Lab Assignment #5

Nick Noel & Liz VIlla

Due March 8, 2023

Instructions

The purpose of this lab is to introduce more advanced regression strategies that were probably not covered in Math 338.

In this lab, we will be working with four datasets. Three (`Boston`, `Carseats`, and `Wage`) are contained in the `ISLR2` package. Information about these datasets can be found by searching R help for them.

The fourth dataset, `RateMyProfessor`, needs to be downloaded from Canvas. This dataset contains the overall average rating from https://www.ratemyprofessors.com/ for over 22,000 professors, as collected by Murray et al. (2020). A data dictionary for the dataset can be found at https://github.com/murrayds/aa_rmp/tree/master/data (note that I removed a bunch of variables so that you're downloading a 2 MB dataset instead of a much larger one).

```
library(ISLR2)
library(ggplot2)
library(dplyr)
library(broom) # See Problem 3b

RateMyProfessor <- read.csv("RateMyProfessor.csv")
```

This lab assignment is worth a total of **15 points**.

Problem 1: Indicator Variables

Part a (Code: 0.5 pts)

Run the code in ISLR Lab 3.6.6. Put each chunk from the textbook in its own chunk.

Part b (Explanation: 1 pt)

Interpret the slope estimate corresponding to `ShelveLocGood` in the model fit in part (a).

Part c (Code: 1 pt; Explanation: 1.5 pts)

Using the RateMyProfessor dataset, fit a linear model predicting the overall rating of a professor (`overall`) from the difficulty rating (`difficulty`), chili pepper rating (`hotness`), and rank (`rank`). What are the reference levels for each categorical variable? How do you know?

Part d (Explanation: 1.5 pts)

Holding difficulty constant, which of the following instructors would be predicted to have the highest overall rating? Which would be predicted to have the lowest overall rating? Explain your reasoning.

- Attractive Assistant Professor
- Attractive Associate Professor
- Attractive Professor
- Less-attractive Assistant Professor
- Less-attractive Associate Professor
- Less-attractive Professor

Problem 2: Interaction Terms

Part a (Code: 0.5 pts)

Run the single line of code in ISLR Lab 3.6.4.

Part b (Explanation: 2 pts)

Notice that `age` is a significant predictor of `medv` in the model without the interaction term (from ISLR Lab 3.6.3 on Lab 4), but it is no longer a significant predictor of `medv` once we add in the interaction term. The p-value is huge (0.971!). What do you think is happening here? Are we okay to remove the `age` variable from the model with the interaction term? Why or why not?

Part c (Code: 1 pt; Explanation: 1.5 pts)

Create a new dataset, `associates`, by `filter`ing the `RateMyProfessor` dataset to include only the Associate Professors.

Next, complete this code chunk to create a graph of overall rating vs. difficulty rating for the associate professors, with "hot" professors shown in red and "cold" professors shown in blue. Remember to delete `eval = FALSE` once you get the code to run!

```
ggplot() + # add appropriate arguments here or in the next line
  geom_point(alpha = 0.25) +
  # if you added arguments to geom_point, add the same arguments to geom_smooth
  # otherwise don't touch these lines - they will run fine
  geom_smooth(method = "lm", se = FALSE) + # regression line, no confidence bands
  scale_color_manual(name = "Chili Pepper Rating",  # make legend nice
                     labels = c(hot = "Attractive", cold = "Less-attractive"),
                     values = c(hot = "red", cold = "blue")) +
```

How does the difficulty of the professor modify the relationship between attractiveness and overall rating?

Part d (Code: 1 pt; Computation and Explanation: 2 pts)

Using the `RateMyProfessor` dataset, fit a linear model predicting overall rating from the difficulty rating (`difficulty`), chili pepper rating (`hotness`), rank (`rank`), and an interaction term between `difficulty` and `hotness`.

Using your results, write out the least-squares regression equation predicting overall rating from difficulty for an attractive associate professor. Also, write out the least-squares regression equation predicting overall rating from difficulty for a less-attractive associate professor. Explain how you obtained each equation.

Do your equations support your conclusions from part (c)? Explain why or why not.

Problem 3: Regression with Nonlinear Transformations of the Predictors

Part a (Code: 0.5 pts)

Run the first four code chunks in ISLR Lab 7.8.1 (up through the point where `fit2b` is created). Put each chunk from the textbook in its own chunk.

Part b (Code: 1 pt)

In the code chunk below, create a data frame with a single variable, `age`, ranging from 18 to 80, then use the `augment` function (in the `broom` package) to obtain the predicted wage, standard error of the mean wage, and the lower and upper bounds of a 95% confidence interval for the population mean wage at each age. (You can use any of `fit`, `fit2`, `fit2a`, or `fit2b` - they should all give the same predictions.)

What is the 95% confidence interval for the population mean wage of 25-year-olds? 50-year-olds?