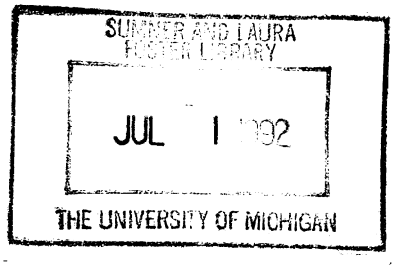
Center for Research on Economic and Social Theory
CREST Working Paper

Consistent Estimation of Linear and Nonlinear Errors-in-Variables Models with Validation Information

Lung-fei Lee
Jungsywan H. Sepanski

April, 1992
Number 92–07

DEPARTMENT OF ECONOMICS
University of Michigan
Ann Arbor, Michigan 48109-1220



Consistent Estimation of Linear and Nonlinear Errors-in-Variables Models with Validation Information

Lung-fei Lee
Department of Economics
The University of Michigan
Ann Arbor, MI 48109-1220

and

Jungsywan H. Sepanski
Department of Statistics and Management Science
School of Business Administration
The University of Michigan
Ann Arbor, MI 48109-1234

Date: April 1992

Abstract

Consistent methods for the estimation of linear and nonlinear regression models with general measurement errors in variables in the presence of validation data information are introduced. The estimation procedures do not rely on any specific specification of auxiliary structural measurement error equations and are robust against any specification of measurement error equations. The estimation procedures can be applied even to situations where validation data alone does not permit possible estimation of the model. When validation data is rich enough for consistent estimation of the model, survey data are still valuable in improving efficiency of the validation data estimates. The estimation procedures can be applied to models and data where direct bias corrections might not be feasible.

JEL classification number: 211

Keywords:
Errors-in-variables, linear regression model, nonlinear regression model, survey data, validation study, bias, consistent estimator, efficiency.

Correspondence Address:
Lung-fei Lee, Department of Economics, Lorch Hall, The University of Michigan, 611 Tappan Street, Ann Arbor, MI 48109-1220.

1. Introduction

In several recent investigations of the quality of survey data with validation studies, there are findings of measurement errors on important labor market variables such as earnings and work hours. From the Panel Study of Income Dynamic Validation, Bound, Brown, Duncan, and Rogers (1989) have found measurement errors in survey reports of earnings and work hours variables. The measurement errors of earnings are negatively correlated with true earnings. The workers with low earnings tend to report higher earnings, and the workers with high earnings tend to under-report their true earnings. Bound and Krueger (1991) compared Current Population Survey data and administrative Social Security payroll tax records. They found that the measurement errors in earning variables are also negatively correlated with true earnings. In another validation study using different data sources (Rodgers and Herzog (1987)), earnings reporting errors are negatively correlated with job tenure and positively correlated with schooling levels. These findings raise concerns on the use of survey data for empirical estimation. It is well known that with measurement errors in explanatory variables, least squares estimates are inconsistent. The measurement errors reported above create more complicated situations. With the measurement errors in dependent variables reported in these studies, the least squares estimates will also be inconsistent even if there are no measurement errors in the explanatory variables. As an illustration, consider the linear regression model

$$y = x\beta_0 + \epsilon,$$

with the reported variables $\tilde{y}$ and z for y and x, respectively. The least squares estimate of regressing $\tilde{y}$ on z will converge in probability to β^*, where

$$\begin{aligned}\beta^* &= [E(z'z)]^{-1}E(z'\tilde{y}) \\ &= \beta_0 + \Delta,\end{aligned}$$

and

$$\Delta = [E(z'z)]^{-1}\{E(z'(\tilde{y}-y)) - E(z'(z-x)\beta_0)\}$$

by using the property $E(z'\epsilon) = 0$. In the classical error-in-variables model, the measurement error $u = z - x$ is assumed to be uncorrelated with x, i.e., $E(x'u) = 0$; and the measurement error $v = \tilde{y} - y$ is assumed to be uncorrelated with z. For these situations, $\Delta = -[E(z'z)]^{-1}E(u'u)\beta_0$. For the simple regression model, such a measurement error of x will bias downward the estimate of the coefficient of x (see Fuller, 1987), but the measurement error of y does not bias this coefficient. With a measurement error u that is correlated with x, the biases will be more complicated. The measurement error equation $z = x + u$ with u correlated with x implies that $E(z'(z-x)) = E(u'u) + E(x'u)$. The bias due to a measurement error in x is

$$\Delta_x = -[E(z'z)]^{-1}\{E(u'u) + E(x'u)\}\beta_0.$$

If u is positively correlated with x, the downward bias will be more severe than the bias of the classical error case. On the other hand, if u and x are negatively correlated, the magnitude of the bias might be reduced, depending on the relative magnitudes of $E(x'u)$ and $E(u'u)$. With a measurement error of y, the measurement error equation $\tilde{y} = y + v$ implies that

$$\tilde{y} - y = E(v|z,y) + \eta,$$

where η is a disturbance with $E(\eta|z,y) = 0$. It follows that $E(z'(\tilde{y}-y)) = E(z'E(v|z,y))$. The bias on β_0 due to the measurement error of y will depend on the covariance of z and the conditional expectation function $E(v|z,y)$, which may be rather complicated. If $E(v|z,y) = \alpha_1 y + z\alpha_2$ happens to be linear in y and z, then the bias due to the measurement error of y is

$$\begin{aligned}\Delta_y &= [E(z'z)]^{-1}\{E(z'y)\alpha_1 + E(z'z)\alpha_2\} \\ &= \alpha_1\beta_0 + \alpha_2.\end{aligned}$$

If $E(v|z,y)$ depends only on y, i.e., $\alpha_2 = 0$, then the estimate of β_0 will be biased proportionally (Bound et al. (1989)). If $\alpha_2 \neq 0$, the bias might also be subject to a horizontal shift.

In summary, measurement error biases would be rather difficult to access accurately without validation data information. One solution is to postulate distributional assumptions on the measurement error. For the case when only the predictors are measured with error, the maximum likelihood estimation has been proposed under normality assumption of the conditional distribution of the true covariate x given the surrogate variable z; see Carroll et al. (1984), Armstrong (1985), Fuller (1987), Schaffer (1987), Burr (1988), etc. Under functional assumptions of the first and second moments of x given z, small measurement error approximations are proposed in Stefanski (1985), Stefanski and Carroll (1985), Whitmore (1988), Gleser (1989), Rosner et al. (1989) and Carroll and Stefanski (1990), etc. Fuller (1987) also introduces the maximum likelihood estimation and the method of moments estimation for the case when both y and x are measured with error under normality assumptions.

In the presence of validation data, one might be able to correct biases and derive consistent estimates from survey data. So far, in the econometric literature, there are few suggestions on how to use validation data information to improve systematically estimates derived from survey data in either the linear or nonlinear regression model except for Bound et al (1989) and Sepanski and Carroll (1991) Bound et al (1989) discussed the use of validation data to correct the bias, Δ, in a linear regression model. Sepanski and Carroll (1991) considered the nonlinear regression model with measurement errors in the explanatory variables only. For the measurement errors-in-variables problem in nonlinear models in econometrics, one can only find a few studies in the econometric literature (e.g., Hsiao (1989) and Hausman et al (1991)). Bias correction in a general nonlinear regression model might be rather difficult. The article by Hausman et al (1991) considered only a polynomial regression model.

In this article, we will propose consistent estimation methods for the error-in-variables linear and/or nonlinear regression models with the presence of validation data information. Several approaches might be pursued. One possible approach is to impose structural measurement error equations as auxiliary equations and consider the estimation problem by some principle such as the method of maximum likelihood. This approach may be subject to severe misspecification error problems, unless we know very well the economic or psychological reasons for reporting errors. The second approach is to estimate the possible measurement error equations by nonparametric methods. The approaches in Carroll and Wand (1991) and Sepanski and Carroll (1991) are nonparametric or semiparametric. In this article, we consider a different approach which provides consistent estimates which are robust with respect to various measurement error structures.

This article is organized as follows. Section 2 considers models with measurement errors in explanatory variables. A procedure which utilizes validation data to implement survey data for estimation is introduced. Consistency and asymptotic distribution issues of the estimator are analyzed. Section 3 considers models with measurement errors in either the dependent variable or in both the independent and dependent variables. Two different estimation methods are introduced. Asymptotic properties of the estimators are derived. Section 4 suggests a procedure for testing the compatibility of validation data with survey data and possible model misspecification. The possibility of pooling both data sources for efficient estimation is considered. Some other efficiency issues are also briefly discussed. The differences of the estimators in Sections 2 and 3 and some bias correction procedures for the linear regression model are pointed out in Section 4. Section 5 provides a conclusion.

2. Measurement Errors in Explanatory Variables

The general (nonlinear) regression model is

$$y = g(x,\beta) + \epsilon, \tag{2.1}$$

where x is a row vector of explanatory variables of dimension k and β is a column vector of unknown parameters of dimension l. In this section, we consider only the case that some or all of the explanatory variables in x are measured with error. The dependent variable y is either accurately measured or measured with error of the classical type, where the measurement error has zero mean and can be absorbed in the overall disturbance ϵ. The survey data are (y_i, z_i), $i = 1, \cdots, n$, where z is a row vector of proxy or instrumental variables of dimension k_1 for x, where $k_1 \geq k$. The true values x_i, $i = 1, \cdots, n$, are not observable in the survey data. However, a validation data set is valuable. The validation data contain sample observations of x and z. Let $(x_{v,j}, z_{v,j})$, $j = 1, \cdots, m$, denote the observations of the validation data.

Let $Y = (y_1, \cdots, y_n)'$ denote the n-dimensional vector of sample observations of y from the survey data. Conformably, Z denotes the $n \times k_1$ matrix of z from the survey data, while X_v denotes the $m \times k$ matrix of x from the validation data, and Z_v denotes the $m \times k_1$ matrix of z from the validation data. Furthermore, let $g(X_v, \beta) = (g(x_{v,1}, \beta), \cdots, g(x_{v,m}, \beta))'$ to simplify notation.

A possible estimation method for β is

$$\min_{\beta \in \Theta} (Y - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\beta))'(Y - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\beta)), \tag{2.2}$$

where Θ is a compact parameter space of β. The intuition behind this estimation method is as follows. For any possible value β of the true parameter vector β_0, a statistical relation of $g(x,\beta)$ and z based on projection can be identified with the validation data. This statistical relation provides a prediction rule which predicts the unobserved value $g(x_i,\beta)$ given the observed z_i in the survey data. The estimate of β is then derived by minimizing the sum of squared residuals of y_i on the predicted value of $g(x_i,\beta)$ given z_i. The error introduced by this prediction rule is asymptotically uncorrelated with z, which is important for the estimator of β_0 to be consistent.

Let

$$Q_n(\beta) = \frac{1}{n}(Y - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\beta))'(Y - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\beta)).$$

The following proposition shows that the estimator from (2.2) is consistent. To be rigorous, the following regularity conditions are assumed:

Assumption 1:

(1) The parameter space Θ is a compact subset of a Euclidean space of dimension l. The true parameter θ_0 is in interior of Θ.
(2) Conditional on x, $E(\epsilon|x) = 0$ and the conditional variance $E(\epsilon^2|x)$ is finite.
(3) $g(x,\beta)$ is a measurable function of x for each β and is twice continuously differentiable in β for each x.

Assumption 2:

(1) The observations (y_i, z_i), $i = 1, \cdots, n$, of the survey data are i.i.d. with finite first and second moments.
(2) The observations $(x_{v,j}, z_{v,j})$, $j = 1, \cdots, m$, of the validation data are i.i.d. The first two moments of (x_v, z_v) exist.
(3) (x_v, z_v) and (x, z) are identically distributed.
(4) ϵ is uncorrelated with z.
(5) $E(z'z)$ is nonsingular.

Proposition 2.1: *In addition to Assumptions 1 and 2, suppose that* $E(\sup_{\beta\in\Theta}|zg(x,\beta)|) \leq \infty$, *and the identification condition that* $E(z[g(x,\beta_0) - g(x,\beta)]) \neq 0$ *for all* $\beta \neq \beta_0$ *holds. Then* $\hat{\beta}$ *from (2.2) is a consistent estimator of* β_0.

Proof: By the law of large number for i.i.d. random variables, $n^{-1}Y'Y \xrightarrow{p} E(y^2)$, $n^{-1}Z'Z \xrightarrow{p} E(z'z)$, and $n^{-1}Z'Y \xrightarrow{p} E(z'y)$. Similarly, since (x_v, z_v) has the same distribution as (x, z), $m^{-1}Z_v'Z_v \xrightarrow{p} E(z'z)$, and

$m^{-1}Z_v'g(X_v,\beta) \xrightarrow{p} E(z'g(x,\beta))$, uniformly in $\beta \in \Theta$, by a uniform law of large number (see, e.g., Amemiya (1985)). It follows that $Q_n(\beta)$ converges in probability uniformly on Θ to $Q(\beta)$, where

$$\begin{aligned} Q(\beta) &= E(y^2) - E(g(x,\beta)z)\,[E(z'z)]^{-1}\,E(z'y) - E(yz)\,[E(z'z)]^{-1}\,E(z'g(x,\beta)) \\ &\quad + E(g(x,\beta)z)\,[E(z'z)]^{-1}\,E(z'g(x,\beta)) \\ &= [E(yz) - E(g(x,\beta)z)]\,[E(z'z)]^{-1}\,[E(z'y) - E(z'g(x,\beta))] + c \\ &= E[(g(x,\beta_0) - g(x,\beta))z]\,[E(z'z)]^{-1}\,E[z'(g(x,\beta_0) - g(x,\beta))] + c, \end{aligned}$$

and $c = E(y^2) - E(yz)\,[E(z'z)]^{-1}\,E(z'y)$ is a constant. $Q(\beta)$ is minimized at $\beta = \beta_0$. The minimum is unique under the identification condition. The consistency of $\hat{\beta}$ follows from the uniform convergence of $Q_n(\beta)$ to $Q(\beta)$, and the unique minimum of $Q(\beta)$ at β_0. Q.E.D.

