Homework 6

Timothy Baker

4/12/2020

1. Describe the data. Who is in this dataset? What are some of the interesting characteristics of this dataset?

- The data has 60 variables with a total of 500 observations for each of the variables. There are a total of 3076 instances of missing data spread across the variables except for Country, Region, DataYear, and ClassGrade. The data appears to describe high school students in Illinois, United States. It captures data describing physical appearance, habitual behaviors, personality traits, and political opinions.

2. Perform the appropriate test to test the null hypothesis that handedness (i.e. the variable named Handed) is independent of favorite season vs the alternative hypothesis that there is some dependence. Perform this test after removing responses that are blank. Do you think it is ok here to remove the blanks? Explain why or why not. Explain your reasoning for the test you chose and state your conclusions.

- It is permissible to remove blank/NA responses for this testing of independence. Provided that this data is from a survey, I am assuming that the data is missing at random, and using logic that handedness and favorite season are not causitive of each other. Given these assumptions, it is okay to remove the observation completely. I chose to use the chi square test since these are nominal variables, and it is an asymmetric contigency table (3x4). At a 5% rejection level, we fail to reject the null hypothesis (p=0.7053) and conclude that these 2 random variables are independent of each other.

```
data <- read.csv('ill_school_data.csv', stringsAsFactors = F, header = T)

hand_season_data <- data %>%
  select(Handed, Favorite_Season) %>%
  filter(Handed != '' & Favorite_Season != '')

chisq.test(table(hand_season_data))
```

```
## Warning in chisq.test(table(hand_season_data)): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(hand_season_data)
## X-squared = 3.7881, df = 6, p-value = 0.7053
```

3. Build a simple linear regression model with height as your response and arm span as your predictor. First, you need to clean the data, then use MICE to impute missing values using a CART model. Estimate the simple linear regression model on each of the compeleted data sets and use Rubin's combining rules to combined estiamtes across imputations. State your final estimates for each of the slope and intercept parameters as well as standard errors for each of these combined estimates.

- I want to explain my reasonining for how I cleaned the data. I chose only Gender and Ageyears as random variables to include during the imputation given that logically they can increase the accuracy

of the imputed values for height and armspan. I ran a roughly uncleaned regression analysis to observe significant predictors for height and armspan. Gender and Ageyears were the only ones that came up significant after including other body measurement predictors like foot length and finger length. From there I began to clean the data by assuming observations in both height and armspan that are greater than 100 were in the correct metric format. Observations that ranged from 50 to roughy 84 were assumed to be inches given that the calculation from inches to centimeters put those values within a reasonable range. Observations that were less than 10, but greater than 2 were assumed to be answers in feet and thus calculated to centimeters. Finally, there was one observation less than 2 to which I assume was in meters and thus converted to centimeters. Provided that the dataset had a large number of observations, I calculated the outliers from IQR boxplot for height and armspan. I removed all observations that fell (roughly 6 observations) into having both height and armspan as outliers; assumed those responses to be erroneus.

- The final estimates for intercept and armspan predictor is 70.9746 and 0.57999, respectively. The final estimates for standard errors for intercept and armspan predictor is 4.8323 and 0.02857, respectively.

4. Repeat the previous problem, but use a random forest for imputation in MICE instead of a cart model.

- The final estimates for intercept and armspan predictor is 73.5466 and 0.5647, respectively. The final estimates for standard errors for intercept and armspan predictor is 5.2605 and 0.0311, respectively.

5. Finally, put your code and results in a github repository. In the final version of your homework that you submit to Sakai, the answer to this part will simply be a link to that github repository.

- https://github.com/tbeaux18/stats_r_code