

Midterm Two - Statistics 153, Spring 2016

Due on April 19, 2016

08 April, 2016

On piazza, you will find five time series datasets: `q1train.csv`, `q2train.csv`, `q3train.csv`, `q4train.csv` and `q5train.csv`. Each of these datasets is of length 536 and gives the google trends data (downloaded on 11 April, 2015) for a particular query from the first week of January, 2004 to the week of 06 April– 12 April, 2014. Your task is to predict the next 104 observations 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.

You need to turn in the following:

1. Your predictions for the datasets that you have worked on. **These are due by 9:30 am on 19 April, 2016.** You are required to turn in the predictions on bcourses as a txt-file for each data set. The text file should contain your predictions for the following 104 time points separated by “,” or “newlines” and it should be named `Q[Number]_[Firstname]_[Lastname]_[SID].txt`. As an example, Soeren has posted a sample submission for the dataset **q1** (this submission contains only 0s) on piazza: it is named `Q1_Soeren_Kuenzel_123456.txt`. We assume that you submit your values in an increasing order:

$$\hat{X}_{537}, \hat{X}_{538}, \dots, \hat{X}_{640}$$

On 14 April, you will have the opportunity to submit a test submission for each data set. This is not required and won't affect your grade, but this is an opportunity for you to see whether your submission is in the right form. Please be aware that your submission on 19 April, 2016 must be of the right form in order to be valid. You are encouraged to also look at Soeren's section code from section 11: `SorenSub.R` which is creating a valid submission for a different data set.

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. **The printed reports are due at the beginning of class at 9:30 am on 19 April, 2016.**

You will be graded on the prediction accuracy as well as your report (**the report will be for 25 points and the prediction accuracy will be graded to a maximum of 10 points**). Here is a description of how your prediction accuracies will be evaluated. Suppose you decide to submit predictions for the dataset **q1**. Let your predictions be denoted by $\hat{X}_{537}, \dots, \hat{X}_{640}$ and let the true values of **q1** (which we will have access to) are X_{537}, \dots, X_{640} . We will first compute the sum of squares

$$\sum_{i=537}^{640} (\hat{X}_i - X_i)^2$$

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

$$5 * \frac{\text{best discrepancy for } \mathbf{q1} \text{ in the class}}{\text{your discrepancy for } \mathbf{q1}}$$

This will be your score for **q1**. Note that the maximum possible value for this score is 5. 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 4.1, 4.8, 3.7, 4.9; then you will get $4.8 + 4.9 = 9.7$ points out of 10 for the prediction part).

You should work completely on your own. 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.