A Brief Guide to a Two-Stage Least Squares Research Design

As we've discussed, there are many possible threats to valid inference when we rely upon purely statistical methods to infer causal relationships. In these notes, we'll consider how a threat to valid inference can arise from an omitted variable. Then, we'll work out how a two-stage least squares research design can help us recover valid inference.

I Omitted Variables Bias

Suppose that the following linear regression model accurately describes the true determinants of a dependent variable y:

$$y_i = x_{1,i}\beta_1 + x_{2,i}\beta_2 + \varepsilon_i,$$

where i indexes observations, x_1 is the first explanatory variable, x_2 is the second explanatory variable, ε is a mean-zero error/noise term, and β_1 and β_2 represent the true effects of x_1 and x_2 respectively upon y. We are assuming that there is no intercept term to make the notation a bit simpler. Since regression 1 represents the true model, we also have that:

$$E\left(x_1\varepsilon\right) = E\left(x_2\varepsilon\right) = 0,$$

where $E\left(\cdot\right)$ denotes the population expectation. In other words, there is no correlation between the true error/noise term ε and the explanatory variables.

If we knew the structure of the model with certainty, we could collect data on all the determinants and estimate the coefficients (the βs) with a high degree of confidence that the estimated effects reflect the true effects. Of course, we rarely know the true structure of the model with certainty. For example, we may not realize that x_2 is an important determinant of y. Alternatively, we may understand the true structure of the model, but we do not have access to data on one of the variables (e.g., x_2).

To see what the effects of such unawareness or missing data, suppose that instead of the true model for y given by regression 1, we postulate and estimate the following model:

$$y_i = x_{1,i}\tilde{\beta}_1 + \tilde{\varepsilon}_i,$$

where the tildes indicate that the parameter and error terms refer to the regression that we actually run. Unlike β_1, $\tilde{\beta}_1$ might not be informative about the true effect of x_1 upon y. Why? The regression model from which it arises is a reflection of our own ignorance or data shortcomings rather than the true regression model. However, it may be that it can still tell us something about the true effect, as we'll now consider.

Furthermore, suppose that x_1 and x_2 are correlated, such that if we had information on x_2, we could write:

$$x_{1,i} = x_{2,i}\gamma + \eta_i,$$

where $\gamma \neq 0$ would indicate that x_1 and x_2 are correlated (positively or negatively) and η represents whatever drivers there are for x_1 which are *un*correlated with x_2. So, we are assuming that $E(x_2\eta) = 0$. Note that there does *not* need to be *any* true causal relationship between x_1 and x_2; all that is required is that they are correlated.

To further simplify the discussion, we will also assume that the sample *is* the population of interest. Then, sample or empirical expectations represent the true population-level expectations. As we know, the ordinary least squares (OLS) estimator for γ in this case would be:

$$\gamma = \frac{E(x_1 x_2)}{E(x_2^2)}.$$

Because we are assuming that we can use the population expectations in the OLS estimator formula, we can ignore concerns about randomness in the sample.

We are now ready to consider what the relationship between regression 1 (the true model) and regression 2 (what we actually estimate) is. The OLS estimator of $\tilde{\beta}_1$ is given by:

$$\tilde{\beta}_1 = \frac{E(x_1 y)}{E(x_1^2)}.$$

If we substitute the true regression model for y into the estimator for $\tilde{\beta}_1$, we can see what the consequences of neglecting the effect of x_2 upon y will be for our inference. We have that:

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{E[x_1(x_1\beta_1 + x_2\beta_2 + \varepsilon)]}{E[x_1^2]} \\
&= \frac{E[x_1^2\beta_1 + x_1x_2\beta_2 + x_1\varepsilon]}{E[x_1^2]} \\
&= \frac{E[x_1^2]}{E[x_1^2]}\beta_1 + \frac{E[x_1x_2]}{E[x_1^2]}\beta_2 + \frac{E[x_1\varepsilon]}{E[x_1^2]} \\
&= \beta_1 + \gamma\beta_2 + \frac{E[x_1\varepsilon]}{E[x_1^2]} \\
&= \beta_1 + \gamma\beta_2,
\end{aligned}$$

where we have used the constancy of the population parameters β_1 and β_2, the definition of γ, and the fact that $E(x_1\varepsilon) = 0$ in the true model so that we can ignore the last term. From this, we can see that $\tilde{\beta}_1$ differs from the true β_1 by a term that is equal to $\gamma\beta_2$, known as the *bias* of $\tilde{\beta}_1$ for β_1. If either γ or β_2 equal zero (no correlation of x_1 and x_2 or no effect of x_2 on y), then we are OK in using $\tilde{\beta}_1$; it will accurately reflect the true effect β_1.

