Steve Pischke
Spring 2000

Lecture Notes on Measurement Error

These notes summarize a variety of simple results on measurement error which I find useful. They also provide some references where more complete results and applications can be found.

Classical Measurement Error We will start with the simplest regression models with one independent variable. For expositional ease we also assume that both the dependent and the explanatory variable have mean zero. Suppose we wish to estimate the population relationship

$$y = \beta x + \epsilon \tag{1}$$

Unfortunately, we only have data on

$$\widetilde{x} = x + u \tag{2}$$

$$\widetilde{y} = y + v \tag{3}$$

i.e. our observed variables are measured with an additive error. Let's make the following simplifying assumptions

$$E(u) = 0 \tag{4}$$

$$\text{plim } \frac{1}{n}(y'u) = 0 \tag{5}$$

$$\text{plim } \frac{1}{n}(x'u) = 0 \tag{6}$$

$$\text{plim } \frac{1}{n}(\epsilon'u) = 0 \tag{7}$$

The measurement error in the explanatory variable has mean zero, is uncorrelated with the true dependent and independent variables and with the equation error. Also we will start by assuming $\sigma_v^2 = 0$, i.e. there is only measurement error in x. These assumptions define the *classical errors-in-variables model.*

Substitute (2) into (1):

$$y = \beta(\widetilde{x} - u) + \epsilon = y_i = \beta\widetilde{x} + (\epsilon - \beta u) \tag{8}$$

The measurement error in x becomes part of the error term in the regression equation thus creating an endogeneity bias. Since $\widetilde{x}$ and u are positively correlated (from (2)) we can see that OLS estimation will lead to a negative bias in $\widehat{\beta}$ if the true β is positive and a positive bias if β is negative.

To assess the **size of the bias** consider the OLS-estimator for β

$$\widehat{\beta} = \frac{cov(\widetilde{x}, y)}{var(\widetilde{x})} = \frac{cov(x+u, \beta x + \epsilon)}{var(x+u)}$$

and

$$\text{plim } \widehat{\beta} = \frac{\beta \sigma_x^2}{\sigma_x^2 + \sigma_u^2} = \lambda \beta$$

where

$$\lambda \equiv \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$$

The quantity λ is referred to as reliability or signal-to-total variance ratio. Since $0 < \lambda < 1$ the coefficient $\widehat{\beta}$ will be biased towards zero. This bias is therefore called *attenuation bias* and λ is the attenuation factor in this case.

The bias is

$$\text{plim } \widehat{\beta} - \beta = \lambda\beta - \beta = -(1-\lambda)\beta = -\frac{\sigma_u^2}{\sigma_x^2 + \sigma_u^2}\beta$$

which again brings out the fact that the bias depends on the sign and size of β.

In order to figure out what happens to the **estimated standard error** first consider estimating the residual variance from the regression

$$\widehat{\epsilon} = y - \widehat{\beta}\widetilde{x} = y - \widehat{\beta}(x+u)$$

Add and subtract the true error $\epsilon = y - \beta x$ from this equation and collect terms.

$$\begin{aligned} \widehat{\epsilon} &= \epsilon - (y - \beta x) + y - \widehat{\beta} x - \widehat{\beta} u \\ &= \epsilon + (\beta - \widehat{\beta}) x - \widehat{\beta} u \end{aligned}$$

You notice that the residual contains two additional sources of variation compared to the true error. The first is due to the fact that $\widehat{\beta}$ is biased towards zero. Unlike in the absence of measurement error the term $\widehat{\beta} - \beta$ does not vanish asymptotically. The second term is due to the additional variance introduced by the presence of measurement error in the regressor. Note that by assumption the three random variables ϵ, x, and u in this

equation are uncorrelated. We therefore obtain for the estimated variance of the equation error

$$\text{plim } \widehat{\sigma_\epsilon^2} = \sigma_\epsilon^2 + (1-\lambda)^2\beta^2\sigma_x^2 + \lambda^2\beta^2\sigma_u^2$$