The estimator $\hat{\beta}$ can also be shown to be asymptotically normally distributed. Its asymptotic distribution will depend on both the sample sizes of the survey data and the validation data.

Proposition 2.2: *In addition to the assumptions in Proposition 2.1, suppose that the uniform integrability conditions* $E(\sup_{\beta\in\Theta} \|\frac{\partial g(x,\beta)}{\partial\beta} z\|) \le \infty$; $E(\sup_{\beta\in\Theta} \|\frac{\partial^2 g(x,\beta)}{\partial\beta\partial\beta_j} z\|) \le \infty$, *for all* $j = 1,\cdots,l$, *hold; and* $E\left(\frac{\partial g(x,\beta_0)}{\partial\beta} z\right)$ *has rank* l. *Then*

$$\begin{aligned} n^{1/2}(\hat{\beta} - \beta_0) = &\left\{E\left(\frac{\partial g(x,\beta_0)}{\partial\beta} z\right)[E(z'z)]^{-1}\,E\left(z'\frac{\partial g(x,\beta_0)}{\partial\beta'}\right)\right\}^{-1} E\left(\frac{\partial g(x,\beta_0)}{\partial\beta} z\right)[E(z'z)]^{-1} \\ &\cdot\left\{\frac{1}{n^{1/2}}\sum_{i=1}^{n} z_i'\left(y_i - z_i\,[E(z'z)]^{-1}\,E(z'y)\right)\right. \\ &\left. - \lambda\frac{1}{m^{1/2}}\sum_{i=1}^{m} z_{v,i}'\left(g(x_{v,i},\beta_0) - z_{v,i}\,[E(z'z)]^{-1}\,E(z'g(x,\beta_0))\right)\right\} + o_p(1), \end{aligned} \tag{2.3}$$

where $\lambda = \lim_{n\to\infty}(n/m)^{1/2}$, *which is assumed to be finite.*

Proof: The estimator $\hat{\beta}$ satisfies the first order condition: $\frac{\partial Q_n(\hat{\beta})}{\partial\beta'} = 0$. By a Taylor series expansion or the mean value theorem,

$$\begin{aligned} 0 = &\left(\frac{\partial g'(X_v,\beta_0)}{\partial\beta} Z_v\right)(Z_v'Z_v)^{-1}Z'\left(Y - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\beta_0)\right) \\ &+\left\{\left[\frac{\partial^2 g'(X_v,\bar{\beta})}{\partial\beta\partial\beta_1} Z_v,\cdots,\frac{\partial^2 g'(X_v,\bar{\beta})}{\partial\beta\partial\beta_l} Z_v\right](Z_v'Z_v)^{-1}Z'\left(Y - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\bar{\beta})\right)\right. \\ &\left. -\left(\frac{\partial g'(X_v,\bar{\beta})}{\partial\beta} Z_v\right)(Z_v'Z_v)^{-1}Z'Z(Z_v'Z_v)^{-1}Z_v'\frac{\partial g(X_v,\bar{\beta})}{\partial\beta'}\right\}(\hat{\beta} - \beta_0). \end{aligned}$$

By the uniform law of large numbers,

$$n^{-1}Z'(Y - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\bar{\beta})) \xrightarrow{p} 0; \quad m^{-1}\frac{\partial g'(X_v,\beta_0)}{\partial\beta} Z_v \xrightarrow{p} E\left[\frac{\partial g(X,\beta_0)}{\partial\beta} z\right];$$

$$m^{-1}\frac{\partial^2 g'(X_v,\bar{\beta})}{\partial\beta\partial\beta_j} Z_v \xrightarrow{p} E(\frac{\partial^2 g(x,\beta_0)}{\partial\beta\partial\beta_j} z); \quad n^{-1}Z'Z \xrightarrow{p} E(z'z); \quad m^{-1}Z_v'Z_v \xrightarrow{p} E(z'z).$$

Hence, it follows that

$$\begin{aligned} n^{1/2}(\hat{\beta} - \beta_0) = &\left\{E\left[\frac{\partial g(x,\beta_0)}{\partial\beta} z\right][E(z'z)]^{-1}E\left[z'\frac{\partial g(x,\beta_0)}{\partial\beta'}\right] + o_p(1)\right\}^{-1} \\ &\cdot\left\{E\left[\frac{\partial g(x,\beta_0)}{\partial\beta} z\right][E(z'z)]^{-1} + o_p(1)\right\}\frac{1}{n^{1/2}}Z'(Y - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\beta_0)). \end{aligned}$$

On the other hand,

$$\begin{aligned}
&\frac{1}{n^{1/2}} Z'(Y - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\beta_0)) \\
&\quad = \frac{1}{n^{1/2}} Z' \left(Y - Z[E(z'z)]^{-1}E(z'g(x,\beta_0))\right) \\
&\qquad - \left(\frac{n}{m}\right)^{1/2} \left(\frac{1}{n}Z'Z\right) \left[\frac{1}{m}Z_v'Z_v\right]^{-1} \frac{1}{m^{1/2}} \left(Z_v'g(X_v,\beta_0) - Z_v'Z_v[E(z'z)]^{-1}E(z'g(x,\beta_0))\right).
\end{aligned}$$

As $\lim_{n\to\infty}(n/m)^{1/2} = \lambda$, the result follows. Q.E.D.

The asymptotic distribution of $n^{1/2}(\hat{\beta} - \beta_0)$ consists of two components of disturbances. The first disturbance can be rewritten as

$$\begin{aligned}
&\frac{1}{n^{1/2}} \sum_{i=1}^{n} z_i' \left(y_i - z_i [E(z'z)]^{-1} E(z'y)\right) \\
&\quad = \frac{1}{n^{1/2}} \sum_{i=1}^{n} z_i'\epsilon_i + \frac{1}{n^{1/2}} \sum_{i=1}^{n} z_i' \left(g(x_i,\beta_0) - z_i [E(z'z)]^{-1} E(z'g(x,\beta_0))\right),
\end{aligned}$$

which reflects the structural disturbance in the regression equation and the error in replacing $g(x,\beta)$ by its population prediction $z[E(z'z)]^{-1}E(z'g(x,\beta_0))$. As

$$\begin{aligned}
y &= g(x,\beta_0) + \epsilon \\
&= z[E(z'z)]^{-1}E(z'g(x,\beta_0)) + \tilde{\epsilon},
\end{aligned} \tag{2.4}$$

$\tilde{\epsilon} = \epsilon + g(x,\beta_0) - z[E(z'z)]^{-1}E(z'g(x,\beta_0))$ is the overall error. The first component of disturbances in $n^{1/2}(\hat{\beta} - \beta_0)$ represents the impact of the overall error of the modified equation (2.4) on the coefficient estimate. The second component of disturbances captures the variations of the sample moments $m^{-1}Z_v'Z_v'$ and $m^{-1}Z_v'g(X_v,\beta)$, which are used to replace $E(z'z)$ and $E(z'g(x,\beta_0))$ in (2.4) for estimation.

The asymptotic distribution of $\hat{\beta}$ can be derived from (2.3) by some central limit theorem. The following corollary provides the asymptotic distribution of $\hat{\beta}$ when the validation data are independent of the survey data.

Corollary 2.1: *Under the assumptions in Proposition 2.2, if* (y_i, z_i), $i = 1,\cdots,n$, *are independent of* $(x_{v,j}, z_{v,j})$, $j = 1,\cdots,m$, *then*

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0,\Omega),$$

where

$$\begin{aligned}
\Omega = &\left\{E\left(\frac{\partial g(x,\beta_0)}{\partial\beta}z\right)[E(z'z)]^{-1}E\left(z'\frac{\partial g(x,\beta_0)}{\partial\beta'}\right)\right\}^{-1} E\left(\frac{\partial g(x,\beta_0)}{\partial\beta}z\right)\cdot[E(z'z)]^{-1}\Sigma \\
&\cdot[E(z'z)]^{-1}E\left(z'\frac{\partial g(x,\beta_0)}{\partial\beta'}\right)\left\{E\left(\frac{\partial g(x,\beta_0)}{\partial\beta}z\right)[E(z'z)]^{-1}E\left(z'\frac{\partial g(x,\beta_0)}{\partial\beta'}\right)\right\}^{-1},
\end{aligned}$$

and

$$\begin{aligned}
\Sigma = &E\left\{[z'y - z'z[E(z'z)]^{-1}E(z'y)]\,[z'y - z'z[E(z'z)]^{-1}E(z'y)]'\right\} \\
&+ \lambda^2 E\left\{[z'g(x,\beta_0) - z'z[E(z'z)]^{-1}E(z'g(x,\beta_0))]\,[z'g(x,\beta_0) - z'z[E(z'z)]^{-1}E(z'g(x,\beta_0))]'\right\}.
\end{aligned}$$

Proof: This result is immediate, since the validation data are independent of the survey data. Q.E.D.

If the validation data were correlated with some of the survey data, the covariance of the two components of disturbances in (2.3) should be taken into account in the asymptotic distribution of $\hat{\beta}$. The rate of

convergence in distribution of $\hat{\beta}$ in Proposition 2.2 is of order $O(n^{-1/2})$ under the situation that $\lim_{n\to\infty} m = \infty$, and $\lim_{n\to\infty} n/m$ exists and is finite. In this case, as n and m grow approximately in proportion, the rate of convergence to distribution of $\hat{\beta}$ is $O(n^{-1/2})$. In the event that $\lim_{n\to\infty} m = \infty$ but $\lim_{n\to\infty} m/n = 0$, the projection errors from the validation data will dominate, and the rate of convergence in distribution of $\hat{\beta}$ can only be of order $O(m^{-1/2})$ but not $O(n^{-1/2})$. By a careful inspection of the proof of Proposition 2.2, the following results are immediate.

Corollary 2.2: *Under the assumptions in Proposition 2.2, if* $\lambda = \lim_{n\to\infty} (n/m)^{1/2}$ *exists and is finite,*

$$m^{1/2}(\hat{\beta} - \beta_0) = \left\{ E\left(\frac{\partial g(x,\beta_0)}{\partial \beta} z\right) [E(z'z)]^{-1} E\left(z' \frac{\partial g(x,\beta_0)}{\partial \beta'}\right) \right\}^{-1} E\left(\frac{\partial g(x,\beta_0)}{\partial \beta} z\right) [E(z'z)]^{-1}$$
$$\cdot \left\{ \frac{1}{\lambda n^{1/2}} \sum_{i=1}^{n} z_i' \left(y_i - z_i [E(z'z)]^{-1} E(z'y)\right) \right.$$
$$\left. - \frac{1}{m^{1/2}} \sum_{i=1}^{m} z_{v,i}' \left(g(x_{v,i},\beta_0) - z_{v,i} [E(z'z)]^{-1} E(z'g(x,\beta_0))\right) \right\} + o_p(1).$$

On the other hand, if $\lim_{n\to\infty} (n/m)^{1/2} = \infty$, *then*

$$m^{1/2}(\hat{\beta} - \beta_0) = -\left\{ E\left(\frac{\partial g(x,\beta_0)}{\partial \beta} z\right) [E(z'z)]^{-1} E\left(z' \frac{\partial g(x,\beta_0)}{\partial \beta'}\right) \right\}^{-1} E\left(\frac{\partial g(x,\beta_0)}{\partial \beta} z\right) [E(z'z)]^{-1}$$
$$\cdot \frac{1}{m^{1/2}} \sum_{i=1}^{m} z_{v,i}' \left(g(x_{v,i},\beta_0) - z_{v,i} [E(z'z)]^{-1} E(z'g(x,\beta_0))\right) \right\} + o_p(1).$$

For the linear regression model $y = x\beta + \epsilon$, the above estimation method is applicable. The estimator $\hat{\beta}$ has a closed form expression:

$$\hat{\beta} = [X_v' Z_v (Z_v' Z_v)^{-1} Z' Z (Z_v' Z_v)^{-1} Z_v' X_v]^{-1} X_v' Z_v (Z_v' Z_v)^{-1} Z' Y. \tag{2.5}$$

This estimator is a least squares estimator found by regressing y on z_M, where $z_M = z(Z_v' Z_v)^{-1} Z_v' X_v$ is a transformed vector of z. The transformation matrix $M = (Z_v' Z_v)^{-1} Z_v' X_v$ is the least squares coefficient estimates found by regressing x_v on z_v. The validation data are used only to construct the transformation matrix M. This matrix is the sufficient statistic from the validation data for the linear regression model in this estimation procedure. This is an advantage of this approach when validation data are confidential in the individual level detail. In the linear regression model, since $\hat{\beta}$ has a closed form expression, some of the regularity conditions in the previous propositions can be relaxed for establishing its asymptotic properties.

Proposition 2.3: *LINEAR REGRESSION MODELS Under the Assumptions 1 and 2, suppose that* $E(x'z)$ *has full row rank* k, *then* $\hat{\beta}$ *is a consistent estimator of* β, *and*

$$n^{1/2}(\hat{\beta} - \beta_0)$$
$$= \{E(x'z)[E(z'z)]^{-1}E(z'x)\}^{-1} E(x'z)[E(z'z)]^{-1} \left\{ \frac{1}{n^{1/2}} \sum_{i=1}^{n} z_i' \left(y_i - z_i[E(z'z)]^{-1}E(z'y)\right) \right.$$
$$\left. - \lambda \frac{1}{m^{1/2}} \sum_{i=1}^{m} z_{v,i}' \left(x_{v,i}\beta_0 - z_{v,i}[E(z'z)]^{-1}E(z'x\beta_0)\right) \right\} + o_p(1).$$

Furthermore, if (y_i, z_i), $i = 1, \cdots, n$, *are independent of* $(x_{v,j}, z_{v,j})$, $j = 1, \cdots, m$, *then*

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, \Omega),$$

where

$$\Omega = \{E(x'z)[E(z'z)]^{-1}E(z'x)\}^{-1} E(x'z)[E(z'z)]^{-1}\Sigma[E(z'z)]^{-1}E(z'x) \left\{E(x'z)\,[E(z'z)]^{-1}\,E(z'x)\right\}^{-1},$$

and

$$\Sigma = E\left\{\left[z'y - z'z[E(z'z)]^{-1}E(z'y)\right]\left[z'y - z'z[E(z'z)]^{-1}E(z'y)\right]'\right\}$$
$$+ \lambda^2 E\left\{\left[z'x\beta_0 - z'z[E(z'z)]^{-1}E(z'x\beta_0)\right]\left[z'x\beta_0 - z'z[E(z'z)]^{-1}E(z'x\beta_0)\right]'\right\}.$$

3. Measurement Errors in Explanatory and Dependent Variables

The last section discussed a consistent estimation method when some or all of the explanatory variables are measured with errors, but the dependent variable y is either measured accurately or its measurement error is of the classical type, that is, independent of all the explanatory or instrumental variables. In this section, we will consider the cases where either the dependent variable or both the dependent and explanatory variables are subject to measurement errors of the general type.

