DATA SCI 415 - Homework 3

Due Friday, February 12, 2024, 11:59 PM

1. I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$.

   (a) Suppose that the true relationship between x and y is linear, i.e. $y = \beta_0 + \beta_1 x + \epsilon$. Consider the training residual sum squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

   (b) Answer (a) using test rather than training RSS.

   (c) Suppose that the true relationship between x and y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

   (d) Answer (c) using test rather than training RSS.

2. This exercise relates to the `Carseats` data set in the `ISLR` package (or it can be found on the book's website). You may use `help(Carseats)` to learn more about the data set.

   (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).

(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`?

(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.

(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.

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

(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?

(g) Test whether the interaction term is needed, and state your conclusion.