

Homework 4

The assignment is due **December 14th at 1159pm**.

Some instructions:

- You should submit your assignment through Blackboard as a nicely formatted pdf document. That is, don't submit a document with just R code.
- Use full sentences to answer your questions.
- Include the code you used to produce your answers as an appendix at the end of the document. Do not include code in your answers to the questions themselves.
- Please include your plots in your answers. Don't put them in the code appendix.
- Your document should be easy to browse: your answers should be easy to identify, and it should be easy to know where an exercise starts and where it ends.
- When in doubt, err on the side of explaining what you did using your own words.

If you have any questions, please let me know (my email is `victor.pena@baruch.cuny.edu`).

For this assignment, you'll need to load the following libraries

```
library(ISLR2)
library(caret)
library(tidyverse)
library(jtools)
library(interactions)
```

In this assignment, we'll work with a dataset that has crime rates in Boston. You can load it with the commands:

```
data(Boston)
```

You can get information about the data with `?Boston`.

We need to transform the data a little bit for this assignment. Please read the instructions below carefully.

The variable `chas` has been inputted as numeric, but it's actually categorical. You can convert it into a factor with the command:

```
Boston$chas = as.factor(Boston$chas)
```

Please run the code above.

Our outcome variable will be a binary variable which tells us whether a suburb has an above median crime rate or not. Here's how you create such a variable:

```
Boston = Boston %>%  
  mutate(highcrim = factor(ifelse(crim > median(crim), "yes", "no")))
```

The goal is predicting whether a suburb has an above median crime rate or not, given other variables.

In your analyses, exclude `crim`, since it's just a transformation of `highcrim`.

Answer the questions below.

1. Create a boxplot of `rad` by `highcrim` and a boxplot of `medv` by `highcrim`. Comment on what you see. Instead of using the variable names `rad`, `medv` and `highcrim`, use descriptions of what those variables actually are in your explanations.
2. Fit a model that predicts `highcrim` given `medv` and `rad`. Interpret the coefficients in the log-odds scale and use `effect_plots` to visualize their effects on the probability scale.
3. Use backward selection with AIC to find a model. Which variables are included? Are `rad` and `medv` included? If so, compare your results to what you found in part 2.
4. Use an estimated probability of 0.5 as a threshold for predicting that a suburb is high crime. Find a confusion table for the model you found in part 3. and answer the following questions.
 - What is the accuracy of the model? Explain what the value of accuracy means in your own words. Think that you're explaining this to a colleague that doesn't know much about logistic regression.
 - What is the sensitivity of the model? Again, explain the meaning of that metric in your own words.
 - What is the specificity of the model? Explain the meaning of the metric in your own words.
 - Do you think the model is doing a good job at predicting `highcrim`? Why or why not?
5. Repeat what you did in the previous part, but now with the model that you fitted in part 2., which predicts `highcrim` given `medv` and `rad` only. Are the results better or worse?