STAT 542, Homework 4

November 30, 2017
Due date: Dec 17 (Sun), 11:59 pm to Compass

Requirements: You should submit your report and R code(s), preferably in separate files. Your report should be in PDF/MS Word format. Font size should be 12pt and plots need to be clearly labeled. Your report should include necessary explanations and should not be a simple output file of the R code. The R code should include comments to help our grading process. There is a 15 page limit to the report. This homework worth 100 points total. Late submission penalty is 5 points for each day (round up) of delay.

Question 1: [15 points] Prove the property used in the block matrix inverse form — an alternative version of the Sherman-Morrison formula: If $A \in \mathbb{R}^{n\times n}$ is an invertible square matrix, $b \in R^n$ is a column vector. If $A - bb^{\mathrm{T}}$ is invertible, show that its inverse is given by:

$$\left(A - bb^{\mathrm{T}}\right)^{-1} = A^{-1} + \frac{A^{-1}bb^{\mathrm{T}}A^{-1}}{1 - b^{\mathrm{T}}A^{-1}b}$$

Hint: to show $A^{-1} = B$, verify that $AB = I$.

Question 2: [30 points] Write your own code to fit the sliced inverse regression, validate it by comparing to the "`dr`" package. Use use 10 as the number of slices. Then perform the following:

a) [10 points] Generate 1000 observations using an underlying model that **can** be detected by SIR. Compare your estimated direction with the truth.

b) [10 points] Generate 1000 observations using an underlying model that **cannot** be detected by SIR. Compare your estimated direction with the truth.

For both questions, you **cannot** use the model that I used in the `SIR.r` file. You should set seed so that the result is replicable.

Question 3: [55 points] Download the tmdb movie dataset from Kaggle.

https://www.kaggle.com/tmdb/tmdb-movie-metadata

Our goal is to predict two variables `revenue` (a regression problem) and whether the `vote_average` is greater than 7 (a classification problem). You cannot use them as one of the covariates to predict each other. We mainly use the "`tmdb_5000_movies.csv`" file. You can choose to ignore the other file, and this choice will not impact your grade. Use all odd `id` as training data and even `id` as testing data. You are not allowed to use any information after the release date, for example, `popularity` and `vote_count`. There is no restriction on what method to use, however, you need to report the following for each outcome:

a) [5 points] An accurate description of your modeling strategy with no formula or code. You should include, for examples, model and feature selections, tuning parameters, feature engineering, etc.

b) [15 points] A detailed summary of your findings that includes at least your final model, prediction errors, and the most important predictor(s).

c) [5 points] All corresponding code should be included in the .rmd file.

[5 points] Finally, the movie "Star Wars: The Last Jedi" is set to release at Dec 15th. Find its information from `imdb.com` so that you can apply your model to predict its total revenue and whether the rating is greater than 7.