ECO 375–Homework 1
University of Toronto Mississauga
Due: 17 October, 2023, 5 PM

Notes:

- Late homework assignments—including assignments submitted late on the day of the deadline—will be subject to a late penalty of 10% per calendar day (including weekends) of the total marks for the assignment. Assignments submitted more than 5 calendar days beyond the due date will be assigned a grade of zero. Homework assignments handed in after the work has been returned to the class or sample solutions have been released cannot be marked for credit. Accommodations due to late registration into the course will NOT be approved.
- For full credit, please **show your work**. The correct answer without a clear explanation will receive little credit.
- All problem sets must be legible; if the grader cannot read your submission, you will receive no credit.
- Provide your do file and log file as part of your submission. If you are using Citrix, remember to save your code and log file on your home computer, not the remote computer! You will need to copy-paste your do and log files into Crowdmark.
- When questions involve the use of Stata, as part of the response to each question, you need to submit a write-up that (a) interprets and explains your computer output and (b) includes the code for that answer answer. Without both components, you will receive a mark of 0. For example, if the question asks for the mean of x, your answer might look like this:
  "The mean of x is 12345.
  Code:
  `sum x`"
- You may take as given anything from the textbook or that we learned in class, but not any other information.
- All work MUST be your own. Using generative AI to answer a question constitutes cheating. Posting questions to an online forum or website constitutes cheating and violates copyright law; looking for the answer in an online forum or website constitutes cheating.

Theoretical Problems

1. Suppose $Y = X^2 + 3$. For this problem, you may use the fact that, if $W \sim N(0,1)$, then $E[W^3] = 0$ and $E[W^4] = 3$. Recall that $N(\mu, \sigma^2)$ indicates a variance of σ^2.

   (a) Suppose $X \sim N(1,3)$.

      i. *(points: 5)* What is the variance of Y? [Hint: recall from class how to standardize a normal distribution.]

      ii. *(points: 5)* What is the covariance of X and Y?

      iii. *(points: 2)* In week 6, you will learn that, if we use OLS to estimate the βs in $Y = \beta_0 + \beta_1 X + \epsilon$, then our estimate $\hat{\beta}_1 \xrightarrow{p} \frac{C[X,Y]}{V[X]}$. What is $plim(\hat{\beta}_1)$ given this data? [Hint: this should be a very easy question if you've answered previous parts correctly.]

   (b) Now, suppose X takes on only two values: with 50% probability, it is 0, and with 50% probability, it is 2.

      i. *(points: 5)* Calculate $E[X]$, $E[X^2]$, $E[X^3]$, and $E[X^4]$.

      ii. *(points: 5)* What is the variance of Y?

      iii. *(points: 5)* What is the covariance of X and Y?

      iv. *(points: 2)* What is the $plim(\hat{\beta}_1)$ given this data?

2. A company wants to improve their online advertising. They therefore develop two advertising campaigns, called A and B, which will be shown to users of a social media company, Y. Their goal is to determine which campaign is better at getting users to click the ad and buy something from their website. Note: to receive full credit for this question, please reference terms and ideas discussed in class in each part of this question. For example, if an idea is bad because it violates an assumption, explicitly state the assumption (e.g. SLR.3) and explain why it is violated.

   (a) *(points: 6)* To test which campaign works better, the company shows campaign A to older users of Y and B to younger users. They then measure whether each person who views the ad clicks it and buys something from their website. They define Y_i as the amount of money person i spends, and D_i as a dummy variable that takes a value of 1 if person i saw campaign A, 0 if person i saw campaign B. They then use OLS to estimate $Y_i = \beta_0 + \beta_1 D_i + u_i$. Will their estimate of β_1 be an unbiased estimate of the effect of A vs B? Explain why or why not.

   (b) *(points: 6)* Now, to test which campaign works better, for one year, the company randomly chooses 3 days a week to show campaign A, and 3 days a week to show campaign B; one day a week, they will use neither campaign. (For example, they might randomly choose Mondays, Tuesdays, and Fridays for campaign A; and Wednesdays, Thursdays, and Saturdays for campaign B. As in the previous problem, they define Y_i as the amount of money person i spends, and D_i as a dummy variable that takes a value of 1 if person i saw campaign A, 0 if person i saw campaign B. They then use OLS to estimate $Y_i = \beta_0 + \beta_1 D_i + u_i$. Will their estimate of β_1 be an unbiased estimate of the effect of A vs B? Explain why or why not.

   (c) *(points: 6)* The company decides to hire you to investigate this question. How would you design a study that would best estimate the effect of A vs B? Explain each step you would take, and why it would work.

3. A research team analyzes a data set and finds the following. Variable X takes on each integer value between 0 and 10, inclusive. (That is, it takes a value of 0, 1, 2, 3, ... , 9, 10.) There are 8 observations for each of these values (that is, 88 total observations). For half of observations where $X = c$, $Y = c$; for the other half, $Y = c + 1$. For example, when $X = 3$, there are four observations where $Y = 3$ and four observations where $Y = 4$. The researchers all want to estimate β_1 in the model $Y = \beta_0 + \beta_1 X + u$.

   (a) *(points: 6)* Researcher A only uses 24 observations: those where $X = 4$, $X = 5$, or $X = 6$. Calculate SST_x, SST_{xy}, and the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ researcher A will estimate.

