Economics 241B
Estimation with Instruments

Measurement Error

Measurement error is defined as the error resulting from the measurement of a variable. At some level, every variable is measured with error. For example, if we measure a person's height as 6 feet 2 inches, it is unlikely that they are exactly 6 feet 2 inches tall. It is simply that they are closer to 6 feet 2 inches than to any other easily measured height. How does mismeasurement of variables affect estimation of a linear regression model? If a variable is mismeasured, then the resulting error is a component of the regression error.

Consider first the dependent variable. In our earliest discussion of the regression error, we said that the best possible outcome is for the regression error to arise from random mismeasurement of the dependent variable. Mismeasurement of the dependent variable does not lead to correlation between the regressors and the error and induces no bias in our OLS estimators. In fact, measurement error of the dependent variable affects only the variance of the regression error. Of course, we would like the error to have as small a variance as possible, so we would like measurement error in the dependent variable to be small.

Mismeasurement of the independent variable is another matter. In general, mismeasurement of the independent variable will lead to correlation between the regressor and the error and so lead to bias in the OLS coefficient estimators. As all regressors may have some measurement error, we treat the problem only when we feel the measurement error in the regressors is large enough relative to the approximation error made in the model under study. For example, survey data is often subject to such large errors that it is wise to treat the regressors as though measured with error. Formally, let the population model (in deviation-from-means form) be

$$Y_t = \beta X_t^* + U_t.$$

The regressor is measured with error, so that we observe X_t, where

$$X_t = X_t^* + V_t$$

and V_t is measurement error. The estimated model is

$$\begin{aligned} Y_t &= \beta X_t + [U_t + \beta (X_t^* - X_t)] \\ &= \beta X_t + [U_t - \beta V_t], \end{aligned}$$

where the error is the term in $[\cdot]$. Because V_t is a component of X_t, the regressor and error are correlated in the estimated model.

For the OLS estimator

$$B = \beta + \frac{\sum_{t=1}^{n} X_t \left[U_t - \beta V_t\right]}{\sum_{t=1}^{n} X_t^2}.$$

Because X_t is uncorrelated with U_t, we have

$$EB = \beta \left[1 - E\left(\frac{\sum_{t=1}^{n} X_t V_t}{\sum_{t=1}^{n} X_t^2}\right)\right].$$

It is difficult to proceed further with exact results, as one cannot condition on X_t while examining V_t. Rather, we have as an approximation

$$EB \approx \beta \left[1 - \frac{C\left(X_t, V_t\right)}{V\left(X_t\right)}\right],$$

where $V\left(X_t\right) = V\left(X_t^*\right) + V\left(V_t\right)$ and $C\left(X_t, V_t\right) = V\left(V_t\right)$. Thus

$$EB \approx \beta \left[\frac{V\left(X_t^*\right)}{V\left(X_t^*\right) + V\left(V_t\right)}\right].$$

Measurement error in the regressor biases the estimator toward zero and the degree of the bias depends upon the variance of the measurement error. The result is intuitive, as the measurement error in the regressor becomes more pronounced, the effect of the regressor is obscured and the estimated coefficient is biased toward zero.

Instruments

In the above, we discussed the problems arising from measurement error. In particular, we found that measurement error in a regressor leads to correlation between a regressor and the error. Of course, regressor and error correlation is generally due to endogenous regressors. Omitted factors that are correlated with the regressor (such as ability in a wage equation with schooling as a regressor) lead to this correlation as well. Correlation between a regressor and the error in turn leads to bias in the OLS coefficient estimator.

If the regressor is endogenous (or mismeasured), how can we correct the problem and reduce (or remove) the correlation between the regressor and the error? The answer is through the use of an additional variable termed an instrument. An instrument is a variable that theory does not suggest belongs in the regression, but

that is correlated with the problem regressor. Because theory does not suggest the instrument as a regressor, the instrument is uncorrelated with the regression error.

