Assignment - 3

BSAN 450 (Spring 2023)

**This assignment is due on February 9, 2023 at 9:00 PM Central. The total points possible are 100 and there are two (2) problems, each carrying equal points. You can form groups to attempt these problems. Each group will submit one copy of the assignment on Canvas, either in word or pdf, and the assignment should clearly include the names of the group members**.

1. The objective is to infer a sensible pricing model for diamond stones based on data pertaining to their weight (in carats), their color (either D, E, F, G, H or I) and clarity (either IF, VVS1, VVS2, VS1 or VS2). Of interest is the relative worth of the different grades of color and clarity and whether differences in prices can be attributed to the 3 different certification bodies (either GIA, IGI or HRD). The data appeared in Singapore’s *Business Times* edition of February 18, 2000. The variables are listed below. The dependent variable is Price.

* Carat = Weight of diamond stones in carat units
* Col = D, E, F, G, H or I
* Clar = IF, VVS1, VVS2, VS1 or VS2
* CB = Certification Body: GIA, IGI or HRD
* Price = price in Singapore dollars

1. The data is in a file named Diamond.csv. Read the data into R Studio and plot the data versus the possible independent variables. Comment on the plots. Which of the independent variables appear to be related to Price? Do the plots suggest any concerns?
2. Based upon the plots in part a) decide upon an appropriate regression model. Perform the diagnostic checks. If there is a need to change the model make any appropriate changes until you have a satisfactory model.
3. After you have a satisfactory model, investigate whether there are outliers or observations that have large Cook’s distance or a large leverage. If there are outliers or influential observations remove them from the data set and re-estimate the model. What difference does removing these observations have on the estimated model?

To check for outliers it is easiest to use the standardized residuals, these are the residuals divided by their standard deviation. If the absolute value of a standardized residual is larger

1

than 3 then the associated data is an outlier. The following R commands can be used to compute the standardized residuals. Note to access the standardized residuals you need to invoke the MASS library.

library(MASS)

*# Here `diam` is the name of the data frame`*

subset(diam,stdres(fit)< -3) subset(diam,stdres(fit) > 3)

In the subset command, diam is the data frame that was used to estimate the regression model and fit is the name of the regression model. The second command prints out the observations which have standardized residuals less than -3 and the third command prints out the observations which have standardized residuals larger than 3.

The following R command will keep the observations for which the standardized residuals are less than 3 and put these observations in a new data frame named diam1. It is a good idea to store the data without the outliers in a new data frame and use that new data frame to re-estimate the model.

diam1=subset(diam,abs(stdres(fit)) < 3)

After you estimate the model using data without the outliers, you should repeat these steps until there are not any standardized residuals that have absolute value larger than 3. When you remove outliers, it will reduce the estimated standard deviation and as a result some additional outliers may be identified.

2. Data is available on the corn yield in six U. S. corn producing states (Iowa, Nebraska, Illinois, Indiana, Missouri and Ohio) recorded each year from 1890 to 1927. The data is in the file named cornyield.csv. The names of the variables are Year = the year, Yield = the average yield in bushels per acre, and Rainfall = the average rainfall in inches. Develop a model to describe Yield as a function of Year and Rainfall.

[*Hint: You may have to consider a polynomial of the variable Rainfall. If you do, justify all your steps and how you ended up choosing the degree of the polynomial.*]

Go with polynomial of order 2. It has the lowest RMSE when done with 5 fold CV. Additionally, higher order polynomial are not statistically significant in improving the model. The only hesitation I would have is that the higher models do have a higher adj. r^2 value. Additionally, no issues with model diagnostics.

2