1 Endogeneity

Consider the following model in the population,

$$\begin{aligned} y &= \beta_1 + x_2\beta_2 + ... + x_k\beta_k + u \\ &= x\beta + u \end{aligned}$$

Above equation requires classical A1-A4 assumptions to be identified, or at least to find the consistent estimator of β, notation of which often comes with hat, $\hat{\beta}$. An explanatory variable x_j is said to be **endogenous** if,

$$E(x_j u) \neq 0$$

That is, **a regressor is endogenous if it is correlated with the error** u. If x_j is uncorrelated with u, we say that x_j is **exogenous**. More specifically, we call this a violation of A3Rsru

If a regressor is endogenous then $\hat{\beta}$, the OLS estimator of β, is not consistent. It is important to understand that endogeneity of a single regressor usually makes the OLS estimator of all k parameters inconsistent. In other words, if one regressor is endogenous, unless all regressors are orthogonal to each other, any type of regression, be it linear multivariate to factor analysis, or even computer-based machine learning models.

The extent to which this occurs depends on the correlation between the endogenous variable and the other regressors. Suppose the last regressor, x_k, is endogenous, $E(x_k u) \neq 0$, but the others are exogenous. Examining the OLS consistency proof we arrive at:

$$\begin{aligned} p\lim\hat{\beta} &= \beta + (p\lim\frac{X'X}{n})^{-1}\ p\lim\frac{1}{n}X'U \\ &= \beta + [E(x'x)]^{-1}\, E(x'u) \\ &= \beta + [E(x'x)]^{-1} \begin{bmatrix} \begin{pmatrix} 0 \\ \vdots \\ E(x_k u) \end{pmatrix} \end{bmatrix} \end{aligned}$$

So unless the first $k-1$ elements in the k_{th} column of $[E(x'x)]^{-1}$ are zero, endogeneity of x_k affects the plim of the other coefficient estimators.

This definition of endogeneity / exogeneity implies that we are concerned only with the consistency of the OLS estimator. **Exogeneity implies consistency but not unbiasedness of OLS**.

In some textbooks, you might find a more strict definition of exogeneity, namely that $E(u|x_j) = 0$ or A3Rmi and, in this case, an exogenous regressor would also imply unbiasedness of OLS. Notice also that the definition of endogeneity / exogeneity refers to a specific model: x_j can be exogenous in one equation but endogenous in another.

Endogeneity cannot be tested directly. Condition $E(x_j u) = 0$ is not directly verifiable because u is not observed. Using the OLS residuals instead of u is pointless because the residuals $\hat{u}$ are always orthogonal to x_j by construction - $\Sigma_{i=1}^{n}\hat{u}_i x_{ij} = 0$ for any $j = 1, ..., k$ - regardless of the correlation between u and x_j.

However, with additional information (on instrumental variables) we can design specification tests which test for $E(x_j u) = 0$, which will be discussed later in this note. In applications, endogeneity usually arises in one of three ways:

1. Omitted variables
2. Simultaneity
3. Measurement error

In this class, we focus on measurement errors.

2 Measurement Error definitions

The data are usually measured with errors. Even though these errors may average to zero, the OLS estimator will be inconsistent when errors of measurement are present. To see this point, let us analyze a simple model using a single regressor

$$y = \alpha + \beta x^* + v \quad E(v|x^*) = 0$$

but the true regressor x^* is not observed, so that this equation cannot be estimated.
Instead we observe x, where

$$x = x^* + \epsilon \quad E(\epsilon|x^*) = 0$$

ϵ is the measurement error, and therefore we say that x^* is measured with error. In the classical error-in-variables (EIV) formulation, it is assumed that

$$\begin{pmatrix} \epsilon_i \\ v_i \end{pmatrix} \sim i.i.d. \left(0, \begin{pmatrix} \sigma_{\epsilon\epsilon} & 0 \\ 0 & \sigma_{vv} \end{pmatrix}\right)$$

In order to understand the effect of using x instead of x^*, we replace x^* with x to obtain an estimable equation,

$$\begin{aligned} y &= \alpha + \beta(x - \epsilon) + v \\ &= \alpha + \beta x - \beta\epsilon + v \\ &= \alpha + \beta x + u, \quad u = v - \beta\epsilon \end{aligned}$$

Hence,

$$E(xu) = E(x^* + \epsilon)(v - \beta\epsilon) = -\beta\sigma_{\epsilon\epsilon} \neq 0$$

in general.

Thus, x is endogenous and OLS would not give a consistent estimator of β. Note that the error-in-variables problem can be interpreted as an omitted variable case: If we add ϵ to the regression then the problem would disappear. In other words, if we observe the measurement error then we can recover the true regressor.

