EKT-816: Problem Set 2

There are 100 points available on this problem set, and it will be graded out of 80. That is, there are 20 bonus points available.

Due Date: Wednesday, March 20

Properties of OLS

1. Consider running a regression of some variable y on a vector of regressors x, where x has dimension $k \geq 1$.

(a) Will the estimated residuals sum to zero if we do not include a constant in the regression? If not, what will $\sum_{i=1}^{n} \widehat{e}_i$ be?

From now on you can assume that x contains a constant.

(b) We showed in class that the coefficient on a particular variable - say x_j - that you get from running a regression of y on x is numerically identical to the one you would get from regressing y on x_j^*, where x_j^* is the residual from a regression of x_j on $x_1, \ldots x_{j-1}, x_{j+1}, \ldots x_k$.

Now, consider regressing y^* on x_j^*, where y^* is the residual from a regression of y on $x_1, \ldots x_{j-1}, x_{j+1}, \ldots x_k$. Call the resulting coefficient estimate $\widetilde{\beta}_j$. What is the relationship between $\widehat{\beta}_j$ and $\widetilde{\beta}_j$? Can you explain why?

[$2 \times 5 = 10$ points]

More Data Manipulation in R

2. Do exercises 2-3 of section 13.4.6 of Wickham and Grolemund (2017).

[$2 \times 5 = 10$ points]

3. Do exercises 1-2 of section 13.5.1 of Wickham and Grolemund (2017).

[$2 \times 5 = 10$ points]

Specification Choice

Examples from Agricultural Economics

4. Recall the setup from class:

- y_i is output per acre on farm i
- x_{i1} is an index of soil quality
- x_{i2} is an index of weather quality
- x_{i3} is an index of pesticide use
- e_i is a measure of insect population density

We know that crop yields are determined as

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + e$$

We are also given that x_1, x_2, and e are mutually independent with mean zero. Now, consider two different assumptions about how pesticide use is determined:

- model A: farmers ignore soil quality (or do not observe it), but they do observe the level of insect populations, and they set $x_3 = -e/\gamma + \eta$ where η is independent of all the x variables.
- model B: farmers set $x_3 = x_1 - e/\gamma_2 + \varepsilon$, where ε is independent of the x variables.

You have access to data on y, x_1, x_2, and x_3.

(a) If model B generates the data, will a regression of crop yields on soil quality - controlling for weather quality **and** pesticide use - deliver a consistent estimate of β_1?
(b) If model B generates the data, will a regression of crop yields on soil quality, controlling for weather quality but **not** pesticide use, deliver a consistent estimate of β_1? What interpretation can you give to the limiting value of $\widehat{\beta}_1$ in this scenario?

[$2 \times 10 = 20$ points]

Measurement Error

5. Suppose you have the following data generating process:

$$y = x_1\beta_1 + x_2\beta_2 + e$$

Here, e is independent of (x_1, x_2), has mean zero and variance σ_e^2. However, you do not observe x_2 but a mismeasured version of it:

$$z_2 = x_2 + u$$

where u is independent of (e, x_1, x_2), has mean zero, and variance σ_u^2.

(a) If you were to run a regression of y on (x_1, z_2), what would the limiting value of $\widehat{\beta}_2$ be?
(b) Same question, but for $\widehat{\beta}_1$. Which factors will tend to make the bias (if any) larger or smaller?

[$2 \times 10 = 20$ points]

Monte Carlo Simulations

6. Write an R script to simulate the sampling distribution of the sample mean for $n = 10, 100, 1000$, and 10000. Generate the underlying data as an iid sample $X_i \sim F$, where F is the distribution you described in Q.3 of problem set 1. Use $B = 1000$ draws from each distribution.

Display a table with the mean, standard deviation, skewness, and kurtosis of the sampling distribution for each n. Also display a plot of the density of $\overline{X}_n$ for each n.

[20 points]

Robust Standard Errors

7. Clone github.com/vikjam/mostly-harmless-replication and replicate Figure 3.1.2 of *Mostly Harmless Econometrics*. Re-run the regression but use robust standard errors. Output your results to a properly formatted table (look up the package `stargazer` to do this).

[10 points]

References

Wickham, Hadley, and Garrett Grolemund. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly. https://r4ds.had.co.nz.