TECHNICAL WORKING PAPER SERIES

BIAS FROM CLASSICAL AND OTHER FORMS OF MEASUREMENT ERROR

Dean R. Hyslop
Guido W. Imbens

Technical Working Paper 257
http://www.nber.org/papers/T0257

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2000

We are grateful to David Card, the editor, an associate editor and a referee for comments and discussions. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

Bias from Classical and Other Forms of Measurement Error
Dean R. Hyslop and Guido W. Imbens
NBER Technical Working Paper No. 257
August 2000
JEL No. C10, C13

ABSTRACT

We consider the implications of a specific alternative to the classical measurement error model, in which the data are optimal predictions based on some information set. One motivation for this model is that if respondents are aware of their ignorance they may interpret the question "what is the value of this variable?" as "what is your best estimate of this variable?", and provide optimal predictions of the variable of interest given their information set. In contrast to the classical measurement error model, this model implies that the measurement error is uncorrelated with the reported value and, by necessity, correlated with the true value of the variable.

In the context of the linear regression framework, we show that measurement error can lead to over- as well as under-estimation of the coefficients of interest. Critical for determining the bias is the model for the individual reporting the mismeasured variables, the individual's information set, and the correlation structure of the errors. We also investigate the implications of instrumental variables methods in the presence of measurement error of the optimal prediction error form and show that such methods may in fact introduce bias. Finally, we present some calculations indicating that the range of estimates of the returns to education consistent with amounts of measurement error found in previous studies. This range can be quite wide, especially if one allows for correlation between the measurement errors.

Dean R. Hyslop
Department of Economics
Bunche Hall
University of California at Los Angeles
405 Hilgard Avenue
Los Angeles, CA 90095

Guido W. Imbens
Department of Economics
Bunche Hall
University of California at Los Angeles
405 Hilgard Avenue
Los, Angeles, CA 90095
and NBER
imbens@econ.ucls.edu

1. Introduction

Many variables used in econometric analyses are recorded with error. These errors may have occurred at various stages of the data collection. They may be the result of misreporting by subjects, miscoding by the collectors of the data, or incorrect transformation from initial reports into a form ready for analyses. Often such errors are ignored. In cases where explicit attention is paid to measurement error, it is typically assumed to be "classical measurement error", where the error is independent, or at least uncorrelated with the true value of the underlying variable (e.g., Klepper and Leamer (1984), Chesher (1999), Schennach (2000); see Grilliches (1987), Angrist and Krueger (2000), and Bound, Brown and Mathiowetz (2000) for surveys). However, when responses have been validated (Bound and Krueger, 1991; Pischke, 1995; Card and Hyslop, 1997) empirical support for classical measurement error has typically been limited.

The implications of deviations from classical measurement error are only rarely considered. Card (1996) and Bollinger (1996) study models with measurement error in binary variables where the classical measurement error assumptions cannot hold. Kane, Rouse and Staiger (1998) investigate categorical response models exploiting the presence of two measures with uncorrelated errors. They do allow the errors to be correlated with both the true and reported values. Horowitz and Manski (1995) study bounds when a fraction of the observations is mismeasured in an unrestricted manner. Bound, Brown and Mathiowetz (2000) survey some of these approaches.

In this paper we explore the consequences of alternative types of measurement error. We argue that if errors occur in reports by agents based on limited information, there are specific alternatives to the classical measurement error model based on the view that respondents are actively choosing a response. Such models have been used before in settings where explicit account was taken of the agent's awareness of the limits on their knowledge and incentives for accurate reporting. Examples include the modelling of preliminary reports of macro-economic aggregates (Mankiw and Shapiro, 1986), in the analysis of the effect of financial incentives on accuracy in surveys (Philipson, 1999), and the analysis of responses

to questions about future events (Manski, 1990; and Das, Dominitz and Van Soest, 1999). The alternatives we consider assume that, in response to the question "What is the value of X?", respondents report their best estimate of the value of interest given their information set. In contrast, under the classical measurement error model, respondents can be viewed as reporting an unbiased, although suboptimal, value.

