen J.; Carroll, Raymond J.; Firth, David (1997) : Polynomial regression and estimation function in the presence of multiplication measurement error, with application to nutrition, SFB 373 Discussion Paper, No. 1997,10, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin, https://nbn-resolving.de/urn:nbn:de:kobv:11-10063707

This Version is available at:
https://hdl.handle.net/10419/66311

POLYNOMIAL REGRESSION AND ESTIMATING FUNCTIONS IN THE PRESENCE OF MULTIPLICATIVE MEASUREMENT ERROR, WITH APPLICATIONS TO NUTRITION

Stephen J. Iturria, Raymond J. Carroll and David Firth *

January 14, 1997

Abstract

In this paper we consider the polynomial regression model in the presence of multiplicative measurement error in the predictor. Consistent parameter estimates and their associated standard errors are derived. Two general methods are considered, with the methods differing in their assumptions about the distributions of the predictor and the measurement errors. Data from a nutrition study are analyzed using the methods. Finally, the results from a simulation study are presented and the performances of the methods compared.

Key Words and Phrases: Asymptotic theory; Bootstrap; Errors-in-Variables; Estimating Equations; Measurement Error; Nonlinear Regression; Nutrition.

Short title: Multiplicative Measurement Error

*Stephen Iturria is a graduate student and Raymond J. Carroll is Professor of Statistics, Nutrition and Toxicology, Department of Statistics, Texas A&M University, College Station, TX 77843–3143. David Firth is Senior Fellow in Statistics for the Social Sciences, Nuffield College, Oxford OX1 1NF. The authors wish to thank Suojin Wang for his generous and helpful comments during the preparation of this article. Iturria and Carroll's research was supported by a grant from the National Cancer Institute (CA–57030). Carroll's research was partially completed while visiting the Institut für Statistik und Ökonometrie, Sonderforschungsbereich 373, Humboldt Universität zu Berlin, with partial support from a senior Alexander von Humboldt Foundation research award.

1 INTRODUCTION

Much work has been done in the estimation of regression coefficients in the presence of additive measurement error in the predictors. A detailed account of the developments for linear regression models can be found in Fuller (1987). Carroll, et al. (1995) summarize much of the recent work for nonlinear regression models. Considerably less work has been done for cases of nonadditive measurement error however. Hwang (1986) derives a consistent estimator for the coefficients of the ordinary linear model under multiplicative measurement error by modifying the usual normal equations of least squares regression. To apply this method, one requires consistent estimates of the moments of the measurement errors. One of the general methods we will consider is a special case of Hwang's estimator. For this method we do not require that any distributional assumptions be made about the unobserved predictor, other than the usual i.i.d. assumptions. We will consider two distributional forms for the measurement errors, and propose methods for estimating their moments. For the second general method we will consider, we model the distribution of the unobserved predictor as well. Fitting this method will require estimating the distribution of the predictor conditional on its mismeasured version. We will apply our methods to a nutrition data set taken from the Nurses Health Survey. We also present the results from a simulation study.

1.1 The Polynomial Regression Model

The polynomial regression model under multiplicative measurement is given by

$$
\begin{aligned}
&Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_i^k + \beta_{p+1}^t Z_i + \epsilon_i, \\
&W_{ij} = X_i U_{ij}, \\
&i = 1, \ldots, n, \quad j = 1, \ldots, r_i,
\end{aligned}
$$

where U_{ij} is the measurement error associated with the jth replicate of the error–prone predictor of X_i, namely W_{ij}, and Z_i is a vector of covariates assumed to be measured without error. Further assumptions are that all elements of (ϵ_i), (U_{ij}), and (X_i) are mutually independent, the (X_i) assume positive values only, the (ϵ_i) have mean zero, and the (U_{ij}) have either mean or median one. We will consider three possible models for the distribution of the (X_i, U_{ij}). No further distributional assumptions will be made about the (Z_i) and (ϵ_i).

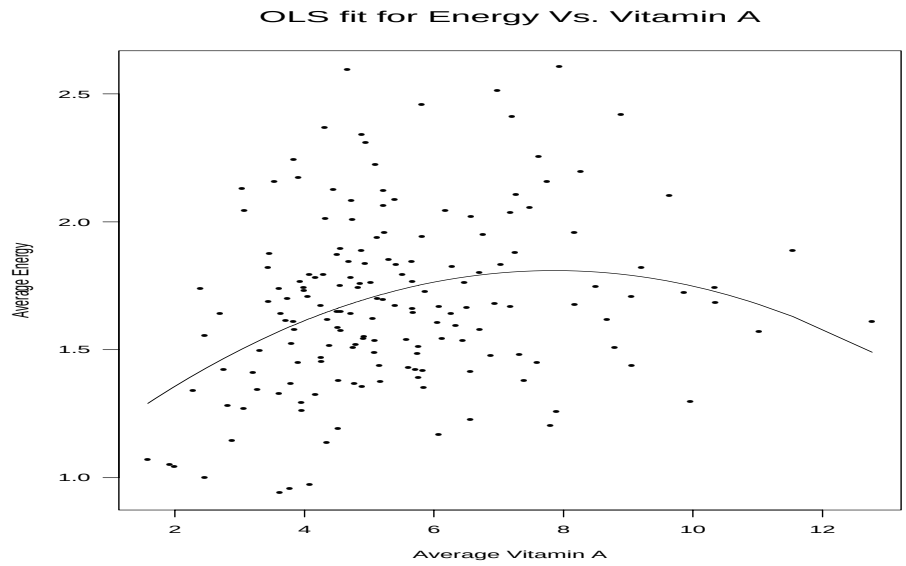

Figure 1: *Least squares quadratic fit for Nurses.*

1.2 Nurses Health Survey

The Nurses Health Survey includes measurements of energy intake and vitamin A intake for 168 individuals calculated from four 7–day food diaries. We will model Y = long–term energy intake as a quadratic function of X = long–term vitamin A intake plus error. No important effects were evident among the possible covariates so we will only consider the regression of Y on X. Food diaries are an imprecise method for calculating long–term nutrient intakes so the reported vitamin A intakes are presumed to be measured with error. Long–term energy intake is also estimated imprecisely when using food diaries, but for the purpose of illustrating our methods we will take such measurement errors to be additive, thus absorbing them into the (ϵ_i). A scatter plot of the averages of the energy replicates against the averages of the vitamin A replicates is given in Figure 1. The p–value for the quadratic term in the ordinary least squares (OLS) fit of the energy replicate averages as a quadratic function of the vitamin A replicate averages is .002.

