Lab 03 STATS 769 Due 4:00pm Monday 19 August

The main point of this lab is to **work effectively in Linux**. We will again generate a literate report that describes a simple analysis, but this time in addition to fitting a simple linear regression predictive model, we will use k-fold cross-validation to evaluate it (see the section at the end of this document) and we will compare it to a polynomial regression model (and we will use a slightly larger data set, though still in a simple format).

The Data

We will work with eleven CSV files, `trips-2018-4.csv` to `trips-2019-2.csv`, which are located in the directory `/course/Labs/Lab02/` on the STATS 769 virtual machines.

Each of these files contains data on 5000 trips on electric bikes or scooters in Austin, Texas, USA. Each row represents a trip and the file `trip-variables.csv` describes the variables measured for each trip.

The Task

1. Use Linux tools to count the number of bicycle trips in each of the CSV files.
2. Use Linux tools to extract just scooter trips from each CSV file. For each original CSV file, generate a new CSV file with just scooter trips in it. For example, for the file `trips-2018-4.csv` in the `/course/Labs/Lab02` directory, generate a file `scooter-trips-2018-4.csv` in the current working directory.
3. Import the scooter trip CSV files into R and combine them into a single data frame.
4. Transform the data by removing trips with a distance or duration that is non-positive, then log both the distance and duration variables.
5. Fit a simple linear regression model via k-fold cross-validation (with k = 10) to predict the trip duration based on trip distance and measure the test MSE.
6. Fit a polynomial regression model via k-fold cross-validation to predict the trip duration based on trip distance and the square of trip distance and measure the test MSE.
7. Provide a plot showing the two models (fitted to all of the data).
8. Write a Makefile so that you can process your report just by typing `make` *and* so that nothing happens if the processed file is newer than the R Markdown file.

The Report

Your submission should consist of an R Markdown document *and* a Makefile submitted via Canvas.

You should write your document and your Makefile so that the markdown document can be processed just by typing `make` anywhere on one of the virtual machines provided for this course (`sc-cer00014-04.its.auckland.ac.nz` or `sc-cer00014-05.its.auckland.ac.nz`).

Please also submit a processed version of your R Markdown document (PDF or HTML) in case the `make` does not work.

Your report should include:

- A description of the data format.
- Shell code to count bicycle trips and extract only scooter trips from the data files.
- How the data were imported to R, and any transformations applied to the data.
- k-fold cross-validation of a simple linear regression model.
- k-fold cross-validation of a polynomial regression model.
- An explanation of your Makefile for processing the report.
- A conclusion summarising the analysis.

Your report should NOT be longer than **10 pages**.

K-fold cross-validation

The "validation set" approach that we have used previously splits the data into a training set and a test set. The model is fit on the training set and then we calculate a "test error" using the test set:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

where y_i are the Y values from the test set and $\hat{f}(x_i)$ are the predictions of the model based on the X values from the test set.

Two weaknesses of this approach are that we do not use all of the data to fit the model *and* we only get one estimate of the test error.

K-fold cross validation attempts to ameliorate both of those problems:

- Divide the data into k (typically 5 or 10) sets.
- For set i, fit a model using all other $k-1$ sets, then calculate the test error using set i.
- Repeat for all k sets, then average the k test errors.

This means that all of the data are used in $k-1$ of the model fits *and* we get k estimates of the test error, so our overall average test error is a more reliable estimate.