We then explore the implications of the alternative models in the context of linear regressions. We find that the standard argument that measurement error in regressors leads to underestimation of the magnitude of the relationship between the true variables can be misleading. In particular, under plausible assumptions, measurement error can lead to over- as well as under-estimation of the underlying relationship. We derive signs of the bias for a number of leading cases. Finally, we present some calculations showing how sensitive regression estimates can be to measurement error under different models in the context of wage regressions, using the PSID validation study (e.g., Bound and Krueger, 1991; and Pischke, 1995) to obtain estimates of the amount of measurement error in reported earnings and the Ashenfelter and Krueger (1994) twins study for estimates of the amount of measurement error in reported years of education. Although in many cases one may not be able to credibly choose between the different types of measurement error, one may be able to assess the amount of measurement error using previously collected data from validation studies. In such cases one can explore the range of parameter values consistent with the amount of measurement error under the various models, as we illustrate in Section 5. These analyses are in the spirit of the sensitivity and bounds analyses of Rosenbaum and Rubin (1983), Leamer (1987), Horowitz and Manski (1995), and Bollinger (1996).

2. A Decomposition of Measurement Error

Let X^* denote the true value of a variable of interest, and X the recorded value. The measurement error is the difference between the recorded and true value:

$$\varepsilon \equiv X - X^*. \tag{1}$$

We decompose ε into three components:

$$\varepsilon = \varepsilon_{cme} + \varepsilon_{ope} + \varepsilon_r.$$

The first component is not predictable by the true value - i.e., the classical measurement error:

$$\varepsilon_{cme} \equiv \varepsilon - E[\varepsilon|X^*] = X - X^* - E[X - X^*|X^*] = X - E[X|X^*]. \tag{2}$$

The second component is not predictable by the reported value, which we refer to as optimal prediction error:

$$\varepsilon_{ope} \equiv \varepsilon - E[\varepsilon|X] = X - X^* - E[X - X^*|X] = -X^* + E[X^*|X]. \tag{3}$$

The third and final component ε_r is defined as the remainder of the error:

$$\begin{aligned}\varepsilon_r \equiv \varepsilon - \varepsilon_{cme} - \varepsilon_{ope} &= X - X^* - \Big(X - E[X|X^*]\Big) - \Big(-X^* + E[X^*|X]\Big) \\ &= E[X|X^*] - E[X^*|X].\end{aligned} \tag{4}$$

This decomposition is definitional in that it does not require any assumptions (beyond finiteness of the appropriate expectations). It is unique, and any assumptions on the measurement error can therefore be formulated as assumptions on the three components.

3. Two Models for Measurement Error

3.1 Classical Measurement Error

The standard Classical Measurement Error (CME) model, assumes that the measurement error is independent of the true value. Assuming that the measurement error has mean zero, this implies $E[\varepsilon|X^*] = 0$. Since by definition $\varepsilon_{cme} = \varepsilon - E[\varepsilon|X^*]$, it follows that for this model to be correct, it must be that $\varepsilon = \varepsilon_{cme}$ and the last two components ε_{ope} and ε_r sum to zero. This model is typically defended by reference to physical measurement models where often passive recording of measurements based on imprecise measuring instruments takes place.

3.2 Optimal Prediction Error

An alternative model, which we refer to as the Optimal Prediction Error (OPE) model, is based on the assumption that the measurement error is independent of the reported value. This implies $E[\varepsilon|X] = 0$ and, since $\varepsilon_{ope} \equiv \varepsilon - E[\varepsilon|X]$, $\varepsilon = \varepsilon_{ope}$ so that $\varepsilon_{cme} + \varepsilon_r = 0$.

An argument in support of this model views the agent reporting the data as fully aware of the lack of precision of the measuring instrument. Suppose the agent is asked to provide the value of some variable. The agent has no way of ascertaining the true value X^* of this variable, but has available a flawed or noisy measure, $\tilde{X} = X^* + \eta_X$, with the measurement error η_X independent of the true value of the variable, exactly as in the CME model. However, suppose that the agent is aware of the lack of precision of the measurement, and corrects for this by reporting the best estimate of the underlying true value X^* based on this measurement $\tilde{X}$. To operationalize this we interpret "best" in terms of a loss function. Assuming a quadratic loss function implies the agent would report the expected value of the true value given the the agent's information set.[3] Then the error $\varepsilon = X - X^*$ should have mean zero given the information set of the agent. Since the reported value is clearly in the information set, this implies that the error has mean zero given the reported value.

Critical in this model is the active role of the respondent. Thus, in order to assess the impact of measurement error, the researcher needs to understand how the respondent views the survey question. If the respondent is aware of not having exact information regarding the value of the variable requested, presumably the question "What is the value of X?" is interpreted as "What is your best estimate of the value of X?". In that case the answer should *not* be the unbiased measurement even if that is the basic piece of information available to the respondent. Although the respondent need not have the exact probability model underlying the unbiased measurement and true value, it is plausible that outliers are adjusted in a way

[3] An alternative would be to assume absolute value loss, in which case the agent would report the median of X^* given the information set. For most of the illustrative calculations below the mean and median will give the same answers because we assume normality. A third possibility arises when there is a constant loss if the reported value differs from the true value. This arises in Philipson's (1999) survey of physicians who, with some probability, get a reward if their response matches administrative records. In that case respondents should report the mode of the distribution.

that leads to some correlation between the true value and the measurement error. One the other hand, this model is less likely to be appropriate if the measurement error is the result of miscoding of survey answers.