Consider the general case that both the dependent and explanatory variables are measured with errors. The survey data are $(\tilde{y}_i, z_i)$, $i = 1, \cdots, n$, where $\tilde{y}$ measures y with error. As in the previous section, z is a vector of instrumental variables for x. Let $(x_{v,j}, z_{v,j})$, $j = 1, \cdots, m$, be the validation data for the explanatory variables, and $(y_{d,j}, \tilde{y}_{d,j}, z_{d,j})$, $j = 1, \cdots, m_d$, be the validation data for the dependent variable. In the later validation data, y_d measures y accurately.

Assumption 3:

(1) The observations $(\tilde{y}_i, z_i)$, $i = 1, \cdots, n$, of the surve y data are i.i.d. with finite first and second moments.
(2) The observations $(y_{d,j}, \tilde{y}_{d,j}, z_{d,j})$, $j = 1, \cdots, m_d$, of the validation data for y are i.i.d. The first two moments of $(y_d, \tilde{y}_d, z_d)$ exist.
(3) (y_d, z_d) and (y, z) are identically distributed.

At least two methods can be introduced to estimate the regression model with this validation information, depending on how we predict the true value y given the variables $\tilde{y}$ and z. The first method is to project the measurement error of the dependent variable, $\tilde{y}_d - y_d$, on to z_d to derive a prediction rule for y_j: $\tilde{y}_j - z_j(Z_d'Z_d)^{-1}Z_d'(\tilde{Y}_d - Y_d)$; where Z_d is the $m_d \times k_1$ data matrix of z_d, and $\tilde{Y}_d$ and Y_d are respectively the m_d-dimensional vectors of $\tilde{y}_d$ and y_d. This prediction rule is based on the intuition that the measurement error $\tilde{y} - y$ might not have zero mean, and it might be correlated with some of the explanatory variables or their instruments. An estimator $\tilde{\beta}_1$ of β can be derived from

$$\min_{\beta \in \Theta}(\hat{Y}_p - Z(Z_v'Z_v)^{-1}Z_v'g(X_v, \beta))'(\hat{Y}_p - Z(Z_v'Z_v)^{-1}Z_v'g(X_v, \beta)), \tag{3.1}$$

where $\hat{Y}_p = \tilde{Y} - Z(Z_d'Z_d)^{-1}Z_d'(\tilde{Y}_d - Y_d)$.

The consistency and the asymptotic distribution of $\tilde{\beta}_1$ can be derived by some uniform law of large numbers and central limit theorems parallel to the proofs of Propositions 2.1 and 2.2 in Section 2.

Proposition 3.1: *Under the Assumptions 1-3, and the assumptions in Proposition 2.1, $\tilde{\beta}_1$ from (3.1) is a consistent estimator of β_0.*

Proof: Denote

$$Q_{d,n} = \frac{1}{n}(\hat{Y}_p - Z(Z_v'Z_v)^{-1}Z_v'g(X_v, \beta))'(\hat{Y}_p - Z(Z_v'Z_v)^{-1}Z_v'g(X_v, \beta)).$$

By the uniform law of large number in Amemiya (1985), $Q_{d,n}(\beta)$ converges in probability to $Q_d(\beta)$ uniformly on Θ, where

$$\begin{aligned}
Q_d(\beta) &= E(\tilde{y}^2) - 2E(\tilde{y}z)[E(z'z)]^{-1}E(z'(\tilde{y} - y)) + E((\tilde{y} - y)z)[E(z'z)]^{-1}E(z'(\tilde{y} - y)) \\
&\quad + E(g(x, \beta)z)[E(z'z)]^{-1}E(z'g(x, \beta)) \\
&\quad - 2E(g(x, \beta)z)[E(z'z)]^{-1}E(z'\tilde{y}) + 2E(g(x, \beta)z)[E(z'z)]^{-1}E(z'(\tilde{y} - y)) \\
&= E((y - g(x, \beta))z)[E(z'z)]^{-1}E(z'(y - g(x, \beta))) + c \\
&= E((g(x, \beta_0) - g(x, \beta))z)[E(z'z)]^{-1}E(z'(g(x, \beta_0) - g(x, \beta))) + c
\end{aligned}$$

since $E(z\epsilon) = 0$, and $c = E(\tilde{y}^2) - E(\tilde{y}z)[E(z'z)]^{-1}E(z'\tilde{y})$ is a constant. Under the identification condition in Proposition 2.1, $Q_d(\beta)$ is uniquely minimized at $\beta = \beta_0$. Q.E.D.

Proposition 3.2: *Under the assumptions in Proposition 3.1 and the assumptions in Proposition 2.2,*

$$
\begin{aligned}
n^{1/2}(\tilde{\beta}_1 - \beta_0) = & \left\{ E\left(\frac{\partial g(x,\beta_0)}{\partial \beta} z\right) [E(z'z)]^{-1} E\left(z' \frac{\partial g(x,\beta_0)}{\partial \beta'}\right) \right\}^{-1} E\left(\frac{\partial g(x,\beta_0)}{\partial \beta} z\right) [E(z'z)]^{-1} \\
& \cdot \left\{ \frac{1}{n^{1/2}} \sum_{i=1}^{n} z_i' \left(\tilde{y}_i - z_i [E(z'z)]^{-1} E(z'\tilde{y})\right) \right. \\
& + \lambda_d \frac{1}{m_d^{1/2}} \sum_{i=1}^{m_d} z_{d,i}' \left(y_{d,i} - \tilde{y}_{d,i} - z_{d,i} [E(z'z)]^{-1} E(z'(y - \tilde{y}))\right) \\
& \left. - \lambda \frac{1}{m^{1/2}} \sum_{i=1}^{m} z_{v,i}' \left(g(x_{v,i},\beta_0) - z_{v,i} [E(z'z)]^{-1} E(z' g(x,\beta_0))\right) \right\} + o_p(1),
\end{aligned}
$$

where $\lambda = \lim_{n\to\infty} (n/m)^{1/2}$ *and* $\lambda_d = \lim_{n\to\infty} (n/m_d)^{1/2}$, *which are assumed to be finite.*

Proof: The estimator $\tilde{\beta}$ satisfies the first order condition: $\frac{\partial Q_{d,n}(\tilde{\beta})}{\partial \beta} = 0$. By a Taylor series expansion or the mean value theorem,

$$
\begin{aligned}
0 = & \left(\frac{\partial g'(X_v,\beta_0)}{\partial \beta} Z_v\right) (Z_v'Z_v)^{-1} Z' \left(\hat{Y}_p - Z(Z_v'Z_v)^{-1} Z_v' g(X_v,\beta_0)\right) \\
& + \left\{ \left[\frac{\partial^2 g'(X_v,\bar{\beta})}{\partial\beta\partial\beta_1} Z_v, \cdots, \frac{\partial^2 g'(X_v,\bar{\beta})}{\partial\beta\partial\beta_l} Z_v\right] (Z_v'Z_v)^{-1} Z' \left(\hat{Y}_p - Z(Z_v'Z_v)^{-1} Z_v' g(X_v,\bar{\beta})\right) \right. \\
& \left. - \left(\frac{\partial g'(X_v,\beta_0)}{\partial \beta} Z_v\right) (Z_v'Z_v)^{-1} Z'Z(Z_v'Z_v)^{-1} Z_v' \frac{\partial g(X_v,\bar{\beta})}{\partial \beta'} \right\} (\tilde{\beta} - \beta_0).
\end{aligned}
$$

Since $E(zy) = E(zg(x,\beta_0))$,

$$
\begin{aligned}
& \frac{1}{n^{1/2}} Z'(\hat{Y}_p - Z(Z_v'Z_v)^{-1} Z_v' g(X_v,\beta_0)) \\
& = \frac{1}{n^{1/2}} Z'(\tilde{Y} - Z[E(z'z)]^{-1} E(z'\tilde{y}) + \frac{1}{n^{1/2}} Z'Z(Z_d'Z_d)^{-1} Z_d'(Y_d - \tilde{Y}_d - Z_d[E(z'z)]^{-1} E(z'(y - \tilde{y}))) \\
& \quad - \frac{1}{n^{1/2}} Z'Z(Z_v'Z_v)^{-1} Z_v'(g(X_v,\beta_0) - Z_v[E(z'z)]^{-1} E(z' g(x,\beta_0))).
\end{aligned}
$$

The results follow from similar arguments in the proof of Proposition 2.2. Q.E.D.

The asymptotic distribution of $\tilde{\beta}_1$ depends on three different projection errors. These errors are mutually independent if the two validation sets come from different sources and are independent of the survey sample. In some situations, the validation data for both the independent and dependent variables might come from one validation study. The following corollary covers these two cases.

Corollary 3.1: *Under the assumptions in Proposition 3.2, if* $(\tilde{y}_i, z_i)$, $i = 1, \cdots, n$, $(x_{v,j}, z_{v,j})$, $j = 1, \cdots, m$, *and* $(y_{d,j}, \tilde{y}_{d,j}, z_{d,j})$, $j = 1, \cdots, m_d$ *are mutually independent samples, then*

$$
n^{1/2}(\tilde{\beta}_1 - \beta_0) \xrightarrow{D} N(0, \Omega),
$$

where

$$
\begin{aligned}
\Omega = & \left\{ E\left(\frac{\partial g(x,\beta_0)}{\partial \beta} z\right) [E(z'z)]^{-1} E\left(z' \frac{\partial g(x,\beta_0)}{\partial \beta'}\right) \right\}^{-1} E\left(\frac{\partial g(x,\beta_0)}{\partial \beta} z\right) [E(z'z)]^{-1} \Sigma \\
& \cdot [E(z'z)]^{-1} E\left(z' \frac{\partial g(x,\beta_0)}{\partial \beta'}\right) \left\{ E\left(\frac{\partial g(x,\beta_0)}{\partial \beta} z\right) [E(z'z)]^{-1} E\left(z' \frac{\partial g(x,\beta_0)}{\partial \beta'}\right) \right\}^{-1},
\end{aligned}
$$

and

$$\Sigma = E\left\{\left[z'\tilde{y} - z'z[E(z'z)]^{-1}E(z'\tilde{y})\right]\left[z'\tilde{y} - z'z[E(z'z)]^{-1}E(z'\tilde{y})\right]'\right\}$$
$$+ \lambda_d^2 E\left\{\left[z'(y-\tilde{y}) - z'z[E(z'z)]^{-1}E(z'(y-\tilde{y}))\right]\left[z'(y-\tilde{y}) - z'z[E(z'z)]^{-1}E(z'(y-\tilde{y}))\right]'\right\}$$
$$+ \lambda^2 E\left\{\left[z'g(x,\beta_0) - z'z[E(z'z)]^{-1}E(z'g(x,\beta_0))\right]\left[z'g(x,\beta_0) - z'z[E(z'z)]^{-1}E(z'g(x,\beta_0))\right]'\right\}.$$

However, if $(x_{v,j}, z_{v,j})$, $j = 1, \cdots, m$, *and* $(y_{d,j}, \tilde{y}_{d,j}, z_{d,j})$, $j = 1, \cdots, m_d$, *are independent of* $(\tilde{y}_i, z_i)$, $i = 1, \cdots, n$, *but the validation data come from the same source; i.e.,* $m_d = m$ *and* $z_{v,j} = z_{d,j}$, *for all* j, *then*

$$\Sigma = E\left\{\left[z'\tilde{y} - z'z[E(z'z)]^{-1}E(z'\tilde{y})\right]\left[\tilde{y}z - E(\tilde{y}z)[E(z'z)]^{-1}z'z\right]\right\}$$
$$+ \lambda^2 E\left\{\left[z'\epsilon - z'\tilde{y} + z'z[E(z'z)]^{-1}E(z'\tilde{y})\right]\left[z'\epsilon - z'\tilde{y} + z'z[E(z'z)]^{-1}E(z'\tilde{y})\right]'\right\}.$$

Proof: When the validation data come from the same source,

$$\lambda_d \frac{1}{m_d^{1/2}} \sum_{i=1}^{m_d} z'_{d,i}\left(y_{d,i} - \tilde{y}_{d,i} - z_{d,i}\left[E(z'z)\right]^{-1}E(z'(y-\tilde{y}))\right)$$
$$- \lambda \frac{1}{m^{1/2}} \sum_{i=1}^{m} z'_{v,i}\left(g(x_{v,i},\beta_0) - z_{v,i}\left[E(z'z)\right]^{-1}E(z'g(x,\beta_0))\right)$$
$$= \lambda \frac{1}{m^{1/2}} \sum_{i=1}^{m} z'_{d,i}\left(\epsilon_{d,i} - \tilde{y}_{d,i} + z_{d,i}\left[E(z'z)\right]^{-1}E(z'\tilde{y})\right),$$

because $E(z\epsilon) = 0$. Q.E.D.

