1 Causality and the omitted variable problem

The most fundamental challenge in empirical economics is to estimate a causal relationship between an explanatory variable, $x$, and an outcome variable, $y$. The primary tool used to do so in econometrics is the linear regression model

$$y = x\beta + \varepsilon$$

(1)

where $\beta$ represents (hopefully) the causal effect of $x$ on $y$, while $\varepsilon$ represents other unobserved (by the econometrician) determinants of $y$. From an economic point of view, it is important to realize that $\varepsilon$ is not simply an innocuous “error” term. The world is very complex and it is unlikely that we are able to correctly measure all the possible determinants of $y$. This means that $\varepsilon$ should be viewed as a set of factors that affect $y$ but that we are, unfortunately, unable to control for. Unless some special conditions are met, the failure to control for these “omitted” factors will generally result in inconsistent estimates of the causal effect $\beta$. It is well known that OLS estimates of $\beta$ are inconsistent in the presence of omitted variables. To fix ideas, say that $\varepsilon$ is a linear function of an unobserved variable $u$, $\varepsilon = \delta u$. It is easily shown (omitted variable bias formula) that:

$$\operatorname{plim} \beta = \beta + \delta \gamma$$

(2)

where $\gamma$ is the coefficient of a regression of $u$ on $x$. ($\gamma = \operatorname{cov}(u,x)/\operatorname{var}(x)$, see more on this covariance formula approach to regression in the appendix at the end of these notes).
This means that the omitted variable \( u \) will not result in a bias if either \( \delta = 0 \) or \( \gamma = 0 \). It is not very plausible for \( \delta \) to be zero, as it would imply that \( \varepsilon = \delta u \) is zero too, i.e. that the R-square of a regression of \( y \) on \( x \) is one. So the key condition is that \( \gamma = 0 \), i.e. that there is no relationship between \( u \) (or \( \varepsilon \)) and \( x \). More specifically, it must be that \( \text{cov}(x, u) = \text{cov}(x, \varepsilon) = 0 \), since \( \gamma = \text{cov}(u, x)/\text{var}(x) \).

The standard assumption used to show that \( \text{plim} \hat{\beta} = \beta \) in the OLS model is \( E(\varepsilon|x) = 0 \), which implies (but not the other way around) that \( \text{cov}(x, \varepsilon) = 0 \) i.e. that there is no omitted variable problem. Thus, it should be clear that the standard zero conditional mean assumption \( E(\varepsilon|x) = 0 \) is a very strong assumption from an economic point of view. It means there is no systematic relationship between \( x \) and the unobserved determinants of \( y \). In labour economics, there are several classic examples where we expect this to be violated:

- Education and earnings: \( y \) is earnings, \( x \) is schooling, \( u \) is unobserved ability. In both signalling and human capital investment models, there are strong reasons to believe that \( u \) and \( x \) are correlated and that \( E(\varepsilon|x) \neq 0 \)

- Labour supply: \( y \) is hours of work, \( x \) is the wage rate, \( u \) is a preference (for leisure) term. But people who work harder for reasons of taste may well get higher wages because of promotions, more effort, etc. invalidating once again the zero conditional mean assumption.

2 Solutions?

So what can we do about this problem? A number of possible solutions exist depending on the nature of the problem.

2.1 More and better data

Of course, if all the unobserved determinants of \( y \) could be measured, we would simply add all the relevant variables in the regression model. In a standard model with cross-sectional data, this means expanding the list of regressors in the hope that we will truly be “holding everything else equal” once we also control for these other factors. This has become more popular lately with the advent of “big data” and machine learning methods imported from computer science, which allow to estimate models with an extremely large number of control variables.
A particularly good way of controlling for unobserved factors is to get panel data where we can get much more information by repeatedly observing the same unit (worker, firm, region, etc.) over time. Let’s add a little bit of notation to better explain what can done here. Our model is now:

\[ y_{it} = x_{it} \beta + \varepsilon_{it} \] (3)

where the subscript \( i \) refers to the observation unit (say a worker) and \( t \) to the time period. Thanks to the repeated observations, we can introduce a richer structure for the error term, for example:

\[ \varepsilon_{it} = \theta_i + \nu_{it} \] (4)

where \( \theta_i \) is usually called a fixed or random effect that does not change over time. Even if \( \theta_i \) represents unobserved determinants of \( y \) that are correlated with \( x \), it is possible to consistently estimate \( \beta \) by first-differencing the data (or using other related “fixed effect” methods). So more data enables us to replace the assumption that \( E(\varepsilon_{it}|x) = E(\theta_i + \nu_{it}|x) = 0 \) by the weaker assumption \( E(\nu_{it}|x) = 0 \).

