University of Pennsylvania
Professor Marc Meredith

Homework Week #1

1. The lecture this week discussed some data on legal financial obligations (LFOs) in Alabama. This question is going to have you explore whether there are racial differences in the amount of LFOs that people convicted of crimes owe to the state of Alabama within the sample of cases included in AlabamaCourt.csv. Please do the following in R:

   (a) Read in the dataset AlabamaCourt.csv

   (b) Construct two new variables:

   - A variable, "black", indicating whether the person convicted of the crime is African-American. The persons's race is represented by the variable "race", with this variable being equal to "B' when the person is African-American.
   - A variable, "amountremain", indicating how much money the person convicted of the crime still owes to the state of Alabama. This is can be constructed by subtracting the variable "amountpaid", which represents how much money the person had paid to date, from the variable "amountdue", which represents the total amount of the assessed LFOs.

   (c) Use Stargazer to create two descriptive statistics tables. One describing the African-American units in the sample and one describing the non-African-Americans units in the data. In this table, include three variables: "amountremain", "amountdue", "amountpaid". Make sure to label the three variables in these tables so that someone could understand the variable that is being described.

   (d) Create three kernel density plots that compare the distributions of the variables "amountremain", "amountdue", "amountpaid", respectively, for African-Americans and non-African-Americans.

   (e) Write a paragraph or two, in which you make conclusions about whether there are racial differences in this sample in the amount of LFOs that people convicted of crimes owe to the state of Alabama based on the data being summarized in the previous parts of this question.

2. This problem is going to have you examine some data on the outcomes of college basketball games. I have a theory for why in certain games that underdogs (i.e., the team that is seen as less likely to win a game before the game starts) will perform better than people expect when the score is higher than people expect, and favorites (i.e., the team that is seen as more likely to win a game before the game starts) will perform better than people expect when the score is lower than people expect. To test this theory, I tracked the outcomes of 241 games between 1/22/2019-3/29/2019 in which I assessed my theory would apply. Data about these games are contained in "CollegeBasketball.csv". We are going to examine these data to see whether the empirical

evidence is consistent with my theory.

Here is a description of the variables contained in "CollegeBasketball.csv":

- "Date" is the date that the game took place
- "Favorite" is the name of the team that gambler's thought would be more likely to win the game
- "Favorite3" is the percentage of three point shots that the favorite had made in all of their games prior to the date that game took place. For those of you unfamiliar with basketball, three point shots are those shots that are taken from more than 20 feet, 9 inches from the basket.
- "Underdog" is the name of the team that gambler's thought would be more likely to lose the game.
- "Underdog3" is the percentage of three point shots that the underdog had made in all of their games prior to the date that game took place. For those of you unfamiliar with basketball, three point shots are those shots that are taken from more than 20 feet, 9 inches from the basket.
- "PredictedDifference" are gamblers' expectations about the median point differential in the game. That is, they expect the favorite to win by more points than the predicted difference at roughly the same rate as they expect the favorite to win by less points than the predicted difference or the favorite to lose the game.
- "PredictedPoints' are gamblers' expectations about the total number of points that will be scored in the game. That is, they expect the combined points scored by the favorite and underdog to be above the predicted points at roughly the same rate as they expect the points scored by the favorite and underdog to be below the predicted points.
- "ActualDifference" is how many more points the favorite scored than the underdog in the game , meaning that it is a negative number when the underdog won the game.
- "ActualPoints" is how many combined points the favorite and the underdog scored.

3. Please answer the following questions using R:

   (a) Make a new variable that is equal to "PredictedDifference" minus "ActualDifference". Create a histogram that shows the distribution of this variable. Also report the mean of the absolute value of this variable.

   (b) Let W represent the event that the favorite won a basketball game by more points than expected, E represent the event that the favorite won a basketball game by exactly the number of points that were expected, and L be the event that the favorite won a basketball game by fewer points than expected or lost a basketball game. Make new variables indicating whether these events occurred in each game and use these variables to calculate within the sample $P(W)$, $P(E)$, and $P(L)$.

(c) Make a new variable that is equal to "PredictedPoints" minus "ActualPoints". Create a histogram that shows the distribution of this variable. Also report the mean of the absolute value of this variable.

(d) Let M represent the event that more combined points were scored than expected, T represent the event that the combined points scored were exactly the number expected, and F represent the event that fewer combined points were scored than expected. Make new variables indicating whether these events occurred in each game and use these variables to calculate within the sample $P(M)$, $P(T)$, and $P(F)$.

(e) Using the conditional probability formula discussed in lecture, calculate within the sample $P(W \mid M)$, $P(L \mid M)$, $P(W \mid F)$, $P(L \mid F)$.

(f) Write a paragraph or two, in which you make conclusions about whether the evidence is consistent with my theory based on the data being summarized in the previous parts of this question.