   (b) *(points: 6)* Researcher B also only uses 24 observations: those where $X = 0$, $X = 5$, or $X = 10$. Calculate SST_x that researcher B will estimate. In order to not waste your time, I will not have you calculate SST_{xy}, $\hat{\beta}_0$, or $\hat{\beta}_1$. However, I hope you can see that $\hat{\beta}_0$ and $\hat{\beta}_1$ are the same for researchers A and B. If not, come to office hours and I will explain!

   (c) *(points: 6)* Calculate the estimated standard error of $\hat{\beta}_1$ for each researcher. [Hint: you may use the fact that both researchers estimate the same value for $\hat{\beta}_0$ and $\hat{\beta}_1$.]

Computer Based Problems

Notes:

- If you are using Citrix, remember to save your code and log file on your home computer, not the remote computer!
- Remember to include the Stata code as part of your writeup. (See "Notes" at the start of this assignment.)

4. **Google Trends**. For this problem, you will analyze data set `eco375hw1data2023.dta` using Stata. This data is gathered from Google Trends; see https://trends.google.com/. The data show the frequency of searches for various terms in provinces and territories of Canada over the past five years.[1] Variable `location` presents the province or territory. All other variables present the search interest in the term in the variable's title. You can refer to this value as the "normalized search interest." For this problem, you can pretend that non-economists understand what this means—that is, a non-economist would understand what it means if "normalized search interest in the word 'warm' increasing by 1 unit."

   (a) *(points: 4)* Present a table showing the mean and standard deviation of normalized search interest in each city name in the data set.

   (b) *(points: 3)* Use OLS to estimate β_1 in:

   $$\text{rain} = \beta_0 + \beta_1 \text{umbrella} + u$$

   Report the estimate $\hat{\beta}_1$ in words a non-economist could understand. Do *not* use causal language. Should we think of this as an estimate of the causal effect of searches for umbrellas on searches for rain? Why or why not?

   (c) *(points: 4)* Create a new variable, `aboveavgcold`, that is equal to 1 if the location has normalized search interest above that variable's average, and 0 otherwise. Use a regression to estimate the difference between normalized search interest for pizza when normalized search interest for cold is above or below average. Is this difference significantly different from 0 at the 5% level?

   (d) *(points: 4)* Use OLS to estimate β_1 in each of the following equations. Report your estimate for each using words a non-economist could understand. Use causal language for this part (even though the results may not actually be causal).

   $$\begin{aligned} \text{poutine} &= \beta_0 + \beta_1 \ln(\text{montreal}) + u \\ \ln(\text{poutine}) &= \beta_0 + \beta_1 \text{montreal} + u \\ \ln(\text{poutine}) &= \beta_0 + \beta_1 \ln(\text{montreal}) + u \end{aligned}$$

   (e) *(points: 2)* Using the data, briefly present and describe one fact that you think is interesting. Your answer to this question must actually use Stata code and the data to find a fact that was not discussed in a previous question.

[1] Google describes the numbers in this way: "Numbers represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term."

5. **Monte Carlo Simulation**. Consider the following model:

$$Y = \beta_0 + \beta_1 X + U$$

where

$$\beta_0 = 2$$

$$\beta_1 = 3$$

$$X \sim N(-1, 1)$$

and

$$U \sim N(0, 1),$$

and X and U are independent.

(a) *(points: 6)* Simulate the model once in Stata. Generate a data set $\{y_i, x_i : i = 1, ..., n\}$ with $n = 400$ observations. Show a scatterplot of X and Y. Use OLS to estimate β_0 and β_1; what are your estimates?

(b) *(points: 6)* Now, simulate this model 200 times in Stata. For each simulation, generate a data set $\{y_i, x_i : i = 1, ..., n\}$ with $n = 200$ observations. Then, for each sample, estimate β_0 and β_1 using OLS and save the results $\hat{\beta}_0$ and $\hat{\beta}_1$.

What are the averages and the standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$ across the simulations? *(Note: most likely, the averages you calculate here are closer to the truth than the estimate from one regression in the previous part. This is related to the Law of Large Numbers, which we will cover later in the course.)* Are $\hat{\beta}_0$ and $\hat{\beta}_1$ close to the true β_0 and β_1? Why should we expect that result? Plot the histograms of $\hat{\beta}_0$ and $\hat{\beta}_1$.

(c) *(points: 6)* With a bit of algebra, we can see that $X = -\frac{2}{3} - \frac{1}{3}Y - \frac{1}{3}U$. Now, again simulate the same model as above ($Y = \beta_0 + \beta_1 X + U$), 200 times, and for each simulation, generate a data set $\{y_i, x_i : i = 1, ..., n\}$ with $n = 400$ observations. This time, though, you should use OLS to estimate γ_0 and γ_1 in the regression $X = \gamma_0 + \gamma_1 Y + V$. (That is, regression X against Y rather than the other way around.) What are the averages and standard deviations of $\hat{\gamma}_0$ and $\hat{\gamma}_1$, the OLS estimates? Are $\hat{\gamma}_0$ and $\hat{\gamma}_1$ usually close to $-\frac{2}{3}$ and $-\frac{1}{3}$? Why should we expect that result?