DSC 423 - Data Analysis and Regression
Assignment 3 - Due on Tuesday October 31st by 11:59 PM
Total points: 41 pts (maximum: 47 pts)

PROBLEM 1 [33 pts]
This problem asks you to build a model for the college dataset (college.csv) that contains the following variables:

school: School name
Private: public/private indicator. YES if university is privare, NO if university is public.
Accept.pct: percentage of applicants accepted
Elite10: Elite schools with majority of students from the top 10% of their high school class
F.Undergrad: number of full-time undergraduate students
P.Undergrad: number of part-time undergraduate students
Outstate: Out-of-state tuition
Room.Board: room and board costs
Books: estimated book costs
Personal: Estimated personal spending
PhD: Percent of faculty with PhD degrees
Terminal: Percent of faculty with terminal degrees
S.F.Ratio: Student/faculty ratio
perc.alumni: Percent of alumni who donate
Expend: Instructional expenditure per student
Grad.Rate: Graduation rate in 4 years

Apply regression analysis techniques to analyze the relationship among the observed variables and build a model to predict Graduation Rates (Grad.Rate). Answer the following questions:

a) Analyze the distribution of Grad.Rate and discuss if the distribution is symmetric, or if you need to apply any transformation. **[1 pt R code, 1 pt distribution plot, 1 pt answer = 3 pts]**
b) Create scatterplots for Grad.Rate vs each of the independent variables. What conclusions can you draw about the relationships between Grad.Rate and the independent variables? (No need to include the scatterplots in your submission, but you can use correlation analysis) **[1 pt R code, 2 pts answer = 3 pts]**
c) Build boxplots to evaluate if graduation rates vary by university type (private vs public) and by status (elite vs not elite). Discuss your findings. **[1 pt R code, 1 pt boxplots, 1 pt answer = 3 pts]**
d) Fit a full model (with all independent variables) to predict Grad.Rate **[1 pt R code, 1 pt full model equation = 2 pts]**
e) Does multi-collinearity seem to be a problem here? What is your evidence? Compute and analyze the VIF statistics. **[1 pt R code, 1 pt VIF statistics, 2 pts answer = 4 pts]**
f) Apply TWO variable selection procedures to find an optimal subset of independent variables to predict Grad.Rate. You can choose any two procedures among the ones we learned in class: backward selection, forward selection, adj-R^2, Cp, stepwise, press. **[2 pts R code for the 2 variable selection procedures, 1 pt answer = 3 pts]**
g) Fit a final regression model M1 for Grad.Rate based on the results in f). Explain your choice. Write down the expression of the estimated model M1. **[1 pt R code for final model, 1 pt explanation, 1 pt expression = 3 pts]**

h) Draw a scatter plot of the studentized residuals against the predicted values. Does the plot show any striking pattern indicating problems in the regression analysis? **[1 pt R code, 1 pt answer = 2 pts]**
i) Analyze normal probability plot of residuals. Is there any evidence that the assumption of normality is not satisfied? **[1 pt R code, 1 pt answer = 2 pts]**
j) Are there any outliers or Influential Points? Compute appropriate statistics. **[1 pt answer, 1 pt R code for statistics = 2 pts]**
k) Analyze the R^2 value for the final model and discuss how well the model explains the variation in graduation rates among the universities. **[1 pt R^2 value, 1 pt answer = 2 pts]**
l) Draw conclusions on graduation rates based on your regression analysis. What are the most important predictors in your model? Does your model show a significant difference in graduation rates between private and public universities? Do "elite" universities have higher graduation rates? **[1 pt conclusion, 1 pt predictors, 1 pt significant difference, 1 pt answer = 4 pts]**

PART 2:

Interaction Terms [8 pts]:

a) You are asked to build a new regression model that includes the following independent variables: Elite10, Accept.pct, Outstate, perc.alumni and Expend, together with the interaction effects of elite10 with each independent variable. Fit the model and analyze if the interaction terms are significant. **[1 pt fitted regression model with R code, 1 pt answer = 2 pts]**
b) Simplify the model and remove interaction terms and additive terms that are not significant. Remember that additive terms included in interaction terms should not be removed. Write down the expression of the final model M2. **[1 pt simplified model, 1 pt expression = 2 pts]**
c) Analyze the parameter estimates of the fitted model and discuss how being an "Elite10" University affects the relationship between Graduation Rates and the four predictors Accept.pct, Outstate, perc.alumni and Expend. **[1 pt R code, 1 pt analysis, 2 pts answer = 4 pts]**

OPTIONAL: Cross-validation [+6 extra credit pts]

d) Apply cross-validation techniques (5-fold cross validation or divide dataset into a training and a testing set) to compute how well your final model M1 in Part 1 predicts new data. Compute the MAPE (mean absolute percentage error) statistic and discuss the results. **[1 pt cross-validation, 1 pt MAPE, 1 pt answer = 3 pts]**

e) Apply the same cross-validation procedure and compute the MAPE statistic for the interaction model M2 computed in Part 2. Compare the predictive power of the models M1 and M2 fitted in Part 1 and Part 2. **[1 pt cross-validation, 1 pt MAPE, 1 pt answer = 3 pts]**

Submission instructions:
Submit the homework at the Course Web page at https:\\d2l.depaul.edu.
Submit your R file for the problem indicating in comments which block of code solves what part of the problem, and your answers in a Word or PDF file. You can also submit an R Markdown file and the HTML/PDF file of the output.