1.3 Effects of Multiplicative Measurement Error on Curvature

One question to consider is whether the curvature exhibited in the OLS fit of the Nurses data accurately reflects the curvature in the underlying relationship between Y and the unobservable X. To see the effect that measurement error can have on curvature, consider the plots given in Figure 2. The top two plots are of Y vs. X and Y vs. $\overline{W}$ for data generated from a *linear* regression model with right–skewed, multiplicative measurement errors. Note the curvature exhibited in the plot of

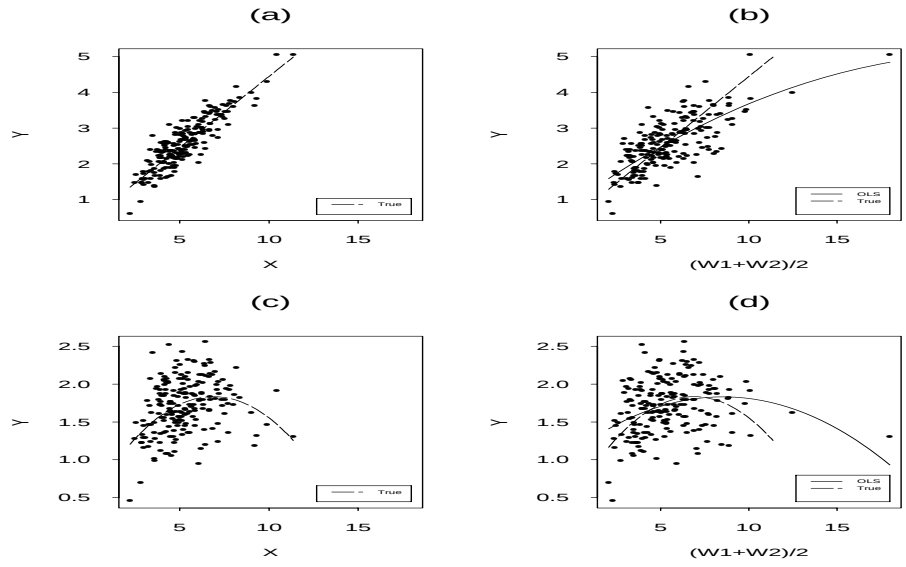

Figure 2: *Plots for two simulated data sets: (a) Y vs X for linear model, (b) Y vs $\overline{W}$ for linear model, (c) Y vs X for quadratic model, (d) Y vs $\overline{W}$ for quadratic model.*

Y vs. $\overline{W}$. Measurement errors of this type can also have the effect of dampening the curvature of the underlying model. The second pair of plots are for data generated from a quadratic regression model with $\beta_2 < 0$. The common feature of the two pairs of plots is that the measurement errors tend to "stretch" the data along the X–axis, giving a distorted view of the true relationship between Y and X.

1.4 Diagnostics for Multiplicative Measurement Error

Measurement error models have been most fully developed for the additive error case, $W = X + U$, with U being either a mean– or median–zero error term that is independent of X. A convenient diagnostic for assessing additivity when X is independent of the mean–zero measurement error term are plots of $|W_{ij} - W_{ik}|$ against $W_{ij} + W_{ik}$ for various $j \neq k$, where W_{ij} is the jth replicate for individual i. In the appendix we show that under the additive model, one would expect to see no correlation in these plots. If, however, the multiplicative model, $W = XU$, is more appropriate, then an additive error model is appropriate when considering the logarithm of W. Plots of $|\log(W_{ij}) - \log(W_{ik})|$ against $\log(W_{ij}) + \log(W_{ik})$ therefore provide a ready diagnostic for multiplicative measurement error.

For our analysis of the Nurses data we will define Y_i to be the average of the four energy replicates for individual i, W_{i1} to be the average of the first two vitamin A replicates for individual i, and W_{i2} to be the average of the third and fourth vitamin A replicates for individual i. The

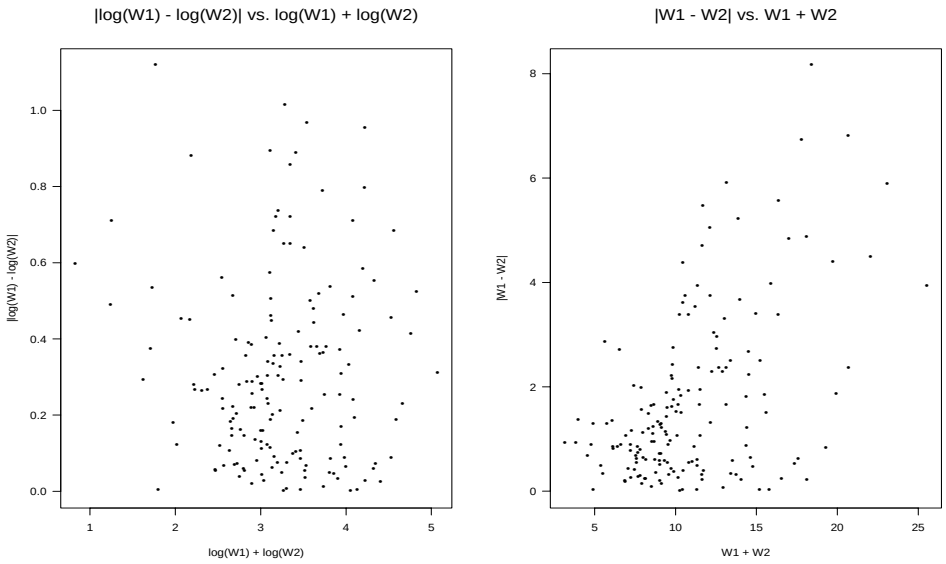

Figure 3: *Measurement error diagnostics for Nurses data.*

diagnostics for the Nurses data are given in Figure 3. The correlation coefficient for the plot of $|\log(W_{i1}) - \log(W_{i2})|$ against $\log(W_{i1}) + \log(W_{i2})$ is $-.02$, suggesting that the measurement errors are additive in the log–scale, and hence multiplicative in the untransformed scale. To see that an additive model is not appropriate for the data in the original scale, note the strength of the correlation in the plot for the untransformed data, which has a corresponding correlation coefficient of .50.

1.5 Models for (X, U)

We will consider two distributional forms for the measurment error, U. The first form is where U can be expressed as $\exp(V)$, where V is mean-zero and symmetric. The second form is a special case of the first, that U is lognormal$(0,\sigma_u^2)$. Note that in both cases we have that W is median–unbiased for X. (The assumption of median as opposed to mean unbiasedness is not really important since there is no way to distinguish between the two cases in practice. The advantage to assuming median–unbiasedness in the case of lognormal measurement error is that it simplifies the identification of parameters.) When working with the first distributional form for U, we do not place any distributional assumptions on X other than that X is nonnegative with finite moments. We call this the *nonparametric case*. For the second distributional form of U, the case of lognormal measurement error, we consider two possibilities for X. The first is where once again we assume only that X is nonnegative with finite moments, which we call the *semiparametric* case. The second form is that X, conditional on Z, is distributed lognormal$(\alpha_0 + \alpha_1^t Z,\sigma_x^2)$, which we will call the

Table 1: Three estimation scenarios.

Model	U	$X\|Z$
Nonparametric	$\exp(V)$, V mean-zero symmetric	nonnegative
Semiparametric	lognormal(0,σ_u^2)	nonnegative
Parametric	lognormal(0,σ_u^2)	lognormal($\alpha_0 + \alpha_1^t Z$,σ_x^2)

parametric case. The three scenarios are summarized in Table 1. Note that the semiparametric model is a special case of the nonparametric model, and that the parametric model is a special case of the other two models. Also note that these names refer only to the assumptions placed on the X and U. For example, the parametric model is not fully "parametric" in that we do not assume anything beyond independence and a zero expectation for the (ϵ_i). We believe this is one of the attractive features of our method.

