DS 301: Homework 8
Due: April 6, 2022 on Canvas by 11:59 pm (CT)

Instructions: Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw `R` code or raw `R` output** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, an R file, text file or .Rmd are all acceptable). You should mark sections of the code that correspond to different homework problems using comments (e.g. `##### Problem 1 #####`).

Problem 1: Bootstrap

We will work with the `Boston` housing data set; it is part of `library(ISLR2)`.

a. Fit a regression model with `medv` as your response and `crim` and `age` as your predictors. Obtain a bootstrapped confidence interval of $\hat{\beta}_{\text{crim}}$. Plot your bootstrapped distribution of $\tilde{F}^{(b)}$.

b. Compare this with the confidence intervals obtained assuming normality using analytical formulas (you can use the `confint()` function). How do they compare your results from (f) compare?

c. Based on this data set, provide an estimate $\hat{\mu}_{\text{med}}$ for the median value of `medv`.

d. We would like to estimate the standard error of $\hat{\mu}_{\text{med}}$. Since there is no simple formula for computing the standard error of the median, bootstrap the standard error. Copy/paste your code and report your standard error here.

e. Using bootstrap, provide a 95% confidence interval for the median of `medv`. Plot your bootstrapped distribution of $\tilde{F}^{(b)} = \frac{\tilde{\mu}^{(b)\text{med}} - \hat{\mu}_{\text{med}}}{se(\tilde{\mu}^{(b)\text{med}})}$.

f. Based on this data set, provide an estimate $\hat{\mu}_{0.1}$, the 10th percentile of `medv`.

g. Use bootstrap to estimate the standard error of $\hat{\mu}_{0.1}$. Comment on your findings.

Problem 2: Email Spam

We will use a well-known dataset to practice classification. You can find it here: `https://archive.ics.uci.edu/ml/datasets/Spambase`. Read the attribute information and download the dataset onto your computer. To load this data into R, use the follow code:

```
spam = read.csv('.../spambase.data',header=FALSE)
```

The last column of the `spam` data set, called `V58`, denotes whether the e-mail was considered spam (1) or not (0).

a. What proportion of emails are classified as spam and what proportion of emails are non-spam?

b. Carefully split the data into training and testing sets. Check to see that the proportions of spam vs. non-spam in your training and testing sets are similar to what you observed in part (a). Report those proportions here.

c. Fit a logistic regression model here and apply it to the test set. Use the `predict()` function to predict the probability that an email in our data set will be spam or not. Print the first ten predicted probabilities here.

d. We can convert these probabilities into labels. If the predicted probability is greater than 0.5, then we predict the email is spam ($\hat{Y}_i = 1$), otherwise it is not spam ($\hat{Y}_i = 0$). Create a confusion matrix based on your results. What's the overall misclassification rate? Break this down and report the false negative rate and false positive rate.

e. What type of mistake do we think is more critical here: reporting a meaningful email as spam or a spam email as meaningful? How can we adjust our classifier to accommodate this?

Problem 3: Weekly data set

This question should be answered using the `Weekly` data set, which is part of the `ISLR2` package. This data is similar in nature to the `Smarket` data we saw in class, except that it contains $1,089$ weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

a. Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be any patterns?

b. Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

c. Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

d. Now fit the logistic regression model using a training data period from 1990 to 2008, with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

Continue onto next page.

Problem 4: Limitations of Logistic Regression

Consider the dataset:

x	y
-2	red
5	blue
-1	red
10	blue
5	blue

a. Plot the data in R **in a single plot by group** (red vs. blue). What do you observe? Are the two groups well-separated?

b. Fit a logistic regression model on the data. What happens? Report any error message here.

c. To understand why this happen, we need to understand conceptually what is happening with our logistic regression model. In our setup Y is binary variable that is either red ($Y = 1$) or blue ($Y = 0$). Our model is estimating:

$$P(Y_i = \text{red}|x_i) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \quad \text{and} \quad P(Y_i = \text{blue}|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

for all $i = 1, 2, 3, 4, 5$. What value(s) of β_0 and β_1 would maximize the likelihood (and therefore be the estimates we would get from fitting this model)? Recall that our likelihood looks like:

$$l(\beta_0, \beta_1, X) = P(Y_1 = \text{red}|\beta_0, \beta_1, x_1) \times P(Y_2 = \text{blue}|\beta_0, \beta_1, x_2) \times \ldots \times P(Y_5 = \text{blue}|\beta_0, \beta_1, x_5).$$

Hint: What is $P(Y_i = \text{blue}|x_i > 4)$? Now what is the $P(Y_2 = \text{blue}|x_2 = 5)$? What values of β_0 and β_1 will get us close to this probability?

d. Putting all this together, explain one limitation of the logistic regression model.