The projection of the measurement error of y, $\tilde{y} - y$, on z in the above method corrects the possible systematic bias of error, which might be explained by the observed instruments z. In some more complicated situation, the measurement error $\tilde{y} - y$ might even be correlated with $\tilde{y}$ in addition to z. In such a situation, it might be appealing to project $\tilde{y} - y$ on $\tilde{y}$, or even some finite order polynomial of $\tilde{y}$, and z in order to correct the systematic bias. This motivates the second estimation method. Let $p = (\tilde{y}, \tilde{y}^2, \cdots, \tilde{y}^r)$, where $r \geq 1$ is the specified order of the polynomial; and $w = (p, z)$. Let W_d denote the $m_d \times (r + k_1)$ matrix of the validation data of w, and let W be the $n \times (r + k_1)$ matrix of the survey data of w. An estimator $\tilde{\beta}_2$ of β_0 is defined from the following minimization:

$$\min_{\beta \in \Theta}(W(W_d'W_d)^{-1}W_d'Y_d - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\beta))'(W(W_d'W_d)^{-1}W_d'Y_d - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\beta)). \quad (3.2)$$

Assumption 4:
(1) The $2r$ moments of $\tilde{y}$ exist.
(2) $E(w'w)$, where $w = (p, z)$, is nonsingular.

Proposition 3.3: *Under the Assumptions 1-4, and the assumptions in Proposition 2.1,* $\tilde{\beta}_2$ *is a consistent estimator of* β_0.

Proof: Denote

$$Q_{w,n}(\beta) = \frac{1}{n}(W(W_d'W_d)^{-1}W_d'Y_d - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\beta))'(W(W_d'W_d)^{-1}W_d'Y_d - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\beta)).$$

$Q_{w,n}(\beta)$ converges in probability to $Q_w(\beta)$ uniformly in β, where

$$\begin{aligned} & Q_w(\beta) \\ &= E(yw)[E(w'w)]^{-1}E(w'y) - 2E(yw)[E(w'w)]^{-1}E(w'z)[E(z'z)]^{-1}E(z'g(x,\beta)) \\ &\quad + E(g(x,\beta)z)[E(z'z)]^{-1}E(z'g(x,\beta)) \\ &= \left(E(g(x,\beta)z) - E(yw)[E(w'w)]^{-1}E(w'z)\right)[E(z'z)]^{-1}\left(E(g(x,\beta)z) - E(yw)[E(w'w)]^{-1}E(w'z)\right)' + c \\ &= \left(E(g(x,\beta)z) - E(g(x,\beta_0)z)\right)[E(z'z)]^{-1}\left(E(g(x,\beta)z) - E(g(x,\beta_0)z)\right)' + c, \end{aligned}$$

and $c = E(yw)[E(w'w)]^{-1}\left(E(w'w) - E(w'z)[E(z'z)]^{-1}E(z'w)\right)[E(w'w)]^{-1}E(w'y)$ is a constant. The last equality holds because $w = (p, z)$ implies that $E(yw)[E(w'w)]^{-1}E(w'z) = E(yz)$ and $E(yz) = E(g(x, \beta_0)z)$. As β_0 minimizes $Q_w(\beta)$, the consistency of $\tilde{\beta}_2$ follows from the identification condition of Proposition 2.1. Q.E.D.

Proposition 3.4: *Under the assumptions in Proposition 3.3 and the assumptions in Proposition 2.2,*

$$\begin{aligned} n^{1/2}(\tilde{\beta}_2 - \beta_0) = &\left\{E\left(\frac{\partial g(x,\beta_0)}{\partial \beta}z\right)[E(z'z)]^{-1}E\left(z'\frac{\partial g(x,\beta_0)}{\partial \beta'}\right)\right\}^{-1}E\left(\frac{\partial g(x,\beta_0)}{\partial \beta}z\right)[E(z'z)]^{-1} \\ &\cdot\left\{\frac{1}{n^{1/2}}\sum_{i=1}^{n}z_i'\left(w_i[E(w'w)]^{-1}E(w'y) - z_i[E(z'z)]^{-1}E(z'y)\right)\right. \\ &+\lambda_d\frac{1}{m_d^{1/2}}\sum_{i=1}^{m_d}z_{d,i}'\left(y_{d,i} - w_{d,i}[E(w'w)]^{-1}E(w'y)\right) \\ &\left.-\lambda\frac{1}{m^{1/2}}\sum_{i=1}^{m}z_{v,i}'\left(g(x_{v,i},\beta_0) - z_{v,i}[E(z'z)]^{-1}E(z'g(x_{v,i},\beta_0))\right)\right\} + o_p(1), \end{aligned}$$

where $\lambda = \lim_{n\to\infty}(n/m)^{1/2}$ *and* $\lambda_d = \lim_{n\to\infty}(1/m_d)^{1/2}$.

Proof: By a Taylor series expansion and a uniform law of large numbers,

$$\begin{aligned} n^{1/2}(\tilde{\beta}_2 - \beta_0) = &\left\{\frac{1}{n}\frac{\partial g(X_v,\beta_0)}{\partial \beta}Z_v(Z_v'Z_v)^{-1}Z'Z(Z_v'Z_v)^{-1}Z_v'\frac{\partial g(X_v,\beta_0)}{\partial \beta'} + o_p(1)\right\} \\ &\cdot\frac{1}{n^{1/2}}\frac{\partial g(X_v,\beta_0)}{\partial \beta}Z_v(Z_v'Z_v)^{-1}Z'\left(W(W_d'W_d)^{-1}W_d'Y_d - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\beta_0)\right). \end{aligned}$$

Since

$$\begin{aligned} &\frac{1}{n^{1/2}}Z'\left(W(W_d'W_d)^{-1}W_d'Y_d - Z(Z_v'Z_v)^{-1}Z_v'g(X_v,\beta_0)\right) \\ &= \frac{1}{n^{1/2}}Z'\left\{W[E(w'w)]^{-1}E(w'y) - Z[E(z'z)]^{-1}E(z'y)\right\} \\ &\quad+\frac{1}{n^{1/2}}Z'W(W_d'W_d)^{-1}W_d'\left\{Y_d - W_d[E(w'w)]^{-1}E(w'y)\right\} \\ &\quad-\frac{1}{n^{1/2}}Z'Z(Z_v'Z_v)^{-1}Z_v'\left\{g(X_v,\beta_0) - Z_v[E(z'z)]^{-1}E(z'y)\right\} \\ &= \frac{1}{n^{1/2}}Z'\left\{W[E(w'w)]^{-1}E(w'y) - Z[E(z'z)]^{-1}E(z'y)\right\} \\ &\quad+\lambda_d E(z'w)[E(w'w)]^{-1}\frac{1}{m_d^{1/2}}W_d'\left\{Y_d - W_d[E(w'w)]^{-1}E(w'y)\right\} \\ &\quad-\lambda E(z'z)[E(z'z)]^{-1}\frac{1}{m^{1/2}}Z_v'\left\{g(X_v,\beta_0) - Z_v[E(z'z)]^{-1}E(z'y)\right\} + o_p(1) \\ &= \frac{1}{n^{1/2}}Z'\left\{W[E(w'w)]^{-1}E(w'y) - Z[E(z'z)]^{-1}E(z'y)\right\} \\ &\quad+\lambda_d\frac{1}{m_d^{1/2}}Z_d'\left\{Y_d - W_d[E(w'w)]^{-1}E(w'y)\right\} - \lambda\frac{1}{m^{1/2}}Z_v'\left\{g(X_v,\beta_0) - Z_v[E(z'z)]^{-1}E(z'y)\right\} + o_p(1), \end{aligned}$$

by using the property $E(z'w)[E(w'w)]^{-1} = [0, I]$, the result follows from these relations. Q.E.D.

Corollary 3.2: *Under the assumptions in Proposition 3.4, if* $(\tilde{y}_i, z_i)$, $i = 1, \cdots, n$, $(x_{v,j}, z_{v,j})$, $j = 1, \cdots, m$, *and* $(y_{d,j}, \tilde{y}_{d,j}, z_{d,j})$, $j = 1, \cdots, m_d$ *are mutually independent samples, then*

$$n^{1/2}(\tilde{\beta}_2 - \beta_0) \xrightarrow{D} N(0, \Omega),$$

where

$$\Omega = \left\{ E\left(\frac{\partial g(x,\beta_0)}{\partial \beta} z\right) [E(z'z)]^{-1} E\left(z' \frac{\partial g(x,\beta_0)}{\partial \beta'}\right) \right\}^{-1} E\left(\frac{\partial g(x,\beta_0)}{\partial \beta} z\right) [E(z'z)]^{-1} \Sigma$$
$$[E(z'z)]^{-1} E\left(z' \frac{\partial g(x,\beta_0)}{\partial \beta'}\right) \left\{ E\left(\frac{\partial g(x,\beta_0)}{\partial \beta} z\right) [E(z'z)]^{-1} E\left(z' \frac{\partial g(x,\beta_0)}{\partial \beta'}\right) \right\}^{-1},$$

and

$$\begin{aligned}
\Sigma = {} & E\left\{ \left[z'w\,[E(w'w)]^{-1} E(w'\tilde{y}) - z'z[E(z'z)]^{-1} E(z'\tilde{y}) \right] \right. \\
& \left. \cdot \left[z'w\,[E(w'w)]^{-1} E(w'\tilde{y}) - z'z[E(z'z)]^{-1} E(z'\tilde{y}) \right]' \right\} \\
& + \lambda_d^2 E\left\{ \left[z'y - z'w[E(w'w)]^{-1} E(w'y) \right] \left[z'y - z'w[E(w'w)]^{-1} E(w'y) \right]' \right\} \\
& + \lambda^2 E\left\{ \left[z'g(x,\beta_0) - z'z[E(z'z)]^{-1} E(z'g(x,\beta_0)) \right] \left[z'g(x,\beta_0) - z'z[E(z'z)]^{-1} E(z'g(x,\beta_0)) \right]' \right\}.
\end{aligned}$$

However, if $(x_{v,j}, z_{v,j})$, $j = 1, \cdots, m$, *and* $(y_{d,j}, \tilde{y}_{d,j}, z_{d,j})$, $j = 1, \cdots, m_d$, *are independent of* $(\tilde{y}_i, z_i)$, $i = 1, \cdots, n$, *but the validation data come from the same source; i.e.,* $m_d = m$ *and* $z_{v,j} = z_{d,j}$, *for all* j, *then*

$$\begin{aligned}
\Sigma = {} & E\left\{ \left[z'w\,[E(w'w)]^{-1} E(w'\tilde{y}) - z'z[E(z'z)]^{-1} E(z'\tilde{y}) \right] \right. \\
& \left. \cdot \left[z'w\,[E(w'w)]^{-1} E(w'\tilde{y}) - z'z[E(z'z)]^{-1} E(z'\tilde{y}) \right]' \right\} \\
& + \lambda^2 E\left\{ \left[z'\epsilon - z'w[E(w'w)]^{-1} E(w'y) + z'z[E(z'z)]^{-1} E(z'y) \right] \right. \\
& \left. \cdot \left[z'\epsilon - z'w[E(w'w)]^{-1} E(w'y) + z'z[E(z'z)]^{-1} E(z'y) \right]' \right\}.
\end{aligned}$$

It is interesting to compare the two different estimators. From Proposition 3.2 and Proposition 3.4, the asymptotic distributions of $\tilde{\beta}_1$ and $\tilde{\beta}_2$ differ from each other only in the terms:

$$\lambda_d \frac{1}{m_d^{1/2}} \sum_{i=1}^{m_d} z'_{d,i} \left(y_{d,i} - \tilde{y}_{d,i} - z_{d,i} [E(z'z)]^{-1} E(z'(y - \tilde{y})) \right)$$

in $n^{1/2}(\tilde{\beta}_1 - \beta_0)$, but

$$\frac{1}{m_d^{1/2}} \sum_{i=1}^{m_d} z'_{d,i} \left(y_{d,i} - w_{d,i} [E(w'w)]^{-1} E(w'y) \right)$$

in $n^{1/2}(\tilde{\beta}_2 - \beta_0)$. However, these terms have some similarity. Since $E(\tilde{y}w)[E(w'w)]^{-1} = (1, 0)$, the latter term can be rewritten as

$$\begin{aligned}
& \frac{1}{m_d^{1/2}} \sum_{i=1}^{m_d} z'_{d,i} \left(y_{d,i} - w_{d,i} [E(w'w)]^{-1} E(w'y) \right) \\
= {} & \frac{1}{m_d^{1/2}} \sum_{i=1}^{m_d} z'_{d,i} \left(y_{d,i} - w_{d,i} [E(w'w)]^{-1} [E(w'(y - \tilde{y})) + E(w'\tilde{y})] \right) \\
= {} & \frac{1}{m_d^{1/2}} \sum_{i=1}^{m_d} z'_{d,i} \left(y_{d,i} - \tilde{y}_{d,i} - w_{d,i} [E(w'w)]^{-1} E(w'(y - \tilde{y})) \right).
\end{aligned}$$

The measurement error $y - \tilde{y}$ is projected to w in (3.2), but it is projected only to z in (3.1). If the projection errors $y_{d,i} - \tilde{y}_{d,i} - w_{d,i} [E(w'w)]^{-1} E(w'(y - \tilde{y}))$ and $y_{d,i} - \tilde{y}_{d,i} - z_{d,i} [E(z'z)]^{-1} E(z'(y - \tilde{y}))$ were conditionally independent of z_d, then the first projection error would have smaller variance than the former term, and $\tilde{\beta}_2$

would be relatively more efficient than $\tilde{\beta}_1$. However, the projection errors are only uncorrelated with z_d; therefore, analytic comparisons of efficiency for the two estimators do not seem to be possible.

