Harvard Economics
Econometrics Math Camp 2020

Break-Out Session #3

I really want to convince you that the law of iterated expectations is important. It shows up everywhere! Please work with your classmates to prove the following theorem. You will need the law of iterated expectations.

The Analysis of Variance Theorem: Let X and Y be random variables. Then:

$$Var[Y] = Var[E[Y|X]] + E[Var[Y|X]]$$

Please prove this result. You will find it very helpful to start with the CEF decomposition theorem from last week's worksheet (also available on GitHub).

Break-Out Session #4

This session consists of two problems that I am asking you to solve by simulation in the programming language of your choice. The first problem concerns the intuition and assumptions behind some of the asymptotic tools we discussed earlier; in particular, the law of large numbers. The second problem is intended to get your hands dirty with matrix algebra (i.e. solving for OLS coefficients by inverting matrices), but also reveals some pretty deep stuff about the CLT and what the distribution of an estimator looks like.

Problem 1: Sample Means and the LLN

This question is intended to make you think about the law of large numbers and when it applies. Please use the programming language of your choice to solve this problem - I'll use R.

1. Let $N = 100$ and and construct two $N \times 1$ vectors drawn i.i.d. from a standard normal distribution, $N(0, 1)$. Call these vectors X_1 and X_2. Construct a new vector from these two vectors as $Y = X_1/X_2$; in other words, Y is the ratio of two normal variables.
2. Compute the sample means of X_1 and Y – that is, $\bar{X} = \frac{1}{N}X_i$ and $\bar{Y} = \frac{1}{N}Y_i$ – for $N = 100$.
3. Repeat (2), but with $N = 1000$, $N = 10000$, and $N = 100000$. What is happening to the sample means of X_1 and Y as you increase N? How is this related to the law of large

numbers? (Hint: something weird is happening with Y)

Problem 2: Repeated Regression and the CLt

1. **Simulate the data**: The advantage of a simulation exercise is that you control (and therefore observe) the data generating process for your data. We will start by creating our data. Let $N = 50$. Generate an $N \times 1$ vector of ones, call it X_1. Generate an $N \times 1$ vector of draws from a standard normal distribution, call it X_2. Form an $N \times 2$ matrix X by concatenating these two column vectors, so the first column of X is X_1 and the second column is X_2. Next, generate another $N \times 1$ vector of standard normal draws, call it e. Finally, construct the $N \times 1$ vector Y as:
$$Y = 1X_1 + 2X_2 + e$$

2. **Generate OLS coefficients**: Use the matrix algebra characterization of $\hat{\beta}$ described in our slides to solve for the OLS estimates of β. What would you expect the coefficients to tend to as the sample grows large? What are they? How can you explain the difference?

3. **Repeated simulation**: Refactor your code so that you run both steps above K times, where we initially will set $K = 100$. Make sure to re-draw the data in each simulation. You will want to save the estimated coefficients in each simulation as the rows of a $K \times 2$ matrix, call it B. Run this simulation exercise for $K = 100$ and $N = 50$ and examine a histogram of the estimated coefficients on X_2 across simulations (i.e. the second element of the coefficient matrix). Now do the same thing, but with $K = 1000$ and $K = 10000$. What is happening to the distribution of estimates as you increase the number of simulations? Why do you think this is?

4. **Varying N**: What happens when you allow N to grow large as well?