CSCI E-106:Assignment 4

Due Date: October 5, 2020 at 7:20 pm EST

Instructions

Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document, word, or html generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

Problem 1

Refer to the Real estate sales data set. Obtain a random sample of 200 cases from the 522 cases in this data set (use set.seed(1023) before selecting the sample). Using the random sample, build a regression model to predict sales price (Y) as a function of finished square feet (X). The analysis should include an assessment of the degree to which the key regression assumptions are satisfied. If the regression assumptions are not met, include and justify appropriate remedial measures. Use the final model to predict sales price for two houses that are about to come on the market: the first has X = 1100 finished square feet and the second has X = 4900 finished square feet. Assess the strengths and weaknesses of the final model. (25 points)

Problem 2

Refer to the Production time data. In a manufacturing study, the production times for 111 recent production runs were obtained. The production time in hours (Y) and the production lot size (X) are recorded for each run. (25 points, 5 points each)

a-) Prepare a scatter plot of the data Does a linear relation appear adequate here? Would a transformation on X or Y be more appropriate here? Why?

b-) Use the transformation $X^{'} = \sqrt{X}$ and obtain the estimated linear regression function for the transformed data.

c-) Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

d-) Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

e-)Express the estimated regression function in the original units.

Problem 3

Refer to the Sales growth data. A marketing researcher studied annual sales of a product that had been introduced 10 years ago. The data are as follows, where X is the year (coded) and Y is sales in thousands. (25 points, 5 points each)

a-) Prepare a scatter plot of the data. Does a linear relation appear adequate here? Use the Box-Cox procedure and standardization to find an appropriate power transformation of Y. Evaluate SSE for $\lambda = .3, .4, .5, .6, .7$. What transformation of Y is suggested?

b-) Use the transformation $Y' = \sqrt{Y}$ and obtain the estimated linear regression function for the transformed data.

c-) Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

d-) Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

e-) Express the estimated regression function in the original units.

Problem 4

The following data were obtained in a study of the relation between diastolic blood pressure (Y) and age (X) for boys 5 to 13 years old. (25 points)

X<-c(5,8,11,7,13,12,12,6) Y<-c(63,67,74,64,75,69,90,60)

a-) Assuming normal error regression model is appropriate, obtain the estimated regression function and plot the residuals e_i against X_i. What does your plot residual plot show? (5 points)

b-) Omit case 7 from the data and obtain the estimated regression function based on the remaining seven cases. Compare this estimated regression function to that obtained in part (a). What can you conclude about the effect of case 7? (10 points)

c-) Using your fitted regression function in part (b), obtain a 99 percent prediction interval for a new Y observation at X = 12. Does observation Y_7 fall outside this prediction interval? What is the significance of this? (10 points)