For the estimate of the variance of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$, call it $\widehat{s}$, we have

$$\text{plim } \widehat{s} = \text{plim } \frac{\widehat{\sigma_\epsilon^2}}{\widehat{\sigma_{\widetilde{x}}^2}} = \frac{\sigma_\epsilon^2 + (1-\lambda)^2\beta^2\sigma_x^2 + \lambda^2\beta^2\sigma_u^2}{\sigma_x^2 + \sigma_u^2}$$

$$= \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}\left(\frac{\sigma_\epsilon^2}{\sigma_x^2}\right) + \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}(1-\lambda)^2\beta^2 + \frac{\sigma_u^2}{\sigma_x^2+\sigma_u^2}\lambda^2\beta^2$$

$$= \lambda\frac{\sigma_\epsilon^2}{\sigma_x^2} + \lambda(1-\lambda)^2\beta^2 + \lambda^2(1-\lambda)\beta^2$$

$$= \lambda s + \lambda(1-\lambda)\beta^2$$

The first term indicates that the true standard error is underestimated in proportion to λ. Since the second term is positive we cannot sign the overall bias in the estimated standard error.

However, the t-statistic will be biased downwards. The t-ratio converges to

$$\frac{\text{plim } t}{\sqrt{n}} = \frac{\text{plim } \widehat{\beta}}{\text{plim } \sqrt{\widehat{s}}} = \frac{\lambda\beta}{\sqrt{\lambda s + \lambda(1-\lambda)\beta^2}}$$

$$= \sqrt{\lambda}\frac{\beta}{\sqrt{s + (1-\lambda)\beta^2}}$$

which is smaller than $\beta/\sqrt{s}$.

Simple Extensions Next, consider measurement error in the dependent variable y, i.e. let $\sigma_v^2 > 0$ while $\sigma_u^2 = 0$. Substitute (3) into (1):

$$\widetilde{y} = \beta x + \epsilon + v$$

Since v is uncorrelated with x we can estimate β consistently by OLS in this case. Of course, the estimates will be less precise than with perfect data.

Return to the case where there is measurement error only in x. The fact that measurement error in the dependent variable is more innocuous than measurement error in the independent variable might suggest that we run

the **reverse regression** of x on y thus avoiding the bias from measurement error. Unfortunately, this does not solve the problem. Reverse (8) to obtain

$$\widetilde{x} = \frac{1}{\beta} y - \frac{1}{\beta}\epsilon + u$$

u and y are uncorrelated by assumption but y is correlated with the equation error ϵ now. So we have cured the regression of errors-in-variables bias but created an endogeneity problem instead. Note, however, that this regression is still useful because ϵ and y are negatively correlated so that $\widehat{1/\beta}$ is biased downwards, implying an upward bias for $\widehat{\beta}_r = 1/\left(\widehat{1/\beta}\right)$. Thus the results from the standard regression and from the reverse regression will bracket the true coefficient, i.e. plim $\widehat{\beta} < \beta <$ plim $\widehat{\beta}_r$. Implicitly, this bracketing result uses the fact that we know that σ_ϵ^2 and σ_u^2 have to be positive. The bounds of this interval are obtained whenever one of the two variances is zero. This implies that the interval tends to be large when these variances are large. In practice the bracketing result is therefore often not very informative. The bracketing result extends to multivariate regressions: in the case of two regressors you can run the original as well as two reverse regressions. The results will imply that the true (β_1,β_2) lies inside the triangular area mapped out by these three regressions, and so forth for more regressors [Klepper and Leamer (1984)].

Another useful fact to notice is that **data transformations** will typically magnify the measurement error problem. Assume you want to estimate the relationship

$$y = \beta x + \gamma x^2 + \epsilon$$

Under normality the attenuation factor for $\widehat{\gamma}$ will be the square of the attenuation factor for $\widehat{\beta}$ [Griliches (1986)].

So what can we do to get consistent estimates of β?

- If either σ_x^2, σ_u^2, or λ is known we can make the appropriate adjustment for the bias in β. Either one of these is sufficient as we can estimate $\sigma_x^2+\sigma_u^2$ (= plim $var(\widetilde{x})$) consistently. Such information may come from validation studies of our data. In grouped data estimation, i.e. regression on cell means, the sampling error introduced by the fact that the means are calculated from a sample can be estimated [Deaton (1985)]. This only matters if cell sizes are small; grouped data estimation yields consistent estimates with cell sizes going to infinity (but not with the number of cells going to infinity at constant cell sizes).