A crucial ingredient in the OPE model is the information set. It may be that the respondent only has a single unbiased measurement of the underlying true variable. Alternatively, other variables which themselves may enter the econometric model of interest may be used to produce this estimate.[4] In the next section we consider, in the context of a linear regression model, two variations of the model that differ in the information set exploited in the calculation of the optimal prediction of the quantity of interest.

Models similar to this OPE model have been used in other contexts where agents are asked to provide information about variables whose values they do not know exactly. The behavior of government agencies reporting macro-economic quantities can be viewed as predicting the underlying variable of interest given the agency's information set, which includes signals of the true value. In this context, the measurement error is expected to be independent of the information set used, which necessarily includes the reported value.[5]

Philipson (1999) carries out experiments to see how the reliability of survey responses varies with incentives. Philipson asks physicians the value of a categorical variable (medical specialization) that he can verify from administrative data. He offers some of the physicians an incentive scheme where with some probability he will check the value of the variable and pay some sum of money if the reported and true values agree.[6] He finds that such incentive schemes increase the probability that the respondents answered the question correctly.

Manski (1990), and Das, Dominitz and Van Soest (1999) analyze data where individuals

[4]For example, Ashenfelter and Krueger (1994) survey twins and ask each sibling both their own education and their sibling's education. To the extent that a respondent is not fully aware of their sibling's education level, but has knowledge of related items, such as occupation, it is plausible that such information would be used to infer the education level.

[5]For example, Mankiw and Shapiro (1986) model the revision in Gross National Product between the preliminary and final reports, and find that the revisions are uncorrelated with the early reports. In addition, the revisions are correlated with the final reports which, if the final reports are assumed to be the truth, is consistent with a OPE model but not the CME model.

[6]As mentioned before, this implies respondents should report the mode of the distribution.

were asked about future events including future income. In that case individuals clearly cannot know the exact value of these variables, and Manski and Dominitz model the qualitative responses as the best predictions (modes) given current information.

Each of these examples suggest that economic agents responding to questions about uncertain quantities can sometimes usefully be modelled as solving a prediction problem rather than as passively reporting noisy measurements. We therefore investigate the implications of this model for the estimation of regression coefficients in a linear model.

4. Implications of Measurement Error in the Linear Regression Model

Let us consider a homoskedastic linear regression model for two scalar variables Y^* and X^*:

$$Y^* = \alpha + \beta \cdot X^* + \nu, \tag{5}$$

where $\nu \perp X^*$, $E[\nu|X^*] = 0$, $\text{VAR}(\nu|X^*) = \sigma_\nu^2$, and the parameter of interest, β, is the ratio of the covariance of Y^* and X^* over the variance of X^*. Possibly mismeasured values Y and X are recorded. We consider the implications for least squares estimates of β based on a random sample from (Y, X) of various properties of the measurement error.

The basic piece of information available to the respondent is assumed to be a pair of noisy measures of the underlying variables:

$$\tilde{X} = X^* + \eta_X, \qquad \tilde{Y} = Y^* + \eta_Y.$$

We assume that the basic measurement errors (η_X, η_Y) are independent of the true value of the regressor X^* and of ν, but potentially correlated with each other. For expositional reasons we also assume joint normality:

$$\begin{pmatrix} X^* \\ \nu \\ \eta_X \\ \eta_Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & 0 & 0 & 0 \\ 0 & \sigma_\nu^2 & 0 & 0 \\ 0 & 0 & \sigma_{\eta_X}^2 & \rho_{\eta_X\eta_Y}\sigma_{\eta_X}\sigma_{\eta_Y} \\ 0 & 0 & \rho_{\eta_X\eta_Y}\sigma_{\eta_X}\sigma_{\eta_Y} & \sigma_{\eta_Y}^2 \end{pmatrix} \right). \tag{6}$$

