Economics 152
Section Notes

OLS Bias and Instrumental Variables

1 OLS Bias

1.1 Structural Equation

From the Statistics Review handout recall the regression equation for an outcome y_i determined by a variable x_i, where i indexes individual observations 1 through N.

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad \text{(Structural Equation)}$$

Call this equation the **structural equation** with the idea that this reflects the true structure of the relationship. There are two unknown parameters, a constant α, which is not of concern here, and the effect of a unit increase of x_i on y_i , β which is the parameter of interest. The "error" term ε_i reflects all other determinants of y_i not captured by the constant or x_i, and which by construction (use of a constant) has $E(\varepsilon_i) = 0$.

Example 1 *Let y_i be the logarithm of earnings, and x_i be years of schooling. A simple model of the return to schooling could be modeled by the above structural equation, assuming everyone has the same discount rate, reflected in a constant β. β is then the marginal rate of return to schooling, which we wish to estimate, α is average earnings with no schooling, and ε_i is determinants of earnings (e.g. "ability") not captured by schooling.*

1.2 OLS Estimation

The ordinary least squares (OLS) estimator is just the sample covariance of y_i and x_i divided by the sample variance of x_i. which as the sample size N gets large should converge on the true covariance and variance. Mathematically this is written as (where the "^" means "estimated")[1]

$$\hat{\beta}_{OLS} = \frac{\widehat{cov}(y_i, x_i)}{\widehat{var}(x_i)} \to \frac{cov(y_i, x_i)}{var(x_i)} \qquad \text{(OLS convergence)}$$

Substituting in for y_i using the structural equation we have that true covariance of y_i and x_i is

$$\begin{aligned} cov(y_i, x_i) &= cov(\alpha + \beta x_i + \varepsilon_i, x_i) \\ &= cov(\alpha, x_i) + cov(\beta x_i, x_i) + cov(\varepsilon_i, x_i) \\ &= 0 + \beta var(x_i) + cov(\varepsilon_i, x_i) \end{aligned}$$

The second line follows from the property of covariance $cov(a+b, c) = cov(a, c) + cov(b, c)$. The first term in the third line follows from the fact that the covariance of a constant with anything is zero. The second term in the third line follows from the fact that β is a constant and can be moved outside of the covariance, and $cov(x_i, x_i) = var(x_i)$. Substituting in this result into (OLS convergence) and simplifying gives

$$\hat{\beta}_{OLS} \to \beta + \frac{cov(\varepsilon_i, x_i)}{var(x_i)}$$

This means that if ε_i is correlated with x_i, then the OLS estimate will not converge on the true value for $\hat{\beta}$, i.e. it is biased. In general this means if the second term is large than our OLS estimates which reflect the overall relationship observed in the population will not be good: they are contaminated by other factors.

Example 2 *The typical "ability bias" problem results when chosen schooling x_i and ability ε_i are correlated. Differences in earnings across people with different levels of schooling do not just reflect the direct causal effect of schooling on earnings, but also the fact that people who would have earned more regardless of schooling also go to school more.*

[1] Technincally, "convergence" here means **convergence in probability** which means that the with probability approaching one the value of the left of the arrow will be arbitrarily close to the value to the right of the arrow as the sample size gets large.

2 Instrumental Variables

2.1 Finding a Good Instrument

Eliminating bias requires finding an alternative estimate to the OLS estimate. One strategy involves trying to find a variable z_i which influences x_i, but otherwise is not correlated with y_i directly, nor indirectly except through x_i. Technically this requirement is that

$$(1)\ cov(x_i, z_i) \neq 0 \qquad \text{(Relevance)}$$
$$(2)\ cov(\varepsilon_i, z_i) = 0 \qquad \text{(Excludability)}$$

a variable that satisfies this requirement is said to be an **instrumental variable**. A common example used here is of an instrumental variable is a binary variable $z_i = 1$ if an observation is "treated" and $z_i = 0$ if an observation is "untreated" or a control, with the idea that treated individuals receive more x_i but not more y_i, and that any systematic difference of y_i in the treatment group is due to the treatment's effect on x_i, and for no other reason.

2.2 The First Stage

An equation of x_i can be written using the instrumental variable on the right hand side, in what is called the first stage equation

$$x_i = \gamma + \delta z_i + \mu_i \qquad \text{(First Stage)}$$

where γ is a constant, , the parameter $\delta = cov(x_i, z_i)/var(z_i)$ relates the effect of z_i on x_i (assumption (1) can be rewritten as $\delta \neq 0$), and μ_i is an error term which by construction $E(\mu_i) = 0$, and $cov(z_i, \mu_i) = 0$.

Taking the interpretation of z_i as a binary variable indicating treatment, the first stage can be interpreted as the result of a "sloppy experiment" where the amount treatment applied x_i, e.g. the number of pills taken, is not completely deterministic. On average the control group gets γ in x's while on average the treatment group gets $\gamma + \delta$ amounts of x's, with $\mu_i \neq 0$ reflecting idiosyncratic differences across subjects. However, so long as the sloppy experiment treatment has an effect on x_i without being correlated with unobserved determinants of y_i, then we can still use the experiment to help us estimate β.

