Homework 3

BANA7038

Write a report to answer the questions in this homework. When answering each question, please organize your answer to have the following 4 parts:

1. What you are trying to do? (Your goal)

2. R code (How do you realize it?).

3. R output (What is the output from R?).

4. Your observations (What do you observe from the output? How do you interpret the output?)

If necessary, you can repeat 1 through 4 for many times to answer one question fully.

**Instructions on the report:**

1. Highlight the R code in one color.

2. Highlight the R output in another color.

3. Use **BLACK color** for the rest of the text, i.e., your goal and your observations and conclusions.

4. Avoid printing large tables. Avoid printing the entire data set or confidence intervals or too many number at once, instead, visualize them in figures and show only a few rows.

5. Avoid plotting large figures. Use par() to plot multiple figures in one panel to save space.

Plot square figure, do not generate “skinny and tall”, and “short and wide” figures.

6. Limit the length of your report. Try to be as concise as possible.

7. Mark the question numbers in your report in **LARGE and BOLD** font.

8. Separately write down your last name, first name (no abbreviations), and M number. For example:

Last Name: Smith

First Name: John

M-number: M12345678

9. The originality and readability of your report plays an important role in grading (10% for originality and 10% for readability).

Generate a report to answer the following questions.

1. Read PGA data into R (PGA.csv). Below is the description of variables.

Source: sportsillustrated.cnn.com

Description: Performance statistics and winnings for 196 PGA participants during, 2004 season.

Variable: Name, Age, Average Drive (Yards), Driving accuracy (percent), Greens on regulation (%), Average # of putts, Save Percent, Money Rank, # Events, Total Winnings ($), Average winnings ($).

2. Visualize the data using scatter plot and histogram.

3. Build a linear regression using Average winnings as response variable and using Age, Average Drive (Yards), Driving accuracy (percent), Greens on regulation (%), Average # of putts, Save Percent, and # Events as covariates.

4. Perform t tests for these coefficient estimates. Obtain t statistics and p values, interpret the results, make a conclusion (i.e. reject or not reject) and explain why. Note: please explain what the null hypothesis is.

5. Use F test to test the significance of the regression. Obtain the F statistic and p value, interpret the results and make a conclusion.

6. Use a partial F test to test for two variables Age and Average Drive (Yards) together. According to your results, what do you conclude? Similarly, use the partial F test to test for three variables Age, Average Drive (Yards), and Save Percent together, what do you conclude?

7. Obtain the interval estimation for all the intercept and slope coefficients.

8. Using the regression in question 3, make a prediction for the case of:

Age = 35,

AverageDrive = 287,

DrivingAccuracy = 64,

GreensonRegulation = 64.9,

AverageNumofPutts = 1.778,

SavePercent = 48,

NumEvents = 26,

The prediction should include fitted value and interval estimation.

9. Similarly, make another prediction for the case of

Age = 42,

AverageDrive = 295,

DrivingAccuracy = 69,

GreensonRegulation = 67.7,

AverageNumofPutts = 1.80,

SavePercent = 54,

NumEvents = 30,

The prediction should again include the fitted value and interval estimation. Compare the interval from question 8, what do you observe? For example, which interval is wider? And why?

10. Obtain the standardized regression coefficients and compare the influence of all variables.