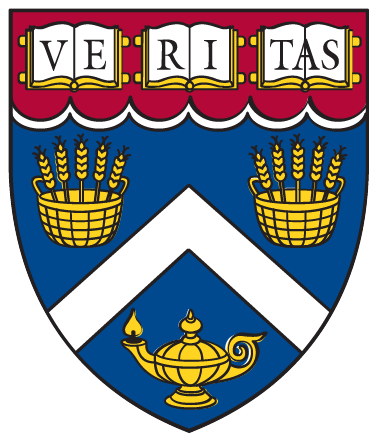
CSCI E-106: Data Modeling



Fall 2019

Dr.Hakan Gogtas

Assignment 9

Due: December 9, 2019 at 11:59pm 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 Employee salaries data. A group of high-technology companies agreed to share employee salary information in an effort to establish salary ranges for technical positions in research and development. Data obtained for each employee included current salary (Y), a coded variable indicating highest academic degree obtained (1 = bachelor's degree, 2 = master's degree; 3 = doctoral degree), years of experience since last degree (X3), and the number of persons currently supervised (X4). (40 pts)
2. Create two indicator variables for highest degree attained: (5pts)



1. Regress Y on X1, X2, X3 and X4, using a first-order model and ordinary least squares, obtain the residuals. and plot them against . What does the residual plot suggest? (5pts)
2. Divide the cases into two groups, placing the 33 cases with the smallest fitted values into group 1 and the other 32 cases into group 2. Conduct the Brown-Forsythe test for constancy of the error variance, using α = .01. State the decision rule and conclusion? (5 pts)
3. Plot the absolute residuals against X3, and against X4. What do these plots suggest about the relation between the standard deviation of the error term and X3, and X4? (5pts)
4. Estimate the. standard deviation function by regressing the absolute residuals against X3 and X4 in first-order form, and then calculate the estimated weight for each case using equation 11.16a on the book. (5pts)
5. Using the estimated weights, obtain the weighted least squares fit of the regression model. Are the weighted least squares estimates of the regression coefficients similar to the ones obtained with ordinary least squares in part (b)? (5 pts)
6. Compare the estimated standard deviations of the weighted least squares coefficient estimates in part (f) with those for the ordinary least squares estimates in pan (b). What do you find? (5 pts)
7. Iterate the steps in parts (e) and (f) one more time. Is there a substantial change in the estimated regression coefficients? If so, what should you do? (10 pts)
8. Refer to the Weight and height. The weights and heights of twenty male 'Students in a freshman class are recorded in order to see how well weight (Y, in pounds) can be predicted from height (X, in inches). Assume that first-order regression is appropriate. (30 pts)
9. Fit a simple linear regression model using ordinary least squares, and plot the data together with the fitted regression function. Also, obtain an Index plot of Cook s distance. What do these plots suggests? (5pts)
10. Obtain the scaled residuals in equation 11.47 and use the Huber weight function (equation 11.44) to obtain case weights for a first iteration of IRLS robust regression. Which cases receive the smallest Huber weights? Why? (10 pts)
11. Using the weights calculated in part (b), obtain the weighed least squares estimates of the regression coefficients. How do these estimates compare to those found in part (a) using ordinary least squares? (5pts)
12. Continue the IRLS procedure for two more iterations. Which cases receive the smallest weights in the final iteration? How do the final IRLS robust regression estimates compare to the ordinary least squares estimates obtained in part (a)? (10 pts)
13. Refer to the Prostate Cancer data set in Appendix C.5 and Homework 7&8. Select a random sample of 65 observations to use as the model-building data set (use set.seed(1023)). Use the remaining observations for the test data. (10 pts)
14. Develop a regression tree for predicting PSA. Justify your choice of number of regions (tree size), and interpret your regression tree. Test the performance of the model on the test data. (5 pts)
15. Compare the performance of your regression tree model with that of the best regression model obtained in HW7. Which model is more easily interpreted and why? (5pts)
16. Refer to Cement composition. The variables collected were the amount of tricalcium aluminate (X1), the amount of tricalcium silicate (X2), the amount of tetracalcium alumino ferrite (X3), the amount of dicalcium silicate (X4), and the heat evolved in calories per gram of cement (Y). (20pts)
17. Fit regression model for four predictor variables to the data. State the estimated regression function. (5pts)
18. Obtain the estimated ridge standardized regression coefficients, variance inflation factors, and R2. Suggest a reasonable value for the biasing constant c (use seq(0,1,by=0.01)) based on the ridge trace, VIF values, and R2 values. (5pts) (hint: use vif.ridge function under library(genride), you can also get MSEs under this library)
19. Transform the estimated standardized ridge regression coefficients selected in part (b) to the original variables and obtain the fitted values for the 13 cases. How similar are these fitted values to those obtained with the ordinary least squares fit ill part (a)? (5pts)
20. Fit Lasso and Elastic net models and compare it against the Ridge regression model results. (5pts) (Hint: Calculate SSEs for each model)

