ECON 251

Discussion Section

Week 9 Solutions

1. In the simple linear regression $Y = \beta_0 + \beta_1 X + U$, what do the following quantities mean and how do they relate to each other, if at all: R^2, t-stat on null hypothesis of zero effect, mean independence condition $E[U|X] = 0$. What do you really care about in all this?

   R^2 gives the fraction of the total variation in the outcome $\sum(Y_i - \overline{Y})^2$ that is explained by our model $\sum(\hat{Y}_i - \overline{Y})^2$. This number is always between 0 and 1, and the closer the cloud of points is to the regression line in a scatterplot, the higher the value of R^2 will be, indicating that the treatment predicts a greater fraction of the variation in the outcome.

   The t-stat on a null hypothesis of zero effect is the ratio of the OLS estimator $\hat{\beta}_1^{OLS}$ to its standard error $\mathrm{SE}(\hat{\beta}_1^{OLS})$, or estimated standard deviation. It tells us whether the estimated line is not flat; if the null hypothesis of zero treatment effect were true, then the slope would be zero and we would get a flat line. Since we only work with a tiny sample of the population, we recognize that there is uncertainty in this decision, and the t-stat helps us quantify it. A higher value (in absolute value) indicates that our estimate of the slope is quite large relative to the variation we would expect from our estimator across many different samples, which we quantify with the estimated standard error, so we can be confident that the true slope in the population is not zero.

   Mean independence requires that the treatment not predict the average value of omitted variables. If this were not true, then inferring the slope of the line by looking at differences in outcomes across units with different levels of the treatment would be fraught with difficulty. A large positive value could indicate that treatment increases the outcome, or

that treatment is associated with something else that leads to changes in the outcome (perhaps due to omitted relevant variables). At the end of the day, if we care about properly estimating causal effects, whether $E[U|X] = 0$ is true is what really matters. A super steep slope (indicated by a big t-stat) or a super tight fit (indicated by a big R^2) is useless if we cannot be sure that all the correlation we documented is truly causal!

2. Heteroskedasticity robust inference, always, but only if your sample is big enough! :)

$$\text{se}\left(\hat{\beta}_j^{OLS}\right)_{\text{HO}} = \sqrt{\frac{\hat{\sigma}^2}{\widehat{\text{Var}}(X_j) \cdot (1 - R_j^2)}} \coloneqq \sqrt{\frac{\frac{1}{N-k-1}\sum_{i=1}^{N} \hat{U}_i^2}{\frac{1}{N-1}\sum_{i=1}^{N}(X_j - \overline{X}_j)^2 \cdot \left(1 - R^2_{\text{reg } X_j \text{ on all } X_k}\right)}}$$

$$\text{se}\left(\hat{\beta}_j^{OLS}\right)_{\text{HR}} = \sqrt{\frac{\hat{\sigma}_{HR}^2}{SSR_j}} \qquad \coloneqq \sqrt{\frac{\sum_{i=1}^{N} \hat{r}_{ij}^2 \cdot \hat{U}_i^2}{\sum_{i=1}^{N} \hat{e}_{ij}^2}}$$

where $\hat{r}_{ij}^2$ denotes the i-th residual and $SSR_j \coloneqq \sum_{i=1}^{N} \hat{e}_{ij}^2$ denotes the sum of squares from reg X_j on all X_k …also known as: Huber, Eicker, White (HEW) standard errors, or "sandwich" standard errors if you write the formula with matrices :)

3. Sensitivity analysis as one answer to intellectual nihilism implied by ubiquitous omitted variables lurking within the natural variation from which we wish to discern "truth"
    - Let's discuss this sentence carefully as a group…
    - Example: how we "know" that smoking tobacco causes an increase in the probability of getting lung cancer (Cinelli 2020 presentation: 1:50 – 7:57)
    - Goal: learn new methods in Cinelli and Hazlet (2020), like using the observed effect of a control to benchmark how strong of an affect unobserved confounders would need to have to overturn an estimated treatment effect

4. Examples to illustrate when *classical* measurement error matters (in the treatment, attenuation bias in slope estimator) and when it doesn't (in the outcome, no bias in OLS). In the simple linear regression $Y = \beta_0 + \beta_1 X + U$ we say the outcome is mismeasured if we only observe $Y^* \coloneqq Y + e$ with $E[e|X] = 0$. Derive the least squares estimator $\hat{b}_1^{OLS}$ when we estimate $Y^* = b_0 + b_1 X + U$ instead of the true model.

$$\begin{aligned}
\hat{b}_1^{OLS} \coloneqq \frac{\widehat{\mathrm{Cov}}(X, Y^*)}{\widehat{\mathrm{Var}}(X)} &= \frac{\sum_{i=1}^N (X_i - \overline{X})\left(Y_i^* - \overline{Y}^*\right)}{\sum_{i=1}^N (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^N (X_i - \overline{X}) \cdot Y_i^*}{\sum_{i=1}^N (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^N (X_i - \overline{X}) \cdot (Y_i + e_i)}{\sum_{i=1}^N (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^N (X_i - \overline{X}) \cdot Y_i}{\sum_{i=1}^N (X_i - \overline{X})^2} + \frac{\sum_{i=1}^N (X_i - \overline{X}) \cdot e_i}{\sum_{i=1}^N (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^N (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{\sum_{i=1}^N (X_i - \overline{X})^2} + \frac{\sum_{i=1}^N (X_i - \overline{X}) \cdot e_i}{\sum_{i=1}^N (X_i - \overline{X})^2}
\end{aligned}$$

$$\begin{aligned}
\Rightarrow E[\hat{b}_1^{OLS}|X_i] &= E\left[\frac{\sum_{i=1}^N (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{\sum_{i=1}^N (X_i - \overline{X})^2} + \frac{\sum_{i=1}^N (X_i - \overline{X}) \cdot e_i}{\sum_{i=1}^N (X_i - \overline{X})^2} \,\middle|\, X_i\right] \\
&= \beta_1 + E\left[\frac{\sum_{i=1}^N (X_i - \overline{X}) \cdot e_i}{\sum_{i=1}^N (X_i - \overline{X})^2} \,\middle|\, X_i\right] \\
&= \beta_1 + \frac{\sum_{i=1}^N (X_i - \overline{X}) \cdot E[e_i|X_i]}{\sum_{i=1}^N (X_i - \overline{X})^2} \\
&= \beta_1
\end{aligned}$$