CS 7200 - Fall 2023 Olga Vitek

Homework 2

Please follow the homework submission instructions provided on Piazza.

Due on Canvas before midnight on Friday, October 13 2023.
Each part of the problems 5 points

1. Please list the members of your project groups: first and last name, their program, and year in the program. Please use 1-2 sentences to describe your project (if you do not know yet, just say so).

2. *[Linear regression libraries (such as lm in R) are not allowed for this question. Use your own implementation from the last homework.]* Consider the Anscombe quartet `https://en.wikipedia.org/wiki/Anscombe%27s_quartet` datasets. For each dataset:

   (a) Check the quality of fit: state the Pearson coefficient of correlation and coefficient of multiple determination R^2. Make residual plots.

   (b) Perform lack of fit test. State the null and the alternative hypotheses, and the conclusions. *[Note: since these datesets do not have replicate observations, group pairs of observations with the closest values of X.] [Hint: is the test applicable to every dataset?]*

   (c) Comment on the effectiveness of the approaches above to diagnose problems of fit, or deviations from model assumptions.

3. *[Linear regression libraries (such as lm in R) are not allowed for this question. Use your own implementation from the last homework.]* This question explores the effect of adjustments for multiple testing. Consider the first dataset in the Anscombe quartet (X_1 vs Y_1). Suppose that we are interested in confidence intervals for $E\{Y_1\}$ for the values $X_1 = 4, 6, 8, 10, 12, 14$.

   (a) Plot confidence intervals for $E\{Y_1\}$ for these values of X_1, using the Bonferroni adjustment for multiple testing. Describe in words the error rate that is controlled by this procedure.

   (b) Plot the Working-Hotelling confidence band for $E\{Y_1\}$. What error rate is controlled by this procedure? Is this procedure more efficient? Explain.

   (c) For a simple linear regression with one predictor, prove that $Cov\{b_0, b_1\} = -\bar{X}Var\{b_1\}$. Why is this result important for multiple testing? What happens when $\bar{X} = 0$? *[Note: Please include a proof for full credit.]*

4. *[Linear regression libraries (such as lm in R)* ***are*** *allowed for this question.]* This question explores the effect of multicollinearity on inference in multiple linear regression. For all the questions, consider sign and absolute values of the parameters, the magnitude of standard errors, R^2.

   Consider the dataset in KNNL Table 7.1 *[Note: the link to the website containing all the datasets in the book is on the course website]* We are interested in predicting body weight as function of triceps, thigh, and midarm.

   (a) Report the variance-covariance matrix between the three predictors, and between the predictor and the response.

   (b) Make three scatterplots of the response versus one of the three predictor at a time. On each plot, overlay the univariate linear regression with the predictor. Use the plots and the summaries of the linear model fit to comment on the strength of the association between the predictor and the response, and on the quality of fit.

   (c) Fit a multivariate linear regression with both triceps and triceps^2 as additive predictors. Comment on why we may want to fit this model, on the strength of the association between the predictor and the response, and on the quality of fit. Explain the differences from the results in (b).

   (d) Fit a multivariate linear regression with both triceps and thigh as additive predictors. Comment on why we may want to fit this model, on the strength of the association between the predictor and the response, and on the quality of fit. Explain the differences from the results in (b).

   (e) Fit a multivariate linear regression with triceps, thigh and midarm as additive predictors. Comment on the strength of the association between the predictor and the response, and on the quality of fit. Explain the differences from the results in (b).

   (f) How would the results of (d)-(e) differ if the predictors were orthogonal?

5. An analyst studies the association of age (predictor) and maturation (response) of middle school kids. The analyst wishes to include number of older siblings in family as a predictor variable in a regression analysis. The number of older siblings ranges from 0 to 2. For each choice below, state the appropriate model, and discuss the assumptions regarding the effect of the number of siblings on the response:

   (a) Include the number of siblings as a single quantitative predictor.

   (b) Include the number of siblings as two indicator variables.

   (c) Include the number of siblings as two indicator variables, and their interactions.