### 2.2 The experimental approach

In natural and medical sciences, a very standard way of controlling for other factors when trying to find the causal effect of \( x \) on \( y \) is to run experiments. This approach is now very popular in economics, though running your own experiment is likely beyond the scope of a 594 paper. When human beings are involved in an experiment, there is of course no way of controlling for all relevant factor since different people will behave differently when confronted with the exact same environment. This difficulty is avoided using randomization. Provided that a control and a treatment group are chosen at random and the two groups are administered different “doses” of \( x \), we are able to estimate the causal effect of \( x \) on \( y \) by simply comparing average outcomes for the two groups. More specifically, for the treatment group, \( T \), we have

\[ y_i = x^T \beta + \varepsilon_i \] (5)

while for the control group, \( C \), we have:
\[ y_i = x^C \beta + \varepsilon_i \]  

The expected value of \( y \) in the treatment and control groups is

\[ E(y|T) = x^T \beta + E(\varepsilon_i|T), \text{ and } E(y|C) = x^C \beta + E(\varepsilon_i|C) \]  

Because the two groups are chosen at random, it follows that \( E(\varepsilon_i|T) = E(\varepsilon_i|C) \) and that

\[ E(y|T) - E(y|C) = (x^T - x^C)\beta \]

or

\[ \beta = \frac{E(y|T) - E(y|C)}{x^T - x^C} \]

The causal parameter \( \beta \) can thus be estimated as a simple difference between average outcomes in the two groups, divided by the assigned difference in \( x \) (under the control of the experimenter) for the two groups. In many applications, \( x \) is simply viewed as a treatment dummy (i.e. do you receive a drug or a training program) so that \( x^T = 1 \) and \( x^C = 0 \).

### 2.3 Natural experiments and instrumental variables

An appealing alternative to controlled experiments is to find cases where “nature” provides us with experiments that are just as good as true experiments for estimating consistently causal effects. A useful way of thinking about natural experiments is that they provide us with very credible instrumental variables \( (IV) \) that can then be used to consistently estimate \( \beta \) using IV/GMM/two-stage least-square methods. Remember that an instrumental variable \( z \) needs to meet two conditions. First it needs to be correlated with the explanatory variable of interest, \( x \) \( (\text{cov}(x,z) \neq 0) \). Second, it should not be correlated with the error term \( (\text{cov}(z,\varepsilon) = 0) \).

Since \( \varepsilon \) represents unobserved determinants of \( y \), the two conditions mean that \( z \) is correlated with one determinant of \( y \), \( x \), but not with the other determinants, \( \varepsilon \) (or \( u \) as we mentioned earlier). This sounds highly arbitrary, in general. While we can check in the actual data whether \( x \) and \( z \) are correlated, we cannot directly check whether \( z \) and \( \varepsilon \) are uncorrelated because \( \varepsilon \) is not observed. So we need some other good reasons to
believe that $z$ and $\varepsilon$ are uncorrelated. This is why natural experiments are so useful to make a credible case for the validity of an instrument. If nature sets $z$ at random, then we don’t expect it to be related to $\varepsilon$.

A classic example of a natural experiment is weather shocks. For instance, if $y$ is the demand for oranges in Canada and $x$ is the price, weather in Florida (where many oranges are produced) should be correlated with the price but there is no obvious reasons why it should be correlated with other unobserved determinants (preferences, Canadian incomes, etc.) of the demand for oranges.

Over the last 25 years, a whole industry of research papers based on natural experiments has emerged in labour economics, and then in other fields. One famous example is the draft lottery paper of Josh Angrist (AER 1990). Here $y$ is civilian earnings in the 1980s, $x$ is whether one served in the U.S. armed forces during the Vietnam War, and the instrumental variable $z$ is a simple function of dates of birth. What happened in the late 1960s is that the U.S. introduced a random dimension to the draft process as the armed forces did not need to draft all young men to fulfill their troop requirements. Birth dates were selected at random, and men born on the first 100 selected days (more or less) were called up. So, for example, men born on August 10 were called up, but those born on August 9 or 11 were not. We would not expect any differences in future outcomes for these two groups except for the fact that those born on August 10 were more likely to serve in the armed forces. So if $z = 1$ for the selected birth dates but $z = 0$ for the others, we expect that $\text{cov}(x, z) \neq 0$ but that $\text{cov}(z, \varepsilon) = 0$.

