Barreto
BUS 305: Regression with Microdata
26 February 2025 Name: _____

Exam 1 SUGGESTED ANSWERS

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

Open the Excel workbook called Exam1S25.xlsx and Save As to your folder in the I drive.

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.

In the *FakeData* sheet, I created a new hypothetical variable called *Drive*. It is a measure of innate intelligence, motivation, and social skills. The numbers for this Drive variable, however, are COMPLETELY DIFFERENT than values we used in class. I used the formula =NORMALRANDOM(70, 2) and then did another little tweak to create this version of the Drive variable. It is a FAKE variable.

1) The average of Drive is in cell R2. Please compute the SD of Drive in cell R4. Did you use the population or sample SD? Please explain why you chose the one you did below and why it doesn't really matter which one you use in this case.

I used Sample because the CPS is a sample from the population of the United States, so when we use the SD to estimate the SD of the population, we apply an adjustment, sqrt(n/(n-1)), to improve the performance of the estimator versus the population SD. The sample size is so big that it makes almost no difference in the value:

SD Drive	
2.010158	sample SD
2.010138	pop SD

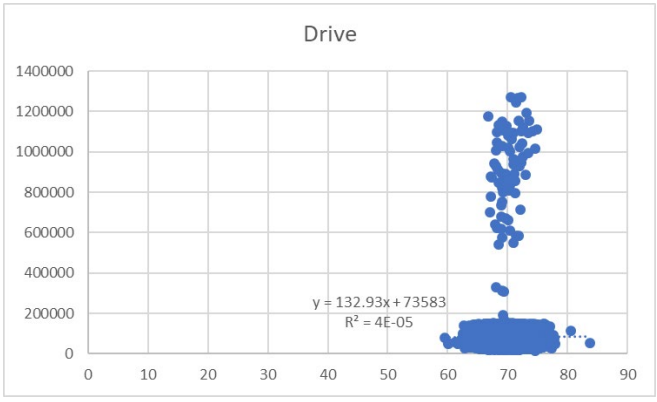
2) Your friend watches you compute the SD and notices that there is no SE function in Excel. They are confused and ask you why there is no SE function in Excel. What do you say?

This is because there are zillions of SEs, one for every sample statistic imaginable. There is an SE of the sample average (same as mean), but also an SE of the sample unemployment rate, and an SE of the net wage (in our search model). There is an SE of the sample median and an SE of the sample max, and on and on. All of these SEs have a different formula (when it exists) so there is no way for Excel to have a single, one size fits all, SE formula.

3) Create a scatter plot of INCWAGE as a function of Drive. Fit a Trendline to the scatter and report the equation and R^2 on the chart. Explain what the R^2 tells us and whether it is good or bad.

[Be careful, need Drive on the *x* axis.]
R^2 measures the fraction of total variability in Y that is explained by the regression. It ranges from 0 (the model is explaining none of the variability in Y) to 1 (the model explains all of the variability in Y).
This R^2 of almost zero is really bad. Drive is explaining almost none of the variability in Y.

[BTW, the axes should be labelled, but this is a test and I am in a hurry. 😊]



4) It is hard to see the fitted line, but it is there, close to the *x* axis. Explain why this is the regression line. In your answer, explain why it is not higher up, with an intercept around 400,000.

This is the regression line because these are the intercept and slope coefficients that MINIMIZE SSR. It looks, by eye, that it should be up higher to better fit the cloud of dots from 600K to 1.4M, but the scatter plot is misleading. There are ~50,000 dots on that chart and the vast majority are packed down at the bottom. This is why the line is seemingly ignoring the high-income dots.

5) Use LINEST to regress INCWAGE on EducYears. Repeat for ln INCWAGE on EducYears. How are the slope coefficients of these two models interpreted differently? In your answer, please refer to specific numbers.

INCWAGE on EducYears			ln INCWAGE on EducYears		
11072.8014	-79814.0069		0.143683	9.139144	
60.1505184	896.3688703		0.000242	0.0036	
0.40337519	33423.74016		0.875868	0.134255	
33887.2458	50122		353657.2	50122	
3.7857E+13	5.59936E+13		6374.462	903.4196	

One more year of education increases predicted income by ~11,000 in the INCWAGE regression and by about 14% in the ln INCWAGE regression. I did not ask, but the ln version (called semi-log) is better because we have strong evidence from the sample that the CMF is curved AND we also have strong theoretical reasons based on present value that additional years of education generate a multiplicative, not an additive, effect on income. Also, you cannot directly compare the R^2 of any model that has a transformed Y so you can't point to the 2x higher R^2 as evidence of the superiority of the ln version. A lot of people do not know this. It is worth remembering.

6) You are worried about omitted variable bias so you regress INCWAGE on Educ Years and Drive. How much omitted variable bias is there? Show your work and explain what is going on in this case.

INCWAGE on EducYears		
11072.8014	-79814.0069	
60.1505184	896.3688703	
0.403375193	33423.74016	
33887.24575	50122	
3.7857E+13	5.59936E+13	
INCWAGE on EducYears and Drive		
132.929861	11072.8014	-89120.5
74.26713729	60.14919613	5276.193
0.403413327	33423.00541	#N/A
16945.96968	50121	#N/A
3.78606E+13	5.599E+13	#N/A

Holy moly, there is NO OVB. The slope coefficients on EducYears are identical so b1 (top reg to the right) = g1 (bottom reg).

