1 Endogeneity

Consider the following model in the population,

$$y = \beta_1 + x_2\beta_2 + ... + x_k\beta_k + u = x\beta + u \tag{9}$$

Above equation requires classical A1-A4 assumptions to be identified, or at least to find the consistent estimator, β. An explanatory variable x_j is said to be **endogenous** if,

$$E(x_j u) \neq 0 \tag{10}$$

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**.

If a regressor is endogenous then b, 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.

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:

$$\begin{aligned} \text{plim} b &= \beta + (\text{plim}\frac{X'X}{n})^{-1}\text{plim}\frac{1}{n}X'U \\ &= \beta + [E(x'x)]^{-1} E(x'u) \\ &= \beta + [E(x'x)]^{-1} \left[\begin{pmatrix} 0 \\ \vdots \\ E(x_k u) \end{pmatrix}\right] \end{aligned}$$

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 $E(u|x_j) = 0$ 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 $E(x_j \cdot u) = 0$ 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. Measurement error
3. Simultaneity

1.1 Omitted Variable Bias (OVB)

Suppose theory tells us that the relevant object of study is the conditional expectation of y given two variables x and q. For some reason, we omit q from the regressions. How is the OLS estimator of the coefficient of x affected?

Why would we omit q from the regression if theory tells us to include it? The typical reason is either that we lack data to measure it or because the variable is intrinsically unmeasurable. For example, when estimating the effect of schooling on wages using household data we would like to control for the effect of the firm in which the individual works because firms may have different wage policies.

Unfortunately, household data sets do not carry information on the employing firm. We would also like to control for the individual's "natural ability" since it certainly affect wages but ability is intrinsically unobserved and, possibly, unmeasurable. Let the model be

$$E(y|x_1, ..., x_k, q) = \beta_1 + x_2\beta_2 + ... + x_k\beta_k + \gamma q \tag{11}$$

where q is the variable that will be omitted from the regression. We are interested in the $\beta_j^{'}$s, which are the partial effects of the observed explanatory variables holding the other explanatory variables constant, including the unobservable q.
Model above in error form is,

$$y = \beta_1 + x_2\beta_2 + ... + x_k\beta_k + \gamma q + v, \quad E(v|x_1, x_2, ..., x_k, q) = 0 \tag{12}$$

If we regress y on the observable variables only we are, in effect, putting the unobservable q into the error term,

$$y = \beta_1 + x_2\beta_2 + ... + x_k\beta_k + u \quad u \equiv \gamma q + v \tag{13}$$

Now, v has zero mean and is uncorrelated with all the $x_j^{'}$s (and q). v is sometimes called the structural error. We can also assume $E(q) = 0$ because an intercept is always included in the regression. Thus, $E(u) = 0$.

But for u to be uncorrelated with each regressor $x_1, ..., x_k$, it must be that q is uncorrelated with each of them. If q is correlated with any of the regressors, then so will u be and we have an endogeneity problem: OLS would not be estimating β consistently.

To understand this omitted variable bias, denote the best linear projection of q onto x by

$$L(q|x) = x\delta \quad \delta = \left[E(x'x)\right]^{-1} E(x'q)$$

We can therefore write

$$q = \delta_1 + \delta_2 x_2 + ... + \delta_k x_k + r \tag{14}$$

where, by definition of a linear projection, $E(r) = 0$, and $Cov(x_j, r) = 0$ for each $j = 1, ..., k$. Then,

$$y = \underbrace{(\beta_1 + \gamma\delta_1)}_{\pi_1} + \underbrace{(\beta_2 + \gamma\delta_2)}_{\pi_2} + ... + \underbrace{(\beta_k + \gamma\delta_k)}_{\pi_k} + \gamma r + v \tag{15}$$

The error term in this equation, $\gamma r + v$, is uncorrelated with all the regressors. That is, the x's are exogenous in this equation, so OLS consistently estimates the coefficients of this regression. Thus, OLS estimator of the coefficients in the model excluding q - which we denote by b^* - would consistently estimate the parameter π and not β, namely

$$\text{plim} b_j^* = \beta_j + \gamma\delta_j = \pi_j \tag{16}$$

which says that the OLS estimator estimates the direct effect of x_j on y plus an indirect effect amounting to the effect of x_j on the unobserved q times the effect of q on y, i.e., $\gamma\delta_j$.

