ESE 542: Statistics for Data Science

Chapter 4 Project

In addition to answering the **bolded** questions on Coursera, also attach your notebook, both as `.ipynb` and `.html`.

This project should be answered using the `Weekly` data set (attached). This data contains 1,089 weekly stock market percentage returns for 21 years, from the beginning of 1990 to the end of 2010. Details about the columns in the data are summarized below:

- Year : The year that the observation was recorded
- Lag1 : Percentage return for previous week
- Lag2 : Percentage return for 2 weeks previous
- Lag3 : Percentage return for 3 weeks previous
- Lag4 : Percentage return for 4 weeks previous
- Lag5 : Percentage return for 5 weeks previous
- Volume : Volume of shares traded (average number of daily shares traded in billions)
- Today : Percentage return for this week
- Direction : A factor with levels Down and Up indicating whether the market had a positive or negative return on a given week

1 Part A

We are first interested in trying to predict the direction of the returns.

1. Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be any patterns?
2. Use the full data set to perform a logistic regression with 'Direction' as the response and the five lag variables as predictors.

3. Use the `summary()` function to print the results. Do any of the predictors appear to be statistically significant? **Answer on Coursera: Which predictors appear to be statistically significant?**

4. Compute the overall fraction of correct predictions. Name this variable `fraction_correct_all`. **Answer on Coursera: What is the overall fraction of correct predictions?**

5. Now fit the logistic regression model using a training data period from 1990 to 2007, with 'Lag2' as the only predictor. Compute the overall fraction of correct predictions for the held out data (that is, the data from 2008, 2009 and 2010) and assign it to a variable called `fraction_correct_test`. **Answer on Coursera: What is the overall fraction of correct predictions?**

2 Part B

Now, we want to develop an investment strategy in which we buy if the returns are greater than 0.5% and sell otherwise.

1. Create a response variable, y_i such that

$$y_i = \begin{cases} 1 & \text{if } Today > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

2. Fit a logistic regression model using a training data period from 1990 to 2008, with the five lag variables and volume as predictors.

3. Use the `summary()` function to print the results. Do any of the predictors appear to be statistically significant? **Answer on Coursera: Which predictors appear to be statistically significant?**

4. Compute the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010). Assign this value to the variable `fraction_correct`. **Answer on Coursera: What is the overall fraction of correct predictions?**