In words, regression 2 potentially suffers from an omitted variable problem, which is a threat to valid inference. There is an important variable x_2 that is omitted and γ is not equal to zero. Consequently, $\tilde{\beta}_1$ may not be very informative about β_1. As the magnitudes of γ and β_2 grow, the bias becomes more prevalent, contaminating our inference.

Notice how the sign of the bias depends upon the signs of γ and β_2. If they have the same sign, then the bias is positive ($\tilde{\beta}_1$ *over*estimates the effect of x_1 upon y). If they have opposite signs, then the bias is negative ($\tilde{\beta}_1$ *under*estimates the effect of x_1 upon y). Depending upon the relative magnitude of the bias to the true effect and their signs, it is even possible for $\tilde{\beta}_1$ and β_1 to have *different* signs.

II Research Design and Two-Stage Least Squares

Is there a solution to the threat to valid inference detailed above? If we have access to a variable known as an *instrument*, then we can use a *two-stage least squares* (TSLS) research design to recover a good estimate of β_1, despite either being unaware of x_2's importance for y or not having data on x_2. An instrument for x_1 is a variable which is correlated with x_1 but is uncorrelated with both x_2 and ε. Let z denote the instrument. It has the following properties:

$$E\left(zx_2\right) = E\left(z\varepsilon\right) = 0 \Rightarrow E\left(z\tilde{\varepsilon}\right) = 0.$$

Moreover, since z is correlated with x_1, then we can decompose x_1 into components related to x_2 and z:

$$x_{1,i} = x_{2,i}\gamma + z_i\delta + \upsilon_i,$$

where $\delta \neq 0$ would indicate that x_1 and z are correlated (positively or negatively) and υ represents whatever drivers there are for x_1 which are uncorrelated with both x_2 and z. Again, note that there does not need to be any true causal relationship between x_1 and z; all that is required is that they are correlated. If we were to regress x_1 on z, we would get:

$$\begin{aligned}
\delta &= \frac{E\left[zx_1\right]}{E\left[z^2\right]} \\
&= \frac{E\left[z\left(x_2\gamma + z\delta + \upsilon\right)\right]}{E\left[z^2\right]} \\
&= \frac{E\left[zx_2\right]}{E\left[z^2\right]}\gamma + \frac{E\left[z^2\right]}{E\left[z^2\right]}\delta + \frac{E\left[z\upsilon\right]}{E\left[z^2\right]} \\
&= \delta,
\end{aligned}$$

since γ and δ are constants and $E\left(zx_2\right) = E\left(z\upsilon\right) = 0$. Our estimate of δ is accurate, because of the properties of the instrument. The fitted values for x_1 from such a regression are $\hat{x}_{1,i} = \delta z_i$. This is the *first stage* result of TSLS.

The *second stage* involves regressing y upon the fitted values from the first stage $\hat{x}_1$. Denote the second stage coefficient by $\widehat{\widehat{\beta}}_1$. It will be:

$$\begin{aligned}
\widehat{\widehat{\beta}}_1 &= \frac{E\left[\hat{x}_1 y\right]}{E\left[\hat{x}_1^2\right]} \\
&= \frac{E\left[\hat{x}_1\left(x_1\beta_1 + x_2\beta_2 + \varepsilon\right)\right]}{E\left[\hat{x}_1^2\right]} \\
&= \frac{E\left[z\delta\left(x_1\beta_1 + x_2\beta_2 + \varepsilon\right)\right]}{E\left[\left(z\delta\right)^2\right]} \\
&= \frac{E\left[zx_1\right]\delta\beta_1}{E\left[z\right]\delta^2} + \frac{E\left[zx_2\right]\delta\beta_2}{E\left[z\right]\delta^2} + \frac{E\left[z\varepsilon\right]\delta}{E\left[z\right]\delta^2} \\
&= \frac{E\left[zx_1\right]\beta_1}{E\left[z\right]\delta} = \frac{\delta\beta_1}{\delta} = \beta_1,
\end{aligned}$$

where we have used the properties of z as an instrument and the definition of δ. The second stage coefficient $\widehat{\widehat{\beta}}_1$ is informative about the true effect of x_1 upon y! The instrument z allows us to disentangle the variability in y that arises from x_2 from its variability that arises from x_1. Consequently, we are able to get a estimate of the true effect of x_1 upon y that is uncontaminated by x_2.

In practice, the toughest part of undertaking a TSLS research design is finding an instrument. It must be correlated with the explanatory variable of interest and uncorrelated with the omitted variable (or more generally, with the error term $\tilde{\varepsilon}$). Empirical economists spend a lot of energy trying to think about instruments and about arguments for and against a particular choice of instrument. The best research designs have convincing arguments about the validity of their choice of instrument.