We consider three cases relating the basic measurements $\tilde{X}$ and $\tilde{Y}$ to the reported values X and Y. The first case is the CME model where the reported value is identical to the

unbiased measurement. In addition we consider two versions of the OPE model. The first (OPE(1)) is where the respondent reports their best estimate based only on the noisy measure of the mismeasured variable itself. The second (OPE(2)) is where the respondent reports their best estimate based on the noisy measures of both variables. Table 1 summarizes the three models. In each of the three cases we consider the sign of the difference between the

Table 1: Three Models for Measurement Error

Reporting Model	X	Y
Classical Measurement Error	$X_{CME} = \tilde{X}$	$Y_{CME} = \tilde{Y}$
Optimal Prediction Error (1)	$X_{OPE(1)} = E[X^*\|\tilde{X}]$	$Y_{OPE(1)} = E[Y^*\|\tilde{Y}]$
Optimal Prediction Error (2)	$X_{OPE(2)} = E[X^*\|\tilde{X}, \tilde{Y}]$	$Y_{OPE(2)} = E[Y^*\|\tilde{X}, \tilde{Y}]$

probability limit of the least squares estimator ($\hat{\beta}$) using the noisy measures (Y,X), and the limit of the least squares estimator (β^*) using the true values (Y^*,X^*): Sign($\hat{\beta} - \beta^*$).

4.1 Measurement Error in the Regressor

First we consider the case with $\sigma^2_{\eta_Y} = 0$, where the measurement error is confined to the regressor. Thus $Y = \tilde{Y} = Y^*$ under all three models. We first briefly review the CME case. The reported value is $X_{CME} = \tilde{X} = X^* + \eta_X$. The least squares estimator therefore underestimates the coefficient in the regression with the true values:

$$\beta_{CME} = \frac{\text{COV}(Y^*, X_{CME})}{\text{VAR}(X_{CME})} = \beta \cdot \frac{\sigma^2_X}{\sigma^2_X + \sigma^2_{\eta_X}},$$

which is less than β in absolute value. This is the standard case of classical measurement error leading to a bias towards zero.

Next consider the OPE(1) case. The reported value X is linear in the unbiased measurement $\tilde{X}$ with coefficient $(1/\sigma^2_{\eta_X})/(1/\sigma^2_X + 1/\sigma^2_{\eta_X})$:

$$X_{OPE(1)} = E[X^*|\tilde{X}] = E[X^*|X^* + \eta_X]$$

$$= \mu_X \cdot \frac{1/\sigma_X^2}{1/\sigma_X^2 + 1/\sigma_{\eta_X}^2} + \tilde{X} \cdot \frac{1/\sigma_{\eta_X}^2}{1/\sigma_X^2 + 1/\sigma_{\eta_X}^2}.$$

To see the bias from this model consider the regression function

$$Y^* = \alpha + \beta \cdot X^* + \nu = \alpha + \beta \cdot X + \tilde{\nu},$$

with the composite error terms $\tilde{\nu}$ equal to

$$\tilde{\nu} = \nu + \beta \cdot (X^* - X).$$

Since by assumption in the OPE(1) model the reporting error $X - X^*$ is independent of the reported value X, the composite error terms $\tilde{\nu}$ is independent of X and there is no bias resulting from the measurement error, or $\beta_{OPE(1)} = \beta$.

Finally, consider the case where the respondent adjusts the report to take into account not just the unbiased measurement $\tilde{X}$ but also the (accurately measured) outcome Y^*:

$$X_{OPE(2)} = E[X^*|\tilde{X}, \tilde{Y}] = E[X^*|\tilde{X}, Y^*].$$

This can be interpreted as estimating X^* based on two noisy measurements, $\tilde{X} = X^* + \eta_X$ and $(Y^* - \alpha)/\beta = X^* + \nu/\beta$, with uncorrelated errors η_X and ν/β. The resulting reported value is therefore a weighted average of the population mean μ and the two unbiased measurements:

$$\begin{aligned} X_{OPE(2)} &= \frac{1/\sigma_X^2}{1/\sigma_X^2 + 1/\sigma_{\eta_X}^2 + \beta^2/\sigma_\nu^2} \cdot \mu_X + \frac{1/\sigma_{\eta_X}^2}{1/\sigma_X^2 + 1/\sigma_{\eta_X}^2 + \beta^2/\sigma_\nu^2} \cdot \tilde{X} \\ &\quad + \frac{\beta^2/\sigma_\nu^2}{1/\sigma_X^2 + 1/\sigma_{\eta_X}^2 + \beta^2/\sigma_\nu^2} \cdot \frac{Y^* - \alpha}{\beta} \\ &= \lambda_1 \cdot \mu_X + \lambda_2 \cdot \tilde{X} + \lambda_3 \cdot \frac{Y^* - \alpha}{\beta}, \end{aligned}$$

with all $\lambda_j \geq 0$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$. We can rewrite this as a linear function of the true value and independent disturbances:

$$X_{OPE(2)} = \lambda_1 \cdot \mu_X + (\lambda_2 + \lambda_3) \cdot X^* + \lambda_2 \cdot \eta_X + \lambda_3 \cdot \frac{\nu}{\beta}.$$