- Any instrument z correlated with x but uncorrelated with u will identify the true coefficient since

$$\widehat{\beta}_{IV} = \frac{cov(y,z)}{cov(\widetilde{x},z)} = \frac{cov(\beta x + \epsilon, z)}{cov(x+u,z)}$$

$$\text{plim } \widehat{\beta}_{IV} = \frac{\beta \sigma_{xz}}{\sigma_{xz}} = \beta$$

  In this case it is also possible to get a consistent estimate of the population $R^2 = \beta^2 \sigma_x^2 / \sigma_y^2$. The estimator

$$\widehat{R^2} = \widehat{\beta}_{IV} \frac{cov(y,\widetilde{x})}{var(y)} = \frac{\widehat{\beta}_{IV}}{\widehat{\beta}_r}$$

  which is the product of the IV coefficient and the OLS coefficient from the reverse regression, yields

$$\text{plim } \widehat{R^2} = \beta \frac{\beta \sigma_x^2}{\sigma_y^2} = R^2$$

- Get better data.

Panel Data Often we are interested in using panel data to eliminate fixed effects. How does measurement error affect the fixed effects estimator? Extend the one variable model in (1) to include a fixed effect:

$$y_{it} = \beta x_{it} + \mu_i + \epsilon_{it} \tag{9}$$

Difference this to eliminate the fixed effect μ_i.

$$y_{it} - y_{it-1} = \beta(x_{it} - x_{it-1}) + \epsilon_{it} - \epsilon_{it-1}$$

As before we only observe $\widetilde{x}_{it} = x_{it} + u_{it}$. Using our results from above

$$\text{plim } \widehat{\beta} = \beta \frac{\sigma_{\Delta x}^2}{\sigma_{\Delta x}^2 + \sigma_{\Delta u}^2}$$

So we have to figure out how the variance in the changes of x relates to the variance in the levels.

$$\sigma_{\Delta x}^2 = var(x_t) - 2cov(x_t, x_{t-1}) + var(x_{t-1})$$

If the process for x_t is stationary this simplifies to

$$\begin{aligned}\sigma^2_{\Delta x} &= 2\sigma^2_x - 2cov(x_t, x_{t-1}) \\ &= 2\sigma^2_x(1-\rho)\end{aligned}$$

where ρ is the first order autocorrelation coefficient in x_t. Similarly, define r to be the autocorrelation coefficient in u_t so we can write

$$\begin{aligned}\text{plim } \widehat{\beta} &= \beta \ \frac{\sigma^2_x(1-\rho)}{\sigma^2_x(1-\rho) + \sigma^2_u(1-r)} \\ &= \beta \ \frac{1}{1 + \frac{\sigma^2_u(1-r)}{\sigma^2_x(1-\rho)}}\end{aligned}$$

In the special case where both x_t and u_t are uncorrelated over time the attenuation bias for the fixed effects estimator simplifies to the original λ. Fixed effects estimation is particularly worrisome when $r = 0$, i.e. the measurement error is just serially uncorrelated noise, while the signal is highly correlated over time. In this case, differencing doubles the variance of the measurement error while it might reduce the variance of the signal. In the effort to eliminate the bias arising from the fixed effect we have introduced additional bias due to measurement error. Of course, differencing is highly desirable if the measurement error $u_{it} = u_i$ is a fixed effect itself. In this case differencing eliminates the measurement error completely. In general, differencing is desirable when $r > \rho$. For panel earnings data $\rho \approx 2r$ [Bound et.al. (1994)], [Bound and Krueger (1991)].