If we suspect that a relevant variable q is omitted from the specification, then it is not surprising that some regressors are endogenous.

The intuition behind is that these regressors may be the result of choices made by individuals or firms and these choices may also be affected by q (which is known to the individual or firm but is not observed by the researcher). In this case, a correlation between the included and excluded variables results (e.g., schooling and ability, inputs and managerial qualities, etc.).

We can use above equation to get an idea of the direction of the bias $\gamma\delta_j$, if not its magnitude. Usually we do have a good prior about the sign of γ. However, in attempting to sign the $\delta_j^{'}$s we should remember that it is not enough to have an idea about the sign of the simple correlation coefficient between q and x_j; δ_j refers to the partial correlation between q and x_j.

The only two cases in which omitting a variable has no effect on the OLS estimates of the included variables are:

1. When the omitted variable has no effect on y, i.e., $\gamma = 0$ so that q is not really an omitted variable. We say that q is not relevant.

2. When the omitted variable is orthogonal to the included variables. That is when $E(x'q) = 0$ because this implies $\delta = 0$.

For completeness, we now show the relationship between the OLS estimator in the regression when q is included (sometimes referred to as the "long" regression) and the OLS estimator from the regression when q is excluded (the "short" regression).

To do this we write $Y = Xb + Q\hat{\gamma} + \hat{V}$, where Q is the $n \times 1$ vector of observations on q, $(b', \hat{\gamma})$ is the $(k+1) \times 1$ OLS estimator and $\hat{V}$ is the $n \times 1$ vector of OLS residuals in the long regression. Now, regressing Y on X only (the sort regression) gives

$$\begin{aligned} b^* &= (X'X)^{-1}X'Y = (X'X)^{-1}X'(Xb + Q\hat{\gamma}\hat{V}) \\ &= b + (X'X)^{-1}X'Q\gamma + (X'X)^{-1}X'\hat{V} \\ &= b + \hat{\delta}\hat{\gamma} \end{aligned}$$

because $X'\hat{V} = 0$ by construction, and $\hat{\delta} = (X'X)^{-1}X'Q$ is the $k \times 1$ vector of OLS estimators of the coefficients in the linear projection of q on $x_1, ..., x_k$(see equation (14)). This relationship is the sample counterpart of (16).

The estimates of the "short" regression estimate the direct (partial) effect of x on y plus the indirect effect on y that operates through the correlation between x and q. Clearly b^* is inconsistent,

$$\text{plim} b^* = \beta + \delta\gamma \equiv \pi \tag{17}$$

which is equal to (16).

1.1.1 Proxy Variables

The effect of omitting a relevant variable can be reduced if we have a proxy variable for the unobserved variable q. A proxy variable z should satisfy two requirements.

$$1. \quad E(y|x, q, z) = E(y|x, q) \tag{18}$$

$$2. \quad L(q|x, z) = L(q|z) = \theta_0 + \theta_1 z \tag{19}$$

Condition 1 means that z does not play a role in explaining y once x and q have been controlled for. This is not a strong assumption since q, and not z, is what affects y (otherwise z would be part of x). In the wage-education example, let q be ability and z be an IQ test score. The model says that it is ability that affects wages; the IQ score should not matter for wages give ability.

Condition 2 is more important and says that q is not correlated with x once we partial out (account for) z. That is the BLP of q on (z, x) in error form is,

$$q = \theta_0 + \theta_1 z + r \tag{20}$$

with $E(r) = 0$ and $Cov(z, r) = 0$ by definition. Condition 2 assumes, in fact, that z accounts for all the possible correlation between q and the x's,

$$Cov(x_j, r) = 0 \quad j = 1, ..., k$$

To see how using the proxy z instead of q affects the estimated coefficients we replace q above to get an estimable equation.

$$y = (\beta_1 + \gamma\theta_0) + x_2\beta_2 + ..., x_k\beta_k + \gamma\theta_1 z + (v + \gamma r)$$