Recall that the OLS estimator of $\hat{\beta}$ is $\hat{\beta} = \beta + \frac{\Sigma_{i=1}^n (x_i - \bar{x})u_i}{\Sigma_{i=1}^n (x_i - \bar{x})^2}$. The plim of b is obtained as follows:

$$\begin{aligned} p\text{lim}\hat{\beta} &= \beta + \frac{p\text{lim}\frac{1}{n}\Sigma_{i=1}^n (x_i - \bar{x})u_i}{p\text{lim}\frac{1}{n}\Sigma_{i=1}^n (x_i - \bar{x})^2} \\ &= \beta + \frac{Cov(x,u)}{V(x)} \\ &= \beta - \frac{\beta\sigma_{\epsilon\epsilon}}{V(x^*) + \sigma_{\epsilon\epsilon}} \\ &= \beta(\frac{V(x^*)}{V(x^*) + \sigma_{\epsilon\epsilon}}) \end{aligned}$$

Note that in this case the OLS estimator underestimates the true parameter if $\beta > 0$, and overestimates β if $\beta < 0$. So it is generally said that classical EIV makes the OLS estimator to be "attenuated" towards zero. This is a powerful result because we not only know that the OLS estimator is inconsistent but also the direction of the inconsistency. In this case, the bias or inconsistency depends on the ratio of error variance to true variance.

Sometimes there is confusion between the EIV formulation and the specification of x as a proxy variable for x^* (which is unobserved). Recall that a proxy variable x satisfies $x^* = \theta_0 + \theta_1 x + r$ and $Cov(x, r) = 0$.

The proxy x and the error r are uncorrelated, but in the EIV model x and the error ϵ are correlated. What differentiates the two cases is the assumed correlation between the observed variable and an error term that is added to the structural error term of the regression (and not whether x is on the left hand side or the right hand side of an equation).

In the event that there are more regressors in the regression (with or without measurement errors), the general result is that all the estimators are potentially inconsistent, even if just only one variable is badly measured.

The OLS estimator of the mismeasured regressor is still attenuated towards zero, but it is harder to sign the direction of the bias in the other (correctly measured) variables as this depends on the correlation between the regressors. What is important to remember is that measurement error in one variable can potentially contaminate the estimators of the coefficients of all other variables.

What happens $y = y^* + \delta$ where y^* is true value, but observed with error δ? Are there conditions on δ under which OLS is consistent?

2.1 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

$$\mathbb{E}(u) = 0 \tag{4}$$
$$\text{plim}\,\frac{1}{n}(y'u) = 0 \tag{5}$$
$$\text{plim}\,\frac{1}{n}(x'u) = 0 \tag{6}$$
$$\text{plim}\,\frac{1}{n}(\epsilon'u) = 0 \tag{7}$$

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 EIV 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

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

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

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

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

2.2 Measurement error for variance

(Mathematical derivation in this subsection is optional for MBA)
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 (even 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

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

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

$$\begin{aligned} \operatorname{plim} \widehat{s} = \operatorname{plim} \frac{\widehat{\sigma}_\epsilon^2}{\widehat{\sigma}_{\widetilde{x}}^2} &= \frac{\sigma_\epsilon^2 + (1-\lambda)^2\beta^2\sigma_x^2 + \lambda^2\beta^2\sigma_u^2}{\sigma_x^2 + \sigma_u^2} \\ &= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \cdot \left(\frac{\sigma_\epsilon^2}{\sigma_x^2}\right) + \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \cdot (1-\lambda)^2\beta^2 + \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \cdot \lambda^2\beta^2 \\ &= \lambda \cdot \frac{\sigma_\epsilon^2}{\sigma_x^2} + \lambda(1-\lambda)^2\beta^2 + \lambda^2(1-\lambda)\beta^2 \\ &= \lambda s + \lambda(1-\lambda)\beta^2 \end{aligned}$$

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.

2.3 Measurement error for t-statistics

(Mathematical derivation in this subsection is optional for MBA)
However, the t-statistic will be biased downwards. The t-ratio converges to

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

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

Given the situations in t-statistics, in the presence of measurement error, which almost always is the case in real life data, we not only lose trust on the estimated value's consistency, but the test statistics are also questioned. Can you still trust any regression analysis?

2.4 Simple Extension in dependent variables

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 $\widehat{\beta}_r = 1/(\widehat{1/\beta})$. Thus the results from the standard regression and from the reverse regression will bracket the true coefficient, i.e. $\operatorname{plim}\widehat{\beta} < \beta < \operatorname{plim}\widehat{\beta}_r$.

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)].

2.5 Simple Extension in data transformation

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 $\sigma_x^2 + \sigma_u^2 = \text{plim}\, var(\widetilde{x})$ 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

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

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

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

- (Obviously) Get better data.

2.6 Simple Extension in multivariate models

(Mathematical derivation in this subsection is optional for MBA)
What happens to the bias if we add more variables to the model? Consider the equation

$$y = \beta x + \gamma w + \epsilon \tag{9}$$

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

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

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

$$\text{plim}\, \widehat{\beta} = \frac{\beta(\sigma_w^2\sigma_x^2 - (\sigma_{xw})^2)}{\sigma_w^2(\sigma_x^2+\sigma_u^2) - (\sigma_{xw})^2} = \beta\lambda' \tag{10}$$

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 $\tilde{x}$ 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

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

where ρ_{xw} is the correlation coefficient between x and w. When $\gamma = 0$, comparisons between the bivariate regression of y on $\tilde{x}$ 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.

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.

