2024 Chicago Booth Big Data Midterm

- This is an **INDIVIDUAL** exam: you must work alone.
- The exam must be submitted before **noon on Wednesday April 24th**.
  This deadline is the same for all sections.
- Questions are ordered by class subject, not by difficulty.
- Coding questions should be asked on Piazza, but you must not give away any answers when you formulate your query. We will only answer questions on R coding – not statistics concepts. We will continue to answer conceptual questions if they pertain to homework solutions or lecture examples.
- You will be graded on your written answers and analysis. Please include and refer to plots to illustrate your answers.
- Please label your graphs and figures (e.g. "Figure 1: Relationship between A and B."), and do not forget to label the axes on these graphs.
- Do not hand in detailed code, but do submit meaningful R output to help us understand your answer (e.g., printed $\hat{\beta}$ values, etc). All but immediately relevant R output should be relegated to an appendix.
- Please be very clear: your answers need to be concise, precise, and easy to understand. Neatness in presentation is expected.

Daily News for Stock Market Prediction

Data and code are on the Canvas site in `midterm.zip`. This includes `midtermStart.R`. Much of the code you will need has already been written for you in this script.

We will study whether stock prices movements are partially attributable to the arrival of news. To this end, we look into the content of 73 606 headlines from the Reddit World News Channel in between 2008-06-08 to 2016-07-01. Only up to the top 25 headlines are considered for each single day. We decompose the headlines into words.

For the same time period, we follow the prices of Dow Jones Industrial Average (DJIA) (downloaded from Yahoo Finance). The data consists of the following files:

REDDITNEWS.CSV is a table with 73606 rows and three columns containing headlines (one row for each headline) for 1989 days in between 2008-06-08 and 2016-07-01. The three columns correspond to

- Id: reddit news id
- Date: day of the headline
- Text of the headline

DJIA.CSV is a table consisting of 1989 rows (one for each day) and 7 columns containing opening, (adjusted) closing, maximal and minimal prices that day.

WORDFREQFINAL.CSV is a simple triplet matrix of 259132 day/word pairings. The table consists of 3 columns

- Day of the headline (a row index of RedditNews.csv)
- Word used (a row index of WordsFinal.csv)
- Times Word: how many times the word occurred in the headlines that day

WORDSFINAL.CSV contains 5271 alphabetically ordered words that occur in the headlines. See the `DataPreparationFile.R` script to see how the data was prepared.

1 Marginal Significance Screening and False Discovery

To begin, we would like to pre-screen words that associate with DJIA returns and volatility. We will look at one-at-a-time regressions (regressing outcome on each word separately) to see if each individual word can predict returns and/or volatility.

We will look at two outcomes:

(1) The first is the **one-day return** R_t defined as

$$R_t = \frac{Y_t - Y_{t-1}}{Y_{t-1}}$$

where Y_t is the closing price (adjusted for splits and dividends) on day t.

(2) The second outcome is the **volatility** V_t defined here as the logarithm of the difference between the maximal and minimal price that day:

$$V_t = \log(\max_s Y_{ts} - \min_s Y_{ts})$$

where Y_{ts} is price at time s during day t.

In the starter script, you will find a code to run the individual regressions (using parallel computing). The code gives you a set of p-values for the marginal effect of each of the words. That is, we fit

$$R_t = \alpha + \beta_j I[x_{tj} > 0] + \epsilon_{tj}$$

for each word j with count x_{tj} in headlines on day t, and return the p-value associated with a test of $\beta_j \neq 0$. We do the same for V_t.

(1.1) Plot the p-values and comment on their distributions (for both outcomes V and R). Is there enough signal to predict prices and volatility? (11 points)

(1.2) What is the alpha value (p-value cutoff) associated with 10% False Discovery Rate? How many words are significant at this level? (Again, analyze both outcomes V and R.) What are the advantages and disadvantages of FDR for word selection? (11 points)

(1.3) Now, focus only on volatility V. Suppose you just mark the 20 smallest p-values as significant. How many of these discoveries do you expect to be false? Are the p-values independent? Discuss. (12 points)

2 LASSO Variable Selection and Bootstrapping

We want to build a LASSO predictor of Returns (R) and Volatility (V) using headlines.

(2.1) Use the LASSO method to come up with a combination of a few words that predict returns R. Pick a lambda and comment on the in-sample R^2 . Is there enough evidence to conclude that headlines predict returns? (11 points)

(2.2) Repeat the analysis from (2.1) to predict Volatility instead of Returns. Next, add one extra predictor, Volatility on a previous day. In other words, fit a LASSO model for predicting today's volatility V_t using word counts and yesterday's volatility V_{t-1}

$$V_t = a + bV_{t-1} + x_t'\beta + \epsilon_t.$$

What is the in-sample R^2 now? Describe the LASSO path and pick the top 10 strongest coefficients. What is the interpretation of the coefficient of word "terrorist" and of V_{t-1}? (11 points)

(2.3) Consider the estimated of lambda selected by AICc in the model from (2.2) (using both words and V_{t-1}). We want to know how variable this estimate is. The starter script has code to bootstrap the sampling distribution for the λ selected by AICc in this regression.

- What is the standard error for the selected λ?
- Find the 95% CI for λ?

(11 points)

3 High-dimensional Controls and Double LASSO

We want to isolate the effect of yesterday's Volatility on today's Volatility, controlling for all relevant words. Our *treatment variable* will be $d = V_{t-1}$.

(3.1) Explore a marginal regression (just a regression of V_t on V_{t-1}) to see if there is any correlation. Predict d from x (headlines words), and comment on the degree of confounding we can expect. Is there any information in d independent of x? (11 points)

(3.2) Isolate the effect of V_{t-1} by running the causal (double) LASSO. Interpret this effect and compare it to the effect obtained from the naive LASSO. (11 points)

(3.3) Can we safely claim that the effect is causal? (11 points)

Bonus Freestyle Analysis

Provide additional analysis of the data. Points will handed out only for insightful use of data mining tools, not for scattershot application of techniques.

USE A MAXIMUM OF TWO PAGES, INCLUDING PLOTS, TO ANSWER.