The composite error $u \equiv v + \gamma r$ is uncorrelated with the regressors under the assumptions made. Condition 1 ensures that z is uncorrelated with v while z is uncorrelated with r by construction. The $x_j^{'}$s are uncorrelated with r. Thus, we know that the OLS regression,

$$y \text{ on } 1, x_2, ..., x_k, z \tag{21}$$

gives consistent estimators of

$$(\beta_1 + \gamma\theta_0), \beta_2, ..., \beta_k, \gamma\theta_1$$

We can estimate β consistently if we use proxy variables (except for β_1 and γ). If one of the $x_j^{'}$s, say x_k, does not satisfy condition 2 then,

$$q = \theta_0 + \lambda_k x_k + \theta_1 z + r \tag{22}$$

Therefore the model can estimate

$$(\beta_1 + \gamma\theta_0), \beta_2, ..., (\beta_k + \gamma\lambda_k), \gamma\theta_1$$

consistently.

1.1.2 Optional: Omitted Variable Effect on Variances

Sometimes it is useful to know which estimator - the one based on the longer regression or the one based on the shorter regression - has a smaller variance. This is not a very interesting question since we are comparing a consistent estimator with an inconsistent one, but something useful will come out from this exercise.

Recall that the variances of b and b^* can be written as

$$\begin{aligned} V(b|X,Q) &= \sigma^2(X'M_qX)^{-1} \quad M_q = I - Q(Q'Q)^{-1}Q' \\ V(b^*|X,Q) &= \sigma^2(X'X)^{-1} \end{aligned}$$

where Q is the $n \times 1$ vector of observations on q. We condition on X and Q to ensure that $\sigma^2 = V(v|x, q)$ is the same in both expressions. This implies

$$V(b|X,Q) - V(b^*|X,Q) = \sigma^2 \left[(X'M_qX)^{-1}(X'X)^{-1}\right]$$

In order to evaluate the sign - in a matrix sense - of this expression we appeal to the following result. Let A, B be two positive definite matrices. If $B - A$ is positve definite then so is $A^{-1} - B^{-1}$. In our case, both matrices in question are positive definite and

$$\begin{aligned} \underbrace{V(b^*|X,Q)^{-1}}_{B} - \underbrace{V(b|X,Q)^{-1}}_{A} &= \sigma^{-2}\left[(X'X) - (X'M_qX)\right] \\ &= \sigma^{-2}X'(I - M_q)X \\ &= \sigma^{-2}X'Q(Q'Q)^{-1}Q'X \end{aligned}$$

Now,

$$X'Q(Q'Q)^{-1}Q'X = \frac{X'QQ'X}{Q'Q}$$

where $Q'Q = \Sigma_{i=1}^{n} q_i^2$ is a positive scalar and the numerator is clearly a positivie definite matrix. Then $V(b^*|X,Q)^{-1} - V(b|X,Q)^{-1}$ is positive definite and therefore so is $V(b) - V(b^*)$, or $V(b) - V(b^*)$ in a matrix sense.

The conclusion is that **the OLS estimator** of β in the **long regression** has a **'larger' covariance matrix** than in the short regression.

Consider the particular case, $y = \alpha + \beta x + \gamma q + v$. The regressors are a constant, the scalar x and the omitted variable q. Let $S_{xx} = \Sigma_{i=1}^{n}(x_i - \bar{x})^2$. It is well known that in the short regression - the regression excluding q - the variance of b^* is

$$V(b^*|x,q) = \frac{\sigma^2}{S_{xx}}$$

while the variance of b in the long regression (including x and q) is

$$v(b|x,q) = \frac{\sigma^2}{\Sigma_{i=1}^{n}\hat{\epsilon}_t^2} = \frac{\sigma^2}{S_{xx}(1 - R_{xq}^2)}$$

where $\hat{\epsilon}$ is the OLS residual in the regression of x on $(1,q)$ and R_{xq}^2 is the R^2 from that regression, $R_{xq}^2 = 1 - \frac{\Sigma_{i=1}^{n}\hat{\epsilon}_t^2}{S_{xx}}$. Because $0 \leq R_{xq}^2 \leq 1$, the variance in the long regression is larger than the variance in the short regression. The higher the correlation between x and q, the larger the variance of b.

The upshot of this discussion is that omitting relevant variables produces biased estimators, but with a variance that is no larger than the one obtained from the long regression.