1.6 Unbiased Estimating Functions for Polynomial Regression under Multiplicative Measurement Error

We derive consistent estimators for the coefficients of the polynomial regression model using the theory of *estimating equations*. An advantage to formulating estimators in terms of estimating equations is that the theory provides a general method for computing asymptotic standard errors. A brief overview of the method is provided in the appendix. A more detailed description can be found in Carroll, et al. (1995). In practice, the estimating function, $\Psi(\cdot)$, is not formulated independently, but rather is a consequence of the estimation method being considered. For example, a maximum likelihood approach would imply taking $\Psi(\cdot)$ to be the derivative of the log–likelihood.

Note that for the polynomial regression model, an unbiased estimating function for $\mathcal{B} = (\beta_0, \beta_{p+1}^t, \beta_1, \ldots, \beta_p)^t$ when the distribution of U is known is

$$\Psi(Y, W, Z, \mathcal{B}) = \begin{pmatrix} (Y - \beta_0 - \beta_{p+1}^t Z - \sum_1^p \beta_k \overline{W}^k / c_k)(1, Z^t)^t \\ (Y - \beta_0 - \beta_{p+1}^t Z)\overline{W}/c_1 - \sum_1^p \beta_k \overline{W}^{k+1}/c_{k+1} \\ \cdots \\ (Y - \beta_0 - \beta_{p+1}^t Z)\overline{W}^p/c_p - \sum_1^p \beta_k \overline{W}^{k+p}/c_{k+p} \end{pmatrix},$$

where $\overline{W}$ is the average of the replicates of W, and c_k is the kth moment of $\overline{U}$. In practice, the distribution of U will be unknown and the c_k will have to be estimated. Unbiased estimating functions for the nonparametric and semiparametric cases can be found by modifying $\Psi(\cdot)$ to incorporate the estimation of the c_k. We take up methods for estimating the c_k in the next section.

For the parametric case, we take an alternative approach that allows us to exploit our knowledge of the distributional form of X. Defining $T_i = {r_i}^{-1} \sum_1^{r_i} \log(W_{ij})$, $i = 1, \ldots, n$, and noting that $E(Y|T, Z) = \beta_0 + \beta_{p+1}^t Z + \sum_1^p \beta_k E(X^k|T, Z)$, a method for estimating $\mathcal{B}$ is to regress the Y_i on the Z_i and on estimates of the $E(X^k|T_i, Z_i)$. Simple calculations give us that the conditional distribution of X given (T, Z) is lognormal with parameters $(\sigma_u^2 \mu_{x|z} + 2\sigma_x^2 T)/(\sigma_u^2 + 2\sigma_x^2)$ and $\sigma_x^2 \sigma_u^2/(\sigma_u^2 + 2\sigma_x^2)$, where $\mu_{x|z} = \alpha_0 + \alpha_1^t Z$. The exact form of the unbiased estimating equation for the parametric case is given in the next section.

2 ANALYSIS OF MEASUREMENT ERROR

2.1 Error Parameter Estimation

Computing estimates of the $E(\overline{U}^k)$ in the nonparametric and semiparametric cases requires that we obtain estimates for the moments of U. Let m_k denote the kth moment of U. An estimator for m_k in the nonparametric case is given by $\widehat{m}_k = \left[\sum_1^n \sum_{j \neq l}^{r_i} \{n r_i (r_i - 1)\}^{-1} (W_{ij}/W_{il})^k\right]^{1/2}$, which follows from the fact that $\left[E\left\{(W_{ij}/W_{il})^k\right\}\right]^{1/2} = m_k$, for all i, j, k, l. For the semiparametric and parametric models, in which U is lognormal(0,σ_u^2), we can take $\widehat{\sigma}_u^2$ to be the mean-square error resulting from an ANOVA on the $\log(W_{ij})$, which is unbiased for σ_u^2. Since the kth moment of lognormal(0,σ_u^2) is $\exp(k^2\sigma_u^2/2)$, an estimator for m_k in the semiparametric case is then given by $\widehat{m}_k = \exp(k^2 \widehat{\sigma}_u^2/2)$. Moments of $\overline{U}$ for the nonparametric and semiparametric cases can be estimated by substituting the $\widehat{m}_k$ into the expansions of the $E(\overline{U}^k)$. For the parametric model, in addition to $\widehat{\sigma}_u^2$, we need estimators for α_0, α_1, and σ_x^2. Estimates for α_0 and α_1 are given by the regression of the $\log(W_{ij})$ on the Z_i. By the independence of X and U, an unbiased estimate for σ_x^2 is given by $\widehat{\sigma}_x^2 = -\widehat{\sigma}_u^2 + \sum_1^n \sum_1^{r_i} (n r_i)^{-1} \{\log(W_{ij}) - \widehat{\alpha}_0 - \widehat{\alpha}_1^t Z_i\}^2$.

2.2 Unbiased Estimating Equations for the case of two replicates

An unbiased estimating function for the nonparametric estimator when $r_i = 2$, $i = 1, \ldots, n$, is given by

$$\Psi^{NP}(Y, W, Z, \mathcal{B}_{NP}) = \begin{pmatrix} (Y - \beta_0 - \beta_{p+1}^t Z - \sum_1^p \beta_k \overline{W}^k / c_k)(1, Z^t)^t \\ (Y - \beta_0 - \beta_{p+1}^t Z)\overline{W}/c_1 - \sum_1^p \beta_k \overline{W}^{k+1}/c_{k+1} \\ \cdots \\ (Y - \beta_0 - \beta_{p+1}^t Z)\overline{W}^p/c_p - \sum_1^p \beta_k \overline{W}^{k+p}/c_{k+p} \\ -m_1^2 + \frac{1}{2}\{(W_1/W_2) + (W_2/W_1)\} \\ \cdots \\ -m_{2p}^2 + \frac{1}{2}\left\{(W_1/W_2)^{2p} + (W_2/W_1)^{2p}\right\} \end{pmatrix},$$

where $\mathcal{B}_{NP} = (\beta_0, \beta_{p+1}^t, \beta_1, \ldots, \beta_p, m_1^2, \ldots, m_{2p}^2)^t$, with the c_k treated as functions of the m_k^2. For the semiparametric estimator, an unbiased estimating function is

$$\Psi^{SP}(Y, W, Z, \mathcal{B}_{SP}) = \begin{pmatrix} (Y - \beta_0 - \beta_{p+1}^t Z - \sum_1^p \beta_k \overline{W}^k / c_k)(1, Z^t)^t \\ (Y - \beta_0 - \beta_{p+1}^t Z)\overline{W}/c_1 - \sum_1^p \beta_k \overline{W}^{k+1}/c_{k+1} \\ \cdots \\ (Y - \beta_0 - \beta_{p+1}^t Z)\overline{W}^p/c_p - \sum_1^p \beta_k \overline{W}^{k+p}/c_{k+p} \\ -2\sigma_u^2 + \{\log(W_1) - \log(W_2)\}^2 \end{pmatrix},$$

where $\mathcal{B}_{SP} = (\beta_0, \beta_{p+1}^t, \beta_1, \ldots, \beta_p, \sigma_u^2)^t$, with the c_k treated as functions of σ_u^2. Finally, an unbiased estimating function in the parametric case is given by