For the linear regression model $y = x\beta + \epsilon$, both estimators are least squares estimators applied to some transformed survey data. The estimator $\tilde{\beta}_1$ from the first approach is

$$\tilde{\beta}_1 = \left[X_v' Z_v (Z_v' Z_v)^{-1} Z' Z (Z_v' Z_v)^{-1} Z_v' X_v\right]^{-1} X_v' Z_v (Z_v' Z_v)^{-1} Z' \left(\tilde{Y} - Z(Z_d' Z_d)^{-1} Z_d' (\tilde{Y}_d - Y_d)\right). \tag{3.3}$$

Here z is transformed to z_M by the transformation $(Z_v' Z_v)^{-1} Z_v' X_v$, and $\tilde{y}$ is adjusted by subtracting from it a linear combination of z, namely, $z(Z_d' Z_d)^{-1} Z_d' (\tilde{Y}_d - Y_d)$. The coefficients of the linear combination are simply the least squares regression coefficient of $\tilde{y}_d - y_d$ on z_d. The least squares coefficients of x_v on z_v and the least squares coefficients of $\tilde{y}_d - y_d$ on z_d are the sufficient statistics from the validation data for these approaches. For the second approach, the estimator $\tilde{\beta}_2$ becomes

$$\tilde{\beta}_2 = \left[X_v' Z_v (Z_v' Z_v)^{-1} Z' Z (Z_v' Z_v)^{-1} Z_v' X_v\right]^{-1} X_v' Z_v (Z_v' Z_v)^{-1} Z' W (W_d' W_d)^{-1} W_d' Y_d. \tag{3.4}$$

A linear combination of w is used as the dependent variable, where the linear coefficients are the least squares estimates of y_d on w_d. In this approach $\tilde{\beta}_2$ is the ordinary least squares estimate of regressing this linear combination of w on z_M.

Finally, let us comment on the case where only the dependent variable is subject to a general measurement error, but the explanatory variables x are measured without error. For such a case, one might attempt to modify the first estimation method to

$$\min_{\beta \in \Theta} (\hat{Y}_p - g(X, \beta))' (\hat{Y}_p - g(X, \beta)),$$

and the second estimation method to

$$\min_{\beta \in \Theta} (W (W_d' W_d)^{-1} W_d' Y_d - g(X, \beta))' (W (W_d' W_d)^{-1} W_d' Y_d - g(X, \beta)).$$

However, these methods might not necessarily be consistent if $g(x, \beta)$ is not linear in x. Consider, for example, the first modified estimation method. Let $\bar{Q}_{d,n} = n^{-1} (\hat{Y}_p - g(X, \beta))' (\hat{Y}_p - g(X, \beta))$, which can be rewritten as

$$\bar{Q}_{d,n} = \frac{1}{n} [(\hat{Y}_p - Y) + (Y - g(X, \beta))]' [(\hat{Y}_p - Y) + (Y - g(X, \beta))],$$

which will converge in probability to $\bar{Q}_d$, where

$$\begin{aligned} \bar{Q}_d = {} & E(\tilde{y}^2) + E(yz)[E(z'z)]^{-1} E(z'y) - E(\tilde{y}z)[E(z'z)]^{-1} E(z'\tilde{y}) \\ & + E(y - g(x, \beta))^2 + 2E[(\tilde{y} - y)(y - g(x, \beta))] - 2E[(\tilde{y} - y)z][E(z'z)]^{-1} E[z'(y - g(x, \beta))]. \end{aligned}$$

From this expression, it is not obvious that β_0 will minimize $\bar{Q}_d$. Another way to understand this possible inconsistency is that, while the prediction error $(\tilde{y} - y) - z[E(z'z)]^{-1} E(z'(\tilde{y} - y))$ is perpendicular to x when z contains x, it is not necessarily perpendicular to $g(x, \beta_0) - g(x\beta)$. If projection were replaced by conditional expectation, this difference would disappear. The conditional expectation approach is possible only within a nonparametric framework. To guarantee consistent estimates, it is desirable to project $g(x, \beta)$ on z and use the projected regression for estimation. Since X in the survey data are observable in this case, the instrument variables z should include x or some of its finite order polynomials for such a projection. Explicitly, the X_v and Z_v in the previous estimation methods can be X and Z from the survey data, or X_d and Z_d from the validation data. It is also possible to pool X from the survey data and X_d from the validation data to form X_v. Similarly, Z and Z_d can be pooled to form Z_v. The asymptotic distributions in Propositions 3.2 and 3.4 are valid, but the explicit asymptotic covariance should take into account the various correlations as the variables are drawn from some common sources. The linear regression is a special case. For the linear regression model, as $Z = X$ and $Z_v = X_v$, $X(Z_v' Z_v)^{-1} Z_v' (X_v \beta) = X\beta$ and Z_v is irrelevant in the prediction of $X\beta$ given X.

4. Some Related Issues

In this section, we will discuss some related issues on the estimation problem with validation data.

4.1 Pooling Estimators from Survey and Validation Data

If a validation data set is available for all the explanatory and dependent variables x and y, it is possible to estimate the unknown parameters in the regression model with the validation data. Let $(y_{v,i}, \tilde{y}_{v,i}, x_{v,i}, z_{v,i})$, $i = 1, \cdots, m$, be the validation data. Since x_v and y_v measure x and y without errors, β_0 can be estimated by the nonlinear least squares method:

$$\min_{\beta \in \Theta} (Y_v - g(X_v, \beta))'(Y_v - g(X_v, \beta)). \tag{4.1.1}$$

The nonlinear least squares estimator $\hat{\beta}_{NL}$ from (4.1.1) will have the following familiar asymptotic property:

$$m^{1/2}(\hat{\beta}_{NL} - \beta_0) = \left\{ E\left(\frac{\partial g(x, \beta_0)}{\partial \beta} \frac{\partial g(x, \beta_0)}{\partial \beta'} \right) \right\}^{-1} \frac{1}{m^{1/2}} \sum_{i=1}^{m} \frac{\partial g(x_{v,i}, \beta_0)}{\partial \beta} \epsilon_{v,i} + o_p(1). \tag{4.1.2}$$

This estimator $\hat{\beta}_{NL}$ and the relevant estimator introduced in the previous sections can be pooled together to provide an asymptotically more efficient estimator (Theil and Goldberger (1961)). As a by-product, the pooling procedure also provides a compatibility test statistic; which can be used to test model misspecification and the compatibility of the validation data with the survey data. Let V be the limiting variance-covariance matrix of $m^{1/2}(\hat{\beta}'_{NL} - \beta'_0, \hat{\beta}' - \beta'_0)'$ where $\hat{\beta}$ is a consistent estimator from the previous sections. Let $\hat{V}$ be a consistent estimate of V, and let $J = [I_k, I_k]'$, where I_k is a k-dimensional identity matrix. A pooled estimator of β is $\hat{\beta}_G = \left(J'\hat{V}^{-1}J\right)^{-1} J'\hat{V}^{-1}(\hat{\beta}'_{NL}, \hat{\beta}')'$, which is asymptotically normal with mean β_0 and asymptotic covariance $m^{-1}\left(J'V^{-1}J\right)^{-1}$. The pooling procedure is a minimum distance procedure. As a minimum distance procedure,

$$m(\hat{\beta}'_{NL}, \hat{\beta}') \left(I_{2k} - J\left[J'\hat{V}^{-1}J\right]^{-1} J'\hat{V}^{-1} \right)' \hat{V}^{-1} \left(I_{2k} - J\left[J'\hat{V}^{-1}J\right]^{-1} J'\hat{V}^{-1} \right)' (\hat{\beta}'_{NL}, \hat{\beta}')'$$

is asymptotically distributed chi-square with k degrees of freedom (see, e.g., Neyman (1949) or Taylor (1953)). This test statistic provides an indirect test of the compatibility of the validation data and the survey data when the regression model is correctly specified. It is possible that the regression model is misspecified, and the test statistic may pick up this misspecification rather than the compatibility of the data. In any case, it is useful for a diagnostic check of the data and the model. An alternative test will be some generalized Hausman type test by comparing the difference of two consistent estimators.

For the case where only the explanatory variables are subject to measurement errors, $m^{1/2}(\hat{\beta}_{NL} - \beta_0)$ could be asymptotically uncorrelated with $m^{1/2}(\hat{\beta} - \beta_0)$ in Corollary 2.2 if $E(\epsilon|x, z) = 0$. The pooled estimator $\hat{\beta}_G$ is a weighted average of $\hat{\beta}_{NL}$ and $\hat{\beta}$ with weights proportional to the inverses of their variances. For the case that both the explanatory and dependent variables are measured with errors and the validation data come from the same source, if the projection error $y_v - \tilde{y}_v - z_v[E(z'z)]^{-1}E(z'(y - \tilde{y}))$ in Proposition 3.2 is distributed independently of z_v and ϵ_v, $\hat{\beta}_{NL}$ will be asymptotically independent of $\tilde{\beta}_1$. Similarly, if the projection error $y_v - w_v[E(w'w)]^{-1}E(w'y)$ in Proposition 3.4 is distributed independently of z_v and ϵ_v, $\hat{\beta}_{NL}$ will be asymptotically independent of $\tilde{\beta}_2$. Pooling the survey data and validation data will improve efficiency of the estimators.

4.2 Asymptotic Efficiency

The proposed estimators in Sections 2 and 3 have utilized only correlation properties of the true variables and the instrumental variables in the model. They are robust with respect to specification of structural relations of measurement errors. For some special measurement error structure, they might also be asymptotically efficient. Consider the case where only the explanatory variables are subject to measurement errors in the linear regression model

$$y = x\beta + \epsilon, \tag{4.2.1}$$

where ϵ is independent of x and is normally distributed $N(0,\sigma_\epsilon^2)$. Suppose the vector of instruments z has the same dimension as x, and the measurement error has the following structure:

$$x = z\Pi + u, \tag{4.2.2}$$

where u is independent of z and is normally distributed $N(0,\sigma_u^2)$. Assume that the survey data (y_i, z_i), $i = 1,\cdots,n$, and the validation data $(x_{v,i}, z_{v,i})$ are independent. It follows from (4.2.1) and (4.2.2) that

$$y = z\Pi\beta + w, \tag{4.2.3}$$

where $w = \epsilon + u\beta$. Let $\pi_1 = \Pi\beta$. With the validation data, Π can be estimated from (4.2.2) by the ordinary least squares procedure. Similarly, π_1 can be estimated with the survey data. These estimates are apparently the maximum likelihood estimates found by pooling the validation and survey data. Let $\hat{\Pi}_m$ and $\hat{\pi}_{1,m}$ denote these two estimates. Since there is a one-to-one correspondence between (Π, π_1) and (Π, β), the maximum likelihood estimator of β is $\hat{\beta}_m = \hat{\Pi}_m^{-1}\hat{\pi}_{1,m}$. Our proposed estimator $\hat{\beta}$ in Section 2 is

$$\begin{aligned}\hat{\beta} &= \left[X_v'Z_v(Z_v'Z_v)^{-1}Z'Z(Z_v'Z_v)^{-1}Z_v'X_v\right]^{-1}X_v'Z_v(Z_v'Z_v)^{-1}Z'Y \\ &= \left[(Z\hat{\Pi}_m)'(Z\hat{\Pi}_m)\right]^{-1}(Z\hat{\Pi}_m)'Y \\ &= \hat{\Pi}_m^{-1}\hat{\pi}_{1,m},\end{aligned}$$

which is exactly the maximum likelihood estimate $\hat{\beta}_m$ of this model. Therefore, $\hat{\beta}$ is asymptotically efficient for this linear regression model with the specific measurement error structure (4.2.2). However, in general, it is not hard to imagine for many other specific measurement error structures on x and z, $\hat{\beta}$ would not be efficient. The proposed estimators in Sections 3 and 4 are not designed for the estimation of structural measurement error models.

4.3 Bias Correction in Linear Regression Models

For the linear regression model $y = x\beta + \epsilon$, when the variables x and/or y are subject to general measurement errors in survey data, the least squares estimate $\hat{\beta}_{LS}$ with the survey data will be inconsistent. If validation data are available for all the variables x and y, it is possible to use the validation data information to correct the asymptotic bias of $\hat{\beta}_{LS}$ (see, for example, Bound et al (1989). In this section, we point out the similarities and differences of our estimation procedures from a direct bias correction procedure in a linear regression model. For nonlinear regression models, bias correction is either impossible or difficult.

Let $(\tilde{y}_i, z_i)$, $i = 1,\cdots,n$, be the survey data, and $(y_{v,j}, x_{v,j}, \tilde{y}_{v,j}, z_{v,j})$, $j = 1,\cdots,m$, be the validation data. Let $\tilde{Y}$, Z, Y_v, X_v, $\tilde{Y}_v$, and Z_v be the corresponding data matrices. Assuming that z measures x, and z and x have the same dimension, the ordinary least squares estimator with the survey data is

$$\begin{aligned}\hat{\beta}_{LS} &= (Z'Z)^{-1}Z'\tilde{Y} \\ &= \beta_0 + (Z'Z)^{-1}Z'(X - Z)\beta_0 + (Z'Z)^{-1}Z'(\tilde{Y} - Y) + (Z'Z)^{-1}Z'E,\end{aligned} \tag{4.3.1}$$

where Y, X, and E are the unobserved true matrices of y, x and ϵ. The asymptotic bias of $\hat{\beta}_{LS}$ is Δ, where

$$\Delta = [E(z'z)]^{-1}\{E(z'x - z'z)\beta_0 + E(z'\tilde{y} - z'y)\}. \tag{4.3.2}$$

As the validation data provide enough information, the asymptotic bias in $\hat{\beta}_{LS}$ can be corrected.

