STAT415/615 REGRESSION PROJECT GUIDELINES

Instructions: The final version of all projects should be submitted as an Rmarkdown file and a Word file to Canvas. The names of all team members should be indicated at the top of the project document.

GRADUATE STUDENTS

Identify a Data Set that consists of one dependent variable and 5 or more independent variables with two or more of the independent variables being categorical. Your Data Set should have at least 50 observational rows. The data set can be extracted from a journal article (no more than 10 years old) or an internet source. Import the data into R studio for processing. This Data set should be appropriate for using the techniques and procedures for Multiple Regression that we explored in class. Do not use an embedded R data table or a data set from your book.

1) Offer a preliminary description of the data set. For example, indicate the size of the data source, describe the variables, and include any other data profile information that would be of interest.
2) Generate relevant data visual plots that explore multicollinearity for the quantitative variables and normality for the quantitative variables as well. Also, use R code to confirm the levels of the categorical variables.
3) Using R code, produce a full Regression Model that consists of quantitative and categorical variables. Make use of the R generated dummy variable matrices
4) Using only the quantitative variables as predictors, produce a model using matrix methods. Also use matrix methods to find the fitted values and the residuals
5) Produce an output summary table to be used to analyze and evaluate the full model (Adjusted R squared, Standard Error, Significance of Variables, ect…)
6) Use procedures and techniques explored in class to produce confidence intervals for the independent quantitative variables of your model. Choose at least two of the quantitative variables to find confidence intervals for.
7) Now produce a reduced model (removing variables of your choice with justification). Use R summary coding for both models and offer justification for choosing one model over the other.
8) Research and apply a model analysis technique not discussed in class to your full model or reduced model. Fully explain the technique or procedure and how it is being applied to your specific model.
9) Offer final summary perspectives about the data and the models that you produce, suggesting how your models or model analysis enhanced your understanding of the data. (4 or 5 sentences)

UNDERGRADUATE STUDENTS

Identify a Data Set that consists of one dependent variable and 6 or more independent variables with two or more of the independent variables being categorical. Your Data Set should have at least 200 observational rows. The data set can be extracted from a journal article (no more than 10 years old) or an internet source. Import the data into R studio for processing. This Data set should be appropriate for using the techniques and procedures for Multiple Regression that we explored in class. Do not use an embedded R data table or a data set from your book.

1) Offer a preliminary description of the data set. For example, indicate the size of the data source, describe the variables, and include any other data profile information that would be of interest.
2) Generate relevant data visual plots that explore multicollinearity for the quantitative variables and normality for the quantitative variables as well. Also, use R code to confirm the levels of the categorical variables.
3) Using R code, produce a full Regression Model that consists of quantitative and categorical variables. Make use of the R generated dummy variable matrices
4) Produce an output summary table to be used to analyze and evaluate the full model (Adjusted R squared, Standard Error, Significance of Variables, ect…)
5) Use procedures and techniques explored in class to produce confidence intervals for the independent quantitative variables of your model. Choose at least two of the quantitative variables to find confidence intervals for.
6) Now produce a reduced model (removing variables of your choice with justification). Use R summary coding for both models and offer justification for choosing one model over the other.
7) Offer final summary perspectives about the data and the models that you produce, suggesting how your models or model analysis enhanced your understanding of the data. (4 or 5 sentences)