MATH 4044 – Statistics for Data Sciences

Assignment 1 SP5 2022

Due Sunday 25 Sep 2022

Instructions

- This assignment is worth 25% of your final mark. It is due no later than **Sunday 25 Sep**, first week after the mid-break.
- You will need to submit your assignment via learnonline.
- The submitted assignment needs to be a **single file**, in either a Microsoft Word (doc or docx) or pdf file format, **25 pages at most excluding any appendices**.
- The assignment is out of 100 marks. To achieve maximum marks for each question, you should aim to:
  - Complete the requested statistical analysis in SAS using appropriate tasks or procedures (40%).
  - Include only the output most relevant to the question and interpret all key results (40%). Do not include every piece of output produced by SAS!
  - Discuss the results more broadly in the context of the given scenario (20%).
- Assignments submitted late, without an extension being granted, will attract a penalty of 5 marks per each working day or any part thereof beyond the due date and time.

Data Description

Bike Sharing Systems

Bike sharing systems are a new generation of bike rentals where the whole process from membership, rental and return has become automatic. Through these systems, a user is able to easily rent a bike from a particular position and return the bike at another position. Currently, there are over 500 bike-sharing programs around the world, with some of the best and largest found in Hangzhou (China), Paris (France), London (England), New York City (US) and Montreal (Canada). Great interest in these systems exists due to their role in addressing traffic congestion, environmental impact and population health issues in big cities.

The data for this assignment comes from one such program, called Capital Bikeshare, operating in Washington in the US. It has over 3000 bicycles that can be rented from over 350 stations across Washington, D.C., Arlington and Alexandria, VA and Montgomery County, MD. Their website encourages users to check out bikes for a trip to work, to run errands, go shopping, or visit friends and family. Users can join Capital Bikeshare for one to three days (casual membership), or for a month or a year (registered membership). Access to the Capital Bikeshare fleet of bikes is available 24 hours a day, 365 days a year. The first 30 minutes of each trip are free.

You will use data derived from Capital Bikeshare trip records to build a statistical model for the purposes of predicting the total number of rentals per day.

References and Data Sources:

- Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository `http:`

`//archive.ics.uci.edu/ml`. Irvine, CA: University of California, School of Information and Computer Science.

- Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg.

- `http://capitalbikeshare.com/system-data`

Data file for this assignment

The data file for this assignment is called `daily.sas7bdat` and contains daily counts of bike rentals for 2011 and 2012, derived from Capital Bikeshare trip history data, with additional weather and seasonal information. The data was downloaded from the UCI Machine Learning Repository. Variables in that file are:

Variable	Description
`instant`	Record index
`dteday`	Date
`season`	winter, spring summer, autumn (northern hemisphere)
`yr`	0=2011, 1=2012
`month`	Month (January to December)
`weekday`	Day of the week (Monday to Sunday)
`workingday`	Working day=1, weekend and public holiday = 0
`temp`	Normalised temperature in degrees Celsius; observed temperature divided by 41 (max)
`atemp`	Normalised 'feels like' temperature in degrees Celsius; values divided by 50 (max)
`hum`	Normalised humidity; observed values divided by 100 (max)
`windspeed`	Normalised wind speed; observed values divided by 67 (max)
`casual`	Count of casual users
`registered`	Count of registered users
`count`	Total count of bike rentals (casual and registered).

Assignment Tasks

Question 1 (20 marks)

(a) **(10 marks)** Use SAS to study the distribution of the number of registered users per day (`registered`) by season. Obtain measures of location, dispersion, skewness and kurtosis. Obtain a boxplot, histogram and a quantile-quantile plot. Also carry out Normal Goodness-of-fit tests. What are the key features of these distributions?

(b) **(10 marks)** Now use SAS to obtain boxplots of `registered` by `season`, and by `yr`, respectively. Similarly, obtain boxplots of `casual` by `season` and `yr`. What do the boxplots suggests about the pattern and trend, if any, of bike rentals?

Question 2 (60 marks)

(a) **(8 marks)** Obtain a Pearson correlation matrix relating variables `registered`, `atemp, temp, hum` and `windspeed`. Also obtain a scatterplot matrix of the same variables. Discuss the relationships.

(b) **(12 marks)** In this question, we investigate observations where `workingday=1`. Fit a simple regression model relating `registered` on working days to `atemp`, with `registered` as the dependent variable. Discuss the fitted relationship and the goodness of fit. Examine residual plots and influence diagnostics and comment on the residual patterns.

(c) **(20 marks)** In this question, we investigate observations where `workingday=1`. Extend your multiple regression model for `registered` on working day by including the numerical and categorical predictors. In building your model consider as many potential explanatory variables as possible (you may need to define additional dummy variables). You can use stepwise selection to help you find the most parsimonious (simplest) model with the highest R-square. Be sure to check for collinearity and keep in mind that neither `casual` nor `count` should be used as explanatory variables for the total number of users.

Summarise how your final model was obtained, including rationale for any modelling decisions you have made, and indicate why that final was considered the 'best'.

Report and interpret your final model in detail, including a discussion of model diagnostics. Are there any observations that may require further inspection due to their influence on the model?

(d) **(20 marks)** In this question, we investigate observations where `workingday=0`. Build a multiple regression model for `registered` on non-working day, similar to question (c).

Summarise how your final model was obtained, including rationale for any modelling decisions you have made. Report and interpret the final model.

Compare and contrast the model with that obtained in question (c), and compare the effects of the predictors on `registered` for working and non-working day.

Question 3 (20 marks)

Write a summary of your findings from Questions 1 and 2. Keep the technical details of the analyses that led you to these conclusions to the absolute minimum. Rather, focus on practical significance and present your findings in non-specialist terms. One to two paragraphs (up to a page) will be sufficient.

Hints:

In order to study the regression for a specific group of observations, we can use the `where` statement. In particular, to build a regression model for working days, we can use

```
proc reg data=... ;
        model ... ;
        where workingday=1;
        run;
```