Department of Statistics
STATS 330: Statistical Modelling
Assignment 2
Semester 1, 2020

Total: 75 marks Due: 3:00pm NZDT, Friday 1 May 2020

Notes:

(i) Write your assignment using R Markdown. Knit your report to either a Word or PDF document.

(ii) Create a section for each question. Include all relevant code and output in the final document.

(iii) 5 presentation marks are available. Please think of your markers - keep your code and plots neat and remember to check your spelling. (`R` has an inbuilt spellchecker!)

(iv) Please remember to upload your Word or PDF document to Canvas by the due date.

(v) Please remember to upload your R Markdown file to Canvas before the deadline, too. If the markers identify an error in your work, being able to run the code you have written can help determine what you did wrong.

Introduction

NYC Open Data is free public data published by New York City agencies and other partners. Bicycle counts[1] are routinely conducted around New York City key locations. To keep count of cyclists entering and leaving Queens, Manhattan and Brooklyn via the East River Bridges, a Traffic Information Management System (TIMS) collects data. Each record represents the total number of cyclists per 24 hours at four New York bridges: Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge and Queensboro Bridge.

[1] See `https://catalog.data.gov/dataset/bicycle-counts-for-east-river-bridges`.

Figure 1: Brooklyn Bridge

You have access to bicycle counts conducted over a seven month period (1st April 2017 to 31st October 2017) You can find the data in a file called `NYBridges.csv` on CANVAS (Assignments > Assignment 2).

The data set `NYBridges.csv` contains the following variables:

- `Date`: the date of the bicycle count (between 1st April and 31st October 2017)
- `Day`: day of the week
- `High.temp`: the maximum temperature recorded on that day (in degrees Fahrenheit)
- `Low.temp`: the minimum temperature recorded on that day (in degrees Fahrenheit)
- `Precipitation`: the rainfall (in inches)
- `Brooklyn`: daily bicycle count for Brooklyn Bridge
- `Manhattan`: daily bicycle count for Manhattan Bridge
- `Williamsburg`: daily bicycle count for Williamsburg Bridge
- `Queensboro`: daily bicycle count for Queensboro Bridge

(1) **Communication.**

The communication of results from statistical analyses relies on using language and terminology that is easy to understand. In order to prepare this dataset, we need to think about how we would like to communicate our findings here in New Zealand.

(a) The temperatures in the data set (`High.temp` and `Low.temp`) are recorded in degrees Fahrenheit, while in New Zealand we tend to talk about temperature in degrees Celsius. Create and calculate two new variables, `HighC` and `LowC`, to represent the maximum and minimum temperatures in degrees Celsius.

(b) The rainfall in the data set (`Precipitation`) is recorded in inches while in New Zealand we are more likely to talk about rainfall in millimetres. Create and calculate a new variable, `Rainmm`, to represent rainfall in millimetres.

(c) The variable `Day` is a factor variable. Recall that, by default in R, the levels of a factor variable are sorted alphabetically, so the baseline level for `Day` is currently Friday. Reorder the levels of `Day` to something more meaningful and provide a justification.

(d) You may have noticed that some of the `Precipitation` observations have been recorded as 'T' which stands for 'trace'. Find out what 'trace' precipitation means. Recode values of 'T' to something more meaningful making sure you justify your actions.

(10 marks)

(2) **Brooklyn Bridge.** A statistician proposes the following four models using the same combination of explanatory variables.

- Model A: a linear model.

```
model.lin.a<-lm(Brooklyn~log1p(Rainmm)+Day+HighC+I(HighC^2),
                data=NYBridges.df)
```

- Model B: a Poisson model.

```
model.pois.b<-glm(Brooklyn~log1p(Rainmm)+Day+HighC+I(HighC^2),
              family=poisson,
                data=NYBridges.df)
```

- Model C: a quasi Poisson model.

```
model.qpois.c<-glm(Brooklyn~log1p(Rainmm)+Day+HighC+I(HighC^2),
              family=quasipoisson,
                data=NYBridges.df)
```

- Model D: a Negative Binomial model.

```
library(MASS)
model.nb.d<-glm.nb(Brooklyn~log1p(Rainmm)+Day+HighC+I(HighC^2),
                data=NYBridges.df)
```

By examining exploratory plots of the data (note: you are *not* required to fit any of the models here), explain why the statistician has:

(a) only included one temperature variable in her models.

(b) used the log1p transformation on the rainfall variable in her models.

(c) included a quadratic effect for temperature in her models.

(10 marks)

(3) Explore the four models described in (2) and assess each for goodness-of-fit using all the relevant methods covered in Handouts 5, 6, 7 and 9. For each model, do the following:

(a) Create a new section with an appropriate heading so that the markers can easily navigate through your report. For example: **'Model A exploration'**.

(b) Include `R` code with any plots or output. Try to keep your plots close together (i.e., explore the `par(mfrow=c(?,?))` options in `R`.

(c) Write a paragraph summarising what the methods tell you about the model.

(40 marks)

(4) Briefly summarise your findings. Which of the four models do you think is best? Why? Do you think your chosen model fits the data well? Are there any additional variables you would consider including in a model that aims to explore the relationship between meteorological information and bicycle activity? Make sure you justify your answers.

(10 marks)

Presentation: (5 marks)