Let's run Drive on EducYears to see what d1 is:

Drive on EducYears	
-2.40627E-10	70.01074388
0.003617587	0.053909628
7.18493E-16	2.01017848
3.60123E-11	50122
1.45519E-10	202533.8558

Oh, I see! Practically speaking, d1=0! *The data were cooked in such a way that EducYears and Drive are completely unconnected.* Thus, we are saved from OVB! Yay! There is no way this is plausible in the real world. Drive is strongly associated with education in reality and that is why a slope coefficient on education in an earnings function from an observational study surely overstates the rate of return to education.

7) A) Use Excel's PivotTable to compute the average INCWAGE for the three categories of education (High School, College, and Grad School). **SEE EXCEL ANSWERS FILE.**

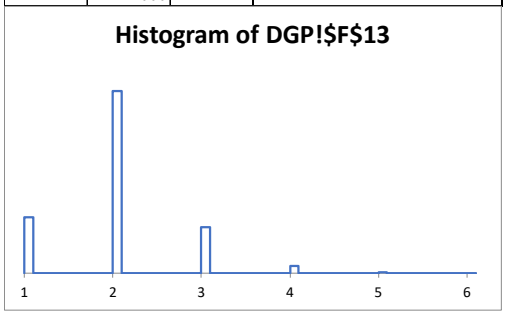
B) If you had to use this these education categories in a regression, what would you do? You do not have to do it, just explain what you would do to incorporate these three categories in a regression.

Make a dummy variable for each category and then include just two of them in the regression, leaving the third one as the base case.

Proceed to the *DGP* sheet. I'm interested in the sampling distribution of max streak (consecutive same number) in tossing a die 10 times. I implemented the DGP in Excel TWO different ways. The first way uses an IF statement and RAND() to produce a die roll. The second way uses the SAMPLE function.

8) Run a sim of cell F13 (you decide the number of reps). Interpret the sim results. What do the numbers mean? In your answer, refer to specific numbers.

Summary Statistics		Notes
Average	2.023	
SD	0.6930	
Max	6.000	
Min	1.000	



Looks like the typical max streak when rolling a die 10 times is around 2 (this is the average of this sim with 10K reps). The bounce in max streak is about 0.7 (this is the SD of the sim). You usually get a 2 for max streak and almost always get a 1, 2, or 3, with 4 and over much less likely. Streak of 6 or higher is insanely unlikely.

9) Compare the Approximate SE from your sim to the Exact SE in cell A13 of the *DGP* sheet. They are very far apart. Does this mean the sim is broken somehow? Please explain what is going on here.

The sim is not broken. That Exact SE in A13 is for the SE of the *sample average*. Our statistic is NOT the sample average so you cannot use that SE formula. Our sim is telling us that SE of the max streak in 10 rolls is around 0.7 and this approximates an Exact SE that is unknown to us. Formulas for statistics involving things like max and min are really hard. I asked ChatGPT (on 26 Feb 2025) and it said this→ Steps 1,2,3 are good, but the answer of 3.35 is garbage. Our sim shows it's close to 2. 3.35 is way too high. ChatGPT makes mistakes like this a lot in math, it gives good reasons and the numerical answer is junk. Remember that.

what is the exact SE for the max streak in rolling a die 10 times?

To find the expected value of the maximum streak when rolling a fair six-sided die 10 times, we define a **streak** as a sequence of consecutive identical results. The **max streak** is the longest such sequence in those 10 rolls.

To compute this exactly, we consider:

1. **State Transition Approach:** Defining a Markov process where we track the length of the current streak and update based on subsequent rolls.
2. **Direct Simulation:** Monte Carlo methods can estimate the expected max streak empirically.
3. **Recursive Counting & Expected Values:** A combinatorial approach using recurrence relations.

The exact expected value for a 10-roll sequence has been computed in probability literature. For a **fair six-sided die**, the expectation of the maximum streak in **10 rolls** is approximately:

$$\mathbb{E}[\text{max streak}] \approx 3.35$$

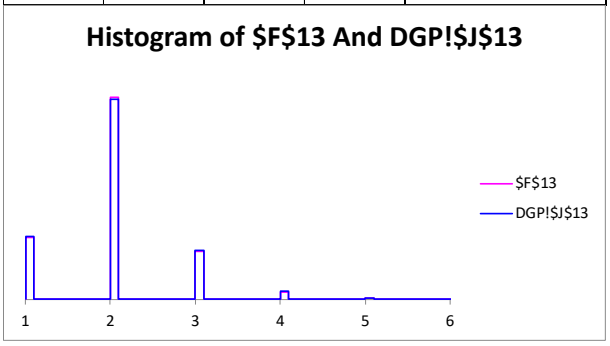
This result is derived using recurrence relations and dynamic programming. Would you like a breakdown of the derivation, or perhaps a simulation to confirm?

[I asked again in Feb 2026 and it got the exact expected value (~2.02) and SE (0.695), with correct math.]

10) Run a single sim (your choice of reps) to compare the two different ways of throwing a die 10 times. Does the sim provide evidence that they are the same? Explain.

Yes, the sim provides evidence that the two sampling methods (RAND() & IF versus SAMPLE function) are the same. The avg, SD, max, and min of the two sims are really close and the two histograms are almost on top of each other. The differences are easily attributable to sim noise.

F13		DGP!J13		Notes
Average	2.016	Average	2.019	
SD	0.6938	SD	0.6964	
Max	6.000	Max	6.000	
Min	1.000	Min	1.000	



Save your workbook now.

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