ST362 Spring 2019 COURSE PROJECT

Multiple Linear Regression: How Much Is Your Car Worth?

Description

Multiple regression is arguably the single most important method in all of statistics. Regression models are widely used in many disciplines. In addition, a good understanding of regression is all but essential for understanding many other, more sophisticated statistical methods.

This Project consists of a set of activities that will enable you to build a multivariate regression model. The model will be used to describe the relationship between the retail price of 2005 used GM cars and various car characteristics, such as mileage, make, model, presence or absence of cruise control, and engine size. The set of activities in this project allows you to work through the entire process of model building and assessment, including

- Applying variable selection techniques
- Using residual plots to check for violations of model assumptions, such as heteroskedasticity, outliers, and non-normality distributed errors
- Transforming data to better fit model assumptions
- Understanding the impact of correlated explanatory variables
- Incorporating categorical explanatory variables into a regression model

In this project, you will use a relatively small subset of the Kelley Blue Book database to describe the association of several explanatory variables (car characteristics) with the retail value of a car. The data set Cars contains the make, model, equipment, mileage, and Kelley Blue Book suggested retail price of several used 2005 GM cars.

You are asked to use the statistical software, R to conduct analysis for the following four different tasks. You should start to work on these activities after the topics have been covered.

Activity 1 A Simple Linear Regression Model

1. Produce a scatterplot from the Cars data set to display the relationship between mileage (Mileage) and suggested retail price (Price). Does the scatterplot show a strong relationship between Mileage and Price?
2. Calculate the least squares regression line, Price = b0 + b1*Mileage. Report the regression model, the R square value, the correlation coefficient, the t-statistics, and p-values for the estimated model coefficients (the intercept and slope). Based on these statistics, can you conclude that Mileage is a strong indicator of Price? Explain your reasoning in a few sentences.
3. The first car in this data set is a Buick Century with 8221 miles. Calculate the residual value for this car (the observed retail price minus the expected price calculated from the regression line).

Activity 2: Comparing Variable Selection Techniques

4. Use the Cars data to conduct a stepwise regression analysis.

a. Calculate seven regression models, each with one of the following explanatory variables: Cy1, Liter, Doors, Cruise, Sound, Leather, and Mileage. Identify the explanatory variable that corresponds to the model with the largest R square value. Call this variable X1.

b. Calculate six regression models. Each model should have two explanatory variables, X1 and one of the other six explanatory variables. Find the two-variable model that has the highest R square value. How much did R square improve when this second variable was included?

c. Instead of continuing this process to identify more variables, use the software to conduct a stepwise regression analysis. List each of the explanatory variables in the model suggested by the stepwise regression procedure.

5. Use the software to develop a model using best subsets techniques for the whole data set. Notice that stepwise regression simply states which model to use, while best subsets provides much more information and requires the user to choose how many variables to include in the model. In general, statisticians select models that have a large R square, and a relatively small number of explanatory variables. Based on the output from best subsets, which several explanatory variables should be included in a regression model?

6. Compare the regression models in Questions 4 and 5.

a. Are different explanatory variables considered important?

b. Did the stepwise regression in Question 4 provide any indication that Liter could be useful in predicting Price? Did the best subsets output in Question 5 provide any indication that Liter might be useful in predicting Price? Explain why best subsets techniques can be more informative than sequential techniques.

Activity 3: Checking the model assumption

7. Using the regression equation calculated in Question 5, create plots of the residuals versus each explanatory variable in the model. Also create a plot of the residuals versus the predicted retail price (often called a residual versus fit plot).

a. Does the size of the residuals tend to change as mileage changes?

b. Does the size of the residuals tend to change as the predicted retail price changes? You should see patterns indicating heteroskedasticity (non-constant variance).

c. Another pattern that may not be immediately obvious from these residual plots is the right skewness seen in the residual versus mileage plot. Often economic data, such as price, are right skewed. To see the pattern, look at just one vertical slice of this plot. With a pencil, draw a vertical line corresponding to mileage equal to 8000. Are the points in the residual plots balanced around the line Y = 0?

d. Describe any patterns seen in the other residual plots.

8. Transform the suggested retail price to $log\ (Price)$ and $\sqrt{Price}$. Create regression models and residual plots for these transformed response variables using the explanatory variables selected in Question 5.

a. Which transformation did the best job of reducing the heteroskedasticity and skewness in the residual plots? Give the R square values of both new models.

b. Do the best residual plots correspond to the best R square values? Explain. While other transformations could be tried, throughout this investigation we will refer to the log-transformed response variable as TPrice.

Activity 4: Outliers and Influential Observations

9. Calculate a regression equation using the explanatory variables suggested in Question 5 and Price as the response. Identify any residuals (or cluster of residuals) that don't seem to fit the overall pattern in the residual versus fit and residual versus mileage plots. Any data values that don't seem to fit the general pattern of the data set are called outliers.

a. Identify the specific rows of data that represent these points. Are there any consistencies that you can find?

b. Is this cluster of outliers helpful in identifying the patterns that were found in the ordered residual plots? Why or why not?

10. Run the analysis with and without the largest cluster of potential outliers (the cluster of outliers corresponds to the Cadillac convertibles). Use Price as the response. Does the cluster of outliers influence the coefficients in the regression line?

Activity 5: Contributing the New Data Sets (Optional)

This is the additional activity that I ask for collecting data sets. Preferably, the data sets should be coming from business, economic, or social science. Additional maximum 5 bonus marks for each group will be given, depending on the quality of the data sets. Each group member will share this bonus mark equally. This activity will be submitted to my email (sxie@wlu.ca) directly with a cover letter that summarize some details of the data set including the source, references, description of the data variables, and a csv data file. The data must be organized into a data frame that can be imported to R directly without error and It can be analyzed using R command, something looks like: lm(y~x1+x2+……). **You must provide an output of any kind to show that it has been done successfully. Failure to do this will end up with no bonus marks.**

Important Note:

1. **The number of a group can only be 2-3, no exception.**
2. **You must organize the results that you obtain for each question and each sub-question into a report.**
3. **The report must cover all topics that you were asked to complete.**
4. **The submitted report must also include a cover page with student names, and IDs for all group members.**
5. **The report must be submitted in class and only hard copy will be accepted.**
6. **R codes for the project must be uploaded to mylearningspace.ca.**
7. **One copy of the report for each group and one submission of the R codes suffice.**