Simple but tedious calculations show that the probability limit of the least squares estimator is equal to

$$\beta_{OPE(2)} = \beta \cdot \left(1 + \frac{1/\sigma_X^2 + 1/\sigma_{\eta_X}^2 + \beta^2/\sigma_\nu^2}{\beta^2/\sigma_\nu^2 + 1/\sigma_{\eta_X}^2 + \sigma_X^2 \cdot (1/\sigma_{\eta_X}^2 + \beta^2/\sigma_\nu^2)^2}\right),$$

which is greater than β in absolute value.[7] In this case the least squares estimator overestimates the magnitude of the regression coefficient, due to the correlation between the reported value and the disturbance ν in the regression, which is induced by the use of Y^* in producing the best estimate of the regressor X^*.

4.2 Measurement Error in the Outcome Variable

In this subsection we consider measurement error in the outcome variable, and assume the regressor is accurately measured: $\sigma_{\eta_X}^2 = 0$, and thus $X = \tilde{X} = X^*$. Under the CME assumption we can write the regression model as

$$Y = \tilde{Y} = Y^* + \eta_Y = \alpha + \beta \cdot X + \nu + \eta_Y.$$

By assumption both components of the composite error term $\nu + \eta_Y$ are independent of X so there is no bias, and $\beta_{CME} = \beta$.

Next, consider the case where the agent reports $Y = E[Y^*|\tilde{Y}]$. The unconditional mean of Y^* is $\alpha + \beta \cdot \mu_X$, with variance $\beta^2 \cdot \sigma_X^2 + \sigma_\nu^2$, so the best estimate of Y^*, based on $\tilde{Y}$, is

$$Y_{OPE(1)} = (\alpha + \beta \cdot \mu_X) \cdot \frac{1/(\beta^2 \cdot \sigma_X^2 + \sigma_\nu^2)}{1/(\beta^2 \cdot \sigma_X^2 + \sigma_\nu^2) + 1/\sigma_{\eta_Y}^2} + \tilde{Y} \cdot \frac{1/\sigma_{\eta_Y}^2}{1/(\beta^2 \cdot \sigma_X^2 + \sigma_\nu^2) + 1/\sigma_{\eta_Y}^2}.$$

[7] To see the form of the coeffient in a regression of Y^* on $X_{OPE(2)}$, consider the coefficient in a regression of a variable Z on a regressor V when V can be written as the sum of K independent variables V_1 through V_K. Simple manipulations show that this is equal to a variance weighted sum of the slope coefficients from the K regressions of Z on V_k:

$$\beta_{Z,V} = \frac{\mathrm{COV}(Z,V)}{\mathrm{VAR}(V)} = \frac{\mathrm{COV}(Z, \sum_{k=1}^K V_k)}{\mathrm{VAR}(\sum_{k=1}^K V_k)} = \frac{\sum_{k=1}^K \mathrm{COV}(Z, V_k)}{\sum_{k=1}^K \mathrm{VAR}(V_k)}$$

$$= \sum_{k=1}^K \beta_{Z,V_k} \cdot \frac{\mathrm{VAR}(V_k)}{\sum_{i=1}^K \mathrm{VAR}(V_i)}.$$

The slope coefficient in a regression of $\tilde{Y}$ on X^* is β, so the slope coefficient in a regression of $Y_{OPE(1)}$ on X is

$$\beta_{OPE(1)} = \beta \cdot \frac{1/\sigma^2_{\eta_Y}}{1/(\beta^2 \cdot \sigma^2_X + \sigma^2_\nu) + 1/\sigma^2_{\eta_Y}},$$

which means $\beta_{OPE(1)}$ is biased towards zero.

