**Final Project, SCMA 632**

**Submission due: 7 PM Wednesday, December 14, 2016** e-mailed to [scma.stat@gmail.com](mailto:scma.stat@gmail.com) R.L. Andrews

This project can be done alone or as a two-person team. If you are not able to finish your project submission by December 16 then let me know to arrange for me to assign an “Incomplete” for the course grade. The purpose for this project is for you to demonstrate your ability to analyze a set of data using methods learned in this class and effectively communicate the results as outlined below. You may confer with others about how to do things but all of the submitted work must be done by those who submitted it.

Give an overview of the problem, phenomenon or situation you want to investigate using multiple regression methods. You **do not need to give a review of literature** on the subject, just a clear description of the population or process of interest for someone who is not an expert in the field.

You are to **attach an Excel file in an** **e-mail sent to** [**scma.stat@gmail.com**](mailto:scma.stat@gmail.com)with a report addressing the items 1 through 11 below, labeled and in order. Put **SCMA632 Project in the subject line** and CC your partner if it is a two person project. Number each part (tab) to correspond with the numbering below. In the same e-mail, include your data in an Excel (tab or separate for a large file), JMP or SPSS file. Indicate if I may use your data in the future for class purposes if I give you credit for your efforts in creating the data set.

1. Describe the data set you will be analyzing and clearly specify the observational unit that will provide the individual rows/cases/records for the data set. Tell how you obtained the data. List the variables and provide precise operational definitions for the units of measurement or measurement categories. The variables are to meet these guidelines unless you get permission from me.

* The response/dependent variable must be a quantitative variable. I want a variable that potentially has **at least 20 different possible values** for the dependent variable.
* There must be **at least 2 categorical variables**. One must have **3 or 4 categories** and the other must have **2 or 3 categories**.
* There must be **at least 4 quantitative or numerical variables** that each potentially has **at least 8 different possible values**.
* There should be **at least 8 total independent or predictor variables**. (The dummy variables created for a categorical variable do not count toward the 8, only the one categorical variable used to create the dummies.) Additional categorical variables can have 2 or more categories each. Additional quantitative variables can have as few as 5 possible values (5-point Likert scale variables can be included).
* You must have at least 50 valid cases (n ≥ 50) and I would strongly prefer at least 100.

2. For each quantitative independent variable state what you think would be the relationship between that independent variable and the dependent variable (Include a statement about whether you think this variable may have a non-linear relationship, such as a quadratic one.) then write the appropriate null and alternate hypotheses that can be tested statistically for each variable individually.

3. For each qualitative/categorical independent variable state what you think would be the relationship between that independent or predictor variable and the dependent or response variable. Define appropriate 0 or 1 dummy (indicator) variables needed to include this variable in a linear regression model. (You may want combine some categories, especially those with just a few observations, to simplify the modeling process.) Write the most appropriate null and alternate hypotheses that can be tested statistically using these dummy variables to determine if the data support what you thought would be the relationship between the categories of this variable and the dependent variable (You may not be able to find a testing procedure with an alternate hypothesis that exactly matches your belief.)

4. Identify any pairs of independent variables for which you think that the effect of one variable on the dependent variable would not be the same for all values of the other variable. (Pick two variables for which you think there may be an interaction effect.) If you do not think that this would be the case for any then pick the most likely pair for which you think this could possibly be the case.

5. Screen the data to determine if there are any data points that you think should be removed for final analysis. Do an initial screening before beginning any analysis and look for such points to surface during the analysis. Tell why you removed data values and list each point with complete data that you removed

6. Build a linear regression model including all of the quantitative variables defined in 2 above. Report the R2 and give the p-value for testing the global utility of this model. Is there any indication of potential collinearity problems using all of the quantitative independent or predictor variables? Give support for your answer. Plot the residuals against each of the quantitative independent variables. Identify any variables for which the residual plots indicate that more than a first-order linear relationship should be used.

7. Give what you believe to be the best first-order model using only the non-transformed quantitative variables and tell why you think it is the best model. (Include the coefficients and significance for each included variable.)

8. Add the dummy variables defined in 3 to the quantitative variables used as the best model in 7 to build a model, and as potential predictor variables. Report the R2 for this model and test to determine if the addition of all the dummy variables made a significant contribution to the model. Then give the best first-order model using both quantitative variables and qualitative dummy variables as potential predictor variables.

9. Create variables to include the potential interaction effects you identified in 4 above. Add these variables to the best model in 8 above for the predictor variables in a regression model. Report the R2. Do you think there is any value to adding these interaction variables to the model in 8? Give a reason for your answer.

10. Determine what you think is the best linear regression model to use to predict the dependent variable using the quantitative variables, dummy variables and any transformed variables as possible predictor variables. Tell what steps you took and the logic you used to arrive at your conclusion to declare this model as being the best model. For this best model report the R2, give the p-value for testing the overall utility of this model and for each variable in the final model give the coefficient values, 2-tail p-values and VIF values. Identify the alternate hypotheses in parts 2 and 3 that are supported by your analysis and tell why. Pick out the point that you think had the most influence on determining the coefficient values in this best model using all the points you decided to retain in the data set for building the final model. Tell why you chose it. Give the value of Cook’s D for this data point.

11. How well could your best model above be used to find confidence intervals for either the mean of Y or new individual value of Y for some given set of X variables for all of the predictors in your model? To answer this question, assess how well all assumptions were met, including Homoscedasticity and Normality. If the assumptions are not met, then the confidence interval could be off the mark. Tell whether you observed anything that would cause you to believe that such intervals would be off the mark in some way or not. Provide a reason or reasons for your conclusion.

**A1P5** (10 pts) Each person will prepare a concise PowerPoint presentation reporting on your project model and creating an Office Mix recording that is at least two minutes but not more than five minutes,. **The target audience is a human resources person for a company that is looking to hire someone who can create linear models for phenomena and communicate the results effectively to a client who is not knowledgeable in regression modeling.** There is no limit to the number of slides but each slide should make a specific point. The objective of the presentation is to concisely present the problem for which you built the model and the findings from your analysis in way that will convince the person that you can effectively communicate analysis results to a non-technical audience.