Econ 400 Problem Set 6 Question 2

Travis Cao

March 8, 2021

(a) From Dis 3, we've learned that β_1's estimate is calculated as

$$\hat{\beta}_1 = \frac{\widehat{Cov}(X_i, Y_i)}{\widehat{Var}(X_i)}$$

With the partially scrambled $\widetilde{X}_i$ being our regressor,

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\widehat{Cov}(\widetilde{X}_i, Y_i)}{\widehat{Var}(\widetilde{X}_i)} \\
&= \frac{\widehat{Cov}(\widetilde{X}_i, \beta_0 + \beta_1 X_i + u_i)}{\widehat{Var}(\widetilde{X}_i)} \qquad \text{(since } Y_i = \beta_0 + \beta_1 X_i + u_i \text{ from the true model)} \\
&= \beta_1 \frac{\widehat{Cov}(\widetilde{X}_i, X_i)}{\widehat{Var}(\widetilde{X}_i)} + \frac{\widehat{Cov}(\widetilde{X}_i, u_i)}{\widehat{Var}(\widetilde{X}_i)} \\
&= \beta_1 \frac{(\frac{1}{n}\sum_i^n \widetilde{X}_i X_i) - \overline{\widetilde{X}}\,\overline{X}}{(\frac{1}{n}\sum_i^n \widetilde{X}_i^2) - \overline{\widetilde{X}}^2} + \frac{(\frac{1}{n}\sum_i^n \widetilde{X}_i u_i) - \overline{\widetilde{X}}\,\overline{u}}{(\frac{1}{n}\sum_i^n \widetilde{X}_i^2) - \overline{\widetilde{X}}^2}
\end{aligned}$$

One thing to notice: since 20% of X_i is scrambled, this means that the position of these data have changed (for example, consider this as we swapping X_{10} with X_{15} – the position of the data changed, but the data point isn't lost), but the sum of all Xs should still be the same, meaning that

$$\overline{\widetilde{X}} = \overline{X} \quad \text{and} \quad \frac{1}{n}\sum_i^n \widetilde{X}_i^2 = \frac{1}{n}\sum_i^n X_i^2$$

This means that

$$\hat{\beta}_1 = \beta_1 \frac{(\frac{1}{n}\sum_i^n \widetilde{X}_i X_i) - \overline{X}^2}{(\frac{1}{n}\sum_i^n X_i^2) - \overline{X}^2} + \frac{(\frac{1}{n}\sum_i^n \widetilde{X}_i u_i) - \overline{X}\,\overline{u}}{(\frac{1}{n}\sum_i^n X_i^2) - \overline{X}^2} \tag{1}$$

so when taking the expectation of $\hat{\beta}_1$:

$$E[\hat{\beta}_1] = \beta_1 E\left[\frac{(\frac{1}{n}\sum_i^n \widetilde{X}_i X_i) - \overline{X}^2}{(\frac{1}{n}\sum_i^n X_i^2) - \overline{X}^2}\right] + E\left[\frac{(\frac{1}{n}\sum_i^n \widetilde{X}_i u_i) - \overline{X}\,\overline{u}}{(\frac{1}{n}\sum_i^n X_i^2) - \overline{X}^2}\right] \tag{2}$$

Consider the second part equation (2) first:

$$
\begin{aligned}
E\left[\frac{(\frac{1}{n}\sum_i^n \widetilde{X}_i u_i) - \overline{X}\overline{u}}{(\frac{1}{n}\sum_i^n X_i^2) - \overline{X}^2}\right] &= E\left[E\left[\frac{(\frac{1}{n}\sum_i^n \widetilde{X}_i u_i) - \overline{X}\overline{u}}{(\frac{1}{n}\sum_i^n X_i^2) - \overline{X}^2}\middle| X\right]\right] \\
&= E\left[\frac{(\frac{1}{n}\sum_i^n \widetilde{X}_i E[u_i|X]) - \overline{X}\frac{1}{n}\sum_{i=1}^n E[u_i|X]}{(\frac{1}{n}\sum_i^n X_i^2) - \overline{X}^2}\right] \\
&= E\left[\frac{(\frac{1}{n}\sum_i^n \widetilde{X}_i 0) - \overline{X}\frac{1}{n}\sum_{i=1}^n 0}{(\frac{1}{n}\sum_i^n X_i^2) - \overline{X}^2}\right] \\
&\qquad \text{(by OLS's zero conditional mean assumption)} \\
&= 0
\end{aligned}
$$