Other popular methods based on natural experiments are difference-in-differences (DiD) designs and regression discontinuity (RD) designs. DiD designs fit nicely in a panel data setting as we are comparing groups before and after a treatment or policy change. RD designs are more closely connected to randomized experiments as they involve the comparison of a treatment and control group on either side of a cutoff point. For instance, we can estimate incumbency effects in politics by comparing politicians who barely won (the treatment group) or barely lost (the control group) a close election.

3 Appendix

Throughout this course we will heavily rely on a simple covariance formula approach to work with regression coefficients and discuss various estimation issues. To see where this is coming from, consider the simple case where $x$ is a scalar variable and where we also
have a constant in the model so that:

\[ y = \beta_0 + x\beta_1 + \varepsilon \tag{10} \]

Under the assumption that \( x \) and \( \varepsilon \) are uncorrelated (\( \text{cov}(x, \varepsilon) = 0 \), or the stronger condition \( \mathbb{E}(\varepsilon|x) = 0 \Rightarrow \text{cov}(x, \varepsilon) = 0 \)), we can write:

\[ \text{cov}(y, x) = \text{cov}(\beta_0 + x\beta_1 + \varepsilon, x) = \text{cov}(x, x)\beta_1 = \text{var}(x)\beta_1 \tag{11} \]

and solve for

\[ \beta_1 = \frac{\text{cov}(y, x)}{\text{var}(x)}. \tag{12} \]

Something known as the “analogy principle” can then be used to obtain an actual estimate of \( \beta_1 \). Up to this point, \( \text{cov}(y, x) \) and \( \text{var}(x) \) are the covariance and variance in the population, or the “true” covariances and variances. To get an estimate of \( \beta_1 \) we can replace the population values with the sample values \( \widehat{\text{cov}}(y, x) \) and \( \widehat{\text{var}}(x) \) where:

\[
\begin{aligned}
\widehat{\text{cov}}(y, x) &= N^{-1} \sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x}), \\
\widehat{\text{var}}(x) &= N^{-1} \sum_{i=1}^{N} (x_i - \bar{x})^2, \\
\bar{y} &= N^{-1} \sum_{i=1}^{N} y_i, \\
\bar{x} &= N^{-1} \sum_{i=1}^{N} x_i.
\end{aligned}
\]

As it turns out, \( \widehat{\beta}_1 = \frac{\widehat{\text{cov}}(y, x)}{\widehat{\text{var}}(x)} \) is exactly equal to the usual OLS estimator. To see this, remember the matrix version of the model

\[ Y = X\beta + \varepsilon, \tag{13} \]

where \( Y \) and \( \varepsilon \) are a \( N \times 1 \) vectors in which all the values of \( y_i \) and \( \varepsilon_i \) are stacked, \( X \) is a \( N \times 2 \) matrix with a column of 1 in the first column and the stacked values of \( x_i \) in the second column, and \( \beta \) is a \( 2 \times 1 \) vector \( (\beta = [\beta_0, \beta_1]^\prime) \). Now consider the well known
formula $\hat{\beta} = (X'X)^{-1}(X'Y)$. We can write

$$
(X'X)^{-1} = \left[ \begin{array}{ccc}
\sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i x_i^2 \\
\sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i^2 x_i \\
\end{array} \right]^{-1}
$$

where $J = N^2[\text{var}(x) + \text{var}(y) - \text{var}(x)]$, and where we have used the fact that $\sum_{i=1}^{N} x_i^2 = N \overline{x}^2 + N \text{var}(x)$.

Since we also have

$$
X'Y = \left[ \begin{array}{c}
\sum_{i=1}^{N} y_i \\
\sum_{i=1}^{N} x_i y_i \\
\end{array} \right] = \left[ \begin{array}{c}
N \overline{y} \\
N \text{cov}(y,x) + N \overline{x} \overline{y} \\
\end{array} \right]
$$

where we used the fact that $\sum_{i=1}^{N} x_i y_i = \sum_{i=1}^{N} [(x_i - \overline{x}) (y_i - \overline{y}) + \overline{x} \overline{y}] = N \text{cov}(y,x) + N \overline{x} \overline{y}$, we end up with

$$
\hat{\beta} = (X'X)^{-1}(X'Y) = J^{-1} \left[ \begin{array}{c}
\cdots \\
-N^2 \overline{x} \overline{y} + N^2 \text{cov}(y,x) + N^2 \overline{x} \overline{y} \\
\end{array} \right]
$$

and the second element of the vector $(\hat{\beta}_1)$ can be written as

$$
\hat{\beta}_1 = \frac{N^2 \text{cov}(y,x)}{N^2 \text{var}(x)} = \frac{\text{cov}(y,x)}{\text{var}(x)},
$$

which is the same as what we got using the analogy principle.