DATA7202 Assignment 2 (Weight: 30%)

Due: Tuesday 12 May 2020.

Part I Re-analysis of the UCI online news popularity dataset

Reconsider the dataset from assignment 1.

Here you will fit a new model to the data to predict number of shares for a new article using the same explanatory variables and data as you used in assignment 1. However, instead you will use a generalised linear model and a non-negative distribution and a suitable link function. Note: due to a question asking about a predictive interval, we recommend you do not attempt to use the quasi-Poisson model.

1. Choose a suitable generalised linear model and suitable link function and explain these decisions. Fit the model to the data and report summary results. [2 marks]

2. Produce a 95% predictive interval for each fitted model (ignore uncertainty with respect to model parameters). Compare the results from using this model with those from using a multiple linear regression model (= general linear model) which you would have run for assignment 1. Include consideration of significant variables, measures of goodness of fit, predictive intervals and related plots. Discuss theoretical and practical differences between the two models. [4 marks]

Part II

Translink is a division of the Queensland government's Department of Transport and Main Roads. They operate the go card system which allows card-based access and payment for South East Queensland public transport systems, including buses, trains, ferries and trams. Some of the data collected via go cards has been analysed by UQ's Dr. Jiwon Kim of the School of Civil Engineering (see https://www.jiwonkim.co/). Translink is interested in monitoring and predicting passenger numbers and demand to help with planning changes to routes and their frequency, and how to best respond to unusual incidents and events.

A subset of the data is available to you via Blackboard covering passenger flow data from 25 Feb to 24 Mar 2013. Your main task is to construct predictive models based on this data for a single high-traffic pair of regions, namely from region 1 to region 5 (Brisbane City to South Brisbane).

You will create two types of model. The first model should be based entirely on reasonable summaries (e.g. averages) from past data, without a stochastic model (see for example https://otexts.com/fpp2/simple-methods.html). The second should involve the fitting of a stochastic time series model (or a comparable model, if given permission by the lecturer). The modelling process, choice of data to use and how to represent it, is up to you.

Please only consider the v0_num_traj column (not the v10 – v60 columns) – assume this is the number of passenger trajectories (trips using all forms of public transport) occurring during the listed time period (hour) for the listed route. The time_id shows the hour since midnight, making the start time and end time redundant, since these are just in minutes since midnight.

The detailed tasks are given below. Assessment will consider clarity of exposition, presentation and statistical reasonableness. Where possible, give reasons for your decisions and some discussion of results.

3. Check for stationarity and seasonality of the time-series data. Explain how you have done this and include relevant graphs and numerical summaries. [4 marks].

Note: you have a number of tricky issues to deal with, including the presence of weekends, and the fact that the range of hours in which people take trips differs from day to day. With respect to the latter, the main interest of Translink is in the busier periods of the day. So, if you wish, you can use only the time periods which are present for every day in the data.

This is linked to the ideas of structural and measured zeroes. In this case, a structural zero is a zero count when it was not possible for any counts to be recorded, e.g. because there were no services running. A measured zero could occur when services were available, but no one used them. In this dataset, zero counts are entirely missing from the dataset. For such missing times, we cannot know whether they were structural or measured zeroes, although presumably Translink staff could find out.

4. Choose and detail each type of model used (including mathematical form and explanation of notation) and some details of how it was fitted. Explain why each model may be suitable for this type of data. [4 marks]

5. Report full details of each fitted model. [2 marks]

6. Discuss the limitations of each model with respect to this dataset. [3 marks]

7. Give and plot model predictions for each model over the observed data range (include the observed values somehow for comparison). Also give 95% predictive intervals for the stochastic model over this range. [3 marks]

8. Evaluate accuracy 1 and 2 hours ahead via final day. [3 marks]

As part of evaluation of the models, evaluate the accuracy of each model's predictions ahead in time 1 and 2 hours. Utilise only the last day of data for testing. However, you should use all the available data available for prediction, so if predicting ahead 1 hr to 3pm, you should train your model on data up to 2pm. Similarly for a 1 hr prediction to 4pm, you should training your model on data up to 3pm. So you will have to produce a number of models for the purpose of testing. It is possible you will find a way to re-use models without full retraining, but either approach is ok. Suggested accuracy measure: mean squared error. Give details of how you calculate this.

