Homework 3, STAT 7630, SPRING 2024

D. Pluta

2024-01-11

Analysis of the Framingham Heart Study Data set

The problems in this homework will be concerned with analyzing the “Framingham Heart Study” data set. It is available on the course website under the **data** folder. Further information on the data set variables and collection procedures is available at https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset (https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset).

1. a. Load the data set into R Studio.

b. Give a list of the names of the variables in the data set.
c. How many observations are there?
d. Propose an appropriate model to quantify the effect of cholesterol on systolic blood pressure from the measurements provided in the given data set. For each covariate, explain whether it is a possible confounder, or a precision variable.

2. a. Provide univariate plots to show the distribution of values of the following covariates: **age, cigs per day, totChol, sysBP, diaBP, BMI, heart rate, gluclose**.

b. Comment on any covariates that have strange or potentially problematic distributions.

3. Produce a nicely formatted table showing the frequency distribution of each of the categorical covariates.

4. Construct boxplots of **BMI, heartRate, glucose, totChol, sysBP, diaBP** stratified by the value of the **TenYearCHD** indicator.

5. Produce a table of summary descriptive statistics of all variables in the data set.

a. For continuous variables, list the mean and standard deviation.
b. For categorical variables, list the raw counts and percentages of each category.

6. Are there any missing values in the data set? If so, remove those observations.

7. Produce a scatterplot of total cholesterol by systolic blood pressure.

8. Produce a scatterplot of total cholesterol by systolic blood pressure, stratified by gender.

9. Produce a scatterplot of total cholesterol by systolic blood pressure, stratified by current smoking status.

10. Conduct any further exploratory data analysis (with plots and tables) that you feel is relevant to the application of your proposed model.

11. Write one or two paragraphs summarizing your findings from the exploratory data analysis.

12. Fit the simple linear regression model with systolic blood pressure as response and total cholesterol as predictor. Provide a nicely formatted table of the regression results.

13. Fit the linear regression model with systolic blood pressure as response, total cholesterol as predictor of interest, and with adjustment variables BMI, currentSmoker. Provide a summary table of the regression output. Briefly comment on how the fit has changed relative to the simple linear regression.

14. Fit the full model you proposed in 1. Provide a nicely formatted table of the regression output.

15. Conduct a goodness of fit test of your full model with the model in 12. Which model is supported by this test?

16. Conduct a diagnostic analysis of your full regression model. Provide the usual plots and comment on any potential violations of the model assumptions.

17. Write one or two paragraphs summarizing your findings. Include comments on the significance of the various terms in your full model, and confidence intervals of the most important effects. What do you conclude regarding the association of cholesterol and blood pressure? Do you feel the model is a sufficiently good fit of the data, or should alternative models be considered?