Barreto
BUS 305: Regression with Microdata
11 April 2025 Name: _____

Exam 2 SUGGESTED ANSWERS

10 questions; each worth 10 points. 6 minutes per question on average. Watch the time.

Open the Excel workbook called Exam2S25.xlsm, enable macros, then Save As to your folder in the I drive.

Install the MCSim add-in: alt, t, i . . .

I put the LMS and OLSReg functions in the Excel file so you must ENABLE MACROS for them to work.

I believe everything in that workbook is correct. If you find a mistake, let me know—you'll get +1. :-)

Maybe one day the exam will be this one question that reveals what you have mastered: "Tell me everything you know about what is in that Excel file." But today is not that day.

Do NOT open any other workbooks or files or look in your book. This is a CLOSED book exam. Do NOT communicate with anyone during the exam. Any violation of these rules is cheating and will result in an immediate F.

The *Dead* sheet is a deadened version of the Xs in the *Live* sheet. Each sim sheet is named for the question number that it is associated with. I am saving you time running sims, but you can run your own if you wish.

1) Sheet Q1 has a sim of cell H3, the sample slope of a regression of Y on X1. The sim results show that the OLS sample slope estimator is performing well. Explain exactly what in the sim results provide evidence of good performance and please include specific numbers in your answer.

The avg of 100K b1s (sample slopes) is 199.959 which is very close to 200, the true parameter value. The fact that the center (technically, it is the expected value) of the sampling distribution of b1 equals the parameter value means it is unbiased and this is what we mean by "good performance" in this context.

2) It might be surprising to someone that the sample slope in cell H3 is performing well since that model is misspecified—X2 is omitted from the regression of Y on X1. Why is omitting X2 not leading to omitted variable bias here? In your answer, define omitted variable bias.

Omitted variable bias occurs when the expected value of the sampling distribution is not equal to the parameter value. There is no OVB in this case because the average of the sim of b1s is very close to 200, which is beta1. This is happening because beta2=0. This means X2 has no effect on Y so it does not bias the b1 estimator. Recall the Omitted Variable Bias rule → g1 = b1 + d1b2. If that last term is zero, g1=b1 and there's no OVB.

This would almost never happen in any real-world application.

3) Let's turn our attention to the estimated SE of b1 in the Y on X1 regression. Sheet Q3 has a sim of both cells H3 and H4 to see how the estimated SE of b1 is performing. The sim shows trouble. Explain what is wrong with the estimated SE of b1 and please include specific numbers in your answer.

The avg of 100K estimated SE of b1 values is around 30, but that is about 8% lower than the Approximate SE of b1 of 32.5. This means the estimated SE of b1 is a biased estimator of its exact SE value (which we don't know, but we are using the Approx SE as a stand in). What is wrong is that the estimated SE is not centered on the Exact SE.

4) You just showed that there is something wrong with the estimated SE of b1. Then why do we routinely use the estimated SE of b1 in many applications of regression analysis?

Because we know that although it is biased, the estimated SE of b1 is a CONSISTENT estimator of the exact SE of b1. In other words, as n increases, the bias gets smaller and smaller. This is why we use it.

5) What is the regression of X1 on X2 useful for and is the R2 value in cell C23 (and also in D18) good or bad? Explain.

That auxiliary regression is used to get the R^2 value in the formula for the exact and estimated SE of b1. The value of 0.1 is good because it is close to zero. The best would be if it was zero because then sqrt(1-0) is 1 and it would not help inflate in the SE. Sqrt(1-0.1) is about 0.95 which is pretty close to 1 so it does increase the SE by 1/0.95 or roughly 6%, but that seems "good" in the sense that it could be much worse.

6) Sheet Q6 has a sim of a race between the OLS and LMS estimator of b1. What is the difference between the OLS and LMS estimators? Please be sure to explain the recipe for these two estimators.

OLS minimizes the SSR, the SUM of squared residuals (which is the same as minimizing the AVERAGE of the squared residuals) while LMS minimizes the MEDIAN of the squared residuals. This means that a single wildly big residual, called an outlier, has much less influence on the fitted LMS line than the OLS line because the median is insensitive to how much bigger than the median a value is, while the average will be pulled up or down by large positive or negative outliers.

7) In sheet Q6, OLS wins. Is this victory a demonstration of the Gauss-Markov Theorem? In your answer, please define and explain the Gauss-Markov Theorem.

The Gauss-Markov Theorem says that when a well-behaved DGP (like measurement error or the classical econometric model with iid errors) is obeyed, then OLS is BLUE. This means that OLS is the best, minimum SE, estimator of all of the linear and unbiased estimators.

However, the Gauss-Markov Theorem does NOT apply to this situation because LMS is not a linear estimator. This is because LMS cannot be written as a linear sum of weights. So, even though OLS beats it, the Gauss-Markov Theorem could not be used to know that this would happen. Gauss-Markov has nothing to say about non-linear estimators like LMS. But OLS did crush it because this is not a dirty data situation with outliers.

8) Sheet Q8 has a sim that shows I reject the null in the Whole Model F test 63% of the time at the 5% level of significance. Is this good or bad? Please explain why.

This is good I guess. I mean, we WANT to reject this null because it is a FALSE NULL since the population slopes are not all zero. So, given the ridiculously small n, I would say it is pretty good that it works more than half the time. You can disagree, but the key to this answer is that *we want to reject this false null*.

9) Inject heteroskedasticity by changing alpha to 2 in cell D8. Run a sim that shows the effect of heteroskedasticity and use your sim results to explain the effect of het. Please include specific numbers in your answer.

I changed alpha to 2 and ran this sim of b1 and its estimated SE → It is easy to see that the avg of the 100K estimated SEs are far from the Approx SE.

H3		Dead!H4		Notes
Average	195.765	Average	1268.008	-33% miss by est SE
SD	1883.4713	SD	735.0666	
Max	8494.481	Max	5960.981	
Min	-8021.417	Min	17.499	

Even with 100K, however, it is hard to see that b1 remains centered on the true beta1 of 200 because 195.765 might make you conclude that there is bias. But there isn't and this deviation is caused by the tremendous variability in b1 of roughly 1883. Het causes bias in the estimated SE, not the slope coefficient estimate.

10) Use the Live sheet to run a sim of b1 (cell H3). Compare your results to the Q1 sheet results. Why is the Approximate SE of b1 (and, therefore, the exact SE of b1) bigger for the Live sheet than the Dead sheet?

Because we changed the DGP. We no longer have "Xs fixed in repeated sampling" because we made the Xs random variables. Their bounce will be passed through to all sample statistics, including the sample slope, and make them more bouncy. This is why the Approx SE jumps from 32.5 to 44. In other words, the usual formula for the exact and estimated SE of b1 no longer apply.

Summary Statistics	
Average	199.799
SD	43.9511
Max	1046.936
Min	-1181.183

Save your workbook now and turn in your exam. You're all done!