9. Make predictions for test day, which will be evaluated by RMSE vs true counts for that day (held by lecturer). [3 marks]

You will need a version of at least one of your models to predict into the future. We provide data below on counts of trips for the first few hours of travel on Monday 25/3/2013. You should make predictions for the remainder of the day. These predictions will be compared against the truth in marking – some marks depend upon the accuracy of these predictions.

The following data is from Monday 25/3/2013 from 5am to 9am:

Date	time_id	start_time	end_time	region_from	region_to	v0_num_traj
25/03/2013	6	300	360	1	5	2
25/03/2013	7	360	420	1	5	18
25/03/2013	8	420	480	1	5	34
25/03/2013	9	480	540	1	5	68

Make predictions of the passenger counts over this route for each hour for the remainder of this day's services (until midnight) using any of your models. You can use different models for different time periods if you wish. Some of the marks for this question will be based on the mean squared error of your predictions. Please list your predictions in a column for time id 10-24 (15 rows, 1 predicted count per line, no other characters) in your report. We will copy them out and evaluate them.

10. Paragraph to Translink staff [2 marks]

Include a paragraph aimed at a member of the Translink planning staff who may not have a statistics background, explaining how your modelling could potentially help them make decisions about how many bus/train/ferry services to run at various times to meet demand.

Notes:

Please store all the R commands you use and submit these via a separate file (This can be within R markdown or a Jupyter notebook if desired – just make sure we can read the R code). Please include your name in the filename for all files submitted. You should not generally give R commands in your main report and should not include any raw output – i.e. just include figures from R (each with a title, axis labels and caption below) and put any relevant numerical output in a table or within the text.

As per http://www.uq.edu.au/myadvisor/academic-integrity-and-plagiarism, what you submit should be your own work. Even where working from sources, you should endeavour to write in your own words. Equations are either correct or not, but you should use consistent notation throughout your assignment, define all of it and ensure that your report flows logically.

You are asked to use the R software environment for this assignment. This is free to install on any of your own computers. Information and downloads are available from http://www.r-project.org/ . Rstudio https://www.rstudio.com/ is a quality free interface for R.

Submit your assignment report as a pdf file via TurnItIn on Blackboard. Any R programs or scripts that you write to answer the assignment should be placed in a separate .zip file and uploaded as a second file (see Blackboard).

References:

A. Agresti, *Categorical Data Analysis*, 2nd edition, Wiley, 2002. (chapter 4 provides a readable introduction to GLMs)

A. J. Dobson, and A. Barnett, *An Introduction to Generalized Linear Models*, 4th edition, CRC Press, 2018. (Main course reference for GLMs)

J. J. Faraway, *Extending the Linear Model with R*, 2nd ed., CRC Press, 2016. (GLMs with R examples and discussion)

K. Fernandes, P. Vinagre, and P. Cortez, *A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News*, in: Pereira F., Machado P., Costa E., Cardoso A. (eds.) Progress in Artificial Intelligence, EPIA 2015, Lecture Notes in Computer Science, vol. 9273, Springer, 2015. (the paper which first analysed the data used in Assignment 1)

R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice* 2nd ed., OTexts, Melbourne, 2018. https://otexts.com/fpp2/ (Accessible text on time series analysis, esp: forecasting)

J. Maindonald and J. Braun, *Data Analysis and Graphics Using R - An Example-Based Approach*, 3rd edition, Cambridge University Press, 2010. (Intro to R with wide range of examples)

R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications with R Examples*, Fourth Edition, Springer, 2017. (Detailed book on time series analysis, with R)

W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, Fourth Edition, Springer, 2002. (Classic book introducing S (the language implemented in R) and applications)

H. Wickham and G. Grolemund, *R for Data Science*, O'Reilly, 2017. http://r4ds.had.co.nz (mainly about Hadley Wickham's R packages, but also a good general introduction to using R for data science)