PA 541 - Midterm - Spring 2021

NAME:\_\_\_\_\_\_\_\_\_\_Alexis Kwan\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

March 8, 2021

## INSTRUCTIONS

The exam will not require any use of R or coding. The exam is **closed-book**. I am using the honor system; I will not use any proctoring software so I trust that you will abide by the rules of the exam.

You are allowed to use a double-sided handwritten sheet of notes (8.5 X 11) for the exam, but no other sources or materials. When you submit your exam, please also submit a copy of your handwritten sheet of notes (maybe the easiest way to do this is to just take a picture on your phone, email the pics, and paste them in at the bottom of your exam). The exam will be a word document. You will simply type your responses directly below the questions on the exam. The exam is designed to be completed in around 2 hours, but you have three hours to complete it. Please turn in your exam by 6pm.

## PART ONE - CONCEPTUAL QUESTIONS

### QUESTION ONE (6 points)

In this class, we discussed randomness and endogeneity as two major challenges that hinder our ability to identify the true causal effect of one variable on another. Provide an intuitive discussion of each of these challenges and an example of when it distorts or biases our estimate of the effect of one variable on another.

### QUESTION TWO (8 points)

For a particular county in the United States *i*, let be the proportion of households that have access to high speed internet. Let be proportion of eligible residents who received the vaccine in that county. Suppose that we have data on these two variables for 300 randomly sampled counties around the country.

1. If you run a simple regression predicting by , do the results provide an unbiased estimate of the effect of the internet on vaccination rates? Why or why not?

A model based on just the proportion of households that have access to high speed internet will be biased because internet access will not likely be the only factor that contributes towards vaccination proportion. Bias is based on the residual term for the model, or the difference between predicted and actual vaccination proportion. If the conditional expectation on the residual is nonzero or if the residual is correlated with another variable not in our model that affects vaccination proportion, then the model will be biased.

1. Suppose you can collect additional background information, such as average income and average education in each county. Would controlling for these factors in your model have a meaningful impact on your estimate of the effect of the internet on the proportion of residents obtaining the vaccine? Explain.

If income and education indeed has an effect on vaccination and is correlated with internet access, the introduction of these factors would affect the estimate of the effect of internet access. Income very likely is correlated with internet access since economically a household requires a certain level of income for internet services to be purchased. Education may not have as strong a correlation with internet access, but there are mechanism rationales that would relate the two, like highly educated people may understand the need for better internet access and therefore value and acquire increased access for themselves or their household. It is also possible that these factors affect vaccination proportion as well for similar reasons. Income would make travel to vaccination sites easier (like being able to have a car) or perhaps allow for the time required for vaccination. Education may motivate people to get vaccinated since it would allow them to reason about and understand the health necessity behind vaccination.

### QUESTION THREE (6 points)

Regression models can be built for prediction and explanation. Describe the difference between these two modeling goals and provide an example of when you would prefer one over the other.

In modeling for prediction we care less about why and how factors explain changes in a dependent variable and more about how well data it produces matches expectations for the future or for effects that have not been observed yet. In modeling for explanation, how much variation is explained by the model and the overall validity of the model matters a lot more. In this case we would like to establish causal links between the independent variables and the dependent one. We would care a lot more about endogeniety in modeling for explanation versus prediction. If I were a stock investor I would likely care more about how stock prices change in the future, so as to maximize profits from selling stock, than why they changed the way they did in the past so prediction would be more important. If I were a program manager and I was testing to see how effective a program was towards changing an outcome, I would prefer to understand what mechanisms were causing the change in outcome and therefore would care about the explanatory power of a model.

### QUESTION FOUR (10 points)

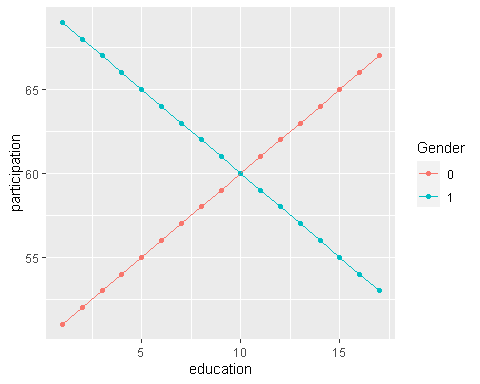
The following model uses several variables to predict a person’s support for President Biden. Assume support is a continuous variable that ranges from -100 to 100. The variable is a dummy variable for political party and takes on a value of 1 if the person is a democrat, and 0 otherwise (yes, I’m lumping republicans and independents together, but for the sake of this question just assume everyone is either a democrat or republican). The variable is the age of the person (in years):

1. What is the null hypothesis we are testing with ?
2. What is the null hypothesis we are testing with ?
3. What do we implicitly assume about the effect of age with regard to political party affiliation?
4. Assume you ran the model that included an interaction between *dem* and *age* and obtained the following values for the coefficients. Sketch a diagram to display the results. (hint, you should have a line for democrats and a line for republicans; please note the coefficient on age is -1. You can make the drawing directly in word using the ‘draw’ tab)

### QUESTION FIVE (6 points)

Suppose that we estimated the following model:

where participation is an index of political participation and then plotted the fitted values for females (*gender* = 1) and males (*gender* = 0) at different levels of education in the plot below. Based on the diagram, answer the following questions. Note: you don’t need to explain anything here, just indicate greater than 0, equal to 0, or less than 0 for each question.