Sometimes it is reasonable to make specific assumptions about the behavior of the measurement error over time. For example, if we are willing to assume that u_{it} is i.i.d. while the x's are correlated then it is possible to identify the true β even in relatively short panels. The simplest way to think about this is in a four period panel. Form differences between the third and second period and instrument these with differences between the fourth and the first period. Obviously

$$\text{plim } \frac{1}{n}(u_4 - u_1)'(u_3 - u_2) = 0$$

by the i.i.d. assumption for u_{it}. The long and short differences for x_{it} will be correlated, on the other hand, since the x's are correlated over time. We have constructed a valid instrument. This example makes much stronger assumptions than are necessary. Alternatively, with four periods and the i.i.d. assumption for u_{it} we can come up with much more efficient estimators since

other valid instruments can be constructed [Griliches and Hausman (1986)]. They also point out that comparing the results from first difference estimates, long difference estimates, and deviations from means estimates provides a useful test for measurement error if $\rho \neq r$ since the attenuation bias varies depending on the specific estimator chosen. But be aware that the same happens if your model is misspecified in some other way, for example there are neglected true dynamics in your x's, so your test only indicates "some misspecification."

Multivariate Models Return to OLS estimation in a simple cross-section and consider what happens to the bias as we add more variables to the model. Consider the equation

$$y = \beta x + \gamma w + \epsilon \tag{10}$$

Even if only $\widetilde{x}$ is subject to measurement error while w is measured correctly both parameters will in general be biased now. $\widehat{\gamma}$ is unbiased when the two regressors are uncorrelated. $\widehat{\beta}$ is still biased towards zero. We can also determine how the bias in $\widehat{\beta}$ in the multivariate regression is related to the attenuation bias in the bivariate regression (which may also suffer from omitted variable bias now). To figure this out, consider the formula for $\widehat{\beta}$ in the two variable case

$$\widehat{\beta} = \frac{var(w)cov(y,\widetilde{x}) - cov(w,\widetilde{x})cov(y,w)}{var(\widetilde{x})var(w) - cov(w,\widetilde{x})^2}$$

Thus we obtain

$$\text{plim } \widehat{\beta} = \frac{\sigma_w^2(\beta\sigma_x^2 + \gamma\sigma_{xw}) - \sigma_{\widetilde{x}w}(\gamma\sigma_w^2 + \beta\sigma_{xw})}{\sigma_w^2(\sigma_x^2 + \sigma_u^2) - (\sigma_{\widetilde{x}w})^2}$$

$$= \frac{\beta\left(\sigma_w^2\sigma_x^2 - \sigma_{\widetilde{x}w}\sigma_{xw}\right) + \gamma\sigma_w^2(\sigma_{xw} - \sigma_{\widetilde{x}w})}{\sigma_w^2(\sigma_x^2 + \sigma_u^2) - (\sigma_{\widetilde{x}w})^2}$$

This does not get us much further. However, in the special case where w is only correlated with x but not with u, this can be simplified because now $\sigma_{xw} = \sigma_{\widetilde{x}w}$ so that

$$\text{plim } \widehat{\beta} = \frac{\beta\left(\sigma_w^2\sigma_x^2 - (\sigma_{xw})^2\right)}{\sigma_w^2(\sigma_x^2 + \sigma_u^2) - (\sigma_{xw})^2} = \beta\lambda' \tag{11}$$

Notice that $\sigma_w^2\sigma_x^2 > (\sigma_{xw})^2$ which proves that $\widehat{\beta}$ is biased towards zero. There are various ways to rewrite (11). I find it instructive to look at the representation of the attenuation factor λ' in terms of the reliability ratio λ

and the R^2 of a regression of $\widetilde{x}$ on w. Since this is a one variable regression the population R^2 is just the square of the correlation coefficient of the variables

$$R^2_{\widetilde{x}w} = \frac{(\sigma_{xw})^2}{\sigma^2_w(\sigma^2_x + \sigma^2_u)}$$

Dividing numerator and denominator in (11) by $(\sigma^2_x + \sigma^2_u)$ yields the following expression for the attenuation factor

$$\lambda' = \frac{\sigma^2_w\lambda - \sigma^2_w R^2_{\widetilde{x}w}}{\sigma^2_w - \sigma^2_w R^2_{\widetilde{x}w}} = \frac{\lambda - R^2_{\widetilde{x}w}}{1 - R^2_{\widetilde{x}w}}$$

