Final Exam

Stats 1 2023 w/ Jim Bisbee

Due 11:59p on Sunday, December 10, 2023

Secret Santa (10 Points)

There are 14 employees at a small video game developer where my buddy Pete works (MDHR, creators of Cuphead). His boss Jared is preparing to organize a Secret Santa gift exchange for their holiday party. However, the boss – while a total sweetheart of a guy – is kinda bonkers, and has come up with a weird, not-totally-fair game: first, they will randomly select 7 people to participate, and then each of the 7 selected will be randomly assigned one of the other six participants to give a gift to in secret.

1. (2 pts) How many different sets of 7 participants are possible?
2. (2 pts) What is the probability that the boss is selected to participate?
3. (2 pts) Conditional on both being selected to participate, what is the probability that Pete's recipient is his boss Jared?
4. (4 pts) Define event A to be: Jared is one of the 7 selected to participate, and event B to be: Pete is one of the 7 selected to participate. Are events A and B mutually exclusive? Are events A and B independent?

The Joys of Coding Event Data (20 Points)

Suppose your colleague solicits your help to assemble a new event dataset about violence during protests. This colleague says that she thinks the average number of violent incidents during a protest is about 25, and so whenever you find a protest with more than 25 violent events, you are to code it as violent. You begin researching protests at random and encounter the following number of violent events in the first twelve protests: 26, 4, 0, 22, 7, 2, 6, 12, 4, 29, 0, 0. You are skeptical that her claim of 25 is correct, though you acknowledge that you've only looked into 12 protests.

5. (10 pts) If you want to be 95% confident that your colleague's hypothesized mean number of violent encounters is too large before you approach her, would your 12 data points be sufficient to convince you to approach her? Carefully explain how you came to this conclusion.

6. (5 pts) If yes, how large would a 13th data point have to be in order to undermine your confidence? If no, how small would a 13th data point have to be in order to give you sufficient confidence?

7. (5 pts) What if it turned out that you started with these 12 protests because they were listed in a news article titled "Protests in Places that are Almost Never Violent." How would that change how confident you were in approaching your colleague? Explain.

Really Simple Matrix Fun (10 points)

Consider the following (admittedly simple) example. A DGP defined by the population model $y = \beta_0 + \beta_1 x + u$ gives rise to the following dataset of 4 observations:

$$\mathbf{y} = \begin{bmatrix} 4 \\ 2 \\ -6 \\ 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{X} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & -1 \\ 1 & 4 \end{bmatrix},$$

where the second column of $\mathbf{X}$ is composed of observations of the variable x. In answering the following questions, be sure to show all your work.

8. (5 points) Show that the OLS estimates $\widehat{\beta}_0 \approx -2.89$ and $\widehat{\beta}_1 \approx 1.57$. You will be glad to know that

$$(\mathbf{X}'\mathbf{X})^{-1} \approx \begin{bmatrix} .536 & -.143 \\ -.143 & .071 \end{bmatrix}.$$

9. (3 points) Show that $\widehat{\sigma} \equiv SEE \approx 3.33$.

10. (2 points) Show that $R^2 \approx .61$.

Bias and Efficiency (10 points)

Consider four random variables W, X, Y and Z, where

$$cov(W,Y) > 0; \quad cov(W,X) = 0;$$
$$cov(Z,Y) = 0; \quad cov(Z,X) < 0,$$
$$\text{and } cov(X,Y) \text{ is unknown.}$$

Say whether the following statements are TRUE or FALSE, and explain why. Assume we have a large number of observations of the joint distribution of all four variables from an i.i.d. random sample.

11. If we estimate the equation

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i,$$

$\widehat{\beta}_1$ is a *biased* estimate of the parameter β_1 due to the omission of w and z.