$$\Psi^{CM}(Y, W, Z, \mathcal{B}_{CM}) = \begin{pmatrix} (Y - \beta_0 - \beta_{p+1}^t Z - \sum_1^p \beta_k v_k)(1, Z^t)^t \\ (Y - \beta_0 - \beta_{p+1}^t Z)v_1 - \sum_1^p \beta_k v_k v_1 \\ \cdots \\ (Y - \beta_0 - \beta_{p+1}^t Z)v_p - \sum_1^p \beta_k v_k v_p \\ \{\log(W_1) + \log(W_2) - 2\alpha_0 - 2\alpha_1^t Z\}\,(1, Z^t)^t \\ -2\sigma_x^2 - 2\sigma_u^2 + \{\log(W_1) - \alpha_0 - \alpha_1^t Z\}^2 + \{\log(W_2) - \alpha_0 - \alpha_1^t Z\}^2 \\ -2\sigma_u^2 + \{\log(W_1) - \log(W_2)\}^2 \end{pmatrix},$$

where we define $v_k = E(X^k|T, Z)$, and $\mathcal{B}_{CM} = (\beta_0, \beta_{p+1}^t, \beta_1, \ldots, \beta_p, \alpha_0, \alpha_1, \sigma_x^2, \sigma_u^2)^t$. We will call the solution to this estimating equation the *conditional mean* estimator, in reference to the conditioning on T and Z. We prefer this name over "parametric" estimator since the latter suggests a likelihood–based estimator. Note that a likelihood estimator would require assuming a distributional form for ϵ, something we wish to avoid.

2.3 Asymptotic Variance Comparisons

Asymptotic variances for the estimators are found by taking one–term Taylor series approximations of $\Psi(\cdot)$ at the estimates, $\widehat{\mathcal{B}}$. An outline of the derivations for the case of quadratic regression without covariates is given in the appendix. The variances are calculated under the assumptions of the parametric model, with the additional assumption of finite and constant variance for the (ϵ_i). We can use these formulae to calculate the asymptotic relative efficiency (ARE) of the conditional mean estimator relative to both the nonparametric and semiparametric estimators for various parameter values. This allows us to assess the gain in efficiency that results from choosing to model X when the parametric model holds. Plots of the AREs for $\widehat{\beta}_2$ are shown in Figure 4. The AREs were computed using the parameter estimates for the Nurses data given in the next section, except that σ_u^2 was allowed to vary, and are plotted as a function of the ratio of the coefficients of variation for U and X. This allows us to see how the efficiency of the conditional mean estimator varies with changes in the

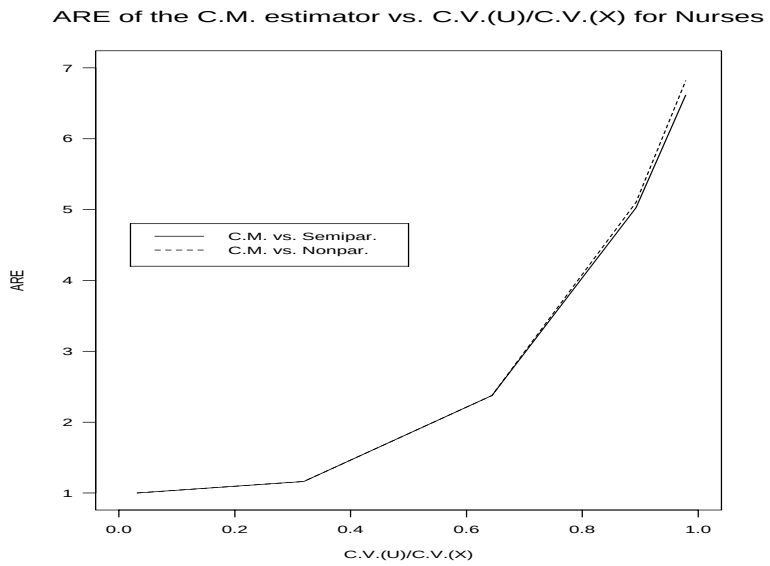

Figure 4: *ARE of C.M. estimator vs. C.V.(U)/C.V.(X) for Nurses.*

relative amount of measurement error. The plot is consistent with our simulation studies in that under the parametric model, the nonparametric and semiparametric methods produce virtually identical estimates for large n. More results from our simulation study are given later.

3 NUMERICAL EXAMPLE

3.1 Diagnostics for U and X for the Nurses Data

In order to determine which of the three methods is the most appropriate for the Nurses data, we must characterize the distributions of U and X. We can assess the lognormality of U by constructing the Q–Q plot for $\log(W_{i1}/W_{i2})$, $i = 1, \ldots, n$. If U is lognormal, this plot should look like that for normally distributed data. If the lognormality assumption for U is valid, a diagnostic for lognormality of X is the Q–Q plot for $\log(W_{i1}) + \log(W_{i2})$, $i = 1, \ldots, n$. For lognormal X, this plot should also look like a Q–Q plot of normally distributed data. Examination of these plots in Figure 5 suggests that the lognormality assumption is reasonable for both X and U. Taken together, the above diagnostics suggest that the conditional mean estimator is reasonable for the Nurses data.

3.2 Regression Fits for the Nurses Data

Plots of the fitted regression functions are given in Figure 6. We computed 95% confidence intervals for the estimates of β_2 using bootstrap percentiles. Confidence intervals for the NP, SP, CM, and

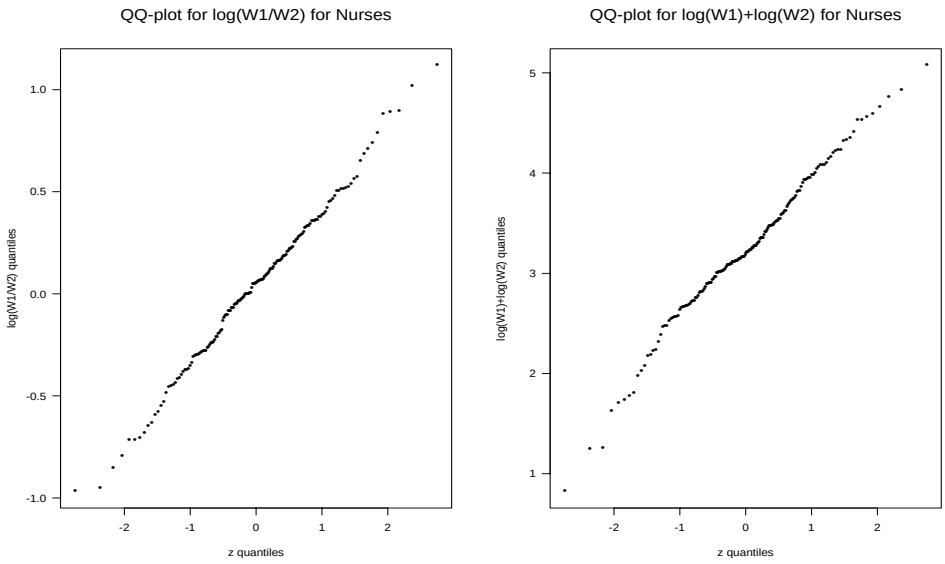

Figure 5: *Q–Q plots for log(W1/W2) and log(W1)+log(W2) for the Nurses data.*

OLS estimators respectively were: (-.121,-.014), (-.165,-.015), (-.051,-.014), and (-.022,-.006). Our simulation results demonstrated that bootstrap percentiles provided the most reliable intervals.