This formula is quite intuitive. It says the following: if there is no omitted variable bias from estimating (1) instead of (10) because the true $\gamma = 0$, then the attenuation bias will *increase* as additional regressors (correlated with x) are added since the expression above is decreasing in $R^2_{\widetilde{x}w}$. What is going on is that the additional regressor w will now serve as a proxy for part of the signal in x. Therefore, the partial correlation between y and $\widetilde{x}$ will be attenuated more, since some of the signal has been taken care of by the w already. Notice that $R^2_{\widetilde{x}w} < \lambda$ because w is only correlated with x but not with u. Hence $0 < \lambda' < \lambda < 1$.

In the special case just discussed, and if x and w are positively correlated, the bias in $\widehat{\gamma}$ will have the opposite sign of the bias in $\widehat{\beta}$. In fact, with the additional assumption that $\sigma^2_x = \sigma^2_w$ we have

$$\text{plim } \widehat{\gamma} - \gamma = \rho_{xw}\left(1 - \lambda'\right)\beta = -\rho_{xw}\left(\text{plim } \widehat{\beta} - \beta\right)$$

where ρ_{xw} is the correlation coefficient between x and w.

When $\gamma \neq 0$, comparisons between the bivariate regression of y on $\widetilde{x}$ and the multivariate model including w are harder to interpret because we have to keep in mind that the bivariate regression is now also subject to omitted variable bias. Some results are available for special cases. If $\beta > 0, \gamma > 0$ and x and w are positively correlated (but w is still uncorrelated with u) then the probability limit of the estimated $\widehat{\beta}$ in the multivariate regression will be lower than in the bivariate regression [Maddala (1977), p. 304-305]. This follows because adding w to the regression purges it of the (positive) omitted variable bias while introducing additional (negative) attenuation bias. This example also makes it clear that no such statements will be possible if the omitted variable bias is negative.

Non-classical Measurement Error We will now start relaxing the classical assumptions. Return to the model (1) and (2) but drop assumption (6) that x and u are uncorrelated. Recall that

$$\widehat{\beta} = \frac{cov(x+u, \beta x + \epsilon)}{var(x+u)}$$

so that we have in this case

$$\text{plim } \widehat{\beta} = \frac{\beta(\sigma_x^2 + \sigma_{xu})}{\sigma_x^2 + \sigma_u^2 + 2\sigma_{xu}}$$

$$= \left(1 - \frac{(\sigma_u^2 + \sigma_{xu})}{\sigma_x^2 + \sigma_u^2 + 2\sigma_{xu}}\right)\beta = (1 - b_{u\widetilde{x}})\beta \tag{12}$$

Notice that the numerator in $b_{u\widetilde{x}}$ is the covariance between $\widetilde{x}$ and u. Thus, $b_{u\widetilde{x}}$ is the regression coefficient of a regression of u on $\widetilde{x}$. The classical case is a special case of this where this regression coefficient $b_{u\widetilde{x}} = 1 - \lambda$. The derivative of $1 - b_{u\widetilde{x}}$ with respect to σ_{xu} has the sign of $\sigma_u^2 - \sigma_x^2$. Starting from a situation where $\sigma_{xu} = 0$ (classical measurement error) increasing this covariance increases the attenuation factor (decreases the bias) if more than half of the variance in $\widetilde{x}$ is measurement error and decreases it otherwise. In earnings data this covariance tends to be negative [Bound and Krueger (1991), they call this *mean reverting measurement error*]. If $\widetilde{x}$ consisted mostly of measurement error then a more negative σ_{xu} implies a lower attenuation factor and may even reverse the sign of the estimated β.