Now let's consider the first part of equation (2):

$$
\begin{aligned}
\beta_1 E\left[\frac{(\frac{1}{n}\sum_i^n \widetilde{X}_i X_i) - \overline{X}^2}{(\frac{1}{n}\sum_i^n X_i^2) - \overline{X}^2}\right] &= \beta_1 E\left[\frac{\widehat{Cov}(\widetilde{X}_i, X_i)}{\widehat{Var}(X_i)}\right] \\
&= \beta_1 E\left[\frac{0.8\widehat{Var}(X_i) + 0.2\widehat{Cov}(X_j, X_i)}{\widehat{Var}(X_i)}\right] \qquad \text{(KEY!)} \\
&= \beta_1 E\left[0.8 + \frac{0}{\widehat{Var}(X_i)}\right] \qquad (\widehat{Cov}(X_j, X_i) = 0 \text{ by i.i.d. assumption}) \\
&= 0.8\beta_1
\end{aligned}
$$

And the (KEY!) line results from the fact that only 20% of X_i is randomly scrambled, meaning that we can consider $\widetilde{X}_i = 0.8X_i + 0.2X_j$, where j represents a different position from i (indicating the scrambling of the data).

Combining the first and second part, we have equation (2) as

$$E[\hat{\beta}_1] = 0.8\beta_1$$

which is the result we're looking for.

(b) An estimator is unbiased if $E[\text{estimator}] = \text{true value}$, so we want to find an estimator for β_1 such that expectation of this estimator $= \beta_1$.

Notice that from (a),

$$
\begin{aligned}
E[\hat{\beta}_1] &= 0.8\beta_1 \\
\frac{1}{0.8}E[\hat{\beta}_1] &= \beta_1 \\
E\left[\frac{\hat{\beta}_1}{0.8}\right] &= \beta_1
\end{aligned}
$$

This means that $\frac{\hat{\beta}_1}{0.8}$ will be an unbiased estimator in this case.

(c) We are now comparing two estimators: $\frac{\hat{\beta}_1}{0.8}$ proposed in part (b), and $\hat{\beta}_1$ but only uses the last 240 observations.

- From (b), we already know that $\frac{\hat{\beta}_1}{0.8}$ is an unbiased estimator.
- For $\hat{\beta}_1$ using the last 240 observations, since these 240 observations are lined up correctly, $E[\hat{\beta}_1] = \beta_1$, meaning that this $\hat{\beta}_1$ is also an unbiased estimator.

  (Intuitively, think of this as if we're using a smaller sample to estimate the true β_1.)

So both estimators are unbiased. Now to tell which estimator is "better", we have to study which estimator yields smaller standard error.

- For the first estimator,

$$Var\left(\frac{\hat{\beta}_1}{0.8}\middle|X\right) = \frac{1}{0.8^2} Var\left(\hat{\beta}_1|X\right) = 1.5625 \times Var\left(\hat{\beta}_1|X\right)$$

- For the second estimator, its variance is $Var\left(\hat{\beta}_1|X\right)$

So obviously, the first estimator yields bigger variance (i.e. bigger standard error), so the second estimator (the one that only uses the last 240 correctly aligned observations) is better.

Side note: You might be concerned with the second estimator using a smaller size sample that might make the standard error of $\hat{\beta}_1$ bigger to begin with, but that effect is dominated by the fact that the first estimator needs to do a $\frac{1}{0.8}$ adjustment factor to make itself unbiased. The $\frac{1}{0.8}$ adjustment has a way bigger effect on standard error than the slightly smaller sample size used in the second estimator.