4 SIMULATION STUDY

4.1 Overview

A simulation study was carried out to assess the relative performance of the three methods under the parametric model without covariates. Generating parameter values were taken from the fit of the conditional mean estimator for the Nurses data. Parameter values used were $\mathcal{B} = (.464, .398, -.029)^t$, $\mu_x = 1.613$, $\sigma_x^2 = .094$, and $\sigma_u^2 = .076$. The (ϵ_i) were taken to be i.i.d. N(0,σ_ϵ^2), with $\sigma_\epsilon^2 = .101$ being the mean of the squared deviations of the data about the conditional mean fit.

4.2 Some Descriptive Statistics

Given in Table 2 are the medians, MADs, and estimated root mean square errors of $\widehat{\beta}_2$ for 5000 simulated data sets. The sampling distributions for the nonparametric and semiparametric estimators, although asymptotically normal, were found to be highly skewed for $n = 168$, making necessary the use of the more robust medians and MADs to assess the bias and standard errors. As one might expect, the OLS estimates were the least variable, but were also the most biased. We see that the conditional mean estimator provided the most favorable tradeoff between bias and

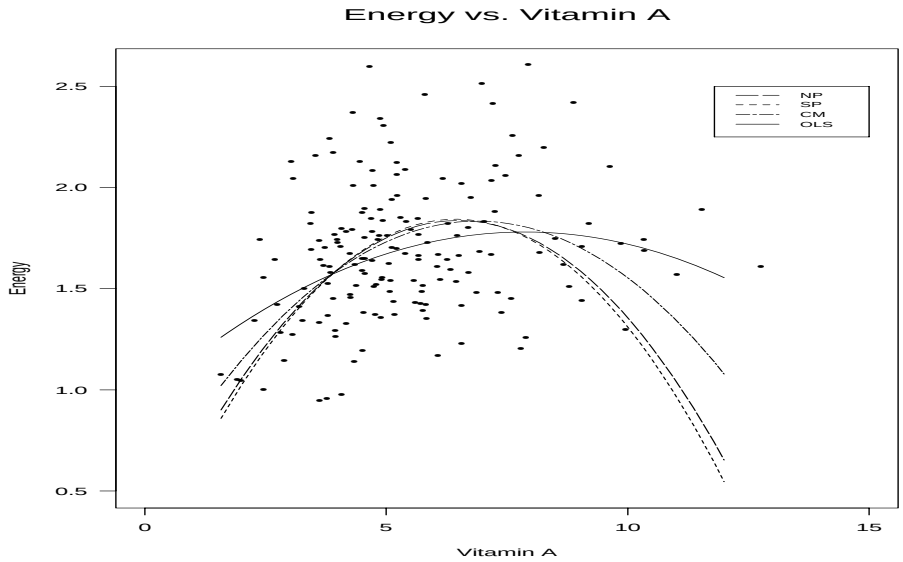

Figure 6: *Nonparametric, semiparametric, conditional mean, and OLS fits for the Nurses data.*

Table 2: Summary statistics for $\widehat{\beta}_2$, $\beta_2 = -.029$.

	median	MAD	sqrt(MSE)
NP	-0.035	0.018	.019
SP	-0.036	0.020	.021
CM	-0.028	0.010	.013
OLS	-0.012	0.004	.016

variance reduction. It is important to note that the nonparametric and semiparametric models both contain the parametric model as a special case, and so are not "incorrect" models for the simulated data. What is evident, however, is that there may be considerable gains to be made if one is willing to model the distribution of the predictor, X.

4.3 Bootstrap–Percentile Confidence Interval Widths and Coverages

The performances of 95% bootstrap–percentile confidence intervals for β_2 were examined by generating 500 data sets at the Nurses parameter estimates and computing bootstrap intervals based on 1000 with–replacement samples. Empirical coverage probabilities and mean confidence interval lengths for the 500 intervals are given in Table 3. We see that only the confidence intervals for the conditional mean estimator provided both accurate coverage and reasonable length. Further simulations showed that as sample size increases, the performances of the nonparametric and

Table 3: Simulated bootstrap confidence interval coverages and mean lengths, $n = 168$.

	NP	SP	CM	OLS
Coverage	.960	.976	.942	.182
Mean length	.470	.663	.051	.020

semiparametric estimators approach that of the conditional mean estimator. Much of the poor performance of the nonparametric and semiparametric methods at moderate values of n appears to be due to highly skewed sampling distributions for the estimators at those sample sizes.

5 GENERALIZATIONS

The methods and results of this paper are easily extended to general estimating functions. In the additive error case, a series of works by Stefanski (1989), Nakamura (1990), Carroll, et al. (1995), and Buzas & Stefanski (1996) have established the method of corrected estimating equations. Under various guises, the basic idea is that in some cases, an estimating function $\Psi(Y, X, Z, \mathcal{B})$ can be expanded as a polynomial

$$\Psi(Y, X, Z, \mathcal{B}) = \sum_{j=0}^{\infty} \Psi_j(Y, Z, \mathcal{B}) X^j.$$

For the special structure of the additive model, expansions can be done either in powers of X as above, powers of $\exp(X)$, or combinations of the two. For the multiplicative model, expanding in powers of X is most convenient. Note that this is equivalent to first replacing X by its logarithm X_*, thus obtaining an additive model, and then expanding the estimating function in terms of powers of exponentials of X_*. For the multiplicative model, if the moments of U are known then under appropriate regularity conditions relating to convergence of the sum, an unbiased estimating function for $\mathcal{B}$ is

$$\Psi^{UB}(Y, W, Z, \mathcal{B}) = \sum_{j=0}^{\infty} \Psi_j(Y, Z, \mathcal{B}) \overline{W}^j / c_j,$$

where c_j is the jth moment of $\overline{U}$. For instance, it is easily seen that for the polynomial regression model, the estimating equations for the nonparametric and semiparametric estimators are of this form up to the nuisance parameters $(m_1, \ldots, m_{2p})^t$ and σ_u^2 respectively, where $m_k = E(U^k)$.

The general equivalent of the parametric approach is described briefly as follows. Suppose that

we can expand both the mean and variance of Y in powers of X, so that

$$E(Y|X,Z,\mathcal{B}) = \sum_{j=0}^{\infty} d_j(Z,\mathcal{B})X^j; \qquad \text{var}(Y|X,Z,\mathcal{B}) = \sum_{j=0}^{\infty} e_j(Z,\mathcal{B})X^j; \tag{1}$$

Then provided that the following sums converge, we have

$$E(Y|W,Z,\mathcal{B}) = \sum_{j=0}^{\infty} d_j(Z,\mathcal{B})v_j;$$

$$\text{var}(Y|W,Z,\mathcal{B}) = \sum_{j=0}^{\infty} e_j(Z,\mathcal{B})v_j + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} d_i(Z,\mathcal{B})d_j(Z,\mathcal{B})(v_{i+j} - v_i v_j); \tag{2}$$

where $v_j = E(X^j|W,Z)$. If we assume a parametric distribution for X and U, the v_j are known up to parameters and we can estimate $\mathcal{B}$ via ordinary quasilikelihood (generalized least squares).