Consider the case that only the explanatory variables are measured with error. For this case, $\tilde{Y} = Y$ and $\Delta = [E(z'z)]^{-1}E(z'x - z'z)\beta_0$. Therefore, β_0 can be consistently estimated by $\hat{\beta}_v = (X_v'X_v)^{-1}X_v'Y_v$, and Δ can be estimated by

$$\Delta_n = (Z_v'Z_v)^{-1}Z_v'(X_v - Z_v)\hat{\beta}_v.$$

This bias can be subtracted from $\hat{\beta}_{LS}$ to derive a bias adjusted estimator: $\hat{\beta}_c = \hat{\beta}_{LS} - \Delta_n$. The adjusted estimator $\hat{\beta}_c$ is consistent. Since

$$\begin{aligned}\hat{\beta}_c - \beta_0 = {} & (Z'Z)^{-1}Z'\left(Y - Z[E(z'z)]^{-1}E(z'x\beta_0)\right) \\ & - (Z_v'Z_v)^{-1}Z_v'\left(X_v\beta_0 - Z_v[E(z'z)]^{-1}E(z'x\beta_0)\right) - (Z_v'Z_v)^{-1}Z_v'(X_v - Z_v)(X_v'X_v)^{-1}X_v'\epsilon_v,\end{aligned}$$

$$
\begin{aligned}
&n^{1/2}(\hat{\beta}_c - \beta_0) \\
&= [E(z'z)]^{-1}\Big\{\frac{1}{n^{1/2}}\sum_{i=1}^{n} z_i'\left(y_i - z_i[E(z'z)]^{-1}E(z'y)\right) - \lambda\frac{1}{m^{1/2}}\sum_{i=1}^{m} z_{v,i}'\left(x_{v,i}\beta_0 - z_{v,i}[E(z'z)]^{-1}E(z'x\beta_0)\right)\Big\} \\
&\quad - \lambda[E(z'z)]^{-1}(E(z'x) - E(z'z))[E(x'x)]^{-1}\frac{1}{m^{1/2}}\sum_{i=1}^{m} x_{v,i}'\epsilon_{v,i} + o_p(1).
\end{aligned}
\tag{4.3.3}
$$

This can be compared with the asymptotic disturbance of $\hat{\beta}$ in (2.5). Since x and z have the same dimension and $E(x'z)$ is invertible, Proposition (2.3) implies that

$$
\begin{aligned}
n^{1/2}(\hat{\beta} - \beta_0) = [E(z'x)]^{-1}\Big\{&\frac{1}{n^{1/2}}\sum_{i=1}^{n} z_i'\left(y_i - z_i[E(z'z)]^{-1}E(z'y)\right) \\
&- \lambda\frac{1}{m^{1/2}}\sum_{i=1}^{m} z_{v,i}'\left(x_{v,i}\beta_0 - z_{v,i}[E(z'z)]^{-1}E(z'x\beta_0)\right)\Big\} + o_p(1).
\end{aligned}
\tag{4.3.4}
$$

Except that an additional disturbance term is introduced in (4.3.3) and $E(z'z)$ is used in place of $E(z'x)$, the expressions in (4.3.3) and (4.3.4) are similar. It is not clear from these expressions whether one estimator might dominate the other. One thing is clear: $\hat{\beta}$ requires only validation data (x_v, z_v) on (x, z), but the bias corrected procedure requires simultaneously the presence of data y_v and (x_v, z_v). Thus, for some validation data, $\hat{\beta}$ could be derived while $\hat{\beta}_c$ might not.

When both the explanatory and dependent variables are measured with errors, the estimate of the asymptotic bias is

$$
\Delta_{b,n} = (Z_v'Z_v)^{-1}Z_v'\{(X_v - Z_v)\hat{\beta}_v + (\tilde{Y}_v - Y_v)\}.
$$

The bias adjusted estimator is $\hat{\beta}_c = \hat{\beta}_{LS} - \Delta_{b,n}$. By simple algebraic manipulations,

$$
\begin{aligned}
\hat{\beta}_c - \beta_0 = &(Z'Z)^{-1}Z'(\tilde{Y} - Z[E(z'z)]^{-1}E(z'\tilde{y})) + (Z_v'Z_v)^{-1}Z_v'\{\epsilon_v - \tilde{Y}_v + Z_v[E(z'z)]^{-1}E(z'\tilde{y})\} \\
&- (Z_v'Z_v)^{-1}Z_v'(X_v - Z_v)(X_v'X_v)^{-1}X_v'\epsilon_v,
\end{aligned}
$$

which implies that

$$
\begin{aligned}
&n^{1/2}(\hat{\beta}_c - \beta_0) \\
&= [E(z'z)]^{-1}\Big\{\frac{1}{n^{1/2}}\sum_{i=1}^{n} z_i'(\tilde{y}_i - z_i[E(z'z)]^{-1}E(z'\tilde{y})) + \lambda\frac{1}{m^{1/2}}\sum_{i=1}^{m} z_{v,i}'(\epsilon_{v,i} - \tilde{y}_{v,i} + z_{v,i}[E(z'z)]^{-1}E(z'\tilde{y}))\Big\} \\
&\quad - \lambda[E(z'z)]^{-1}(E(z'x) - E(z'z))[E(x'x)]^{-1}\frac{1}{m^{1/2}}\sum_{i=1}^{m} x_{v,i}'\epsilon_{v,i} + o_p(1).
\end{aligned}
\tag{4.3.5}
$$

For the estimator $\tilde{\beta}_1$ in (3.3), as z and x have the same dimension, and the validation data come from the same resource, Proposition 3.2 implies that

$$
\begin{aligned}
n^{1/2}(\tilde{\beta}_1 - \beta_0) = [E(z'x)]^{-1}\cdot\Big\{&\frac{1}{n^{1/2}}\sum_{i=1}^{n} z_i'(\tilde{y}_i - z_i[E(z'z)]^{-1}E(z'\tilde{y})) \\
&+ \lambda\frac{1}{m^{1/2}}\sum_{i=1}^{m} z_{v,i}'(\epsilon_{v,i} - \tilde{y}_{v,i} + z_{v,i}[E(z'z)]^{-1}E(z'\tilde{y}))\Big\} + o_p(1).
\end{aligned}
\tag{4.3.6}
$$

The asymptotic distribution of $n^{1/2}(\hat{\beta}_c - \beta_0)$ differs from the asymptotic distribution of $n^{1/2}(\tilde{\beta}_1 - \beta_0)$ with an additional disturbance term and with $E(z'z)$ in place of $E(z'x)$.

For the case where only the dependent variable is measured with error, since $Z = X$ and $Z_v = X_v$, the bias adjusted least squares estimator is $\hat{\beta}_c = (X'X)^{-1}X'\tilde{Y} - (X_v'X_v)^{-1}X_v'(\tilde{Y}_v - Y_v)$. The estimator in

(3.3) can be simplied to $\tilde{\beta}_1 = (X'X)^{-1}X'(\tilde{Y} - X(X_v'X_v)^{-1}X_v'(\tilde{Y}_v - Y_v))$. The two estimators in this case are identical.

5. Simulation Results

To evaluate the performance of the proposed methods, simulations were run for the linear model $y = \beta_0 + \beta_1 x + \epsilon$, where ϵ has a standard normal distribution and $(\beta_0, \beta_1) = (0.5, 1)$. Five hundred simulated data set were generated for each of the validation data sizes $m = 50, 100, 200$ and for each of the survey data sizes $n = 100, 200, 300$.

First case considered is when only the explanatory variable x is erroneously measured and therefore observations on (y, z) are measured in the survey data. The validation data set containing (x, z) or (y, x, z) were generated such that $z = x + \delta u$, where x is normally distributed with mean 4 and a standard deviation 1 and u has a standard normal distribution. In addition to the proposed estimates (Errorx), three other estimates were computed. Naive estimates (Naive) were obtained by ignoring the measurement error using the survey data. The Naive estimates are the ordinary least squares estimates. If the response y is observed in the validation data set, unbiased estimators (Valid) of parameters can be calculated by linear regression of y on x based on the validation data. The pooled estimated (Pool) described in section 4.1 were also computed by using the validation data to obtain an estimate $\hat{V}$ of V. The results are given in Tables 1.1 and 1.2 for $\delta = 0.5$ and 0.75 respectively.

We next consider the case when both the response y and the explanatory variable x are measured with error. In this case, the survey data contains observation on $(\tilde{y}, z)$ and the validation data set contains observation on $(\tilde{y}, y, x, z)$. There were two sampling situations for the validation data. First, the surrogate variables $\tilde{y}$ and z were generated such that $z = x + \delta u$, where x has a normal distribution with mean 4 and standard deviation 1 and u is a standard normal variable, and $\tilde{y} = 0.5y + \delta_1 v$, where v is a standard normal variable. Second, the surrogate z was generated as in the previous case and $\tilde{y}$ was generated such that $\tilde{y} = 0.5y + 0.2y^2 + \delta_1 v$.

In addition to the naive estimates (Naive) computed by regression of $\tilde{y}$ on z using the survey data, the estimates (Projer) described in (3.3) and the estimates (Errorxy) in (3.4) with $w = (\tilde{y}, z)$ were also computed. For the second sampling situation, the estimates (Quard) in (3.4) were calculated with $w = (\tilde{y}, \tilde{y}^2, z)$. Since the true response y and the true predictor x are both observed in the validation data, the Valid estimates were assessed. Using the validation data, it is possible to estimate the covariance matrix $\hat{V}$ in section 4.1 and the pooled estimates (Pool) were therefore computed. For the second plan the pooled estimates (Poolq) were calculated by pooling the estimates Quard and Valid together, since the Quard estimates were more efficient than Errorxy in this case. Tables 2.1–2.4 give the results based on the mode $\tilde{y} = 0.5y + \delta_1 v$ for each combination of $\delta = 0.5, 0.75$ and $\delta_1 = 0.5, 0.75$. Tables 3.1–3.4 give the results based on the mode $\tilde{y} = 0.5y + 0.3y^2 + \delta_1 v$.

Since these simulations were run for linear model, we also consider the case when only the response y is observed with error in the survey data. Tables 4.1 and 4.2 give the results for the case when the validation data set containing $(\tilde{y}, y, x)$ were generated such that x has a normal distribution with mean 4 and standard deviation 1 and $\tilde{y} = 0.5y + \delta_1 v$, where v is a standard normal variable and $\delta_1 = 0.5, 0.75$. Tables 5.1 and 5.2 give the results for the case when the validation data set were generated such that x has a normal distribution with mean 4 and standard deviation 1 and $\tilde{y} = 0.5y + 0.2y^2 + \delta_1 v$. Naive estimates, the estimates (Projer), the estimates (Errory) in (3.4), the Valid estimates and the pooled estimates (Pool) were all computed.

The naive estimates are obviously bias for all cases. The proposed estimates Errorx, Errory, Errorxy, and Projer perform well in terms of bias. The biases are small and they are almost as good as the unbiased Valid estimates. For the linear model $\tilde{y} = 0.5y + \delta_1 v$, the proposed two estimates Projer and Errorxy (or Errory) for the measurement error cases in the response y are in general compatible. When the measurement errors of y have larger variance $\delta_1 = 0.75$, the Errory (Errorxy) gives slightly more efficient estimates than the Projer estimates. When $\delta_1 = 0.5$, the Projer estimates are slightly more efficient. When the error model is $\tilde{y} = 0.5y + 0.2y^2 + \delta_1 v$, projection on $(\tilde{y}, \tilde{y}^2, z)$ or $(\tilde{y}, \tilde{y}^2, x)$ can capture the nonlinearity of the model and therefore produced more efficient estimates than projection on $(\tilde{y}, z)$ or $(\tilde{y}, x)$.

When the Valid estimate using only the correctly measured variables in the validation data is possibly obtained, pooling the proposed estimate and the Valid estimate together does yield a more efficient estimate. Indeed, depending on the sample sizes of both the survey and validation data and the variances of the

measurement errors, our proposed estimates Errorx, Errorxy, Errory, Quard, and Projer can even be more efficient than the Valid estimates. When the variances δ^2 or δ_1^2 of measurement errors are not too large and the sample size of the survey data are large relative to the sample size of the validation data, the proposed estimators can be more efficient.

6. Conclusion

This article has introduced consistent methods for the estimation of linear and nonlinear regression models with measurement errors in variables in the presence of validation data information. The measurement errors can be correlated with the true variables in the model. Validation data provide information on some properties of measurement errors. By exploiting the general correlation of the true variables and the instrumental variables with validation data, it is possible to provide consistent estimates of coefficients even in nonlinear regression models. The estimation procedures can be applied to models with measurement errors in the explanatory variable, the dependent variable, or both the explanatory and dependent variables. The estimation procedures do not rely on any particular specification of auxiliary structural measurement error equations. While these consistent estimation procedures are not nonlinear two-stage estimation procedures (Amemiya (1974)), they share some common features in terms of projection. However, the nonlinear two-stage least squares method is not defined with the survey data because x is unobserved.

The estimation procedures can be applied even to situations where validation data alone do not permit possible estimation of the model. When validation data are rich enough for consistent estimation of the model, survey data are still valuable in improving efficiency of the validation data estimates. The estimation procedures can be applied to models and data where direct bias corrections might not be feasible. For the linear regression model, when direct bias correction is feasible, the proposed estimators share some similarity with certain direct bias correction estimates.

Alternative methods based on nonparametric regression methods seem possible for the estimation of the linear and nonlinear regression models with general measurement errors in variables. Such methods have not been considered here but will be considered in another article. One may expect that for some cases, nonparametric approaches might be more efficient than the proposed approaches in this article. However, there are computational advantages in the proposed approaches in this article. The proposed estimation methods will in general be computationally simpler than any nonparametric method. For the estimation of linear regression models, validation data information can be summarized in terms of some simple statistics for the implementation of our estimation procedures, thus the proposed methods may be feasible even when validation data are confidential in the individual level details (Duncan and Pearson (1991)).

Acknowledgement

Lung-fei Lee is grateful for financial support from the NSF under grant no. SES-9296071 for his research. J. H. Sepanski is grateful for a development fund from the Rackham Gradute School of The University of Michigan for her research.

REFERENCES

Amemiya, T. (1974), "The nonlinear two stage least squares estimator," *Journal of Econometrics*, 2, 105-110.

Amemiya, T. (1985), *Advanced Econometrics*, Harvard University Press, Cambridge, MA.