Finally, consider the case where the respondent reports the best estimate of Y given $\tilde{Y}$ and X^*. Based on X^* alone the best estimate of Y^* would be $\alpha + \beta \cdot X^*$. Knowledge of both X^* and $\tilde{Y}$ can be interpreted as knowledge of both $\alpha + \beta \cdot X^*$ and $\tilde{Y} - \alpha - \beta \cdot X^* = \eta_Y + \nu$. Hence we can write

$$\begin{aligned} Y_{OPE(2)} &= E[Y|\tilde{Y}, X^*] = \alpha + \beta \cdot X^* + E[\nu|X^*, \tilde{Y}] \\ &= \alpha + \beta \cdot X^* + E[\nu|X^*, \eta_Y + \nu] = \alpha + \beta \cdot X^* + E[\nu|\eta_Y + \nu], \\ &= \alpha + \beta \cdot X^* + (\eta_Y + \nu) \cdot \frac{1/\sigma^2_\nu}{1/\sigma^2_\nu + 1/\sigma^2_{\eta_Y}}. \end{aligned}$$

Because ν and η_Y are independent of X^* again there is no bias from regressing $Y_{OPE(2)}$ on X^*.

4.3 Measurement Error in Both the Regressor and Outcome Variable

In this subsection we consider the case where both regressor and outcome are measured with error. In each case the individual reporting the variables has available an unbiased measurement,

$$\tilde{X} = X^* + \eta_X, \qquad \tilde{Y} = Y^* + \eta_Y,$$

with possibly correlated errors,

$$\begin{pmatrix} \eta_X \\ \eta_Y \end{pmatrix} \Big| X^*, \nu \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} \sigma^2_{\eta_X} & \rho_{\eta_X \eta_Y} \sigma_{\eta_X} \sigma_{\eta_X} \\ \rho_{\eta_X \eta_Y} \sigma_{\eta_X} \sigma_{\eta_X} & \sigma^2_{\eta_Y} \end{pmatrix} \right).$$

We look at the bias resulting from the three models considered before, CME, OPE(1), and OPE(2). In general, with the errors in $\tilde{X}$ and $\tilde{Y}$, η_X and η_Y respectively, potentially correlated, the biases from measurement error cannot be signed. If the correlation between

Table 2: Measurement Error Bias in Slope Coefficient

	σ_{η_X}	σ_{η_Y}	Reporting Model: CME $X=\tilde{X}$, $Y=\tilde{Y}$	OPE(1) $X=E[X^*\|\tilde{X}]$, $Y=E[Y^*\|\tilde{Y}]$	OPE(2) $X=E[X^*\|\tilde{X},\tilde{Y}]$, $Y=E[Y^*\|\tilde{X},\tilde{Y}]$
No Error	0	0	no bias	no bias	no bias
Error in Regressor Only	>0	0	towards zero	no bias	away from zero
Error in Outcome Only	0	>0	no bias	towards zero	no bias
Error in Both (zero correlation)	>0	>0	towards zero	towards zero	away from zero

the measurement errors is zero, the direction of the bias follows intuitively from the previous calculations. These results, combined with those of Sections 4.1 and 4.2 are reported in Table 2. If the correlation between η_X and η_Y is close enough to one the bias will always be upward, and if it is close enough to negative one, the bias will always be downward. To see how big these effects can be we report in the Section 5 some numerical calculations, based on numbers relevant for wage regressions.

4.4 Instrumental Variables Estimation

One standard approach to dealing with classical measurement error is to use instrumental variables methods (see the Bound, Brown and Mathiowetz (2000) survey for a general discussion). Here we explore what instrumental variables methods do when the measurement error is of the OPE variety. We maintain the linear model structure above, and assume Y^* is observed without error, but X^* is measured with error and two noisy measures are available. We also assume the two reports are optimal predictions based on unbiased and independent measurements:

$$X_1 = E[X^*|\tilde{X}_1], \qquad X_2 = E[X^*|\tilde{X}_2],$$

where $\tilde{X}_1 = X^* + \eta_1$, $\tilde{X}_2 = X^* + \eta_2$, $(\eta_1, \eta_2) \perp X^*$, and $\eta_1 \perp \eta_2$. From Section 4.1 we know that regressing Y^* on X_1 (or X_2) leads to unbiased estimates of β because the measurement error is uncorrelated with the reported value. If instead we use the second measure as an instrument for the first one, we estimate β as

$$\hat{\beta}_{iv1} = \frac{\text{Cov}(Y^*, X_2)}{\text{Cov}(X_1, X_2)} = \beta \cdot \frac{\sigma^2_{\eta_2} + \sigma^2_X}{\sigma^2_X}.$$

Instrumenting for the mismeasured regressor now leads to a bias away from zero, proportional to the inverse of the reliability ratio of the noisy measure. Note that in this case, as before, the data are not informative about the nature of the measurement error. The finding that, as in the Ashenfelter-Krueger (1994) study, instrumenting leads to considerably higher estimates that ordinary least squares estimates, is consistent with both the classical measurement error story as well as with the optimal prediction error model. The interpretation of the results is very different, however, under the two models.