Measurement error in the dependent variable that is correlated with the true y or with the x's can be analyzed along similar lines. A general framework for this is provided by [Bound et.al. (1994)]. Make X an $n \times k$ matrix of covariates, β a k vector of coefficients, etc. so that (1) becomes

$$y = X\beta + \epsilon$$

Then

$$\begin{aligned}\widehat{\beta} = (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{y} &= (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'(\widetilde{X}\beta - U\beta + v + \epsilon) \\ &= \beta + (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'(-U\beta + v + \epsilon)\end{aligned}$$

and

$$\text{plim } \widehat{\beta} = \beta + \text{plim}(\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'(-U\beta + v)$$

Collecting the measurement errors in a matrix

$$W = [U \mid v]$$

yields

$$\text{plim } \widehat{\beta} = \beta + \text{plim}(\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'W \begin{bmatrix} -\beta \\ 1 \end{bmatrix} \tag{13}$$

so that the biases in more general cases can always be thought of in terms of regression coefficients from regressing the measurement errors on the mismeasured covariates. Special cases like (12) are easily obtained from (13). These regression coefficients of the measurement errors on the mismeasured covariates are therefore what validation studies ought to focus on. Even when the measurement errors are correlated with the true variable or with other regressors or with the dependent variable we can get consistent estimates by using instrumental variables as long as the instruments are only correlated with true X's but not with any of the measurement errors.

A Small Detour: Group Aggregates Occasionally, we run into the following problem. We wish to estimate the standard regression model

$$y_{it} = \beta x_{it} + \epsilon_{it}$$

but instead of x_{it} we only observe the group or time average x_t. For example, we may wish to estimate a wage curve, where y_{it} are individual level wages over time and x_t is the aggregate unemployment rate, or x_{it} might be class size in school, but you only know class size at the school level and not at the individual level. Obviously, x_t is an error ridden version of the true regressor x_{it}. Typically, x_t will be the mean of x_{it}, often from a larger sample or in the population so that $x_{it} = x_t + u_{it}$, and u_{it} will be uncorrelated with x_t. If this is the case, the OLS estimator of β is consistent. It is easy to see that this is true:

$$\widehat{\beta} = \frac{cov(y_{it}, x_t)}{var(x_t)} = \frac{cov(\beta x_{it} + \epsilon_{it}, x_t)}{var(x_t)} = \frac{cov(\beta(x_t + u_{it}) + \epsilon_{it}, x_t)}{var(x_t)}$$

so that

$$\text{plim } \widehat{\beta} = \frac{\beta\sigma^2_{x_t}}{\sigma^2_{x_t}} = \beta$$

While this looks similar to a classical measurement error problem, it is not. In the classical case the observed regressor x_t equals the true regressor plus noise that is uncorrelated with the truth. Here, the true regressor x_{it} equals the observed regressor plus noise that is uncorrelated with the observed regressor. In terms of the notation we developed above, the covariance between the true x and the measurement error $\sigma_{xu} = -\sigma^2_u$. The

negative covariance of the measurement error with the true regressor just cancels the effect of the measurement error, or $b_{u\widetilde{x}} = 0$ in (12). Therefore, our estimates are consistent. Moreover, OLS using the group average will yield correct standard errors. These will be larger, of course, than in the case where the micro level regressor x_{it} is available.

Measurement Error in Dummy Variables There is an interesting special case of non-classical measurement error: that of a binary regressor. Obviously, misclassification of a dummy variable cannot lead to classical measurement error. If the dummy is one, measurement error can only be negative; if the dummy is zero, it can only be positive. So the measurement error is negatively correlated with the true variable. This problem has enough structure that it is worthwhile looking at it separately. Consider the regression

$$y_i = \alpha + \beta d_i + \epsilon_i \tag{14}$$

where $d_i \in \{0, 1\}$. For concreteness, think of y_i as wages, $d_i = 1$ as union members and $d_i = 0$ as nonmembers so that β is the union wage differential. It is useful to note that the OLS estimate of β is the difference between the mean of y_i as $d_i = 1$ and the mean as $d_i = 0$. Instead of d we observe a variable $\widetilde{d}$ that misclassifies some observations. Take expectations of (14) conditional on the observed value of d_i:

$$E(y_i \mid \widetilde{d}_i = 1) = \alpha + \beta P(d_i = 1 \mid \widetilde{d}_i = 1)$$
$$E(y_i \mid \widetilde{d}_i = 0) = \alpha + \beta P(d_i = 1 \mid \widetilde{d}_i = 0)$$

The regression coefficient for the union wage differential is the sample analogue of the difference between these two, so it satisfies

$$\text{plim } \widehat{\beta} = \beta \left[P(d_i = 1 \mid \widetilde{d}_i = 1) - P(d_i = 1 \mid \widetilde{d}_i = 0) \right] \tag{15}$$

Equation (15) says that β will be attenuated because some (high wage) union members are classified as nonmembers while some (low wage) nonmembers are classified as members.