Formally, consider the regression model in deviation-from-means form

$$Y_t = \beta_1 X_{t1} + \beta_2 X_{t2} + U_t,$$

in which X_{t1} is exogenous but X_{t2} is endogenous. As we have seen, the OLS estimator is a method-of-moments estimator that uses the population moment

$$E\left(X_t' U_t\right) = 0 \text{ with } X_t = \left(X_{t1}, X_{t2}\right).$$

(The value 0 takes the appropriate dimension in all equations. Here it is a 2 by 1 vector.) That is, the OLS estimator is the value B such that

$$\sum_{t=1}^{n} X_t' \left(Y_t - X_t B\right) = 0.$$

If the regressor and the error are correlated, then $E\left(X_t U_t\right) \neq 0$. As a result, the OLS estimator does not equate the population moment and the sample analog (because the sample analog is set to zero) and so is biased. (In essence, when the regressor and error are correlated, the parameter is not identified. The presence of an instrument solves the identification problem.)

An improved estimator is obtained by again equating population moments and their sample analogs. Let Z_{t2} be an instrument for the endogenous regressor. The full instrument vector is $Z_t = (X_{t1}, Z_{t2})$ (note that X_{t1} serves as an instrument for itself). Although the regressor and the error are correlated, the regressor and the instrument Z_t are uncorrelated

$$E\left(Z_t' U_t\right) = 0.$$

The instrumental variable estimator is the method-of-moments estimator B_{IV} for which

$$\sum_{t=1}^{n} Z_t' \left(Y_t - X_t B_{IV}\right) = 0. \tag{0.1}$$

From (0.1) it follows that

$$B_{IV} = \left(\sum Z_t' X_t\right)^{-1} \sum Z_t' Y_t.$$

for which

$$EB_{IV} = \beta + E\left[\left(\sum Z_t' X_t\right)^{-1} \sum Z_t' U_t\right] \neq \beta.$$

Despite the fact that $E(Z_t' U_t) = 0$, the estimator is biased because we cannot condition on X_t when treating U_t as random. (If X_t and U_t are correlated, then we must consider either both to be fixed or both to be random.) In general, the IV estimator does not have a finite moment! Kinal (1980) shows that if the endogenous variables have homoskedastic Gaussian distributions with expectations linear in the exogenous variables, then the number of moments of the IV estimators is one less than the number of (over)identifying restrictions. Thus for the exactly identified case here, the IV estimator does not have a finite mean.

The bias of B_{IV} does diminish as the sample grows, so B_{IV} is a consistent estimator of β, while the OLSE is inconsistent. The two requirements for an instrument are clearly seen. The instrument must be uncorrelated with the error to allow for consistent estimation. The instrument must be correlated with the regressor to avoid an ill-defined estimate. If the instrument is uncorrelated with the regressor and the error, then the instrument is also uncorrelated with the dependent variable and $B_{IV} \approx \frac{0}{0}$.

One would like to know if a proposed random variable is a valid instrument. Unfortunately, one cannot test the condition that $E(Z_t U_t) = 0$ because U_t is not observed. One can, however, test the condition that $E(Z_t X_t) \neq 0$. To do so, perform the (first-stage) regression

$$X_{t2} = \alpha_1 X_{t1} + \alpha_2 Z_{t2} + V_t.$$

We must be able to reject the null hypothesis that $\alpha_2 = 0$. It is not enough that X_{t2} and Z_{t2} be correlated, they must be partially correlated after controlling for X_{t1}. If all the correlation was contained in X_{t1}, then Z_{t2} would not provide additional identifying information. The first-stage regression produces $X_{t2}^P = Z_t A$. To estimate β we estimate the (second-stage) regression

$$Y_t = \beta_1 X_{t1} + \beta_2 X_{t2}^P + U_t' \text{ with } U_t' = U_t + \beta_2 V_t^P.$$

With one endogenous regressor and one instrument, the two-stage least squares (2SLS) estimator is identical to the IV estimator. The 2SLS method is more commonly used, because it naturally provides a test of the condition $E(Z_t X_t) \neq 0$ in the first stage.

