Assignment #4

Linear Regression

Problem 1 (10 marks) File: MALL. XLS

A national chain of women's clothing stores with locations in the large shopping malls thinks that it can do a better job of planning more renovations and expansions if it understands what variables impact sales. It plans a small pilot study on stores in 25 different mall locations. The data it collects consist of monthly sales, store size (sq. ft), number of linear feet of window display, number of competitors located in mall, size of the mall (sq. ft), and distance to nearest competitor (ft).

1. Find a multiple regression model for the data.

   See Figure 1.1 "Parameter Estimates" Table. We can see the values for each Beta coefficients corresponding to each variable. Therefore, the multiple regression model found for Mall Sales is as follows:

   Sales= - 0.6768553*Competitors – 0,00090285*Mall_Size + 2.09589*Nearest_Competitor + 0.91937*Size +9.07598*Windows + 1506.80179

2. Interpret the values of the coefficients in the model.

   Coefficients for each independent variable represent the size of the effect that each variable has on the dependent variable Sales and the sign of each coefficient represent the direction of this effect. Therefore, negative coefficients decrease the value of sale and positive coefficients increase sales:

   - Mall_size has a minimal effect on Sales with a negative direction with a factor of -0,00090285
   - Competitors has the 2nd lowest effect on Sales also with a negative direction with a factor of - 0.6768553
   - Size has the 3rd lowest effect on Sales and positive direction with a factor of 0.91937
   - Nearest_Competitor has the 2nd highest effect on Sales with a positive direction with a factor of +2.09589
   - Windows has the highest effect on Sales with a positive direction with a factor of +9.07598
   - The intercept is the value that Sales has when all of the independent variables are ZERO.

3. Test whether the model as a whole is significant. At the 0.05 level of significance, what is your conclusion?

   As shown in Figure 1.1 “Analysis of Variance” table, p value < 0.0001, therefore we can conclude that there is significant association between Sales and the model.

4. Use the model to predict monthly sales for each of the stores in the study.

   As shown in Figure 1.2 “Output Statistics” table, we can see the Predicted Values versus Dependent Variable and Residuals, Student Residuals, Std Error Mean Predict and Cook’s D value.

5. Plot the residuals versus the actual values. Do you think that the model does a good job of predicting monthly sales? Why or why not?

   In Figure 1.1 the “Residual by Regressors for Sales” Plots do not show any pattern or systematic deviation about the zero line which is desirable outcome for a good fit in the regression model.

6. Find and interpret the value of $R^!$ for this model.

   As shown in Figure 1.2 “Analysis of Variance” the R-square or coefficient of determination is 0.8349, which represents the proportion of variability in the dependent variable that can be explained by the regression model. Therefore, this model explains the variability in Sales at 83.49%, which is high enough.

7. Do you think that this model will be useful in helping the planners? Why or why not?

   Yes, this model will help the planners because explains the variability in Sales with an R Square value of 0.8349, which is good enough to predict sales with confidence.

8. Test the individual regression coefficients. At the 0.05 level of significance, what are your conclusions?

   As shown in Figure 1.2 “Parameter Estimates”, variables Nearest_Competitor and Windows have p values greater than 0.05 significance level. Therefore, we accept the null hypothesis and conclude that these variables are not significant enough to predict Sales.

9. If you were going to drop just one variable from the model, which one would you choose? Why?

   I would drop “Windows” because its p value is 0.75, because using the backward elimination model, it has the highest p value considering a 0.05 significance level or 0.10 significance level.

   The store planners for the women’s clothing chain want to find the best model that they can for understanding what store characteristics impact monthly sales.

10. Use stepwise regression to find the best model for the data.

    See Figure 1.3 to show the results from running the Stepwise Selection method for the following model:

    Sales= -71.03060*Competitors +0.00079216*Mall_size+
    1.04482*Size+1769.60574

11. Analyze the model you have identified to determine whether it has any problems.

    As seen in figure 1.3 the model with 3 variables has an R Square (determination coefficient) of 0.8155 which is less accurate to predict than the original model with 5 variables.

12. Write a memo reporting your findings to your boss. Identify the strengths and weaknesses of the model you have chosen.

    Dear Manager,

    After following the company’s regular practice to conduct a stepwise selection method to determine the best prediction model based on multiple regression for our 25 Store Sales, I have concluded that the following model includes the variables that are the most significant in predicting sales. However, the accuracy of prediction is lower (0.8155 determination coefficient) than using a model that includes all variables (0.8349 determination coefficient). Therefore, further statistical analysis must be made by using 2 other methods like forward and backward regression to compare models resulting from all 3 methods and select the best model as well .

Problem 2 (10 marks)
The File NFLValues.xlsx show the annual revenue ($ millions) and the estimated

team value ($ millions) for the 32 teams in the National Football League.

1. Develop a scatter diagram with Revenue on the horizontal axis and Value on the vertical axis. Does it appear that there are any outliers and/or influential observations in the data?

   In Figure 2.1 we can see that there is one outlier at least 4 potential outliers in the data, which could be influential because might change the fit of the regression model if they are removed.

2. Develop the estimated regression equation that can be used to predict team value given the value of annual revenue.

   As shown in Figure 2.2 we can see the Parameter Estimates and Analysis of Variance of the following model and has an R Square of 0.7673, which shows moderate-strong accuracy to explain the variability of Variable Value based on Revenue.

   Value= 5.83167*Revenue -252,07830

3. Use residual analysis to determine whether any outliers and/or influential observations are present. Briefly summarize your findings and conclusions.

   Figure 2.3 shows Studentized Residuals to identify outliers or influential values thru the Cook'd method. When sorting the last Table, we can see that the bottom 4 observations are outliers and also when looking at the Studentized and Cook's D for value Bar Chart/Table, we can see the threshold for cooks'd and for studentized residuals and conclude that obs 9 is an influential point and outlier and obs 32 is an influential point (4/n=o.125).
   Indeed ,in the fit Plot for Value, we can see the outliers an influential points with more clarity

   For the 2 influential values, my recommendation is to analyze further whether these points are error or what caused them to be so out of range. If the points are errors then they need to be corrected and if the points are not errors, then an analysis of this observations must be done before deciding to remove them because if removed the influential, the regression model could not be significant enough.