In our formulation of the conditional mean estimator for polynomial regression, we did not specify a model for $\text{var}(Y|X,Z,\mathcal{B})$, but rather worked only with $E(Y|X,Z,\mathcal{B})$. Since we are not directly specifying a variance model, for the purposes of estimation we have computed the ordinary least squares estimate of $\mathcal{B}$, given estimates of the v_j. This is in effect a solution to a generalized estimating equation with a homoscedastic "working" variance function (Zeger, et al., 1988). Modeling the variance of Y given (X,Z) as in (1) and using (2) as the observed variance function may lead to a more efficient estimator, but as seen in Figure 4, our working parametric solution is already reasonably efficient relative to the nonparametric and semiparametric estimators. We do wish to reemphasize, however, that the gains in efficiency come from correctly modeling the distribution of X.

CONCLUDING REMARKS

In this paper we have considered two general approaches to fitting polynomial regression models in the presence of multiplicative measurement error in the predictor. The approaches differed in that for one we did not make any distributional assumptions for the predictor beyond the usual i.i.d. assumption, and for the other we assumed a distributional form. In our analysis we found that the latter approach, though less robust, can in some cases lead to a substantial increase in efficiency, particularly for small to moderate sample sizes. We also found that these gains in efficiency increase with the degree of the measurement error. Much of the gain in efficiency appears due to the slow convergence to normality of the less parametric approach.

REFERENCES

Buzas, J. S. & Stefanski, L. A. (1996). A note on corrected score estimation. *Statistics & Probability Letters*, 28, 1–8.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models.* London: Chapman and Hall.

Hwang, J. T. (1986). Multiplicative errors in variables models with applications to the recent data released by the U.S. Department of Energy. *Journal of the American Statistical Association*, 81, 680-688.

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

Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, 77, 1, 113-116.

Stefanski, L. A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Communications in Statistics, Series A*, 18, 4335–4358.

Zeger, S. L., Liang, K. and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44, 1049–1060.

6 APPENDIX

6.1 Justifications for the Measurement Error Diagnostics

For the additive model, $\mathrm{Cov}(|W_1 - W_2|, W_1 + W_2) = \mathrm{E}\{|U_1 - U_2|(U_1 + U_2)\}$, which is $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |s - t|(s+t) f_{U_1}(s) f_{U_2}(t)\ \mathrm{d}s\ \mathrm{d}t$. By a change of variable, this is $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |s+r|(s-r) f_{U_1}(s) f_{U_2}(r)\ \mathrm{d}s\ \mathrm{d}r$, which is 0. Similarly, for the multiplicative model, $\mathrm{Cov}\{|\log(W_1) - \log(W_2)|, \log(W_1) + \log(W_2)\}$ is 0.

6.2 Estimating Functions

A function $\Psi(Y, X, \mathcal{B})$ is an *unbiased estimating function* for $\mathcal{B}$ if $E\{\Psi(Y, X, \mathcal{B})\} = 0$. Given such a function, $\Psi(\cdot)$, one possible estimator for $\mathcal{B}$ is the solution, $\widehat{\mathcal{B}}$, of $n^{-1}\sum_1^n \Psi(Y_i, X_i, \mathcal{B}) = 0$. Under a set of mild regularity conditions on Ψ, one can show that $\widehat{\mathcal{B}}$ is a consistent estimator of $\mathcal{B}$. The limiting distribution of $\widehat{\mathcal{B}}$ can be found by taking a first-order Taylor series approximation of $n^{-1}\sum_1^n \Psi(Y_i, X_i, \widehat{\mathcal{B}})$ about $\mathcal{B}$, and then applying Slutsky's Theorem and the CLT. One finds that asymptotically $n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B})$ has mean 0 and covariance $A^{-1}BA^{-t}$, where $A = E\{(\partial/\partial\mathcal{B}^t)\Psi\}$, $B = E\{\Psi(Y, X, \mathcal{B})\Psi^t(Y, X, \mathcal{B})\}$, and $A^{-t} = (A^{-1})^t$.

6.3 Asymptotic Variance of the Nonparametric Estimator

An unbiased estimating equation for the nonparametric estimator in the quadratic regression case with two replicates and without covariates is

$$\Psi^{NP}(Y,W,\mathcal{B}_{NP}) = \begin{pmatrix} (Y-\beta_0-\beta_1\overline{W}/c_1-\beta_2\overline{W}^2/c_2) \\ (Y-\beta_0)\overline{W}/c_1-\beta_1\overline{W}^2/c_2-\beta_2\overline{W}^3/c_3 \\ (Y-\beta_0)\overline{W}^2/c_2-\beta_1\overline{W}^3/c_3-\beta_2\overline{W}^4/c_4 \\ -m_1^2+\frac{1}{2}\{(W_1/W_2)+(W_2/W_1)\} \\ -m_2^2+\frac{1}{2}\left\{(W_1/W_2)^2+(W_2/W_1)^2\right\} \\ -m_3^2+\frac{1}{2}\left\{(W_1/W_2)^3+(W_2/W_1)^3\right\} \\ -m_4^2+\frac{1}{2}\left\{(W_1/W_2)^4+(W_2/W_1)^4\right\} \end{pmatrix},$$

where $\mathcal{B}_{NP} = (\beta_0,\beta_1,\beta_2,m_1^2,m_2^2,m_3^2,m_4^2)^t$, with the c_k treated as functions of the m_k^2. In determining A_{NP} and B_{NP}, it can be shown that

$$A_{NP} = -\begin{pmatrix} E(C) & E(F) \\ \mathbf{0}_{4\times 3} & I_4 \end{pmatrix},$$

where $C_{ij} = \overline{W}^{i+j-2}/c_{i+j-2}$, and F has (i,j)–element $(Y-\beta_0)\overline{W}^{i-1}c_{i-1,j}/c_{i-1}^2-\beta_1\overline{W}^i c_{i,j}/c_i^2-\beta_2\overline{W}^{i+1}c_{i+1,j}/c_{i+1}^2$, with $c_{i,j}$ defined to be $\partial c_i/\partial m_j^2 = 2^{-i}\binom{i}{j}(m_{i-j}/m_j)$ for $j\le i$, and 0 otherwise. Taking expectations, we have that $E(C)$ has (i,j)–element μ_{i+j-2}, and $E(F)$ has (i,j)–element $(\beta_1\mu_i+\beta_2\mu_{i+1})c_{i-1,j}/c_{i-1}-\beta_1\mu_i c_{i,j}/c_i-\beta_2\mu_{i+1}c_{i+1,j}/c_{i+1}$, where we define $\mu_k = E(X^k)$.