To construct the 2SLS estimator, we replaced X_{t2} with X_{t2}^P. Yet for the IV estimator, we did not replace X_{t2} with Z_{t2} in the regression model. It is natural

to ask, why can we replace X_{t2} with X_{t2}^P when we could not replace X_{t2} with Z_{t2}? To see this, consider first the second stage of 2SLS

$$Y_t = \beta_1 X_{t1} + \beta_2 X_{t2}^P + \left[U_t + \beta_2 V_t^P\right].$$

By construction, the residuals are uncorrelated with the regressors and so the second stage regressors are uncorrelated with V_t^P. The equation makes clear that it is vital to include the exogenous regressors in the first-stage. Failure to do so could lead to correlation between X_{t1} and V_t^P, rendering the 2SLS estimators inconsistent. If one replaced X_{t2} with Z_{t2}, then the second stage would be

$$Y_t = \beta_1 X_{t1} + \beta_2 Z_{t2} + \left[U_t + \beta_2 \left(X_{t2} - Z_{t2}\right)\right].$$

At the very least

$$n^{-1} \sum Z_{t2} \left(X_{t2} - Z_{t2}\right) \rightarrow C\left(X_2, Z_2\right) - V\left(Z_2\right) \neq 0.$$

(It could also be the case that X_{t1} is correlated with the compound error.) In essence, Z_t is not from a linear projection (while X_t^P is) and so $X_t - Z_t$ is correlated with Z_t.

An interesting example arises from attempts to estimate the returns to schooling. The goal is to determine how wages are affected by years of education (ed). As years of experience (ex) are also quite important, a sensible model could be

$$\ln(wage) = \beta_0 + \beta_1 ex + \beta_2 ex^2 + \beta_3 ed + U.$$

Yet education is likely correlated with the error. First, consider omitted ability. High ability individuals will likely earn higher wages and also seek out more education, inducing a positive correlation between the regressor and the error. In similar fashion, individuals receiving a high quality education will likely seek out more education. Finally, individuals with strong family backgrounds are likely both to do well in the job market and to seek out more education. For all of these reasons, we expect ed to be correlated with the error.

To find an instrument for ed, consider first the education of the mother. It is likely the case that mother's education affects the education of the child, in a positive way. Hence the proposed instrument is correlated with the endogenous regressor. Unfortunately, highly educated mothers may be more likely to have high ability children, or more likely to send their children to high quality education institutions. Hence, mothers education is not uncorrelated with the error. Next,

consider the last digit of the individual's social security number. Clearly the number is randomly assigned, and so uncorrelated with the error. Yet the number is also uncorrelated with *ed*. Angrist and Kruger (1991) proposed quarter of birth as an instrument. Because individuals must be 16 years old to leave school, those born in the first quarter of the year (who are some of the oldest students in a given class) are able to leave school earlier. Hence quarter of birth is correlated with *ed*. While A&K argue that the timing of births are random, and so uncorrelated with the error, other researchers question whether relative age effects (older students do better) induce correlation between the proposed instrument and the error. Moreover, if the return to education is not constant across individuals, then the IV estimator is capturing a local effect: that is, the impact of additional schooling on individuals who are on the margin of dropping out early.

Finally, in certain research problems one may also be interested in the reduced form, in which the dependent variable is regressed on all exogenous variables (the exogenous regressors and the instruments). With only one regressor and one instrument (the relationship holds if there are additional instruments and covariates) the reduced form is

$$Y_t = \gamma Z_t + V_t$$

and the reduced-form estimator is $G = \frac{\sum Z_t Y_t}{\sum Z_t^2}$. While G is clearly an inconsistent estimator of β, it does estimate the simple correlation between Y and Z. We could also estimate a first-stage equation

$$X_t = \alpha Z_t + V_t'$$

with the estimator of the first-stage coefficient as $A = \frac{\sum Z_t X_t}{\sum Z_t^2}$. One can quickly see that the 2SLS (IV) estimator satisfies

$$\frac{G}{A} = \frac{\sum_{t=1}^n Z_t Y_t}{\sum_{t=1}^n Z_t X_t} = B_{2S}.$$

Over-Identification

What if we have a set of possible instruments $\{Z_{t,2}, \ldots, Z_{t,J}\}$? As all instruments are uncorrelated with the regressor by definition, it seems we should select the instrument with the highest degree of correlation with the regressor. While the intuition is correct, there is no need to restrict our choice to single elements

