Midterm Examination 1
November 5, 2020

This exam has two questions. You must answer both of them. We wrote the exam so that most students could answer the exam with a time commitment of at most three hours. We anticipate that most of you will need less time.

For this exam you are allowed to use the textbook, class notes, and other reference material. You are **not** allowed to communicate or share any information about the exam. All questions about the exam should be directed to the class instructors. If you get stuck on the R code in one of the early parts of a question that is preventing you from answering a later part, email one of the instructors for assistance.

We expect that everyone will act in accordance with the Faculty Statement on Academic Integrity and Honor Code. Students found to have cheated or plagiarized on this exam can expect to receive a failing grade on the exam and will be referred to the Dean's Discipline process.

The exam is due by 5pm ET on Friday November 6, 2020. The exam should be submitted to "Exam 1" under the Assignment tab in CourseWorks. Late submissions will receive a grade penalty.

Question 1. How well did the polls predict Tuesday's presidential election results? **(35 pts)**

For this question you are going to analyze how well the pre-election state polls predicted the initial state level votes [these are the vote shares as of mid-day Thursday November 5th]. The polling data are in the file *polls_2020.csv*. The state election data are in the file *vote_2020.csv*. The variables included in the files are described below. *pollsvotes.R* is a script file containing commands similar to those presented on page 132 of the QSS textbook. You will use these datasets and the script file to answer the questions below.

a) Late Pre-Election State Polls versus Early State Election Returns (15 pts)

i. In the *pollsvotes.R* script fill in the command line for creating a variable **polls$margin**, which is the difference between support for Trump and Biden for each poll (i.e., **polls$trump** minus **polls$biden**). Put the command under line in the script that reads *#Question 1.a.i Enter command to create variable for the poll margin here.* (3 pts)

ii. The *pollsvotes.R* script creates an object **poll.pred** which contains an average of **polls$margin** for all the latest state polls for each state. The script also creates the variable **pres$margin**, which is the difference between Trump and Biden's actual vote shares. Create a scatterplot of the latest poll margins (**poll.pred)** against the actual vote margins (**pres$margin).** (Hint: This is similar to the figure on page 135 of the textbook, but for the 2020 election.) The scatterplot should have a title and the axes should be labeled. Include the scatterplot and the commands used to make the scatterplot with your answer. [Successfully completing part **1.a.i** is required to answer this part. If you were not able to complete part **1.a.i**, email the instructors for assistance.] (9 pts)

iii. Instead of containing the average of the latest polls for each state, make **poll.pred** contain an average of all polls in the last 10 days before the election for each state. Create a scatterplot of the new **poll.pred** object against **pres$margin**. The scatterplot should have a title and the axes should be labeled. Does the scatterplot look any different? Include the scatterplot and revised R script with your answer. (3 pts)

b) Prediction Error (20 pts)

i. Calculate the prediction error, where the state electoral margins are the actual outcome and the poll margins in **poll.pred** from **1.a.iii** above is the predicted outcome. Create a histogram of the prediction error. Make sure to include the correct labels on the figure. (4 pts)

ii. Compute the average prediction error and the root mean-squared error. (4 pts)

iii. How good were the state polls at predicting the outcome of the state elections? Discuss at least two potential survey sampling biases that may (or may not) have led the polls to differ from the actual vote margins described above. [Your response should be less than 250 words] (12 pts)

Description of variables included in *polls_2020.csv*. The data, which includes polling data going as far back as 2018, come from FiveThirtyEight.com.

Variable	Description .
state	abbreviated name of the state in which the poll was conducted
biden	predicted support for Biden (percentage)
trump	predicted support for Trump (percentage)
middate	middate of the period when the poll was conducted

Description of variables included in *vote_2020.csv*

Variable	Description .
state	abbreviated name of the state in which the poll was conducted
biden	Biden vote share (percentage)
trump	Trump vote share (percentage)

Question 2. Do Black candidates increase turnout among Black voters? **(65 pts)**

In an article titled "Candidates or Districts? Reevaluating the Role of Race in Voter Turnout" the author, Bernard Fraga, examines a common claim that minority voters are more likely to vote in elections when one of the candidates is a co-ethnic.[1] For this question you will analyze data related to this article, which is in the file *blackturnout.csv*. A description of the variables include in the file is listed below. You will be examining whether Black voter turnout is higher when there is a Black candidate running for office.

a) Answer the following questions about the data **(6 pts)**
   i. How many observations are in the dataset? How many variables are in the dataset? (2 pts)
   ii. How many districts have Black candidates? (2 pts)
   iii. How many states are included in the dataset? (2 pts)

b) Distribution of Black voter turnout across districts **(15 pts)**
   i. Provide the basic descriptive statistics for the central tendency of the *turnout* variable (i.e., mean and median). Do these statistics differ for elections with a Black candidate versus those without a Black candidate? What is the difference in the average turnout in elections with a Black candidate compared those without a Black candidate? (5 pts)
   ii. Provide the following basic descriptive statistics of the spread of the *turnout* variable: variance, standard deviation, min, max, and IQR. Do these statistics differ for elections with a Black candidate versus those without a Black candidate? (5 pts)
   iii. Create a boxplot that compares the turnout in elections with and without a Black candidate. Make sure to provide appropriate labels to the figure. (hint: We are looking for something similar to one of the boxplots on page 93 of the textbook.) (5 pts)

c) Consider the following linear regression model:

$$turnout_i = \alpha + \beta\, candidate_i + \varepsilon_i.$$

turnout is Black turnout in the election. *candidate* is an indicator variable for whether one of the candidates in the election is Black. Estimate the parameters of the model using the lm( ) function in R. **(32 pts)**

   i. What is the outcome (dependent) variable in this regression? What is the explanatory (predictor or independent) variable? (3 pts)
   ii. What is the estimate of α using the lm( ) function? What is the substantive interpretation of this parameter? (6 pts)

[1] Fraga, Bernard. 2015. "Candidates or Districts? Reevaluating the Role of Race in Voter Turnout," *American Journal of Political Science* 60:1:97-122.

iii. What is the estimate of β using the lm( ) function? What does this estimate tell us about Black voter turnout when there is a Black candidate in the election? How does this estimate of β relate to our answer regarding the differences in the means in **2.b.i** above? (8 pts)

iv. Should we interpret the estimate of β as the causal effect of Black candidates on the turnout of Black voters? Why or why not? [Your response should be less than 200 words] (10 pts)

v. How well does this regression model fit the data (i.e., roughly how much of the variation in the outcome (dependent) variable is explained by the model)? (5 pts)

d) Now include the variable *CVAP* in the regression model. *CVAP* measures the proportion of a district's voting-age population that is Black. The regression model is:

$$Turnout_i = \alpha + \beta_1\ candidate_i + \beta_2\ CVAP_i + \varepsilon_i.$$

(12 pts)

i. What is the estimate of β_2 using the lm( ) function? What does this tell us about Black turnout when the share of the Black voting-age population in a district increases by 0.10? (6 pts)

ii. What is the estimate of the coefficient on *candidate*, β_1, using the lm( ) function? How does the estimate of β_1 compare to the estimate of β in **2.c.iii** above? Explain why the estimates of β_1 and β differ? (6 pts)

Description of the variables in *blackturnout.csv*. These data were provided by Princeton University Press.

Name	Description.
year	Year the election was held
state	State in which the election was held
district	District in which the election held (unique within state but not across states)
turnout	Proportion of the Black voting-age population in a district that votes in the general election
CVAP	Proportion of a district's voting-age population that is Black
candidate	Binary variable coded "1" when the election includes a Black candidate; "0" when the election does not include a Black candidate