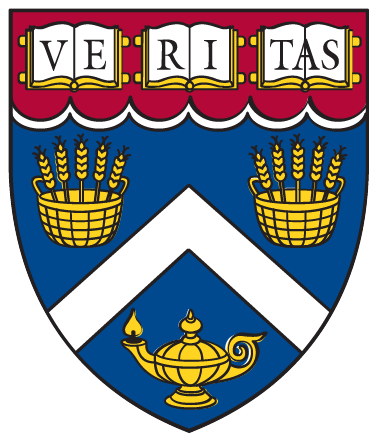
CSCI E-106: Data Modeling



Fall 2019

Dr.Hakan Gogtas

Assignment 4

Due: Monday, 10/21/19 at 7pm 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 generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

1. Refer to the Production Time data set.
2. 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?
3. Use the transformation and obtain the estimated linear regression function for the transformed data.
4. Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?
5. Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?
6. Express the estimated regression function in the original units.
7. Refer to Solution Concentration data set.
8. Fit a linear regression function. Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?
9. Prepare a scatter plot of the data. What transformation of Y might you try,

to achieve constant variance and linearity?

1. Use the Box-Cox procedure and standardization (3.36) to find an appropriate power transformation by using λ = -.2, -.1,0, .1, .2. What transformation of Y is suggested?
2. Use the transformation Y' = log Y and obtain the estimated linear regression function for the transformed data.
3. Plot the estimated regression line and the transformed data Does the regression line appear to be a good fit to the transformed data?
4. Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?
5. Express the estimated regression function in the original units.
6. Refer to Crime rate data set.
7. Fit a linear regression function. Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?
8. Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into the two groups, X≤69, X > 69, and use α= .05. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (a)?
9. Conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use α= .05. State the alternatives. decision rule, and conclusion. Is your conclusion consistent with your preliminary findings in part (a and b)?
10. Refer to Plastic Hardness dataset.
11. Fit a linear regression function. Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?
12. Obtain Bonferroni joint confidence intervals for β0 and β1, using a 90 percent family confidence coefficient. Interpret your confidence intervals.
13. Are bo and b1 positively or negatively correlated here? Is this reflected in your joint confidence intervals in part (b)
14. Management wishes to obtain interval estimates of the mean hardness when the elapsed time is 20, 30, and 40 hours, respectively. Calculate the desired confidence intervals using the Bonferroni procedure and a 90 percent family confidence coefficient. What is the meaning of the family confidence coefficient here?
15. The next two test items will be measured after 30 and 40 hours of elapsed time, respectively. Predict the hardness for each of these two items, using the most efficient procedure and a 90 percent family confidence coefficient.
16. Refer to the CDI data set. Consider the regression relation of number of active physicians to total population.
17. Obtain Bonferroni joint confidence intervals for β0 and β1 using a 95 percent family confidence coefficient.
18. An investigator has suggested that β0 should be -100 and β1 should be .0028. Do the joint confidence intervals in part (a) support this view? Discuss.
19. It is desired to estimate the expected number of active physicians for counties with total population of X = 500, 1000, 5000 thousand with family confidence coefficient .90. Which procedure, the WoIking-Hotelling or the Bonferroni, is more efficient here?
20. Obtain the family of interval estimates required in part (c), using the more efficient procedure. Interpret your confidence intervals.
21. Refer to the SENIC data set. The average length of stay in a hospital (Y) is anticipated to be related to infection risk, available facilities and services, and routine chest

X-ray ratio.

1. Regress average length of stay on each of the three predictor variables. State the estimated regression functions.
2. For each of the three fitted regression models, obtain the residuals and prepare a residual plot against X and a normal probability plot. Summarize your conclusions.
3. Obtain the fitted regression function for the relation between length of stay and infection risk after deleting cases 47 (X47 = 6.5, Y47 = 19.56) and 112 (X112 = 5.9, Y112 = 17.94). From this fitted regression function obtain separate 95 percent prediction intervals for new Y observations at X = 6.5 and X = 5.9, respectively. Do observations Y47 and Y112 fall outside these prediction intervals? Discuss the significance of this.