1. Is greater than, equal to or less than zero?
2. Is greater than, equal to or less than zero?
3. Is greater than, equal to or less than zero?

## PART TWO - APPLICATIONS AND INTERPRETATION

### QUESTION SIX (10 points)

A university was concerned that there may be differences in faculty and staff pay based on a variety of demographic attributes, in particular race and gender. The university compiled a random sample of 400 faculty and staff employed at the university. The dependent variable is the average hourly earnings (*wage*). The independent variables are (i) years of education (*educ*), (ii) years with university (*tenure*), (iii) a dummy variable to indicate if the person is white or non-white (*white*, where being white =1), (iv) a dummy variable to indicate if the person is male (*Male*, where being male = 1), and (v) a variable of the squared education level (*educ.sq*). Please answer the following questions based on the output below.

## ================================  
## Model 1   
## --------------------------------  
## (Intercept) -2.591 (0.821) \*\*   
## tenure 0.488 (0.054) \*\*\*  
## white 1.882 (0.306) \*\*\*  
## male 0.227 (0.480)   
## educ 0.321 (0.054) \*\*\*  
## educ.sq -0.006 (0.002) \*\*   
## --------------------------------  
## R^2 0.372   
## Adj. R^2 0.364   
## Num. obs. 400   
## ================================  
## \*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05

1. Is there any evidence from the model results to support the claim that there is discrimination based on race? Why or why not? If yes, does this mean that the university is guilty of racial discrimination? Explain.
2. Is there any evidence to support the claim that there is discrimination based on gender? Why or why not. If yes, does this mean that the local government agencies are guilty of gender discrimination? Explain.
3. What does the coefficient value and significance of education squared tell us? Draw a graph depicting the relationship between education and hourly earnings. Make the y-axis hourly earnings and the x-axis education. (again, you can use the drawing tools in word to do this)

### QUESTION SEVEN (10 points)

Some research suggests that race has a differential effect on wages for men and women. Thus, a new model was run to include an interaction between *male* and *white*. Assume that *male* is moderating the effect of *white*.

## ================================  
## Model 1   
## --------------------------------  
## (Intercept) -2.777 (0.966) \*\*   
## tenure 0.490 (0.054) \*\*\*  
## white 2.195 (0.909) \*   
## male 0.415 (0.704)   
## educ 0.322 (0.054) \*\*\*  
## educ.sq -0.006 (0.002) \*\*   
## white:male -0.353 (0.964)   
## --------------------------------  
## R^2 0.372   
## Adj. R^2 0.363   
## Num. obs. 400   
## ================================  
## \*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05

1. Ignoring whether it is significant or not, interpret the coefficient on *male*.
2. Ignoring whether it is significant or not, interpret the coefficient on *white*.
3. Ignoring whether it is significant or not, interpret the coefficient on the interaction term (*male:white*). Recall, from above, that *male* is moderating the effect of *white*.

### QUESTION EIGHT (10 points)

Finally, the university examining discrimination within its system decided to control for the particular campus the employee works on. The university has four campuses. An updated model was run to include a set of dummy variables to control for campus. The base category is south (for the south campus) and thus dummy indicators were entered into the model for the north-central campus, northeast campus, and west campus. [Note: the interaction term was dropped from this model]

## ================================  
## Model 1   
## --------------------------------  
## (Intercept) -2.137 (0.888) \*   
## tenure 0.480 (0.054) \*\*\*  
## white 1.910 (0.304) \*\*\*  
## male 0.189 (0.477)   
## educ 0.317 (0.053) \*\*\*  
## educ.sq -0.006 (0.002) \*\*   
## northcen -0.670 (0.424)   
## northeast -0.743 (0.397)   
## west 0.641 (0.473)   
## --------------------------------  
## R^2 0.392   
## Adj. R^2 0.379   
## Num. obs. 400   
## ================================  
## \*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05

1. Ignoring significance, interpret the coefficient on the indicators for northcen and west.
2. Because none of the dummy variables for the west, north central, or northeast campus were statistically significant at the 5% level, it was a decided to use an F-test to test the exclusion restriction of dropping these three indicators for university campus. Below is the formula for the F-test. The test relies on the SSR from both the restricted and unrestricted model. To appropriately implement the F-test, what variables (i.e., predictors) would be in the restricted model?
3. What are the values for q, n, and k for the F-test discussed above?
4. Below are the results of the F-test. What do we conclude?

## Analysis of Variance Table  
##   
## Model 1: hourly.wage ~ tenure + white + male + educ + educ.sq  
## Model 2: hourly.wage ~ tenure + white + male + educ + educ.sq + northcen +   
## northeast + west  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 394 3468.5   
## 2 391 3360.2 3 108.27 4.1994 0.006088 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1