We need some further notation. Let q_1 be the probability that we observe somebody to be a union member when he truly is, i.e. $q_1 \equiv P(\widetilde{d}_i = 1 \mid d_i = 1)$, and similarly $q_0 \equiv P(\widetilde{d}_i = 1 \mid d_i = 0)$. Thus $1 - q_1$ is the probability that a member is misclassified and q_0 is the probability that a nonmember is misclassified. Furthermore, let $\pi \equiv P(d_i = 1)$ be the true membership rate. Notice that the estimate of π given by $\widehat{\pi} = N^{-1} \sum \widetilde{d}_i$ satisfies

$$\text{plim } \widehat{\pi} = \pi q_1 + (1 - \pi) q_0$$

Return to equation (15). By Bayes' Rule we can write the terms that appear in this equation as

$$P(d_i = 1 \mid \widetilde{d}_i = 1) = \frac{P(\widetilde{d}_i = 1 \mid d_i = 1) \cdot P(d_i = 1)}{P(\widetilde{d}_i = 1)} = \frac{\pi q_1}{\pi q_1 + (1-\pi)q_0}$$

and

$$P(d_i = 1 \mid \widetilde{d}_i = 0) = \frac{\pi(1-q_1)}{\pi(1-q_1) + (1-\pi)(1-q_0)}$$

and substituting back into (15) yields

$$\text{plim } \widehat{\beta} = \beta \left[\frac{\pi q_1}{\pi q_1 + (1-\pi)q_0} - \frac{\pi(1-q_1)}{\pi(1-q_1) + (1-\pi)(1-q_0)} \right] \tag{16}$$

Absent knowledge about q_1 and q_0 we cannot identify the true β and π from our data, i.e. from the estimates $\widehat{\beta}$ and $\widehat{\pi}$. In a multivariate regression, no simple formula like (16) is available, although β and π can still be identified if q_1 and q_0 are known [Aigner (1973)].

If we have panel data available and we are willing to impose the restriction that q_1 and q_0 do not change over time all coefficients will be identified. In fact, even with just a two period panel there is already one overidentifying restriction. To see this, notice that there are now not two states for union status but four possible transitions (continuous union members, continuous nonmembers, union entrants and union leavers). The key is that there have to be some switchers in the data. Then we can observe separate changes in y over time for each of the four transition groups. Furthermore, we observe three independent transition probabilities. This makes a total of seven moments calculated from the data. From these we have to identify β, q_1, q_0, and the three true transition probabilities, i.e. only six parameters. The algebra is much messier [Card (1996)]. See [Krueger and Summers (1988)] for results on measurement error in multinomial variables (e.g. industry classifications).

Instrumental Variables Estimation of the Dummy Variable Model
Suppose we have another binary variable z_i available, which has the same properties as the mismeasured dummy variable $\widetilde{d}_i$. Can we use z_i as an instrument in the estimation of (14)? Instrumental variables estimation will not yield a consistent estimate of β in this case. The reason for this is simple. Recall that the measurement error can only be either -1 or 0 (when $d_i = 1$), or 1 or 0 (when $d_i = 0$). This means that the measurement errors in two mismeasured variables will be positively correlated.

