ECO 375–Homework 2
University of Toronto Mississauga
Due: 14 November, 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. Consider data drawn from the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U \tag{1}$$

where X_1, X_2, and U are each iid. Assume that MLR.4 is satisfied. Your goal is to estimate β_1.

(a) *(points: 6)* Suppose you use OLS to estimate

$$Y = \beta_0 + \beta_1 X_1 + U \tag{2}$$

Write the asymptotic mean squared error (MSE) of the estimator $\hat{\beta}_1$ you would get from using OLS to estimate Equation 2. Write your answer in terms of β_1, β_2, and:

- $V_1 \equiv V[X_1] > 0$ and $V_2 \equiv V[X_2] > 0$
- $\sigma^2 \equiv V[U] > 0$
- $r \equiv corr[X_1, X_2] \neq 0$
- n is the sample size

(b) *(points: 6)* Write the asymptotic MSE of the estimator $\hat{\beta}_1$ you would get from using OLS to estimate Equation 1. Write your answer in terms of β_1, β_2, and the variables defined in the previous question.

(c) *(points: 8)* Suppose we have 10 observations, and $\beta_1 = \beta_2 = V_1 = V_2 = \sigma^2 = 1$. If the researcher wants to minimize MSE, for which values of r should the researcher use the short regression, for which the long regression, and for which would they have the same MSE?

2. A retailer wants to sell laptop computers. To learn about the market, they gather data on 90 computers and estimate the following equation with OLS:

$$price = \beta_0 + \beta_1 screen + \beta_2 weight + \beta_3 screen \times weight + u$$

where *price* is the price of the computer in dollars, *screen* is the screen size in inches, and *weight* is the weight in pounds. Here are their estimates (with standard errors in parentheses):

$$\begin{aligned} \hat{\beta}_0 &= 600 \ (100) \\ \hat{\beta}_1 &= 20 \ (4) \\ \hat{\beta}_2 &= 50 \ (20) \\ \hat{\beta}_3 &= -3 \ (1) \end{aligned}$$

Further assume that we estimate $cov(\hat{\beta}_i, \hat{\beta}_j) = .1$ for all $i \neq j$. The R^2 from this regression is .3.

(a) *(points: 6)* What is the point estimate for the difference in price based on an additional inch of screen size for a 4-pound computer? What about a 5-pound computer?

(b) *(points: 6)* Test the null hypothesis that the association between screen size and price is not related to a computer's weight at the 5% level. Do not use an asymptotic assumption here. Clearly state the test statistic and the critical value; you should look up critical values in the back of the textbook (which is available in MindTap). If the critical value you are interested in is not in the table, choose a critical value with degrees of freedom as close as possible to the one you are interested in, but note the correct degrees of freedom and the degrees of freedom you have chosen. (For example, you might say: "I wanted to find a critical value for a chi-squared distribution with 31 degrees of freedom, but I chose one for a chi-squared with 30 degrees of freedom.")

(c) *(points: 8)* The retailer believes that an additional inch of screen size for a 4-pound computer is associated with a \$5 higher price. State this null hypothesis in terms of the β's, then calculate the test statistic and 5% critical value (following the instructions for finding critical values above).

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.)

3. **Data from the US**. For this problem, you will analyze data set `eco375hw2data2023` using Stata.[1] All questions below refer to this sample. In that data set, each observation describes data about a person who lived in the United States in 2023. Variables in this data set are:

Variable	Definition
age	Age in years
sex	Sex (1=male, 2=female)
region	Region of the US; 1=Northeast, 2=Midwest, 3=South, 4=West
indly	Industry worked in; see https://cps.ipums.org/cps/codes/ind_2020_codes.shtml
incwage	Labor income, in dollars
own	Whether household owns home (as opposed to renting); 1=yes, 0=no
uhrsworkly	Usual hours worked per week

Do NOT use heteroskedasticity-robust standard errors for this problem.

(a) *(points: 5)* Estimate the following model using OLS:

$$\ln(wage) = \beta_0 + \beta_1 age + u \tag{3}$$

State the estimated $\hat{\beta}_1$ using words a non-economist could understand; do not use causal language. Answer twice: once using an approximation, and once using an exact answer.

(b) *(points: 4)* What is the t-statistic when testing whether β_1=.02 in Equation 3?

