Assignment - 2

BSAN 450 (Spring 2023)

**This assignment is due on February 2, 2023 at 9:00 PM Central. The total points possible are 100 and there are three (3) problems, each carrying equal points. You can form groups of 3 or less 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. In this example we are interested in predicting the gas consumption of an automobile based on characteristics of the automobile. The data consists of 38 cars with measurements on fuel efficiency, weight of the car, engine displacement, and number of cylinders. The data is in the file gasconsumption.csv. The names of the variables are GPM = gallons used per 100 miles, WT = weight of the car in 1000 pounds, DIS = displacement of the engine in cubic inches, and NC = the number of cylinders in the engine. You are to develop a model for GPM based on the characteristics of the car, WT, DIS, and NC.

1. Read the data into R Studio and plot scatter plots of GPM versus all the X variables. Comment on these plots. Is there a relationship between the GPM and the possible X variables? How would you describe the relationship?

WT – positive, moderately strong, linear

DIS – Positive, non-linear, non-constant variance

NC – Positive, does not look linear

1. Estimate a multiple regression model with the Y variable equal to GPM and the X variables equal to WT, DIS, and NC. Print out a summary of this model.
2. Use the output for this model to test the null hypothesis: *β*1 = *β*2 = *β*3 = 0 versus the alternative hypothesis: not all *βi*’s are equal to 0. What does your result mean?

F-stat, an intercept model is not idea bc p-value < .05

1. Based upon the output, are all of the 3 variables WT, DIS, and NC needed in the model? Justify your answer.

Based on the saturated model, number of cylinders is not a significant predictor of gallons used per 100 miles since it’s p-value is .06 (this would be the same as running an anova test on the two models). There is not significant evidence that adding NC as a predictor improves the model. However, the p-value is barely over .05, so we should fail to reject with caution. Especially since the saturated model has a higher adj. r^2.

Since the p-value is near .05, we can also compare the models using cross validation. After running 5-fold CV repeated 3 times, the RMSE of both models is nearly the same, .4. Since they have roughly the same RMSE, and given all the information, I would be in favor of choosing the simpler model.

1. Perform the usual diagnostic checks for this model. Based upon your analysis is there any problems with this model?

No, no outliers, non-constant variance, or issues with normality

1. Is there anything that you would do to change the model? If yes, then make the appropriate changes and estimate a new model.

No obvious changes besides choosing the simpler model. I tried making NC categorical since cars typically only have a certain number of set cylinders, but that did not produce a better model according to the adj. r^2.

2. House Prices. This data set includes prices and characteristics of 128 houses in a major metropolitan area. The variables include Price (sales price in dollars), SqFt (size in square feet), Bed (number of bedrooms), Bath (number of bathrooms), Offers

(number of offers the house has received while on the market), Brick (whether it is brick construction: Yes/No) and Nbrhood(East/North/West). The objective is to explain the sale price of a house as a function of its characteristics. The data is in a file named HousePrices.csv.

For this example note that there are two categorical variables: Brick and Nbrhood. Thus, if the variable Brick is included in a multiple regression model, R will create an indicator variable that takes on the value 1 if the variable Brick = “Yes” and 0 if the variable Brick = “No”.

The variable Nbrhood is more complex because there are three groups and as a result there must be 2 indicator variables created for the regression. R creates the variable NbrhoodNorth which is equal to 1 if the house is in the North neighborhood and 0 if the house is not in the North neighborhood. It also creates a second variable NbrhoodWest which is equal to 1 if the house is in the West neighborhood and 0 if the house is not in the West neighborhood. If the house is in the East neighborhood then NbrhoodNorth = 0 and NbrhoodWest = 0. Thus, knowing the values of these two indicator variables is the same as knowing the neighborhood.

1. Read the data into R Studio. Plot the following scatter plots: Price vs. SqFt, Price vs. Bed, Price vs. Bath, and Price vs. Offers. Comment on these plots. Use the following command to read in the data: price=read.csv("HousePrices.csv")
2. Since the variables Brick and Nbrhood are categorical variables, the appropriate plots are boxplots. The commands to plot boxplots in R are the following.

boxplot(price$Price~price$Brick,main="Brick (Yes/No)") boxplot(price$Price~price$Nbrhood,main="Neighborhood")

In boxplots, the median of the data is the bold middle line, the lower edge of the box is the

25th percentile (25 percent of the data is below this value), and the upper edge of the box is the 75th percentile (75 percent of the data is below this value). The lines extending vertically from the boxes are called whiskers. If there are no outliers or potential outliers, then the ends of the whiskers are respectively the minimum and the maximum values in the data.

Plot box plots of Price grouped by Brick and Price grouped by Nbrhood. Based on these plots, which variables appear to have a relationship with Price? Do the model assumptions for the multiple regression model appear to be correct? Do these plots suggest any problems?

1. Fit a multiple regression model with Price as the dependent variable and all the other variables as independent variables. Print out a summary of the model and perform the diagnostic checks for this model. Do the diagnostic checks suggest any problems with this model? Justify your answer.
2. For the model you fit in part c), answer the following questions. Are all the variables needed in the model? Why or why not? What is the difference between the average selling price of house with brick construction compared to the average selling price of a house without brick construction?
3. What is the difference in the average selling price between houses in the North neighborhood compared to houses in the East neighborhood? What is the difference in the average selling price between houses in the West neighborhood compared to houses in the East neighborhood? What is the difference in the average selling price between houses in the North neighborhood compared to houses in the West neighborhood?

3. Meadowform is a small plant found growing in moist meadows of the U. S. Pacific Northwest. It has been domesticated at Oregon State University for its seed oil, which is unique among vegetable oils for its long carbon strings. Like oil from sperm whales, it is non-greasy and highly stable. The data for this problem was obtained to find out how to elevate meadowform production to a profitable crop. The data is in the file meadowform.csv the data names are Flowers, Time, and Intensity. The variable Flowers is the average number of flowers per meadowform plant. The variable Intensity is the light intensity used to grow the plants. The variable Time is a categorical variable: Early means the light intensity was started at an early time period and Late means the light intensity was started at a later time period.

1. Analyze this data to determine if a regression model can be found to predict the variable Flowers from the two variables Intensity and Time. From this data, what should be done to maximize the number of flowers?

Early with little to no intensity

1. As part of your preliminary analysis, you should plot the variable Flowers versus the variable Intensity and use a different symbol for the early and late time. The following R commands will produce this plot.

flower=read.csv("meadowform.csv") group=ifelse(flower$Time=="Early","E","L") plot(flower$Flowers~flower$Intensity,pch=group)