12. The estimate of the parameter β_1 we obtain from the estimated equation

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

will be *more efficient* than the estimate of the parameter β_1 obtained from the equation

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{\beta}_2 z_i.$$

13. The estimate of the parameter β_1 we obtain from the estimated equation

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

will be *more efficient* than the estimate of β_1 obtained from the equation

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{\beta}_2 w_i.$$

Interpreting a Regression (50 Points)

Imagine an alternate universe in which the South took the lead in public policy to combat COVID, implementing a vaccine mandate at the state level. Specifically, imagine that Tennessee, Kentucky, Georgia and Alabama all implemented a state-wide vaccine mandate, though allowed counties to set the level of the fine for individuals failing to get vaccinated. Imagine also that we have data at the county level in these four states on the number of active COVID-19 cases per 100,000 residents in the year following the mandate, the size of the fine that the county set in dollars, and we also know which of the four states the counties reside in. Suppose we run the regression:

$$CASES_i = \alpha + \beta FINE_i + \delta_1 KY_i + \delta_2 GA_i + \delta_3 AL_i + \epsilon_i$$

14. (3 pts) Why are dummy variables included for only three of the four states for which we have county-level data?

15. (4 pts) What is $\frac{\partial CASES}{\partial FINE}$? Why would we calculate this?

16. (5 pts) Suppose we estimate $\delta_1 = 1100$, $\delta_2 = -150$, and $\delta_3 = 3400$. What would each of these estimates tell us?

17. (2 pts) Suppose β is estimated to be -10.5. What would this tell us about the relationship between the size of the fine and the number of COVID cases?

18. (4 pts) Knowing that α is also estimated to be 5200, what would the expected number of COVID cases per 100,000 people be in a Tennessee county that implemented a $150 fine? What about for such a county in Kentucky?

19. (3 pts) Suppose the 95% confidence interval for β is $(-10.7, -10.3)$. What would this mean for your conclusion about the relationship between the size of the fine and the number of COVID cases?

20. (3 pts) If you were to plot the relationship between $FINE$ and $CASES$ for each of the four states, would the lines be parallel or not? How do you know?

21. (5 pts) If we reran the regression with the same data but changed the states we included to be:

$$CASES_i = \alpha + \beta FINE_i + \delta_1 KY_i + \delta_2 GA_i + \delta_3 TN_i + \epsilon_i$$

what about our conclusions and interpretation would change?

22. (5 pts) Suppose we think that the average income in the county interacts with the size of the fine, so that fines reduce COVID cases in general but are less effective at doing so in wealthy counties. We collect county-level income data to add to our dataset and run the following regression:

$$CASES_i = \alpha + \beta_1 FINE_i + \beta_2 INC_i + \beta_3 FINE_i * INC_i + \delta_1 KY_i + \delta_2 GA_i + \delta_3 AL_i + \epsilon_i$$

Now, to examine the relationship between the size of the fine and the number of cases, what information will we need to look at?

23. (3 pts) How would we interpret β_2 from this model?

24. (3 pts) How would we interpret α from this model?

25. (5 pts) How would we use the results to calculate the expected difference in cases between Georgia and Kentucky?

26. (5 pts) What is an example of a set of regression coefficients that we could find that, if statistically significant, support our hypothesis about the interacted relationship between the size of the fine, income, and the number of cases? (Make up one (of many!) set of numbers for α, β_1, β_2, β_3, δ_1, δ_2, δ_3 that would be consistent with the hypothesis).

Extra Credit

EC1) (2 pts) Suppose I run the first regression above ($CASES_i = \alpha + \beta FINE_i + \delta_1 KY_i + \delta_2 GA_i + \delta_3 AL_i + \epsilon_i$), but in actuality the true data generating process is

$$CASES_i = \alpha + \beta FINE_i^2 + \delta_1 KY_i + \delta_2 GA_i + \delta_3 AL_i + \epsilon_i$$

where

$$\epsilon_i \sim N(0, 2 * FINE_i).$$

Which assumption(s) about the data generating process that are required for inference would be violated?

EC2) (4 pts) Use the methods of generating simulated data to show how you could detect these violated assumptions with the diagnostic plots available when you plot a regression output. Specifically, make a dataset where your dependent variable $CASES$ is a function of made up variables $FINE$, KY, GA, AL, TN, and true coefficients, say, $\alpha = 5000$, $\beta = 5$, $\delta_1 = 1200$, $\delta_2 = -170$, and $\delta_3 = 2400$. Generate $CASES$ according the to the true data generating process in EC1, using errors generated according to the error equation in EC1. Then run the regression as if you thought the assumptions of the linear model were met and as if you thought the true data generating process were perfectly described by the regression equation from 22-28. Does your regression return the true coefficients on average? How can you detect that there is a problem using the regression diagnostic plots?