(c) *(points: 5)* Now, estimate the following model using OLS:

$$\ln(wage) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 age^3 + u \tag{4}$$

Using an approximation, based on these results, how much does income increase or decrease with age when someone is 25? What about when they are 65?

(d) *(points: 4)* Test whether log income varies non-linearly with age in Equation 4. Can you reject the null of linearity at the 5% level?

[1] This data set is based on data is used with permission from Flood et al. (2023). You may not redistribute this data. However, if you wish, you may (and are encouraged to!) download data on your own from https://cps.ipums.org/. Full citation:
Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Megan Schouweiler and Michael Westberry. IPUMS CPS: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2023. https://doi.org/10.18128/D030.V11.0.

(e) *(points: 6)* Now, estimate the following model using OLS:

$$\ln(wage) = \beta_0 + \beta_1 age + \beta_2 own + \beta_3 age \times own + u \tag{5}$$

Should estimated β's be interpreted causally? Why or why not? Refer to the assumption(s) we discussed in class. Then, state the estimated $\hat{\beta}_3$ using words a non-economist could understand; use an approximation, and use causal language if and only if the estimates should be interpreted causally.

(f) *(points: 5)* How much faster or slower do wages rise with age for people in the construction industry versus those in all other industries? (Use an approximation here.)

(g) *(points: 5)* By how many standard deviations is log income higher for men than for women in this data? (You must use a standardized variable in Stata for this problem.)

(h) *(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.

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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$$

where

$$\begin{aligned} \beta_0 &= 3, \\ \beta_1 &= 4, \\ \beta_2 &= 2, \\ X_1 &\sim N(0,1) \\ X_2 &= X_1 + V, \text{ where } V \sim N(0,4), \\ U &\sim Uniform(-6,6), \end{aligned}$$

and X_1, V, and U are independent. [Hint 1: a $N(0,4)$ random variable has standard deviation of 2. In Stata, you can generate a normal random variable with mean 0 and standard deviation of 2 using the command `rnormal(0,2)`.] [Hint 2: $Uniform(-6,6)$ indicates a uniform distribution between -6 and 6. In Stata, you can generate a uniform distribution between A and B using the command `runiform(A,B)`.]

Simulate this model 10,000 times in Stata. [Hint: get your code working with fewer simulations, so you can correct errors quickly. Once your code is ready, change it to simulate 10,000 times.] For each simulation, generate a data set with $n = 100$ observations. Then, for each sample, estimate β_0, β_1, and β_2 with OLS, calling your estimates β_0, β_1, and β_2. Create two sets of estimates for each simulation: one based on the first 5 observations in the data set, and one based on the full 100 observations.

Do NOT use heteroskedasticity-robust standard errors for this problem.

(a) *(points: 8)* Define a new variable $\gamma \equiv \beta_1 \times \beta_2$. Estimate γ using $\hat{\gamma} = \hat{\beta}_1 \times \hat{\beta}_2$ within each simulation, for each set of estimates. Create a table like the following with average estimates of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\gamma}$ based on 5 observations or 100 observations.

	5 obs	**100 obs**
$\hat{\beta}_1$		
$\hat{\beta}_2$		
$\hat{\gamma}$		

(b) *(points: 5)* Which of the values in this table are close to the true values of β_1, β_2, and γ? For each value (whether close or not), based on what we've learned in class, do we have a good reason to expect that the estimate would be close to the truth? If so, what is that reason?

(c) *(points: 6)* Now, for each regression, test $H_0 : \beta_1 = 4$ against $H_1 : \beta_1 \neq 4$. Conduct this test in two ways: first with a small-sample t-test, and second with the asymptotic t-test. For each test, how often do we reject the null at a 5% significance level? Create tables like the following with the critical values for the tests and the fraction of observations in which we reject the null hypothesis.

Critical values

	5 obs	**100 obs**
Small sample		
Asymptotic		

Fraction of rejections

	5 obs	**100 obs**
Small sample		
Asymptotic		

(d) *(points: 5)* Which of the values in this table are close to the 5% that we hope for? (That is, do the tests have the correct size?) For each value (whether close or not), based on what we've learned in class, do we have a good reason to expect that the estimate would be close to that value? If so, what is that reason?