To evaluate B_{NP}, first note that the upper-left 3×3 matrix of $\Psi_{SP}\Psi_{SP}^t$ is given by $(DY-C\mathcal{B})(DY-C\mathcal{B})^t = DD^tY^2-D\mathcal{B}^tCY-C\mathcal{B}D^tY+C\mathcal{B}\mathcal{B}C$, where D is $(1,\overline{W}/c_1,\overline{W}^2/c_2)^t$. Taking expected values, we get that for $1\le i\le 3$, $1\le j\le 3$, the (i,j)–element of B_{SP} is

$$\sigma_\epsilon^2\frac{c_{i+j-2}}{c_{i-1}c_{j-1}}\mu_{i+j-2}+\frac{c_{i+j-2}}{c_{i-1}c_{j-1}}\sum_{k=1}^3\sum_{l=1}^3\beta_{k-1}\beta_{l-1}\mu_{i+j+k+l-4}$$

$$-\sum_{k=1}^3\beta_{k-1}\frac{c_{i+j+k-3}}{c_{i+k-2}c_{j-1}}\sum_{l=1}^3\beta_{l-1}\mu_{i+j+k+l-4}-\sum_{k=1}^3\beta_{k-1}\frac{c_{i+j+k-3}}{c_{i-1}c_{j+k-2}}\sum_{l=1}^3\beta_{l-1}\mu_{i+j+k+l-4}$$

$$+\sum_{k=1}^3\sum_{l=1}^3\beta_{k-1}\beta_{l-1}\frac{c_{i+j+k+l-4}}{c_{i+k-2}c_{j+l-2}}\mu_{i+j+k+l-4}.$$

Next note that for $1\le i\le 3$, $1\le j\le 4$,the $(i,3+j)$–element of $\Psi_{NP}\Psi_{NP}^t$ can be shown to be $(1/2)\left\{(Y-\beta_0)\overline{W}^{i-1}/c_{i-1}-\beta_1\overline{W}^i/c_i-\beta_2\overline{W}^{i+1}/c_{i+1}\right\}\left\{(W_1/W_2)^j+(W_2/W_1)^j\right\}$, which has expectation $(\beta_1\mu_i+\beta_2\mu_{i+1}/c_{i-1})\,g_{i-1,j}-(\beta_1\mu_i/c_i)\,g_{i,j}-(\beta_2\mu_{i+1}/c_{i+1}/)\,g_{i+1,j}$, where $g_{i,j} = E\left\{\overline{U}^i(U_1/U_2)^j\right\} = 2^{-i}\sum_0^i\binom{i}{k}m_{j+k}m_{i-(j+k)}$.

Finally, we have that for $1\le i\le 4$, $1\le j\le 4$, the $(3+i,3+j)$–element of $\Psi_{NP}\Psi_{NP}^t$ is given by $m_i^2m_j^2-2m_i^2m_j^2+(1/4)E\left\{(W_1/W_2)^{i+j}+(W_2/W_1)^{i+j}+(W_1/W_2)^{|i-j|}+(W_2/W_1)^{|i-j|}\right\}$, which is $(m_{i+j}^2+m_{|i-j|}^2)/2-m_i^2m_j^2$.

6.4 Asymptotic Variance of the Semiparametric Estimator

An unbiased estimating equation for the semiparametric estimator in the quadratic regression case with two replicates and without covariates is

$$\Psi^{SP}(Y,W,\mathcal{B}_{SP}) = \begin{pmatrix} (Y-\beta_0-\beta_1\overline{W}/c_1-\beta_2\overline{W}^2/c_2) \\ (Y-\beta_0)\overline{W}/c_1-\beta_1\overline{W}^2/c_2-\beta_2\overline{W}^3/c_3 \\ (Y-\beta_0)\overline{W}^2/c_2-\beta_1\overline{W}^3/c_3-\beta_2\overline{W}^4/c_4 \\ -\sigma_u^2+\frac{1}{2}\left\{\log(W_1)-\log(W_2)\right\}^2 \end{pmatrix},$$

where $\mathcal{B}_{SP} = (\beta_0,\beta_1,\beta_2,\sigma_u^2)^t$, and the c_k are treated as functions of σ_u^2. We note that $A_{SP} = E\left\{\frac{\partial}{\partial\mathcal{B}_{SP}^t}\Psi^{SP}(\mathcal{B}_{SP})\right\}$ and $B_{SP} = E\left\{\Psi^{SP}(\mathcal{B}_{SP})\Psi^{SP}(\mathcal{B}_{SP})^t\right\}$, and

$$A_{SP} = \begin{pmatrix} -1 & -\mu_1 & -\mu_2 & \beta_1\mu_1 c_1^{(1)}/c_1+\beta_2\mu_2 c_2^{(1)}/c_2 \\ -\mu_1 & -\mu_2 & -\mu_3 & -\beta_1\mu_2(c_1^{(1)}/c_1-c_2^{(1)}/c_2)-\beta_2\mu_3(c_1^{(1)}/c_1-c_3^{(1)}/c_3) \\ -\mu_2 & -\mu_3 & -\mu_4 & -\beta_1\mu_3(c_2^{(1)}/c_2-c_3^{(1)}/c_3)-\beta_2\mu_4(c_2^{(1)}/c_2-c_4^{(1)}/c_4) \\ 0 & 0 & 0 & -1 \end{pmatrix},$$

where μ_k is the kth moment of X, $c_k = 2^{-k}\sum_{i=0}^k \binom{k}{i}\exp\{\sigma_u^2(k^2-2ik+2i^2)/2\}$, and $c_k^{(1)}$ is the derivative of c_k with respect to σ_u^2, namely $2^{-k}\sum_{i=0}^k \binom{k}{i}(k^2-2ik+2i^2/2)\exp\{\sigma_u^2(k^2-2ik+2i^2)/2\}$.

To evaluate B_{SP}, first note that the upper-left 3×3 matrix of $\Psi_{SP}\Psi_{SP}^t$ is the same as for B_{NP} given previously.
For $1\le i\le 3$, the $(i,4)$–element of $\Psi_{SP}\Psi_{SP}^t$ can be shown to be

$$\left\{(Y-\beta_0)\overline{W}^{i-1}/c_{i-1}-\beta_1\overline{W}^i/c_i-\beta_2\overline{W}^{i+1}/c_{i+1}\right\}\left[\left\{\log(U_1)-\log(U_2)\right\}^2/2-\sigma_u^2\right].$$

Taking expectations, we get $(\beta_1\mu_i+\beta_2\mu_{i+1})h_{i-1}/c_{i-1}-\beta_1\mu_i h_i/c_i-\beta_2\mu_{i+1}h_{i+1}/c_{i+1}$, where h_k is the expected value of $\overline{U}^k\left\{\log(U_1)-\log(U_2)\right\}^2$. Noting that $E(U^k)=\exp(k^2\sigma_u^2/2)$, $E\left\{\log(U)\right\}=0$, $E\left\{\log^2(U)\right\}=\sigma_u^2$, and that $E\left\{U^k\log^r(U)\right\}$ is $\exp(k^2\sigma_u^2/2)$ times the rth moment of $\mathrm{N}(k\sigma_u^2,\sigma_u^2)$, we have that h_k is

$$\begin{aligned} & E\left[2^{-k}\sum_{i=0}^k\binom{k}{i}U_1^iU_2^{k-i}\left\{\log^2(U_1)-2\log(U_1)\log(U_2)+\log^2(U_2)\right\}\right] \\ = \;& 2^{-k}\sum_{i=0}^k\binom{k}{i}\left\{(k^2-4ik+4i^2)\sigma_u^4+2\sigma_u^2\right\}\exp\left\{(k^2-2ik+2i^2)\sigma_u^2/2\right\}, \end{aligned}$$

Finally, we have that the $(4,4)$–element of B_{SP} is $\sigma_u^4-\sigma_u^4+(1/4)E\left[\left\{\log(U_1)-\log(U_2)\right\}^4\right]$, which is $(1/4)E\left[\left\{\sqrt{2\sigma_u^2}Z\right\}^4\right]=3\sigma_u^4$, where $Z\sim\mathrm{N}(0,1)$.

