Queen Mary
University of London

MTH783P Time Series Analysis for Business Spring 2022
Project

The data for this project has been extracted from the dataset *Absenteeism at Work* in the UCL Machine Learning Repository, which was created from records of absenteeism at work at a courier company in Brazil.

This dataset contains the absenteeism time (in hours), together with information on transportation expense, distance from residence to work, service time and some personal details. Data are provided for about three years, from July 2007 to July 2010. The description of the variables is available in Table 1.

Table 1: Description of the variables

month	month (from 1: January to 12: December)
day	day (2: Monday, 3: Tuesday, 4: Wednesday, 5: Thursday, 6: Friday)
season	season (1: summer, 2: autumn, 3: winter, 4: spring)
transexp	transportation expense
distance	distance from residence to work (km)
servtime	service time (years)
age	age (years)
children	number of children
bmi	body mass index
absenttime	absenteeism time (hours)

The dataset is available on QMplus as *absenteeismatwork.csv*. Use R to analyse the dataset and address the following tasks.

1. (5 points) Split the data into two datasets: a training dataset and a test dataset. The training dataset should include the first 600 observations and the test dataset the remaining 137.

2. (25 points) Explore the training dataset: plot and produce summary statistics to identify the key characteristics of the data and produce a report of your main findings. The topics that you might choose to discuss include: possible issues with the data collection; identification of possible outliers or mistakes in the data; role of missing data (if any); distribution of the variables provided; relationships between variables.

3. Fit a statistical model to the training data and use it to forecast the absenteeism time for the test dataset.

   (a) (20 points) How did you decide which model to fit? Include details of other models that you tried, if any.

(b) (10 points) What are the underlying assumptions of the model that you have chosen? Carry out a residual analysis to ensure that the assumptions are satisfied.

(c) (10 points) Forecast the absenteeism time for the test dataset and discuss the results.

(d) (10 points) Discuss any weaknesses of this analysis.

4. (10 points) All tables and plots that you include in your report should be reproducible. Therefore, include in your submission on QMplus a text file with the R commands that can be used to reproduce your results, including tables and plots. This text file should include only those lines of code used to produce results presented in the report, and it should be written in a clear and readable way.

5. (10 points) Marks will be given for the overall presentation of the project, and the quality of figures and writing.

All modelling and forecasting choices and assumptions must be justified.

Requirements for the project submission:

- The submission deadline is **3pm on Friday 29 April**.
- The submission should include a pdf file containing the answers to Questions 2 and 3 (with a three-page limit, including figures and discussions) and a text file with the R code used for the results presented in the report. Minimum fontsize is 12.
- While discussing the project with your classmates is encouraged, the submission must be your own independent work. Every submission will be checked for plagiarism using an automated system. Please refer to the QMUL Academic Regulations for information about the definition of plagiarism and related penalties at `https://qmplus.qmul.ac.uk/mod/book/view.php?id=1723497&chapterid=135331`.
- The policy for late submissions in the School of Mathematical Sciences will be used. You can read the policy at `https://qmplus.qmul.ac.uk/mod/book/view.php?id=1723497&chapterid=135327`.