of the instrument set. Rather we select the linear combination of instruments that is most highly correlated with the regressor. Let us consider the case of two instruments. The vector of instruments is now $Z_t = (X_{t1}, Z_{t2}, Z_{t3})$. Because a linear regression maximizes the correlation between the dependent variable and the regressors, the proposed instrument is obtained as the predicted value from the first-stage regression

$$X_{t2} = \alpha_1 X_{t1} + \alpha_2 Z_{t2} + \alpha_3 Z_{t3} + V_t.$$

As before, we use the first stage to test the condition that the instruments have parital correlation with the endogenous regressor. For the case at hand, we are testing $H_0 : \alpha_2 = \alpha_3 = 0$ against $H_1 :$ not H_0. If the errors are homoskedastic, the F statistic can be used.

Let the predicted value from the first stage be X_{t2}^P. The predicted values are $\hat{X}_t = \left(X_{t1}, X_{t2}^P\right)$ where X_{t1} serves as instrument for itself. The 2SLS estimator is

$$B_{2S} = \left(\sum \hat{X}_t' \hat{X}_t\right)^{-1} \sum \hat{X}_t' Y_t.$$

Recall, $\hat{X}_t = Z_t \left(\sum Z_t' Z_t\right)^{-1} \sum Z_t' X_t$, so the 2SLS estimator is equivalently written as

$$\begin{aligned} B_{2S} &= \left[\sum X_t' Z_t \left(\sum Z_t' Z_t\right)^{-1} \sum Z_t' X_t\right]^{-1} \left[\sum X_t' Z_t \left(\sum Z_t' Z_t\right)^{-1} \sum Z_t' Y_t\right] \\ &= \beta + \left[\sum X_t' Z_t \left(\sum Z_t' Z_t\right)^{-1} \sum Z_t' X_t\right]^{-1} \left[\sum X_t' Z_t \left(\sum Z_t' Z_t\right)^{-1} \sum Z_t' U_t\right]. \end{aligned}$$

To establish consistency, multiply each summation by n^{-1}, apply laws of large numbers and the Slutsky theorem. To establish asymptotic normality, apply a central limit theorem for $n^{-\frac{1}{2}} \sum Z_t' U_t$. If we assume the errors are conditionally homoskedastic

$$E\left(U_t^2 Z' Z\right) = \sigma^2 E\left(Z' Z\right)$$

(note, this assumption is implied by $E\left(U_t^2 | Z\right) = \sigma^2$), then

$$n^{\frac{1}{2}} \left(B_{2S} - \beta\right) \Rightarrow N\left(0, V\right)$$

where

$$V = \sigma^2 \left\{E\left(X' Z\right) \left[E\left(Z' Z\right)\right]^{-1} E\left(Z' X\right)\right\}^{-1}.$$

To estimate σ^2, construct the residuals from two-stage least squares (note, to construct the residuals we use X_t, not X_t^P, as the predicted values of the regressors are simply statistical controls)

$$U_t^P = Y_t - X_t B_{2S}.$$

The estimator of the error variance is $S^2 = n^{-1} \sum \left(U_t^P\right)^2$. The estimator of V is then obtained either by using sample averages in the above formula or with the equivalent

$$S^2 \left(\sum \hat{X}_t' \hat{X}_t \right)^{-1}.$$

With two instruments and only one regressor, we have more moment conditions than we need

$$E\left[Z_{1,t} U_t\right] = 0 \text{ and } E\left[Z_{2,t} U_t\right] = 0.$$

We have over-identifying information. We can test the over identification in the following way. These residuals are the best estimates of the error. To see if the instruments are uncorrelated with the error, we regress the residuals on the instruments

$$U_t^P = \alpha_0 + \alpha_1 Z_{t,1} + \alpha_2 Z_{t,2} + V_t.$$

If the instruments are valid, then they should be uncorrelated with U_t and the estimators of the coefficients should be zero. The (J) test statistic is

$$J = mF \qquad J \sim \chi^2_{m-k}$$

where F is the statistic that tests the joint null hypothesis $H_0 : \alpha_1 = \alpha_2 = 0$, m is the number of instruments (in this case 2) and k is the number of endogenous regressors (in this case 1).