6.5 Asymptotic Variance of the Conditional Mean Estimator

For X distributed as lognormal(μ_x,σ_x^2), $X|T$ was shown to be lognormal with parameters $(\sigma_u^2\mu_x+2\sigma_x^2T)/(\sigma_u^2+2\sigma_x^2)$ and $\sigma_x^2\sigma_u^2/(\sigma_u^2+2\sigma_x^2)$. As a consequence,

$$E(X|T)=\exp\left\{\frac{\sigma_u^2\mu_x+2\sigma_x^2T}{\sigma_u^2+2\sigma_x^2}+\frac{\sigma_x^2\sigma_u^2}{2(\sigma_u^2+2\sigma_x^2)}\right\},\quad E(X^2|T)=\exp\left\{\frac{2\sigma_u^2\mu_x+4\sigma_x^2T}{\sigma_u^2+2\sigma_x^2}+\frac{2\sigma_x^2\sigma_u^2}{\sigma_u^2+2\sigma_x^2}\right\}.$$

For notational convenience we define $v_i = E(X^i|T)$, $i = 0, 1, 2$. Notice that we can express v_i as $k_i(W_1W_2)^{i\lambda} = k_i X^{2i\lambda}(U_1U_2)^{i\lambda}$, where $k_0 = 1$, $k_1 = \exp\{(2\sigma_u^2\mu_x + \sigma_x^2\sigma_u^2)/(2\sigma_u^2 + 4\sigma_x^2)\}$, $k_2 = \exp\{(2\sigma_u^2\mu_x + 2\sigma_x^2\sigma_u^2)/(\sigma_u^2 + 2\sigma_x^2)\}$, and $\lambda = \sigma_x^2/(\sigma_u^2 + 2\sigma_x^2)$. An unbiased estimating equation for the conditional mean estimator in the quadratic regression case with two replicates and without covariates is

$$\Psi^{CM}(Y, W, \mathcal{B}_{CM}) = \begin{pmatrix} (Y - \beta_0 - \beta_1 v_1 - \beta_2 v_2) \\ (Y - \beta_0)v_1 - \beta_1 v_1^2 - \beta_2 v_1 v_2 \\ (Y - \beta_0)v_2 - \beta_1 v_1 v_2 - \beta_2 v_2^2 \\ -\sigma_u^2 + \frac{1}{2}\{\log(W_1) - \log(W_2)\}^2 \\ -\mu_x + \frac{1}{2}\{\log(W_1) + \log(W_2)\} \\ -\sigma_x^2 - \sigma_u^2 + \frac{1}{2}\left[\{\log(W_1) - \mu_x\}^2 + \{\log(W_2) - \mu_x\}^2\right] \end{pmatrix},$$

where $\mathcal{B}_{CM} = (\beta_0, \beta_1, \beta_2, \sigma_u^2, \mu_x, \sigma_x^2)^t$. Note that the first three elements of Ψ^{CM} can be expressed as $\mathbf{v}Y - \mathbf{v}\mathbf{v}^t\mathcal{B}$, where $\mathbf{v} = (v_0, v_1, v_2)^t$. To determine A_{CM} and B_{CM}, note that

$$A_{CM} = -E\begin{pmatrix} \mathbf{vv^t} & & \mathbf{D} \\ & 1 & 0 & 0 \\ \mathbf{0_3} & 0 & 1 & 0 \\ & 1 & 0 & 1 \end{pmatrix},$$

where $\mathbf{0_3}$ is a 3×3 matrix of zeros, and $\mathbf{D}$ is the derivative of $\mathbf{vY} - \mathbf{vv^t}\mathcal{B}$ with respect to $(\sigma_u^2, \mu_x, \sigma_x^2)$, which can be evaluated directly. The elements of the expectation of $\mathbf{D}$ can be expressed as sums in terms of the form $f(a, b, m, n) = E\left\{X^a(U_1U_2)^b\log^m(W_1)\log^n(W_2)\right\}$. Expanding this expression, we get

$$f(a, b, m, n) = \sum_{i=0}^{m}\sum_{j=0}^{n}\binom{m}{i}\binom{n}{j} E\left\{X^a\log^{i+j}(X)\right\} E\left\{U^b\log^{m-i}(U)\right\} E\left\{U^b\log^{n-j}(U)\right\}.$$

Defining $g_x(k, l) = E\left\{X^k\log^l(X)\right\}$ and $g_u(k, l) = E\left\{U^k\log^l(U)\right\}$, one can show that $g_x(k, l) = \exp(k\mu_x + k^2\sigma_x^2/2)\sum_0^l \sigma_x^i\xi^i(\mu_x + k\sigma_x^2)^{l-i}$, where ξ^i is the ith moment of standard normal.

To evaluate B_{CM}, first note that the upper-left 3×3 matrix of Ψ_{CM} can be written as $\mathbf{v}(\mathbf{x}^t\mathcal{B}+\epsilon) - \mathbf{v}\mathbf{v}^t\mathcal{B} = \mathbf{v}(\mathbf{x}^t-\mathbf{v}^t)\mathcal{B}+\mathbf{v}\epsilon$, and so $\Psi_{CM}\Psi_{CM}^t = \mathbf{v}(\mathbf{x}^t-\mathbf{v}^t)\mathcal{B}\mathcal{B}^t(\mathbf{x}^t-\mathbf{v})\mathbf{v}^t+\epsilon^2\mathbf{v}\mathbf{v}^t$, ignoring the terms that have expectation zero. This matrix has (i,j)–element $v_{i-1}v_{j-1}\left\{\epsilon^2 + \sum_0^2\sum_0^2\beta_k\beta_l(X^k - v_k)(X^l - v_l)\right\}$. Finding the expectations of these terms requires that we take expectations of the form $\mathrm{E}\{X^i v_j v_k v_l v_m\}$. Remembering that $v_i = k_i(W_1W_2)^{i\lambda} = k_i X^{2i\lambda}(U_1U_2)^{i\lambda}$, we can write this expression in the form $c_* X^{\lambda_{1*}}(U_1U_2)^{\lambda_{2*}}$, which has expectation $c_*\exp(\lambda_{1*}\mu_x + \lambda_{1*}^2\sigma_x^2/2 + \lambda_{2*}^2\sigma_u^2)$. The remaining elements of B_{CM} can be expressed as sums in $f(a, b, m, n)$ terms.