5. Measurement Error in Wage Regressions

In this section we look at the regression of the logarithm of wages on education where both may be measured with error, and interest is in the regression coefficients based on the regression with the true values.[8] We calculate some of the moments of hourly wages and education levels from NLSY data.[9] The earnings measure used is the logarithm of the usual weekly wage, and the education measure is years of completed schooling. The estimated regression function based on these data is

$$\begin{array}{rcccc} \hat{Y}_i & = & 5.16 & + & 0.061 \quad X_i. \\ & & (0.09) & & (0.006) \end{array}$$

The standard deviations of the log wage is $\sigma_Y = 0.43$, and the standard deviation of the education level is $\sigma_X = 2.2$.

[8] In some cases one can argue that interest should be in the regression on perceived values. For example, if individuals do not know their own income with certainly, one may argue that their estimated income is more relevant for consumption decisions than true income. Here we would argue that in answering a survey an individual may have insufficient incentive to carefully check his or her records, and that if the value of the variable is needed for making economically meaningful decisions, one might acquire the relevant information.

[9] See Hellerstein and Imbens (1999) for a discussion of the particular subsample used.

To find appropriate numbers for the measurement error variances we turn to some of the validation studies.[10] For the measurement error in the education level we take our numbers from the Ashenfelter and Krueger (1994) study. Ashenfelter and Krueger asked twins about their own education as well as their twin sibling's level of education. Using those data they estimate a reliability ratio of approximately 90%, implying that the variance of the measurement error is approximately ten percent of the variance of education. We therefore use $\sigma_{\eta_X} = \sqrt{0.1} \times \sigma_X = 0.63$. For log wages we take our numbers from Bound and Krueger (1991) and Pischke (1995) who analyze the validation study of the PSID. Their numbers suggest a reliability ratio of 75%, and hence $\sigma_{\eta_Y} = \sqrt{0.25} \times \sigma_Y = 0.3$. Based on

Table 3: Returns to Education and Percentage Bias in the Presence of Measurement Error: Estimated Return to Education is 0.061

				Reporting Model					
				CME		OPE(1)		OPE(2)	
				$X = \tilde{X}$		$X = E[X^*\|\tilde{X}]$		$X = E[X^*\|\tilde{X}, \tilde{Y}]$	
	σ_{η_X}	σ_{η_Y}	$\rho_{\eta_X \eta_Y}$	$Y = \tilde{Y}$		$Y = E[Y^*\|\tilde{Y}]$		$Y = E[Y^*\|\tilde{X}, \tilde{Y}]$	
No Error	0.00	0.00	–	0.061	(0%)	0.061	(0%)	0.061	(0%)
Error in Regressor	0.63	0.00	–	0.069	(-12%)	0.061	(0%)	0.055	(9%)
Error in Outcome	0.00	0.30	–	0.061	(0%)	0.077	(-27%)	0.061	(0%)
Error in Both	0.63	0.30	-0.90	0.108	(-76%)	0.109	(77%)	0.076	(24%)
	0.63	0.30	-0.50	0.090	(-48%)	0.095	(-55%)	0.068	(-11%)
	0.63	0.30	0.00	0.069	(-12%)	0.077	(-26%)	0.057	(6%)
	0.63	0.30	0.50	0.047	(22%)	0.060	(1%)	0.046	(25%)
	0.63	0.30	0.90	0.030	(50%)	0.046	(24%)	0.036	(40%)

[10] Although these validation studies are obviously different from the NLSY in the way individuals were selected and in the formulation of the questions, and the estimates are all based on the CME assumption, they may be informative about the relative amount of measurement error for the earnings and educations measures.

these error variances and the distribution of the observed variables we calculate the true parameter values β^*, and percentage bias, $(\hat{\beta} - \beta^*)/\beta^* \times 100\%$, under different measurement error scenarios. Table 3 summarizes the results. The results in the first three rows, with measurement error in at most one variable, reflect the qualitative results in Sections 4.1 and 4.2. For example, in the second row, with only measurement error in the regressor, comparing the estimated parameter of 0.061 with the true parameter value of 0.069, implies that the estimated value is biased downward by 12%. The largest bias in these three rows is on the order of 27%. When both variables are measured with error and with the errors correlated the bias can get much larger. With zero correlation the bias for the classical measurement error model is 12%. Allowing the correlation between measurement errors to go to -0.90, the bias goes to 76%, and with the correlation up to 0.90, the bias goes to 50%. Similarly for the other reporting models the bias goes up considerably, also not quite as much as under the CME model.