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

Bound, J., C. Brown, G.J. Duncan, and W.L. Rodgers (1989), "Evidence the validity of cross-sectional and longitudinal labor market data," Manuscript, University of Michigan, Ann Arbor, Michigan.

Bound, J. and A.B. Krueger (1991), "The extent of measurement error in longitudinal earnings Data: Do Two Wrongs Make a Right?," *Journal of Labor Economics*, 9, 1-24.

Burr, D. (1988), "On errors-in-variables in binary regression-Berkson case," *Journal of the American Statistical Association*, 83, 739-743.

Carroll, R. J., Spiegelman, C., Lan, K. K. G., Bailey, K. T. and Abbott, R. D. (1984), "On errors-in-variables for binary regression models," *Biometrika,* 71, 19–2 6.

Carroll, R. J. and Stefanski, L. A. (1990), " Approximate quasilikelihood estimation in models with surrogate predictors," *Journal of the American Statistical Association,* 85, 652–663.

Carroll, R. J. and Wand, M. P. (1991), "Semiparametric estimation in logistic measurement error models," *Journal of the Royal Statistical Association, Series B,* to appear.

Chamberlain, G. (1982), "Multivariate regression models for panel Data," *Journal of Econometrics,* 18, 5-46.

Duncan, G. and D. Hill (1985), "An investigation of the extent and consequences of measurement error in labor-economics survey data,"" *Journal of Labor Economics,* 3, 508-532.

Duncan, G.T. and R.W. Pearson (1991), "Enhancing access to microdata while protecting confidentiality," *Statistical Science* 6, 219-239.

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

Gleser, L. J. (1990), "Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models," *Statistical Analysis of Measurement Error Models and Application,* P. J. Brown and W. A. Fuller, editors. American Mathematics Society, Providence.

Hausman, J., H. Ichimura, W, Newey, and J. Powell (1991), "Identification and estimation of polynomial errors-in-Variables models", *Journal of Econometrics,* 50, 273-295.

Hsiao, C. (1989), "Consistent estimation for some nonlinear errors-in-variables models," *Journal of Econometrics,* 41, 159-185.

Neyman, J. (1949), "Contribution to the theory of χ^2 test," In, J. Neyman, ed., *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability,* pp. 239-273, University of California Press, Berkeley.

Rodgers, W.L. and A.R. Herzog (1988), "Covariances of measurement errors in survey responses," *Journal of Official Statistics,* 3, 403-418.

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,* 8, 1051-1070.

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

Sepanski, J.H. and R.J. Carroll (1991), "Semiparametric quasilikelihood and variance function estimation in measurement error models," Manuscript, Department of Statistics and Management Science, School of Business Administration, The University of Michigan, Ann Arbor, Michigan; forthcoming in *Journal of Econometrics.*

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.

Taylor, W.F. (1953), "Distance functions and regular best asymptotically normal estimates," *Annals of Mathematical Statistics,* 24, 85-92.

Theil, H. and A.S. Goldberger (1961), "On pure and mixed statistical estimation in economics," *International Economic Review,* 2, 65-78.

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

Table 1.1.

Model:

$y = \beta_0 + \beta_1 x + \epsilon$, where $x \sim N(4,1), \epsilon \sim N(0,1)$, and $(\beta_0, \beta_1) = (0.5, 1)$
$z = x + \delta u$, where $u \sim N(0,1)$.

$\delta = 0.5$	m=50		m=100		m=200	
n=100	Mean	SD	Mean	SD	Mean	SD
Naive	0.792	0.096	0.804	0.098	0.792	0.099
Errorx	0.997	0.141	1.012	0.136	0.990	0.128
Valid	1.005	0.157	0.996	0.106	1.003	0.070
Pool	1.006	0.113	1.004	0.086	0.999	0.062
n=200	Mean	SD	Mean	SD	Mean	SD
Naive	0.808	0.070	0.797	0.069	0.804	0.071
Errorx	1.013	0.119	1.001	0.099	1.005	0.093
Valid	0.996	0.140	1.001	0.104	0.998	0.070
Pool	1.009	0.093	1.002	0.073	1.001	0.058
n=300	Mean	SD	Mean	SD	Mean	SD
Naive	0.799	0.054	0.802	0.059	0.800	0.059
Errorx	1.002	0.102	1.009	0.092	1.002	0.082
Valid	1.005	0.150	0.994	0.104	0.999	0.073
Pool	1.009	0.088	1.004	0.070	1.000	0.055

Table 1.2.

$\delta = 0.75$	m=50		m=100		m=200	
n=100	Mean	SD	Mean	SD	Mean	SD
Naive	0.633	0.090	0.645	0.094	0.632	0.093
Errorx	1.003	0.181	1.020	0.170	0.990	0.155
Valid	1.005	0.157	0.996	0.106	1.003	0.070
Pool	1.009	0.126	1.005	0.092	1.000	0.065
n=200	Mean	SD	Mean	SD	Mean	SD
Naive	0.647	0.068	0.637	0.065	0.644	0.068
Errorx	1.019	0.161	1.005	0.126	1.007	0.117
Valid	0.996	0.140	1.001	0.104	0.998	0.070
Pool	1.010	0.107	1.004	0.081	1.001	0.061
n=300	Mean	SD	Mean	SD	Mean	SD
Naive	0.639	0.051	0.642	0.057	0.639	0.056
Errorx	1.005	0.137	1.013	0.120	1.001	0.104
Valid	1.005	0.150	0.994	0.104	0.999	0.073
Pool	1.012	0.105	1.005	0.080	1.000	0.059

Table 2.1.

Model:

$y = \beta_0 + \beta_1 x + \epsilon$, where $x \sim N(4,1), \epsilon \sim N(0,1)$, and $(\beta_0, \beta_1) = (0.5, 1)$
$z = x + \delta u$, where $u \sim N(0,1)$.
$\tilde{y} = 0.5y + \delta_1 v$, where $v \sim N(0,1)$.

$\delta = 0.5, \delta_1 = 0.5$	m=50		m=100		m=200	
n=100	Mean	SD	Mean	SD	Mean	SD
Naive	0.403	0.065	0.402	0.068	0.402	0.066
Projer	1.005	0.147	1.006	0.120	1.009	0.099
Errorxy	1.005	0.158	1.006	0.128	1.009	0.105
Valid	0.997	0.138	1.007	0.101	1.001	0.069
Pool	1.002	0.123	1.007	0.092	1.005	0.067
n=200	Mean	SD	Mean	SD	Mean	SD
Naive	0.397	0.047	0.404	0.046	0.397	0.045
Projer	0.998	0.131	1.007	0.102	0.990	0.084
Errorxy	0.998	0.135	1.008	0.107	0.990	0.088
Valid	0.996	0.140	1.000	0.105	0.996	0.070
Pool	0.998	0.117	1.005	0.090	0.994	0.064
n=300	Mean	SD	Mean	SD	Mean	SD
Naive	0.401	0.038	0.398	0.037	0.399	0.039
Projer	1.004	0.131	1.001	0.095	1.000	0.076
Errorxy	1.005	0.134	1.002	0.099	1.001	0.081
Valid	1.002	0.157	0.998	0.106	1.001	0.071
Pool	1.005	0.130	1.000	0.088	1.001	0.062

Table 2.2.

$\delta = 0.75, \delta_1 = 0.5$	m=50		m=100		m=200	
n=100	Mean	SD	Mean	SD	Mean	SD
Naive	0.322	0.061	0.321	0.063	0.323	0.061
Projer	1.006	0.171	1.006	0.139	1.011	0.115
Errorxy	1.007	0.190	1.005	0.153	1.011	0.128
Valid	0.997	0.138	1.007	0.101	1.001	0.069
Pool	1.002	0.127	1.007	0.096	1.005	0.068
n=200	Mean	SD	Mean	SD	Mean	SD
Naive	0.317	0.043	0.323	0.043	0.317	0.042
Projer	1.001	0.153	1.007	0.118	0.990	0.097
Errorxy	1.001	0.162	1.009	0.127	0.989	0.107
Valid	0.996	0.140	1.000	0.105	0.996	0.070
Pool	1.000	0.122	1.004	0.094	0.994	0.066
n=300	Mean	SD	Mean	SD	Mean	SD
Naive	0.321	0.035	0.318	0.034	0.319	0.036
Projer	1.008	0.151	1.002	0.109	1.001	0.088
Errorxy	1.009	0.159	1.003	0.117	1.001	0.097
Valid	1.002	0.157	0.998	0.106	1.001	0.071
Pool	1.007	0.135	1.001	0.091	1.002	0.064

Table 2.3.

$\delta = 0.5, \delta_1 = 0.75$	m=50		m=100		m=200	
n=100	Mean	SD	Mean	SD	Mean	SD
Naive	0.404	0.081	0.402	0.085	0.403	0.084
Projer	1.005	0.187	1.006	0.152	1.012	0.128
Errorxy	1.003	0.161	1.006	0.124	1.009	0.097
Valid	0.997	0.138	1.007	0.101	1.001	0.069
Pool	1.000	0.131	1.007	0.097	1.005	0.069
n=200	Mean	SD	Mean	SD	Mean	SD
Naive	0.397	0.059	0.405	0.059	0.395	0.057
Projer	0.997	0.165	1.009	0.130	0.987	0.106
Errorxy	0.999	0.144	1.005	0.112	0.990	0.086
Valid	0.996	0.140	1.000	0.105	0.996	0.070
Pool	0.998	0.127	1.003	0.098	0.994	0.067
n=300	Mean	SD	Mean	SD	Mean	SD
Naive	0.401	0.048	0.398	0.046	0.399	0.048
Projer	1.004	0.160	1.003	0.119	1.001	0.096
Errorxy	1.005	0.149	1.002	0.104	1.002	0.079
Valid	1.002	0.157	0.998	0.106	1.001	0.071
Pool	1.004	0.142	1.000	0.095	1.002	0.066

Table 2.4.

$\delta = 0.75, \delta_1 = 0.75$	m=50		m=100		m=200	
n=100	Mean	SD	Mean	SD	Mean	SD
Naive	0.322	0.075	0.320	0.078	0.323	0.076
Projer	1.006	0.214	1.004	0.173	1.014	0.146
Errorxy	1.004	0.189	1.005	0.147	1.011	0.116
Valid	0.997	0.138	1.007	0.101	1.001	0.069
Pool	1.001	0.134	1.007	0.101	1.005	0.070
n=200	Mean	SD	Mean	SD	Mean	SD
Naive	0.317	0.054	0.324	0.053	0.316	0.052
Projer	1.000	0.189	1.008	0.148	0.987	0.120
Errorxy	1.002	0.168	1.006	0.131	0.989	0.101
Valid	0.996	0.140	1.005	0.105	0.996	0.070
Pool	0.999	0.131	1.003	0.100	0.994	0.069
n=300	Mean	SD	Mean	SD	Mean	SD
Naive	0.321	0.044	0.318	0.042	0.319	0.044
Projer	1.007	0.182	1.004	0.135	1.001	0.109
Errorxy	1.007	0.170	1.004	0.120	1.003	0.093
Valid	1.002	0.157	0.998	0.106	1.001	0.071
Pool	1.006	0.144	1.001	0.097	1.002	0.067

Table 3.1.

Model:

$y = \beta_0 + \beta_1 x + \epsilon$, where $x \sim N(4,1), \epsilon \sim N(0,1)$, and $(\beta_0, \beta_1) = (0.5, 1)$
$z = x + \delta u$, where $u \sim N(0,1)$.
$\tilde{y} = 0.5y + 0.2y^2 + \delta_1 v$, where $v \sim N(0,1)$.

$\delta = 0.5, \delta_1 = 0.5$	m=50		m=100		m=200	
n=100	Mean	SD	Mean	SD	Mean	SD
Naive	1.840	0.245	1.845	0.256	1.841	0.240
Projer	1.006	0.446	1.002	0.387	0.997	0.336
Errorxy	1.006	0.157	1.008	0.144	1.002	0.125
Quard	0.999	0.140	1.003	0.130	1.003	0.117
Valid	0.997	0.138	1.007	0.101	1.001	0.069
Poolq	1.000	0.107	1.007	0.082	1.002	0.061
n=200	Mean	SD	Mean	SD	Mean	SD
Naive	1.829	0.166	1.847	0.162	1.834	0.164
Projer	0.997	0.412	1.023	0.306	0.983	0.263
Errorxy	1.004	0.134	1.009	0.104	0.993	0.097
Quard	0.999	0.115	1.005	0.091	0.990	0.088
Valid	0.996	0.140	1.006	0.105	0.996	0.070
Poolq	0.995	0.102	1.006	0.077	0.994	0.057
n=300	Mean	SD	Mean	SD	Mean	SD
Naive	1.841	0.139	1.833	0.137	1.834	0.145
Projer	1.017	0.376	0.997	0.283	0.986	0.243
Errorxy	1.011	0.118	1.000	0.095	0.997	0.087
Quard	0.997	0.104	0.996	0.085	0.994	0.079
Valid	1.002	0.157	0.998	0.106	1.001	0.070
Poolq	1.003	0.113	0.998	0.077	0.999	0.055

Table 3.2.