3 Non-classical Measurement Error

(Mathematical derivation in this section is optional for MBA)
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

$$\begin{aligned}\operatorname{plim}\widehat{\beta} &= \frac{\beta(\sigma^2_x + \sigma_{xu})}{\sigma^2_x + \sigma^2_u + 2\sigma_{xu}} \\ &= \left(1 - \frac{(\sigma^2_u + \sigma_{xu})}{\sigma^2_x + \sigma^2_u + 2\sigma_{xu}}\right)\beta = (1 - b_{u\tilde{x}})\beta \end{aligned} \tag{11}$$

Notice that the numerator in $b_{u\tilde{x}}$ is the covariance between $\widetilde{x}$ and u. Thus, $b_{u\tilde{x}}$ 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 $b_{u\tilde{x}} = 1 - \lambda$.

The derivative of $1 - b_{u\tilde{x}}$ 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

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

Collecting the measurement errors in a matrix

$$W = [U|v]$$

yields

$$\operatorname{plim}\widehat{\beta} = \beta + \operatorname{plim}(\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'W\begin{bmatrix}-\beta \\ 1\end{bmatrix} \tag{12}$$

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.

What happens when we do instrumental variables in this case? For simplicity, focus on the one regressor case.

$$\begin{aligned}\widehat{\beta}_{IV} &= \frac{cov(y,z)}{cov(\widetilde{x},z)} = \frac{cov(\beta x + \epsilon, z)}{cov(x+u,z)} \\ \operatorname{plim}\widehat{\beta}_{IV} &= \frac{\beta\sigma_{xz}}{\sigma_{xz} + \sigma_{zu}}\end{aligned}$$

This demonstrates that we can still 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, i.e. the term $\sigma_{zu} = 0$ above. On the other hand, this condition is much more challenging in this case, since we have $\sigma_{xu} \neq 0$ and we need $\sigma_{zu} = 0$ and $\sigma_{zx} \neq 0$. Think, for example, about the case where $z = x + \eta$ is a second independent report of the same underlying x.

In this case, $\sigma_{zu} = \sigma_{xu} + \sigma_{\eta u}$. Hence, even if the errors were uncorrelated, i.e. $\sigma_{\eta u} = 0$, we still have $\sigma_{zu} = \sigma_{xu} \neq 0$ [Black, Berger, and Scott (1998)]. The upshot from this is that the instruments most likely to be helpful are the types of instruments we would be using anyway for other reasons (say to cure selection bias). For example, quarter of birth in [Angrist and Krueger (1991)] is much less likely to be correlated with the measurement error in schooling than is a sibling's report of ones schooling [Ashenfelter and Krueger (1994)].

4 Measurement Error in Dummy Variables

(Mathematical derivation in this section is optional for MBA)
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

$$y_i = \alpha + \beta d_i + \epsilon_i \tag{13}$$

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 (15) conditional on the observed value of d_i:

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

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

$$\operatorname{plim} \widehat{\beta} = \beta \cdot \left[P(d_i = 1|\widetilde{d} = 1) - P(d_i = 1|\widetilde{d} = 0)\right] \tag{14}$$

Equation (16) 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. $q_1 \equiv P(\widetilde{d}_i = 1|d_i = 1)$, and similarly $q_0 \equiv P(\widetilde{d}_i = 1|d_i = 0)$. 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

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

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

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

and

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

and substituting back into (15) yields

$$\begin{aligned}
\operatorname{plim} &= \beta \cdot \left[\frac{\pi q_1}{\pi q_1 + (1 - \pi) q_0} - \frac{\pi(1 - q_1)}{\pi(1 - q_1) + (1 - \pi)(1 - q_0)}\right] \\
&= \beta \cdot \left[\frac{\pi q_1}{\widehat{\pi}} - \frac{\pi(1 - q_1)}{1 - \widehat{\pi}}\right] \\
&= \beta \cdot \frac{(1 - \widehat{\pi})\pi q_1 - \widehat{\pi}\pi(1 - q_1)}{\widehat{\pi}(1 - \widehat{\pi})} \\
&= \beta \cdot \frac{\pi[(1 - \widehat{\pi}) q_1 - \widehat{\pi}(1 - q_1)]}{\widehat{\pi}(1 - \widehat{\pi})} \\
&= \beta \cdot \frac{\pi(q_1 - \widehat{\pi})}{\widehat{\pi}(1 - \widehat{\pi})}
\end{aligned} \tag{15}$$

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 is available, although β and π can still be identified if q_1 and q_0 are known [Aigner (1973)].

4.1 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 (16)? 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 $h_1 \equiv P(z_i = 1|d_i = 1)$ and $h_0 \equiv P(z_i = 1|d_i = 0)$. The IV estimator in this case is simply the Wald estimator so that

$$\operatorname{plim} \widehat{\beta}_{IV} = \frac{\mathbb{E}(y_i|z_i = 1) - \mathbb{E}(y_i|z_i = 0)}{\mathbb{E}(\widetilde{d}_i|z_i = 1) - \mathbb{E}(\widetilde{d}_i|z_i = 0)} \tag{16}$$

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:

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

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

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

With some elementary algebra this simplifies to

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

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.