One conclusion is that the classical measurement error model may overstate the biases associated with measurement error, as well as understate them. A second point is that although classical measurement error alone in the dependent variables does not lead to bias, if correlated with measurement error in the regressors it can affect the results considerably.

6. Conclusion

Whereas the classical measurement error model views the individual as passively reporting a flawed but unbiased measurement, the optimal prediction model implies the individual may interpret the question "what is the value of this variable?" as "what is your best estimate of the value of this variable?" and adjust the raw measurement accordingly. This leads to measurement error that is uncorrelated with variables in the individuals' information set, and therefore by necessity correlated with the true value of the variable of interest.

In the linear regression framework, this implies that, under plausible alternatives to the classical measurement error model, measurement error in the regressor can lead to over- as well as under-estimation of the coefficients of interest. In addition, when both regressor and

outcome variables are measured with potentially correlated errors, biases can be away from zero even in the classical measurement error model. Critical for determining the bias is the model for the individual reporting the mismeasured variables, the content of the individuals' information set, and the correlation structure of the errors. We present some calculations indicating that the range of values consistent with amounts of measurement error found in wages and years of schooling. This range can be quite wide, especially if the measurement errors are correlated.

Whether the classical measurement error model, a version of the optimal prediction error model, or a hybrid model is appropriate depends on the specific context, but there is no reason to believe that the classical measurement error assumption is generally applicable.

References

Ashenfelter, O., and A. Krueger, (1994), "Estimates of the Economic Return to Schooling from a New Sample of Twins", *American Economic Review*, Vol. 84, No. 5, 1157-1173.

Angrist, J., and A. Krueger, (2000), "Empirical Strategies in Labor Economics", in *Handbook of Labor Economics*, Ashenfelter and Card (eds.), Vol. 3.

Bollinger, C., (1996), "Bounding Mean Regressions when a Binary Regressor is Mismeasured" *Journal of Econometrics*, Vol 73, No 2, 387-400.

Bound, J., and A. Krueger, (1991), "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics*, Vol. 9, No. 1, 1–24.

Bound, J., C. Brown, and N. Mathiowetz, (2000), "Measurement Error in Survy Data", forthcoming *Handbook of Econometrics*, Heckman and Leamer (eds.), Vol. 5.

Card, D., (1996), "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis", *Econometrica*, Vol. 64, No. 4, 957-979.

Card, D., and D. Hyslop, (1997), "Does Inflation 'Grease the Wheels of the Labor Market'?", in Christina Romber and David Romer (eds), *Reducing Inflation: Motivation and Strategy*, University of Chicago Press.

Das, M., J. Dominitz, and A. van Soest, (1999), "Comparing Predictions and Outcomes: Theory and Application to Income Changes", Vol. 94, No. 445, 75-85.

Chesher, A., and , (1999),

Griliches, Z., (1987), "Economic Data Issues", in Grilliches and Intrilligator (eds.) *Handbook of Econometrics*, Vol. 3, 1465–1514.

Horowitz, J., and C. Manski, (1995), "Identification and Robustness with Contaminated and Corrupted Data", *Econometrica*, Vol. 63, No 2, 281-302.

Kane, C. Rouse, and D. Staiger, (1998), "Estimating Returns to Schooling When Schooling is Mismeasured", unpublished manuscript, Princeton University.

Klepper, S., and E. Leamer, (1984), "Consistent Sets of Estimates for Regressions with Errors in All Variables", *Econometrica*, Vol. 52, No. 1, 163-183.

Leamer, E., (1987), "Errors in Variables in Linear Systems", *Econometrica*, Vol. 55, No. 5, 893–909.

Mankiw, N. G., and M. Shapiro, (1986) "News of Noise. An Analysis of GNP Revisions", *Survey of Current Business*, 66, 20-25.

Manski, C., (1990), "The Use of Intentions Data to Predict Behavior: A Best-Case Analysis", *Journal of the American Statistical Association*, Vol 85, 934-940.

Philipson, T. (1999), "Missing Data and Incentives", mimeo, Department of Economics, University of Chicago.

Pischke, S., (1995), "Measurement Error and Earnings Dynamics: Some Estimates from the PSID Validation Study", *Journal of Business and Economic Statistics*, Vol. 13, No 3, 305-314.

Rosenbaum, P., and D. Rubin, (1983), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society, Series B*, 45, 212-218.

Schennach, S., (2000), PhD Thesis, Department of Economics, MIT.