First we need to compute an estimate of δ, the effect of the treatment on x_i, which is given by the OLS estimator

$$\hat{\delta}_{OLS} = \frac{\widehat{cov(x_i, z_i)}}{\widehat{var(z_i)}} \qquad \text{(First Stage OLS)}$$

which according to our model will converge to δ as $cov(\mu_i, z_i) = 0$, i.e. $\hat{\delta}_{OLS} \to \delta$. In the case where z_i is binary the OLS estimate is just the simple difference estimator $\hat{\delta}_{OLS} = \hat{\delta}_D = \bar{x}_1 - \bar{x}_0$ where $\bar{x}_1$ is the average x_i when $z_i = 1$ and $\bar{x}_0$ is the average x_i when $z_i = 0$. This converges as

$$\hat{\delta}_D = \bar{x}_1 - \bar{x}_0 \to E[x_i | z_i = 1] - E[x_i | z_i = 0] = (\gamma + \delta) - \gamma = \delta.$$

where $E[x_i | z_i = z]$ is the conditional expectation of x_i given that $z_i = z$.[2]

2.3 Reduced Form Equation

Substituting in the first stage equation for x_i into the structural equation we get an expression for y_i in terms of z_i.

[2] Letting $\Pr(z_i = 1) = p$ as with a standard Bernoulli variable, $var(z_i) = p(1-p)$ and more subtly

$$\begin{aligned} cov(x_i, z_i) &= E(x_i z_i) - E(x_i) E(z_i) \\ &= pE[x_i | z_i = 1] \cdot 1 + (1-p) E(x_i | z_i = 0) \cdot 0 - \{pE[x_i | z_i = 1] + (1-p) E[x_i | z_i = 0]\} p \\ &= p(1-p) \{E[x_i | z_i = 1] - E[x_i | z_i = 0]\} \end{aligned}$$

Therefore the two are equivalent as $\delta = cov(x_i, z_i)/var(z_i) = E[x_i | z_i = 1] - E[x_i | z_i = 0]$

$$\begin{aligned} y_i &= \alpha + \beta(\gamma + \delta z_i + \mu_i) + \varepsilon_i \\ &= (\alpha + \gamma\delta) + (\beta\delta) z_i + (\beta\mu_i + \varepsilon_i) \end{aligned}$$

The equation compounds the first stage and structural coefficients resulting in a compounded constant $\phi = (\alpha + \beta\gamma)$ a compounded error term $\eta_i = (\beta\mu_i + \varepsilon_i)$ and a compounded coefficient $\pi = \beta\delta$ on z_i. Using this shorthand we can rewrite this equation in a more reduced manner known as the **reduced form**

$$y_i = \phi + \pi z_i + \eta_i \qquad \text{(Reduced Form)}$$

Since z_i is uncorrelated with ε_i and μ_i it follows that it is uncorrelated with the reduced form error term η_i as

$$cov(z_i, \eta_i) = cov(z_i, \beta\mu_i + \varepsilon_i) = \beta cov(z_i, \mu_i) + cov(z_i, \varepsilon_i) = \beta \cdot 0 + 0 = 0$$

Therefore an OLS estimate of this equation will give us a good estimate of $\pi = \beta\delta$

$$\hat{\pi}_{OLS} = \frac{\widehat{cov}(y_i, z_i)}{\widehat{var}(z_i)} \to \frac{cov(y_i, z_i)}{var(z_i)} = \pi \qquad \text{(RF OLS)}$$

In the case where z_i is binary, a simple difference estimator $\hat{\pi}_{OLS} = \hat{\pi}_D = \bar{y}_1 - \bar{y}_0$, where the $\bar{y}_1$ is the average y_i when $z_i = 1$, and $\bar{y}_0$ is the average y_i when $z_i = 0$. This works for similar reasons given for $\hat{\pi}_D$.

2.4 Instrumental Variable Estimator

Now that we have established good ("consistent") estimators for δ and $\pi = \beta\delta$, which converge to the true values, it suffices to take their ratio to get an estimate which converges to the true β. This results in the instrumental variable IV estimator

$$\hat{\beta}_{IV} = \frac{\hat{\pi}_{OLS}}{\hat{\delta}_{OLS}} \to \frac{\pi}{\delta} = \frac{\beta\delta}{\delta} = \beta$$

which converges to the true value of β.[3] Taking the ratio of the OLS estimators you can see that the $\widehat{var}(z_i)$ terms cancel out so

$$\hat{\beta}_{IV} = \frac{\widehat{cov}(y_i, z_i)}{\widehat{cov}(x_i, z_i)} \qquad \text{(IV estimate)}$$

In the case where z_i is binary and then the IV estimate is known as the **Wald Estimate** as $\hat{\pi}_{OLS} = \bar{y}_1 - \bar{y}_0$ and $\hat{\delta}_{OLS} = \bar{x}_1 - \bar{x}_0$ and so

$$\hat{\beta}_{Wald} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0} \qquad \text{(Wald Estimate)}$$

This estimate is in fact quite intuitive: take the of the change in x due to z and use it to divide the change in y due to z. If our assumptions are true all of the change in y is due to the change in z occurs through the channel of x, than this is a good estimate of the effect of x on y.

[3] The fact that $\hat{\pi}_{OLS} \to \pi$ and $\hat{\delta}_{OLS} \to \delta$ implies $\hat{\pi}_{OLS}/\hat{\delta}_{OLS} \to \pi/\delta$ is not a trivial result and is known as "Slutsky's theorem."