Introduction to Statistical Methods

Spring 2021 | Exam 3

Total Score: 25 Points

1. The past twenty years of the US labor market witnessed a dramatic growth of programming-intensive jobs—jobs that involve tasks related to coding, computer programming, and dealing with large data sets. A team of researchers constructed an score to measure the level of programming-intensity of jobs.

   (a) (3pts) Let Y be the *programming intensity score* of an individual's job. Evidence suggests that men's jobs tend to have higher programming intensity score than women's jobs. Now, suppose you want to examine whether this gender gap is larger in 2015 than in 1995. Specify what dummy variable you will need to construct and write down the appropriate regression equation.

   Dummy variables: D_{female} and D_{2015}.

   Regression equation: $Y = \beta_0 + \beta_1 D_{female} + \beta_2 D_{2015} + \beta_3 D_{female} \cdot D_{2015} + \epsilon$

   (b) (4pts) The results from your regression are given in the table below:

Outcome: Programming Intensity Score		
Variable	Coefficient	Sig.
Female	-18.8	***
Year 2015	9.2	***
Female × Year 2015	-3.6	***
Constant	34.5	***

   What is the predicted mean programming intensity score for *women* in 1995 and 2015, respectively?

   $\hat{Y}_{female,1995} = 34.5 - 18.8 = 15.7$

   $\hat{Y}_{female,2015} = 34.5 - 18.8 + 9.2 - 3.6 = 21.3$

   (c) (4pts) What is the predicted *gender gap* in programming intensity score in 1995 and 2015, respectively?

   $\hat{Gap}_{1995} = 18.8$

$\hat{Gap}_{2015} = 18.8 + 3.6 = 22.4$

(-18.8 and -22.4 are also correct answers.)

(d) (2pts) The constant term is equal to the predicted mean of a subgroup of population. Which subgroup is this?

Men in 1995.

(e) (4pts) Now, suppose you are interested in whether the gender gap in programming intensity score differs by education. You will focus on the 2015 data for this question.

Education is measured as a categorical variable that contains three categories, represented by three dummies: (1) high school or less ($\mathbf{D_{hs}}$); (2) some college ($\mathbf{D_{somecol}}$); (3) BA or higher ($\mathbf{D_{BA}}$).

Suppose you want to test whether the gender gap differs between *those with some college* and *those with BA or higher*, write down the regression equation that allows you to test this, and indicate which coefficient in this equation you will focus on. (HINT: Be sure to use the appropriate reference group in your regression.)

Regression equation:

$$Y = \beta_0 + \beta_1 D_{female} + \beta_2 D_{hs} + \beta_3 D_{BA} + \beta_4 D_{female} \cdot D_{hs} + \beta_5 D_{female} \cdot D_{BA} + \epsilon$$

I will focus on β_5, because this coefficient captures the difference in predicted gender gap between *those with some college* and *those with BA or higher*.

An alternative regression equation would be:

$$Y = \beta_0 + \beta_1 D_{female} + \beta_2 D_{hs} + \beta_3 D_{somecol} + \beta_4 D_{female} \cdot D_{hs} + \beta_5 D_{female} \cdot D_{somecol} + \epsilon$$

2. A job candidate comes to the Puck Seminar and describes a study designed to estimate the effect of high quality schooling during childhood on the probability of unemployment as an adult. The job candidate uses evidence from a large-scale survey dataset and presents results showing that attending a high quality school during childhood reduces the probability of adult unemployment substantially. The argument is based on a regression of adult unemployment on a measure of the quality of the child's high school, along with control variables for the child's parents' education, the child's race/ethnicity, the child's parents' age, and the child's age.

(a) (4pts) An audience member raises her hand and argues that the estimated effect of high quality schooling is biased because the candidate hasn't measured the income of the child's family. Assume that the candidate can't do anything about the critique because a measure of family income is not available in the data. If the audience member is right that income is a confounder, how would the candidate's estimate of the effect of high quality schools likely change if he had included family income in the regression equation? Draw a diagram to show why.

Answer: If family income is included as a control, the negative effect of school quality on unemployment will become smaller, because part of the association between school quality and unemployment is due to the fact that children high-income family tend to go to high-quality schools and these children also tend to have low unemployment rate. (The diagram should show FAMILY INCOME as a confounder that positively affects school quality and negatively affects unemployment.)

(b) (4pts) Another audience member raises his hand and asks the following question: "From what I know about the dataset you're using, it was based on a survey conducted in a small number of schools that were selected from different states across the country. Have you considered the state-level confounders that may affect the causal interpretation of your results?" The candidate frowns and says that he did not.

Question: What do you think would be a possible *state-level* confounder? Use 1-2 sentence(s) to justify your answer. Include relevant diagrams if you like.

Answer: e.g. poverty rate, median household income, % college educated, etc.