OPIM 5603 — Statistics in Business Analytics
Fall 2019, University of Connecticut

Homework 10 - v1

Instructions: Please complete the following questions and submit them as an RNotebook (as an `Rmd` file) via the submission link on HuskyCT. You must submit the assignment by the time and due date listed on the course syllabus. Failure to submit a file by the deadline will result in a score of 0 on the assignment. You must show all work and clearly explain any results.

Set the heading of the RNotebook as an `html_document`, with a table of contents and without numbered sections. Add your name and a date to the header as well. The solution to each problem should be a separate section (specified by #), and each subproblem should be set as a subsection (specified as ##). For example, for Problem 2, you should have a section titled Problem 2, specified by:

```
# Problem 2
```

in your RNotebook. Also, for subproblem b in Problem 2, you should have a subsection, specified by:

```
## Problem 2b
```

As with all course material, the problems appearing in this homework assignment are taken from the instructor's real-world experiences, from other courses taught at the University of Connecticut, and from the sources listed in the course syllabus.

Note that R code submitted should work independent of the data that sits in the data structure. For example, suppose there was a vector `r_vec` with the values $(1, 2, 6)$ and the problem asks for you to create R code to create a vector `answer` which doubles each element of `r_vec`. The answer

$$\texttt{answer} \leftarrow \mathrm{c}(2, 4, 12)$$

would be given no credit. The answer

$$\texttt{answer} \leftarrow 2^{*}\texttt{r\_vec}$$

would be an appropriate answer.

You must show all steps in your solution. For example, if a problem asks for the expected value of a random variable that is binomially distributed with $n = 10$ and $\pi = 0.3$, and you simply write

$$3,$$

this will be given no credit. However,

$$10 * 0.3$$

would be given credit.

If you have any questions, please submit them via email to the instructor and/or the teaching assistant prior to submitting your solution.

For each problem, unless otherwise specified, please assume that the confidence level is 95%.

Problem 1 (50 points)

In this problem we will generate data and evaluate how well a linear model can be used to model a relationship. Don't worry if the assumptions required to use linear regression aren't true—even if the assumptions are false, it is still OK to use a linear model for prediction, however some of the statistical tests on the model that we will discuss in the next class will no longer be appropriate.

a. Generate a vector X with 30 values, where each value is generated from a normal distribution with mean 30 and standard deviation 5. Generate a vector Y with 30 values where each value is 10 times the corresponding value in X plus a random value, generated from a normal distribution with mean 3 and standard deviation 1. Now, fit a linear model.

   1. What are the coefficients?
   2. What do you predict the Y value will be if $X = 32.8$?
   3. What is the R^2 value?

b. Generate a vector X with 30 values, where each value is generated from a uniform distribution with minimum 20 and maximum 30. Generate a vector Y with 30 values where each value is 10 times the corresponding value in X plus a random value, generated from a normal distribution with mean 3 and standard deviation 1. Now, fit a linear model.

   1. What are the coefficients?
   2. What do you predict the Y value will be if $X = 32.8$?
   3. What is the R^2 value?

c. Generate a vector X with 30 values, where each value is generated from a normal distribution with mean 30 and standard deviation 5. Generate a vector Y with 30 values where each value is 10 times the corresponding value in X squared, plus a random value, generated from a normal distribution with mean 3 and standard deviation 1. Now, fit a linear model.

   1. What are the coefficients?
   2. What do you predict the Y value will be if $X = 32.8$?
   3. What is the R^2 value?

d. Compare the three models. Why do you think the R^2 values compare as they do? Generating scatter plots may be useful.

Problem 2 (50 points)

Consider the data set `mtcars` which is available in base `R`.

a. Build a linear regression model predicting `mpg` using `hp`.

b. What is the regression equation?

c. What would you predict the `mpg` of a car to be if `hp` is 110 using the regression equation?

d. Run a statistical test to determine if there is a linear relationship between `hp` and `mpg`. Formally write the null and alternative hypotheses. What is the p-value of the test?

e. As briefly mentioned in class, there is a statistical test that can be run for the correlation between two variables. If you have variable X and variable Y, the null hypothesis is that there is no correlation between the two variables, and the alternative hypothesis is that there is non-zero correlation between the two variables. The command is `cor.test` which takes two arguments, that are the variables you are testing for correlation. Before running the test, what is the correlation between `hp` and `mpg`?

f. Run a statistical test to determine if there is correlation between the two variables using `cor.test(mtcars$mpg,mtcars$hp)`. This has various outputs. What is the confidence interval? Note: This represents a 95% confidence interval for the correlation between the two variables.

g. What is the p-value for the statistical test in the previous part? Do you reject the null hypothesis?

h. In part d. you found a p-value for whether or not there is a statistically significant linear relationship between `hp` and `mpg`, and you found a p-value. How does that p-value compare with the p-value found in part g.? Explain the relationship.

i. There are other variables in the data set. Try adding some variables to the regression model. Do any result in a better adjusted r^2? What is the best collection of variables that you can identify for predicting `mpg` that has the highest adjusted r^2?