ECON 251

Discussion Section

Week 7 Solutions

1. Prove that OLS is unbiased under independence, and consistent under exogeneity.

$$\hat{\beta}_1^{OLS} \coloneqq \frac{\widehat{\mathrm{Cov}}(X,Y)}{\widehat{\mathrm{Var}}(X)} = \frac{\sum_{i=1}^{N}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{N}(X_i-\overline{X})^2} = \frac{\sum_{i=1}^{N}(X_i-\overline{X})\cdot Y_i}{\sum_{i=1}^{N}(X_i-\overline{X})^2}$$

$$= \frac{\sum_{i=1}^{N}(X_i-\overline{X})\cdot(\beta_0+\beta_1 X_i+U_i)}{\sum_{i=1}^{N}(X_i-\overline{X})^2}$$

$$= \frac{\beta_0\sum_{i=1}^{N}(X_i-\overline{X})+\beta_1\sum_{i=1}^{N}(X_i-\overline{X})X_i+\sum_{i=1}^{N}(X_i-\overline{X})U_i}{\sum_{i=1}^{N}(X_i-\overline{X})^2}$$

$$= \frac{\beta_0\cdot 0+\beta_1\sum_{i=1}^{N}(X_i-\overline{X})^2+\sum_{i=1}^{N}(X_i-\overline{X})U_i}{\sum_{i=1}^{N}(X_i-\overline{X})^2}$$

$$= \beta_1 + \frac{\sum_{i=1}^{N}(X_i-\overline{X})U_i}{\sum_{i=1}^{N}(X_i-\overline{X})^2}$$

$$\Rightarrow E\left[\hat{\beta}_1^{OLS}\middle|X_i\right] = E\left[\beta_1 + \frac{\sum_{i=1}^{N}(X_i-\overline{X})U_i}{\sum_{i=1}^{N}(X_i-\overline{X})^2}\middle|X_i\right]$$

$$= \beta_1 + \frac{\sum_{i=1}^{N}(X_i-\overline{X})E[U_i|X_i]}{\sum_{i=1}^{N}(X_i-\overline{X})^2} = \beta_1$$

$$\operatorname*{plim}_{N\to\infty}\hat{\beta}_1^{OLS} \coloneqq \operatorname*{plim}_{N\to\infty}\frac{\widehat{\mathrm{Cov}}(X,Y)}{\widehat{\mathrm{Var}}(X)} = \frac{\operatorname*{plim}_{N\to\infty}\widehat{\mathrm{Cov}}(X,Y)}{\operatorname*{plim}_{N\to\infty}\widehat{\mathrm{Var}}(X)} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$$

$$= \frac{\mathrm{Cov}(X,\beta_0+\beta_1X+U)}{\mathrm{Var}(X)} = \frac{\mathrm{Cov}(X,\beta_0)+\mathrm{Cov}(X,\beta_1X)+\mathrm{Cov}(X,U)}{\mathrm{Var}(X)}$$

$$= \frac{0+\beta_1\cdot\mathrm{Var}(X,X)+\mathrm{Cov}(X,U)}{\mathrm{Var}(X)} = \beta_1 + \frac{\mathrm{Cov}(X,U)}{\mathrm{Var}(X)} = \beta_1$$

2. Relate the statistical concepts of unbiasedness and consistency with the econometric concepts of selection bias (under the potential outcomes/treatment effects framework) and omitted variable bias (under the linear regression framework).

In the previous question, we showed that the independence assumption $E[U|X] = 0$ implies OLS is unbiased $E\left[\hat{\beta}_1^{OLS}\right] = \beta_1$ while the exogeneity assumption $\text{Cov}(X, U) = 0$ implies OLS is consistent $\underset{N\to\infty}{\text{plim}}\, \hat{\beta}_1^{OLS} = \beta_1$. In the potential outcomes framework, the independence assumption means that we can identify the average treatment effect $ATE \coloneqq E[Y_i(1) - Y_i(0)]$ with simple comparisons[1] as there is no selection bias on average between treated and control units. Thus, we write

$$\overline{Y}_1 - \overline{Y}_0 = ATT + SB = ATE$$

where the first equality is <u>always</u> true (mean differences equal causal effect of treatment among the treated plus selection bias $SB \coloneqq E[Y_i(0)|X_i = 1] - E[Y_i(0)|X_i = 0]$) and the second equality is <u>only</u> true under independence $X_i \perp Y_i(0), Y_i(1)$ which is equivalent to $E[U|X] = 0$. In the linear regression framework, we can imagine what might happen if we omit an important (for determining the outcome) and relevant (for correlating with treatment) variable from our model. The probability limit of OLS is always given by

$$\underset{N\to\infty}{\text{plim}}\, \hat{\beta}_1^{OLS} = \beta_1 + \frac{\text{Cov}(X, U)}{\text{Var}(X)} = \beta_1 + \gamma^{X\to V} \cdot \delta^{V\to Y}$$

where the second equality follows from omitted variable bias (OVB) formula, with $\gamma^{X\to V}$ giving the linear effect of the treatment on the omitted variable V and $\delta^{V\to Y}$ giving its effect on the outcome. The exogeneity assumption $\text{Cov}(X, U) = 0$ rules out such omitted random variables which are both important and relevant, and delivers consistency!

[1] Recall that OLS is equivalent to a simple comparison (mean difference) when treatment is binary!

3. In *Stata*, review interpretation of simple and multiple linear regression tables. Discuss functional form issues (and possibilities) with linear regression: logarithm of outcome or treatment, binary variables and groups, quadratic forms, interactions.

4. Discuss "On Binscatter" by Cattaneo et al. (2022) as a rigorous method to visually inspect assumptions SLR1/MLR1, that linear regressions are actually linear!

   The linearity assumption requires that the conditional expectation function relating the outcome to the treatment be linear $E[Y|X = x] = g(x) = \beta_0 + \beta_1 x$. Sometimes this assumption is overlooked[2] in practice, but an easy check is to simply plot the data! If the scatterplot displays non-linear patterns, then we could either consider some transformations to make things linear (such as taking logarithms of skewed variables in our analysis or adding the square of treatment to our specification). A quite practical limitation of this simple and intuitive check is when the sample size is large! In these situations, it can be quite difficult to discern the shape of the conditional expectation function, so a simple solution involves making a scatter plot of the treatment vs the average value of the outcome within certain intervals of the treatment distribution. Cattaneo et al. (2022) present methods on implementing this in practice, which involves several important technical steps like figuring out the right way to specify the bins across treatment to average the outcome and creating confidence intervals properly.

[2] This is not without justification: we can show that even if $g(x)$ is non-linear, estimating slopes and an intercept with least squares produces the best linear predictor of the conditional expectation function.