1.1.3 Optional: Inclusion of "Irrelevant" Regressors

Suppose that the true model is

$$E(y|x-1,...,x_k,q) = \beta_1 + x_2\beta_2 + ... + x_k\beta_k + \gamma q$$

but now $\gamma = 0$. That is, we assume in fact that $E(y|x,q) = x\beta$, the variable q is irrelevant or ignorable in this model. Nevertheless we regress y on x and on q.

There is no problem with incldluing the irrelevant variable q. Our estimate of γ, the coefficient of q should be close to zero. Indeed, we know that $E(\hat{\gamma}) = 0$ because we know that in this particular case, the value of γ happens to be zero. σ^2 is also an unbiased estimator of σ^2.

Hence, there is nothing wrong with the inclusion of irrelevant variables in terms of bias. We simply are not using some correct information about the value of γ. In contrast, when omitting relevant variables, we impose incorrect information into our estimation, i.e., we assume that $\gamma = 0$ when this is in fact not true.

Nevertheless, there should be something that stops us from adding variables to the regression model. Otherwise, we should just keep adding regressors. As seen, the cost to adding irrelevant regressors is in terms of the precision the estimator: the variance of the estimator when q is included is larger than when q is omitted.

1.2 Measurement Errors

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

$$y = \alpha + \beta x^* + v \quad E(v|x^*) = 0 \tag{23}$$

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

$$x = x^* + \epsilon \quad E(\epsilon|x^*) = 0 \tag{24}$$

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

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

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

$$\begin{aligned} y &= \alpha + \beta(x - \epsilon) + v \\ &= \alpha + \beta x - \beta\epsilon + v \\ &= \alpha + \beta x + u, \quad u = v - \beta\epsilon \end{aligned} \tag{25}$$

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

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

Note that inthis 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 in consistency. In this case, the bias or inconsistency depends on the ratio of error variance to true variance.

Sometimes there is confusion between the EIV formuatlion 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.

Exercise: Consider the case of the dependent variable being measured with error, $y = y^* + \delta$ where y^* is true value, but we observe it with error η.Are there conditions on η under which OLS is consistent? What are they?

1.3 Simultaneity

Simultaneity arises when at least one of the explanatory variables is partially determined by y. If, say, x_k is determined partly as a function of y, then we will show that this usually induces a correlation between x_k and u, which makes x_k endogenous in the regression for y.

For example, if y is a city's crime rate and x_k is the size of its police force, size of the police force is partly determined by the crime rate. If the amount of labor determines production but production also determines the demand for labor, then both production and labor are simultaneously determined.
Suppose the equation of interest is

$$y = x\beta + u \tag{26}$$

but the last regressor is partly determined by y (and other regressors z which can also include the other $k-1$ number of x's),

$$x_k = \alpha y + z\delta + \epsilon \tag{27}$$

Is x_k uncorrelated with u? Guess increasing u. This increase in u increases y directly through the first structural equation. But an increase in y also affects x_k through the other structural equation, when $\alpha \neq 0$. Thus, u and x_k will be correlated when y helps to determine x_k. This is the simultaneity problem. Below is an example for Supply and Demand.

Example 1.

$$\begin{gathered}\text{D: } q = y\beta + \alpha p + u^D \\ \text{S: } p = w\gamma + \delta q + u^S\end{gathered}$$

Assume that $E(u^D) = E(u^S) = 0$ and that $E(u^D y) = E(u^S y) = 0$ and $E(u^D w) = E(u^S w) = 0$. That is y (income) and w (wage) are exogenous in both the demand and supply equations. We will show that the price regressor (p) in the demand eqution is endogenous.

$$\begin{aligned} E(pu^d) &= E\left[(w\gamma + \delta q + u^S)u^D\right] \\ =&\delta E(qu^D) + E(u^S u^D) \\ =&\delta E\left[(y\beta\alpha p + u^D)u^D\right] + E(u^S u^D) \\ =&\delta E(pu^D) + \delta V(u^D) + Cov(u^S, u^D) \\ =&E(pu^D) = \frac{\delta V(u^D) + Cov(u^S, u^D)}{1 - \delta\alpha} \neq 0 \end{aligned}$$

even if the demand and supply errors (shocks) are uncorrelated. It is as easy to show that the quantity regressor in the supply equation is also endogenous.