$\delta = 0.75, \delta_1 = 0.5$	m=50		m=100		m=200	
n=100	Mean	SD	Mean	SD	Mean	SD
Naive	1.472	0.233	1.472	0.239	1.476	0.227
Projer	1.011	0.544	1.003	0.455	1.001	0.400
Errorxy	1.008	0.186	1.008	0.172	1.004	0.152
Quard	1.002	0.182	1.004	0.160	1.004	0.144
Valid	0.997	0.138	1.007	0.101	1.001	0.069
Poolq	1.001	0.115	1.007	0.088	1.003	0.063
n=200	Mean	SD	Mean	SD	Mean	SD
Naive	1.460	0.152	1.475	0.153	1.468	0.584
Projer	1.000	0.494	1.028	0.368	0.982	0.326
Errorxy	1.007	0.168	1.010	0.129	0.993	0.122
Quard	0.990	0.149	1.006	0.115	0.990	0.113
Valid	0.996	0.140	1.000	0.105	0.996	0.070
Poolq	0.998	0.112	1.004	0.084	0.995	0.061
n=300	Mean	SD	Mean	SD	Mean	SD
Naive	1.475	0.132	1.465	0.131	1.468	0.139
Projer	1.036	0.468	0.999	0.342	0.985	0.296
Errorxy	1.020	0.153	1.002	0.120	0.997	0.109
Quard	1.003	0.139	0.997	0.111	0.994	0.101
Valid	1.002	0.157	0.998	0.106	1.001	0.071
Poolq	1.005	0.125	0.999	0.085	1.000	0.059

Table 3.3.

$\delta = 0.5, \delta_1 = 0.75$	m=50		m=100		m=200	
n=100	Mean	SD	Mean	SD	Mean	SD
Naive	1.841	0.249	1.845	0.261	1.842	0.246
Projer	1.006	0.461	1.001	0.398	1.000	0.347
Errorxy	1.005	0.158	1.007	0.143	1.003	0.124
Quard	1.000	0.144	1.005	0.131	1.004	0.116
Valid	0.997	0.138	1.007	0.101	1.001	0.069
Poolq	1.000	0.109	1.007	0.083	1.003	0.061
n=200	Mean	SD	Mean	SD	Mean	SD
Naive	1.829	0.170	1.848	0.165	1.833	0.167
Projer	0.997	0.425	1.028	0.314	0.980	0.270
Errorxy	1.004	0.134	1.009	0.104	0.992	0.096
Quard	0.991	0.119	1.006	0.093	0.990	0.089
Valid	0.997	0.140	1.000	0.105	0.996	0.070
Poolq	0.996	0.104	1.004	0.079	0.994	0.058
n=300	Mean	SD	Mean	SD	Mean	SD
Naive	1.841	0.142	1.832	0.139	1.834	0.149
Projer	1.018	0.386	0.998	0.294	0.986	0.250
Errorxy	1.011	0.119	1.001	0.095	0.997	0.087
Quard	1.000	0.108	0.998	0.088	0.995	0.080
Valid	1.002	0.157	0.998	0.106	1.001	0.071
Poolq	1.004	0.116	0.999	0.079	1.000	0.056

Table 3.4.

$\delta = 0.75, \delta_1 = 0.75$	m=50		m=100		m=200	
n=100	Mean	SD	Mean	SD	Mean	SD
Naive	1.472	0.237	1.472	0.244	1.476	0.231
Projer	1.011	0.561	1.001	0.466	1.004	0.409
Errorxy	1.008	0.195	1.008	0.171	1.005	0.150
Quard	1.003	0.184	1.005	0.160	1.006	0.143
Valid	0.997	0.138	1.007	0.101	1.001	0.069
Poolq	1.001	0.117	1.007	0.089	1.003	0.063
n=200	Mean	SD	Mean	SD	Mean	SD
Naive	1.460	0.156	1.476	0.155	1.466	0.161
Projer	0.999	0.508	1.030	0.377	0.979	0.332
Errorxy	1.007	0.168	1.010	0.128	0.992	0.120
Quard	0.993	0.152	1.007	0.117	0.989	0.113
Valid	0.996	0.140	1.000	0.105	0.996	0.070
Poolq	0.999	0.113	1.004	0.086	0.995	0.061
n=300	Mean	SD	Mean	SD	Mean	SD
Naive	1.475	0.134	1.465	0.132	1.468	0.141
Projer	1.035	0.479	1.000	0.352	0.985	0.305
Errorxy	1.019	0.154	1.002	0.119	0.998	0.108
Quard	1.006	0.142	0.999	0.113	0.995	0.101
Valid	1.002	0.157	0.998	0.106	1.001	0.071
Poolq	1.006	0.126	1.000	0.086	1.000	0.060

Table 4.1.

Model:

$y = \beta_0 + \beta_1 x + \epsilon$, where $x \sim N(4,1), \epsilon \sim N(0,1)$, and $(\beta_0, \beta_1) = (0.5, 1)$
$\tilde{y} = 0.5y + \delta_1 v$, where $v \sim N(0,1)$.

$\delta_1 = 0.5$	m=50		m=100		m=200	
n=100	Mean	SD	Mean	SD	Mean	SD
Naive	0.495	0.072	0.500	0.071	0.497	0.071
Projer	0.994	0.117	1.004	0.100	0.995	0.089
Errory	0.993	0.120	1.004	0.100	0.994	0.089
Valid	1.001	0.147	1.002	0.104	0.993	0.071
Pool	0.997	0.117	1.003	0.088	0.994	0.070
n=200	Mean	SD	Mean	SD	Mean	SD
Naive	0.503	0.052	0.504	0.050	0.497	0.052
Projer	0.998	0.111	1.013	0.087	0.997	0.076
Errory	0.999	0.112	1.013	0.087	0.997	0.076
Valid	0.994	0.146	1.006	0.100	0.999	0.075
Pool	0.997	0.116	1.009	0.084	0.998	0.066
n=300	Mean	SD	Mean	SD	Mean	SD
Naive	0.500	0.041	0.501	0.041	0.498	0.040
Projer	1.008	0.105	1.004	0.083	0.995	0.064
Errory	1.008	0.106	1.004	0.084	0.995	0.064
Valid	1.012	0.147	1.001	0.099	0.997	0.074
Pool	1.010	0.114	1.002	0.082	0.996	0.060

Table 4.2.

$\delta_1 = 0.75$	m=50		m=100		m=200	
n=100	Mean	SD	Mean	SD	Mean	SD
Naive	0.493	0.091	0.499	0.091	0.496	0.091
Projer	0.991	0.147	1.006	0.126	0.994	0.112
Errory	0.994	0.129	1.004	0.100	0.993	0.085
Valid	1.001	0.147	1.002	0.104	0.993	0.071
Pool	0.998	0.128	1.003	0.094	0.993	0.071
n=200	Mean	SD	Mean	SD	Mean	SD
Naive	0.505	0.066	0.506	0.064	0.497	0.065
Projer	0.998	0.141	1.017	0.109	0.997	0.095
Errory	0.998	0.126	1.013	0.093	0.998	0.076
Valid	0.994	0.146	1.006	0.100	0.999	0.075
Pool	0.996	0.128	1.010	0.091	0.999	0.070
n=300	Mean	SD	Mean	SD	Mean	SD
Naive	0.501	0.052	0.501	0.051	0.498	0.051
Projer	1.010	0.133	1.006	0.106	0.994	0.081
Errory	1.012	0.120	1.003	0.091	0.995	0.067
Valid	1.012	0.147	1.001	0.099	0.997	0.074
Pool	1.012	0.126	1.002	0.089	0.996	0.066

Table 5.1.

Model:

$y = \beta_0 + \beta_1 x + \epsilon$, where $x \sim N(4,1), \epsilon \sim N(0,1)$, and $(\beta_0, \beta_1) = (0.5, 1)$
$\tilde{y} = 0.5y + 0.2y^2 + \delta_1 v$, where $v \sim N(0,1)$.

$\delta_1 = 0.5$	m=50		m=100		m=200	
n=100	Mean	SD	Mean	SD	Mean	SD
Naive	2.228	0.260	2.301	0.251	2.288	0.259
Projer	0.986	0.363	0.997	0.300	0.994	0.269
Errory	0.999	0.119	1.001	0.105	0.994	0.104
Quard	0.989	0.100	0.996	0.096	0.994	0.096
Valid	1.001	0.147	1.002	0.104	0.993	0.071
Pool	0.993	0.093	0.998	0.075	0.993	0.065
n=200	Mean	SD	Mean	SD	Mean	SD
Naive	2.306	0.180	2.307	0.183	2.287	0.187
Projer	1.016	0.310	1.001	0.246	0.985	0.220
Errory	1.006	0.094	1.005	0.084	0.995	0.080
Quard	0.991	0.077	1.001	0.071	0.994	0.071
Valid	0.994	0.146	1.006	0.099	0.999	0.075
Pool	0.992	0.084	1.003	0.066	0.996	0.056
n=300	Mean	SD	Mean	SD	Mean	SD
Naive	2.292	0.149	2.305	0.156	2.298	0.148
Projer	0.998	0.298	1.008	0.230	0.996	0.198
Errory	1.006	0.087	1.005	0.077	0.998	0.067
Quard	0.991	0.069	0.997	0.062	0.994	0.057
Valid	1.012	0.147	1.001	0.099	0.997	0.074
Pool	1.000	0.085	0.998	0.063	0.995	0.049

Table 5.2.

$\delta_1 = 0.75$	m=50		m=100		m=200	
n=100	Mean	SD	Mean	SD	Mean	SD
Naive	2.286	0.267	2.300	0.255	2.287	0.263
Projer	0.982	0.377	0.998	0.311	0.994	0.274
Errory	0.998	0.120	1.002	0.104	0.993	0.102
Quard	0.991	0.104	0.998	0.097	0.994	0.094
Valid	1.001	0.147	1.002	0.104	0.993	0.071
Pool	0.995	0.096	0.999	0.077	0.993	0.067f
n=200	Mean	SD	Mean	SD	Mean	SD
Naive	2.308	0.184	2.309	0.187	2.287	0.191
Projer	1.016	0.324	1.005	0.254	0.986	0.227
Errory	1.005	0.097	1.007	0.084	0.995	0.079
Quard	0.995	0.085	1.004	0.074	0.995	0.072
Valid	0.994	0.146	1.006	0.100	0.999	0.075
Pool	0.994	0.089	1.005	0.069	0.996	0.057
n=300	Mean	SD	Mean	SD	Mean	SD
Naive	2.293	0.152	2.306	0.158	2.297	0.151
Projer	0.990	0.314	1.009	0.240	0.995	0.205
Errory	1.007	0.089	1.005	0.078	0.997	0.067
Quard	0.997	0.076	0.999	0.066	0.994	0.059
Valid	1.012	0.147	1.001	0.099	0.997	0.074
Pool	1.003	0.090	1.000	0.066	0.996	0.051

Recent Crest Working Papers

91–01: Mark Bagnoli, Stephen W. Salant, Joseph E. Swierzbinski, "Price Discrimination and Intertemporal Self-Selection," June, 1990.

91–02: Howard Doran, Jan Kmenta, "Multiple Minima in the Estimation of Models with Autoregressive Disturbances," April 1991.

91–03: Ted Bergstrom, Mark Bagnoli, "Courtship as a Waiting Game," April 3, 1991.

91–04: Ted Bergstrom, Jeffrey K. MacKie-Mason, "Peak-Load Pricing-With and Without Constrained Rate of Return," April 2, 1991.

91–05: Laura E. Kodres, Daniel P. O'Brien, "The Existence of Pareto Superior Price Limits and Trading Halts," February 1991.

91–06: Theodore Bergstrom, David Lam, "The Effects of Cohort Size on Marriage Markets in Twentieth Century Sweden," November 1989.

91–07: Theodore Bergstrom, David Lam, "The Two-Sex Problem and the Marriage Squeeze in an Equilibrium Model of Marriage Markets," April 1991.

91–08: Greg Shaffer, "Capturing Strategic Rent: Full-line Forcing, Maximum Resale Price Maintenance, Brand Discounts and Aggregate Rebates," March 1991.

91–09: Jeffrey K. MacKie-Mason, Roger H. Gordon, " Taxes and the Choice of Organizational Form," April 3, 1991.

91–10: Hal R. Varian, "A Solution to the Problem of Externalities When Agents are Well-Informed," September 1991.

91–11: Hal R. Varian, "The Economics of User Interfaces," September 1991.

91–12: Eduardo Ley, Hal R. Varian, "A Note on the Dow–Jones Index," September 1991.

91–13: Hal R. Varian, "Goodness of Fit for Revealed Preference Tests," September 1991.

91–14: Hal R. Varian, "Sequential Provision of Public Good," September 1991.

91–15: Ken Binmore, "A Liberal Leviathan," September 1991.

91–16: Ken Binmore, "DeBayesing Game Theory," September 1991.

91–17: Ken Binmore, "Bargaining and Morality," September 1991.

91–18: Ted Bergstrom, "When Non–Transitive Relations Take Maxima and Competitive Equilibria Can't be Beat."

92–01: Lung-fei Lee, "Asymptotic Distribution of the Maximum Likelihood Estimator for a Stochastic Frontier Function Model with a Singular Information Matrix," May, 1992.

92–02: Stephen Salant, Karen Kalat, Ana Wheatcroft, "Surviving Winter: A Fitness-Based Explanation of Hoarding and Hibernation," February, 1992.

92–03: Gérard Gaudet, Stephen Salant, "The Limits of Monopolization Through Acquisition: Further Results," February, 1992.

92–04: Jonathan Cave, Stephen W. Salant, "Cartel Quotas Under Majority Rule," February, 1992.

92–05: Mark Bagnoli, Stephen W. Salant, Joseph E. Swierzbinski, "Intertemporal Self-Selection with Multiple Buyers Under Complete and Incomplete Information," April, 1992.

92–06: Daniel P. O'Brien, Greg Shaffer, "Non-linear Contracts, Foreclosure, and Exclusive Dealing," April, 1992.

92–07: Lung-fei Lee, Jungsywan H. Sepanski, "Consistent Estimation of Linear and Nonlinear Errors-in-Variables Models with Validation Information," April, 1992.

92–08: Lung-fei Lee, "Semiparametric Minimum-distance Estimation," November, 1991.

92–09: Michelle J. White, "Corporate Bankruptcy as a Filtering Device," March 29, 1992.

92–10: Ted Bergstrom, Robert Schoeni, "Income Prospects and Age at Marriage," March 4, 1992.

92–11: James Andreoni, Ted Bergstrom, "Do Government Subsidies Increase the Private Supply of Public Goods?," February 6, 1992.