STATS 415, Homework 2

Due at 11:59pm on Thursday, Sep 17, 2020

Turn in a pdf scan of your homework on Canvas. Please limit your answer to Q2 to 8 pages, organized into a coherent typed data analysis report. Answers to Q1 may be either typed or handwritten. Please clearly write your name, your UMID, and your GSI/lab number on the homework.

1. Suppose you are trying to predict starting salary in a certain position (y, in thousands of dollars) from undergraduate GPA (x_1, from 0 to 4), gender (with only two options recorded, $x_2 = 1$ if female, 0 if male), and relevant prior experience (x_3, in years). Suppose you fit a linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$$

and obtained the OLS estimate $\hat{\beta} = (40, 3, -2, 1.5, -0.5, -0.1)^T$. [50 points]

   (a) Predict the starting salary for a female employee with a GPA of 3.5 and 2 years experience. [5 points]

   (b) For a given value of GPA and years of experience, do male or female employees earn more, on average? By how much? Show the work you did to obtain your answer to each of these two questions, or explain why it is not possible to answer. [10 points]

   (c) The coefficient for the interaction between GPA and gender is negative ($\hat{\beta}_4 = -0.5$). What is the interpretation of that? [10 points]

   (d) The smallest (in absolute value) coefficient is $\hat{\beta}_5 = -0.1$. Does that mean that there is no interaction between GPA and experience? Explain. [10 points]

   (e) For someone with no prior experience ($x_3 = 0$), sketch the two lines showing the relationship between income and GPA for males and females. Write down the equations for both lines. [10 points]

   (f) What hypothesis would you need to test to decide whether the population lines in the previous question should be parallel? [5 points]

2. This exercise relates to the `Carseats` data set in the `ISLR` package. You may use `help(Carseats)` to learn more about the data set. [50 points]

   (a) Fit a multiple regression model to predict `Sales` using all other variables in the model with no interactions (the full model) Report the values of coefficients, and how well the model fits (using R^2). Include diagnostic residual plots and comment on any interesting features. [10 points]

   (b) Which variables have significant p-values? What is the hypothesis corresponding to the p-value which appears in the summary table for the variable `Urban`? [6 points]

   (c) Drop all the variables that are not significant in the full model (Note: this is not the best way to do model selection; we will study better ways later). Fit the linear model with the remaining variables and no interactions (the reduced model). It will include one categorical variable, `ShelveLoc`. Compare the fit of the reduced model to the fit of the full model using R^2. [5 points]

   (d) Use the `anova()` command to formally compare the full and reduced models and state your conclusion. Comment on the difference between their R^2 in light of your conclusion. [6 points]

   (e) Write out the reduced model in equation form and interpret the coefficients. Be careful with the coefficients of the categorical variable. [6 points]

   (f) Add an interaction term between the categorical variable `ShelveLoc` and the variable `Price` to the reduced model. Report the estimated coefficients, and interpret the coefficients of the interaction term. Do the corresponding p-values suggest the interaction term is necessary? [10 points]

   (g) Use `anova()` to formally test whether the interaction term is needed, and state your conclusion. [7 points]