Final Project

Due December 19, 2022 by 11:59 p.m.

Aruni Jayathilaka

This is the final group project for DSCC/CSC/STAT 262. You are assigned to work in pairs (your assigned group member is on Blackboard). You are only allowed to discuss the final project with your assigned group member. You are not permitted to talk with anyone outside of your group aside from the professor and course teaching assistants. You are permitted to use any approved course material for this project. You are NOT allowed to use online resources outside of what is available through Blackboard (i.e. no Chegg, stack exchange/overflow type sites, etc.). If the submitted report (including code and answers) is similar (either partially or fully) to that of another group, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.

Your group needs to submit a single file for both of you (i.e. you do not all upload the document, if one group member does it, that's sufficient). Include a description of the contribution of each member. All group members are expected to contribute equally to the final project, and you are responsible for knowing everything that is going on in the project (i.e. one person can't be solely responsible for one question and not know what happened in the other questions. If I were to email you and ask you to explain to me what your group did for any questions, you should be comfortable with the submission).

The project must be completed in RMarkdown. You will upload your knitted PDF document through Gradescope. Only one person will upload it, but ensure that all group member's names are on the knitted PDF. **Please put your group number on the document as well**. Please take the time to ensure your interpretations are thorough and thoughtful.

You can use the following `template` for your report:
Group number:
Name of member 1:
Names of member 2:
Contribution of each group member:

Please keep an eye out for any Blackboard announcements I may post relating to the project and procedures for submission.

Consider the diabetes dataset containing the following variables:
- id: Subject ID
- stab.glu: Stabilized Glucose
- ratio: Cholesterol to HDL ratio
- glyhb: Glycosylated Hemoglobin
- location: Location (county) of subject
- age: Age
- gender: Gender
- frame: Body frame
- bmi: Body mass index
- whip: Waist to Hip ratio

1. Begin by reading the data file in R.

2. Examine the variable `location`.

    a. Create absolute and relative frequency tables of location.
    b. Create barplots of absolute and relative frequency tables of location, making each of the three bars a different color. Include appropriate titles.
    c. From previous knowledge, we assume that proportion of location that the data is coming from are equally likely. Conduct an appropriate hypothesis test to see if the proportions of locations are equal to one another. Make sure to report the test statistic and p-value, and interpret the results within the context of the problem.

3. Examine the variable `bmi`

    a. Calculate the five-number summary of bmi.
    b. Create a modified boxplot of bmi, labeling the y-axis as "BMI". Comment on what the plot tells us about the bmi variable.
    c. Create a histogram of bmi, and color the bars using color `cadetblue`. Add a red vertical line on the histogram at the median of bmi, and report the calculated value of the median in red next to the line using the form $\tilde{x}$. Briefly comment on the distribution.
    d. Calculate the standard deviation of bmi.
    e. Calculate the skewness of bmi and comment on that.
    f. Construct a one-sided upper-bound 90% confidence interval for the mean bmi. Interpret the interval within the context of the problem.
    g. Evaluate how well the empirical rule applies to bmi. In particular, calculate the proportion of bmi that falls within one, two, and three standard deviations of the mean. Compare these values to the theoretical percentages as stated by the empirical rule. Overall, does the empirical rule do a good job at describing this variable? Briefly justify your response.
    h. Construct a normal qq plot for bmi (include the 1:1 line). Comment on the plot.
    i. Use a Box-Cox power transformation to determine an appropriate transformation of bmi. Report the recommended transformation.

j. Apply the exact Box-Cox recommended transformation (round it to four decimal places) to bmi. Create a histogram of the transformed data. Comment on the plot.

4. Examine stabilized glucose `stab.glu`as a function of body frame `frame`.

   a. Create side-by-side boxplots of stab.glu for each of the three different frames and color each box plot a different color. Comment on any differences you notice.
   b. Does the mean of stabilized glucose differ by body frame? To test this, construct an ANOVA table for testing at $\alpha = 0.05$ significance level. State the hypotheses, value of the test statistic and the p-value, and interpret the results within the context of the problem.
   c. Further explore the results from question 4b using a Bonferroni multiple comparison procedure with an overall familywise error rate of 0.05 (If needed). Comment which means are significantly different from each other if you perform the test.

5. Examine the body frame `frame` by gender `gender`.

   a. Create a two-way table with marginal totals of the joint distribution of `frame` and `gender`.
   b. Construct a stacked barplot with a bar for each body frame. Each bar should be broken into two sections: male and female. Color the female sections green and color the male sections orange. Label the x-axis to say "frame". Add a legend (it is fine if the legend covers part of the plot itself, as long as the plot can still be reasonably read).
   c. Conduct an appropriate test at the $\alpha = 0.05$ significance level to determine if `frame` and `gender` are associated with one another. Make sure to state the hypotheses, report the test statistic and p-value, and interpret the results in the context of the problem.

6. Examine stabilized glucose `stab.glu` as a function of `gender`.

   a. At the $\alpha = 0.01$ significance level, test whether the variance of the stab.glu is different for gender. Report the test statistic and p-value, and interpret the results within the context of the problem.
   b. Conduct an appropriate two-sample t-test at the $\alpha = 0.05$ significance level to determine if the mean stab.glu of females is less than the mean stab.glu of males (Use your results from question 6a to determine whether or not to assume equal variances). State the hypothesis, report the test statistic and p-value, and interpret the results within the context of the problem.

7. Examine stabilized glucose `stab.glu` as a function of other variables.

   a. Construct a scatterplot of stab.glu over glyhb, plotting points as $\times$ instead of the default $\circ$. Color points based on which body frame they are. Also, report correlation between stab.glu over glyhb. Briefly comment whether linear regression seems appropriate using plot and numeric value.
   b. Conduct a test at the $\alpha = 0.05$ significance level to determine if the Pearson correlation between stab.glu and glyhb from 0 (i.e. $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$). Report the test statistic and p-value, and interpret the results within the context of the problem.
   c. Fit a linear regression to model stab.glu as a function of glyhb. Call this model `lm1`. Report the regression equation.
   d. Construct a two-sided 99% confidence interval for the coefficient of glyhb from `lm1`. Interpret the interval.
   e. What is the value of the coefficient of determination for `lm1`? Interpret this value within the context of the question.
   f. Fit a linear regression to model the stab.glu function of glyhb, bmi and frame including interaction term for bmi and frame. Report the all regression lines separately. Comment about intercept and slope for these lines.
   g. Compare R^2 and Adjusted R^2. If they are different, give a brief explanation as to why.
   h. Use an F test at the $\alpha = 0.05$ significance to determine whether bmi and frame significantly predicts stab.glu, once we have accounted for glyhb. Make sure to state your hypotheses, report the test statistic, and interpret the results in the context of the problem.
   i. Construct diagnostic plots for second model (model in part f), and briefly comment on the fit of the model with respect to linearity, homoscedasticity and normality assumptions.