In order to study this case, define $h_1 \equiv P(z_i = 1 \mid d_i = 1)$ and $h_0 \equiv P(z_i = 1 \mid d_i = 0)$. The IV estimator in this case is simply the Wald estimator so that

$$\text{plim } \widehat{\beta}_{IV} = \frac{E(y_i \mid z_i = 1) - E(y_i \mid z_i = 0)}{E(\widetilde{d}_i \mid z_i = 1) - E(\widetilde{d}_i \mid z_i = 0)}. \tag{17}$$

The numerator has the same form as (15) with z_i replacing $\widetilde{d}_i$. The terms in the denominator can also easily be derived:

$$\begin{aligned}
E(\widetilde{d}_i \mid z_i = 1) &= P(\widetilde{d}_i = 1 \mid z_i = 1) \\
&= \frac{P(\widetilde{d}_i = 1, z_i = 1)}{P(z_i = 1)} \\
&= \frac{P(\widetilde{d}_i = 1, z_i = 1 \mid d_i = 1)P(d_i = 1) + P(\widetilde{d}_i = 1, z_i = 1 \mid d_i = 0)P(d_i = 0)}{P(z_i = 1 \mid d_i = 1)P(d_i = 1) + P(z_i = 1 \mid d_i = 0)P(d_i = 0)} \\
&= \frac{q_1 h_1 \pi + q_0 h_0 (1 - \pi)}{h_1 \pi + h_0 (1 - \pi)}
\end{aligned}$$

and similarly for $E(\widetilde{d}_i \mid z_i = 0)$. Substituting everything into (17) yields

$$\text{plim } \widehat{\beta}_{IV} = \frac{\beta \left[\frac{\pi h_1}{h_1 \pi + h_0 (1-\pi)} - \frac{\pi (1-h_1)}{(1-h_1)\pi + (1-h_0)(1-\pi)} \right]}{\frac{q_1 h_1 \pi + q_0 h_0 (1-\pi)}{h_1 \pi + h_0 (1-\pi)} - \frac{q_1 (1-h_1)\pi + q_0 (1-h_0)(1-\pi)}{(1-h_1)\pi + (1-h_0)(1-\pi)}}.$$

With some elementary algebra this simplifies to

$$\text{plim } \widehat{\beta}_{IV} = \frac{\beta}{q_1 - q_0}.$$

The IV estimate of β is biased by a factor $1/(q_1 - q_0)$. This has some interesting features. The bias only depends on the misclassification rates in the variable $\widetilde{d}_i$ which is being used as the endogenous regressor. This is because more misclassification in the instrument will lead to a smaller first stage coefficient. Since generally $1 > q_1 > q_0 > 0$, IV will be biased upwards. Hence, OLS and IV estimation could be used to bound the true coefficient.

However, the true coefficient is actually identified from the data, using an idea analogous to the panel data case above [Kane, Rouse, and Staiger (1999)]. There are seven sample moments which can be computed from the data. There are four cells defined by the cross-tabulation of $\tilde{d}_i$ and z_i. The mean of y_i can be computed for each of these cells. In addition, we have three independent sampling fractions for the cross-tabulation. This makes a total of seven empirical moments. From these moments we have to identify α, β, π, q_0, q_1, h_0, and h_1, i.e. seven parameters. These parameters are indeed just identified and can be estimated by method of moments.

References

[Aigner (1973)] "Regression With a Binary Independent Variable Subject to Errors of Observation," *Journal of Econometrics* 1, 49-60.

[Bound et.al. (1994)] "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data," *Journal of Labor Economics* 12, 345-368

[Bound and Krueger (1991)] "The Extend of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics* 9, 1-24.

[Card (1996)] "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica* 64, 957-979

[Deaton (1985)] "Panel Data from Time Series of Cross-Sections," *Journal of Econometrics* 30, 109-126

[Griliches (1986)] "Data Problems in Econometrics," in: Zvi Griliches and Michael Intriligator, eds., *Handbook of Econometrics,* vol. 3, Amsterdam: North Holland, 1465-1514

[Griliches and Hausman (1986)] "Errors in Variables in Panel Data" *Journal of Econometrics* 31, 93-118

[Kane, Rouse, and Staiger (1999)] "Estimating Returns to Schooling When Schooling is Misreported," NBER Working Paper No. 7235

[Klepper and Leamer (1984)] "Consistent Sets of Estimates for Regressions with Errors in All Variables," *Econometrica* 52, 163-183

[Krueger and Summers (1988)] "Efficiency Wages and the Inter-Industry Wage Structure," *Econometrica* 56, 259-293.

[Maddala (1977)] *Econometrics,* New York: McGraw Hill