Midterm Two - Statistics 153, Fall 2017

Due on November 16, 2017

November 2, 2017 (Updated November 7, 2017)

On piazza, you will find five time series datasets: `q1_train.csv`,`q2_train.csv`,`q3_train.csv`,`q4_train.csv` and `q5_train.csv`. Each of these datasets is of length 584 and gives the google trends-like data for queries from the first week of January, 2004 to the second week of March, 2015. Your task is to predict the next 104 observations (2 years) of these time series.

You are encouraged but not required to work on all the five datasets. You are required to work on at least two of the datasets.

One member of your team must turn in the following. Please make sure the same team member turns in all parts:

1. Your predictions for the datasets that you have worked on. **These are due by 11:59PM on November 16, 2017**. You are required to turn in a txt-file for each data set. It should be named `Q[Number]_[Firstname]_[Lastname]_[SID].txt`. For example, `Q1_Alex_DAmour_123456.txt`.

   The text file should contain your predictions and a 95% prediction interval for each of the following 104 time points. The file should have 104 lines. Each line should contain three numerical values separated by ",". The first value should be the lower bound of your predictive interval, followed by your prediction, followed by the upper bound of your predictive interval. We assume that you will submit your values in an increasing time order:

   $$\hat{X}_{585}, \hat{X}_{586}, \ldots, \hat{X}_{689}$$

   Please be aware that your submission must be of the right form in order to be valid.

   You will submit your work via bCourses. We will provide more details on the submission protocol as the due date nears.

2. A report describing your analysis. For one of the datasets that you have worked on, write a clean report describing your analysis attaching the relevant plots and R output. Include your R code as an Appendix to the report. Do not write a report for each of the datasets that you worked on. Just write it for one of those datasets and include your R work for the other datasets as an Appendix. The length of the report including the relevant plots and R output (and excluding the R code) cannot exceed 8 pages. **Make sure to include both team members' names on the report.**

You will be graded on prediction accuracy, interval coverage, and your report (**the report will be for 30 points, prediction accuracy will be graded to a maximum of 8 points,a nd coverage will account for 2 points**).

Here is a description of how your prediction accuracies and interval coverages will be evaluated. Suppose you decide to submit predictions for the dataset **q1**. Let your predictions be denoted by $\hat{X}_{585}, \ldots, \hat{X}_{689}$ and let the true values of **q1** (which we will have access to) are $X_{585}, \ldots, X_{689}$. We will first compute the sum of squares

$$\sum_{i=585}^{689} \left(\hat{X}_i - X_i\right)^2$$

This result will measure your discrepancy for **q1**. From here, we will compute:

$$4 * \frac{\text{best discrepancy for } \mathbf{q1} \text{ in the class}}{\text{your discrepancy for } \mathbf{q1}}$$

This will be your score for **q1**. Note that the maximum possible value for this score is 4. The minimum possible score is 0 (this will be the case if the best in-class discrepancy is zero). We will similarly compute your score for each of the datasets that you submit predictions for. To get your final points for the prediction part, we will take the sum of your highest two scores. (For example, if you submit predictions for four datasets and your scores are 3.1, 3.8, 2.7, 3.9; then you will get 3.8 + 3.9 = 7.7 points out of 8 for the prediction part).

For the interval coverage portion, we will assess whether your prediction interval contains the true values as frequently as it should. Note that if the prediction interval is valid, we would expect 95% of the true values to lie within the corresponding prediction interval. Let your upper- and lower-bound estimates be denoted $U_{585}, \ldots U_{689}$ and $L_{585}, \ldots, L_{689}$, respectively. We will compute the absolute difference the coverage rate of your interval with 0.95.

$$1 - \left| \left(\frac{1}{104} \sum_{i=585}^{689} \mathbf{I}\{L_i < X_i < U_i\} \right) - 0.95 \right|.$$

This will be your score for **q1**. Note that the maximum possible value for this score is 1, and the minimum possible score is 0.05. As in the last part, we will add together your coverage scores from your two highest-scoring question submissions. The submissions chosen to compute your prediction score and your coverage score may be different. We reserve the right to dock additional points here if your intervals are clearly gaming the scoring system.

As you can tell by the relative weights of the questions, the most important part of the assignment is the writeup, followed by predictive accuracy. The coverage check is to incentivize you to keep from making overconfident predictions.

You may work with a partner, but your group **may not** collaborate with other groups. You are allowed to use code from the lectures and the section without explicit citation. You are also allowed to consult books or online resources for your analysis but you must credit all such sources in your report. Anyone caught cheating (which includes copying code, reports etc.) risks failing the class and being referred to the Office of Student Conduct.