STSCI 3740/5740 Machine Learning and Data Mining

Dr. Nayel Bettache **Homework 2, due October 31, 11:59pm**

Problem 1 (6 points)

Solve problem 3 on page 121 in the textbook Introduction to Intro to Statistical Learning with R (second edition).

Problem 2 (16 points)

This question should be answered using the Carseats data set, which is contained in the package ISLR.

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

(b) Provide an interpretation of each coefficient in the model. Be careful–some of the variables in the model are qualitative!

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$? Use the significance level 0.05 for the hypothesis test.

(e) On the basis of your response to question (d), fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

(f) What are the values of R^2 for models in (a) and (e)? Does larger R^2 mean that it is a better model?

(g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).

(h) Fit linear regression models in (e) with interaction effects. Provide an interpretation of each coefficient in the model.

Problem 3 (12 points)

Solve problem 4 on page 122 in the textbook Introduction to Intro to Statistical Learning with R (second edition).

Problem 4 (6 points) (Required for 5740, bonus for 3740.)

Assume that we have i.i.d data $(x_1, y_1), ..., (x_n, y_n)$, where y_i is the response variable and x_i is the covariate. For simplicity, let us consider the case that x_i is univariate. We fit a simple linear regression **without** intercept,

$$y_i = \beta x_i + \varepsilon_i, \tag{1}$$

where β is the coefficient parameter. We know that the least square estimator is derived by regressing y on x. It has the following form

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}. \tag{2}$$

What happens if we regress x on y (rather than y on x)? Would the least square estimate for the regression of x onto y be a good estimate of $1/\beta$? We will address these questions in the following two steps.

(a) Conduct a simulation in R. We first generate $n = 500$ data points $(x_1, y_1), ..., (x_n, y_n)$ from (**??**). You can choose the distribution of x, ε and the value of β by yourself. Given the data $(x_1, y_1), ..., (x_n, y_n)$, regress x on y. You may try lm(x∼y-1,data) or something similar. We record the least square estimate of the coefficient. We now repeat the process 100 times, that is we regenerate data from the same setup and compute the estimate of the coefficient. To do this, you may need to write a for loop. Now, we have 100 estimates of the coefficient. Finally, we can compare the mean of the 100 estimates of the coefficient with $1/\beta$. Please answer the professor's question: Would the coefficient estimate for the regression of x onto y be a good estimate of $1/\beta$ in the model (**??**)?

(Please note that your answer could be sensitive to how you generate the data. In statistics, we often consider the so-called sensitivity analysis. That is you can vary the setting and parameters in the data generating process and see if your conclusion still holds or not)

(b) Please provide a proof to show whether the coefficient estimate for the regression of x onto y is a consistent estimator of $1/\beta$. For this question, you can start from the least square formula in (**??**). (Hint: The proof needs the law of large numbers to show the convergence of the estimator.)