

Time Series (F70TS)
Academic Year 2021-22
Assessed Project

INSTRUCTIONS

Please read the instructions carefully before you commence doing the project.

In this project the student is required to use the *R* programming language to simulate and fit time series models together with some related issues including model selection criteria.

This project has a total of 30 marks and will count towards 30% of your final grade. The marks are apportioned as follows.

- Question 1 - 10 marks
- Question 2 - 14 marks
- Report Writing - 6 marks

REPORT WRITING

Your findings should be presented in the form of a report, which should:

- contain the answers to the each of the questions, in order, with the part and question numbers labelled;
- include details of your mathematical calculations (if any) and the statistical methods used so that your results could be reproduced by another statistician;
- include clearly labelled and correctly referenced tables and diagrams, as appropriate;
- do not include your R code, you will need to hand it in separately;
- maximum page limit of four (4) pages [2 sheets - front and back] (11-point font, A4 size). Figures/plots do not count towards the page limit.
- include citation and referencing for any material (books, papers, websites etc) used. When possible, use reliable sources, produced by respected and well-known authors, published by recognised publishers and associated with well established government, academic, or educational institutions. Note that some webpages, YouTube videos, blog posts, and Wikipedia pages might include errors.

A total of 6 marks is available for these aspects of your report. The 6 marks available for the exposition of your report will be awarded according to the scale below:

0-1 Marks

- Lack of clear and logical structure.
- Conclusions missing or not suitable for a non-statistician.
- Statistical calculations and Methodology not clearly set out for the reader.
- Tables and figures unclear, badly labelled or not correctly referred to.
- *R* codes not included, or no comments included in it.
- Sources used not clearly referenced.

2-4 Marks

- Clear and logical structure.
- Conclusions somewhat suitable for a non-statistician.
- Statistical calculations and Methodology generally set out clearly for the reader.
- Tables and figures often clear and correctly referred to.
- *R* codes included with some comments.
- Sources used clearly referenced.

5-6 Marks

- Clear and logical structure.
- Conclusions suitable for a non-statistician.
- Statistical calculations and Methodology set out clearly for the reader.
- Tables and figures clear, correctly referred to and easy to interpret.
- *R* codes included with comments.
- Sources used clearly and correctly referenced.

CODING

Part of this project are coding tasks in the *R* programming language. Aside from the report, we ask you to also submit an *.R-file that performs all of the required computations. Part of your submitted *.R-file will be marked automatically. Thus, it is necessary that you follow precisely the guidance given below; not following this guidance can lead to a deduction of marks:

- the *.R-file needs to be runnable;
- you may not use any (additional) libraries/packages;
- the *R* code needs to be carefully commented;
- for Q1a, Q2b, and Q2c, you are instructed to name the objects (vectors/lists/arrays/matrices) in your code as Q1a, Q2b, and Q2c, respectively. You need to follow this naming convention precisely.

IMPORTANT NOTES

- **Plagiarism** is a serious academic offence and carries a range of penalties, some very serious. Copying a friend's report or code, or copying text into your report from another source (such as a book or website) without citing and referencing that source, is plagiarism.

Collusion is also a serious academic offence. You must not share a copy of your report (as a hard copy or in electronic form) or your computer code with anyone else. Penalties for plagiarism or collusion can include voiding of your mark for the course.

Please refer to, <https://www.hw.ac.uk/students/studies/examinations/plagiarism.htm>, for more details.

- Your report should be a **.pdf** file while your code should be an **.R** file.

These files have to be submitted through CANVAS by **3.30 pm (UK time for Edinburgh based students)** and **3.30 pm (MALAYSIA time for Malaysia based students)** on or before **Friday 25 March 2022**. There are separate submission pages for the report (*.pdf) and the code (*.R). They are available through the 'Assignment' tab on the course CANVAS page.

For late submissions, 30% will be deducted for work submitted at most 5 days late while submissions that are more than 5 days late will receive 0 marks.

- Students can expect feedback after about 15 working days.

[PROJECT QUESTIONS]

1. Consider the time series model $X_t = \phi X_{t-12} + \epsilon_t$, where $\epsilon_t \sim WN(0, 1)$. Select the value of ϕ based on the last digit of your student ID number as follows.

| Last digit of student ID | Value of ϕ |
|--------------------------|-----------------|
| 0 | +0.1 |
| 1 | -0.1 |
| 2 | +0.2 |
| 3 | -0.3 |
| 4 | +0.4 |
| 5 | -0.5 |
| 6 | +0.6 |
| 7 | -0.7 |
| 8 | +0.8 |
| 9 | -0.9 |

- (a) Use the `polyroot(...)` command in *R* to compute the roots of the *AR* polynomial & name the output as **Q1a** and thereby determine whether this model is CAUSAL or NOT CAUSAL.

[2 marks]

- (b) Find the values of the theoretical ACF and theoretical PACF up to a maximum lag of 60 lags and plot them.

[2 marks]

- (c) Simulate 120 observations from the above model and plot the simulated series.

[2 marks]

- (d) Plot the sample ACF and sample PACF of the simulated series up to a maximum lag of 60 lags. Compare them with the theoretical ACF and PACF and give your comments.

[4 marks]

[10 marks]

[PLEASE TURN OVER]

2. The table below contains five(5) different files of 1000 sample time series observations. In this question, you would select one of the following files based on the last digit of your student ID number to answer Q2.

| Last digit of student ID | File name |
|--------------------------|--------------------|
| 0 or 1 | assign_data_01.csv |
| 2 or 3 | assign_data_23.csv |
| 4 or 5 | assign_data_45.csv |
| 6 or 7 | assign_data_67.csv |
| 8 or 9 | assign_data_89.csv |

- (a) Estimate the sample mean \bar{x} and the sample autocovariance function (ACVF) $\hat{\gamma}(k)$ of $\{x_1, x_2, \dots, x_{1000}\}$ for lags $k \in \{0, \dots, 19\}$. Plot the sample ACVF.

[2 marks]

- (b) Fit an ARMA(p, q) model for each of the 16 combinations (p, q), where $p, q \in \{1, 2, 3, 4\}$ and determine $Q(p, q) := -\ell$, where ℓ is equal to the `log likelihood` value of the estimated model. Then, use formula (76) in the lecture notes to compute $BIC(p, q)$. When fitting these models, please use `arma(..., optim.control = list(maxit = 1000))` instead of `arma(...)`. Your *R* code should generate a 4x4 numeric array named **Q2b**, where **Q2b[p,q]** contains the value of $BIC(p, q)$ for $p, q \in \{1, 2, 3, 4\}$.

[5 marks]

- (c) Repeat the instruction in (b) to find the BIC values also for fitted ARIMA($p, 1, q$) models, again for all 16 combinations (p, q), where $p, q \in \{1, 2, 3, 4\}$. Your *R* code should generate a 4x4 numeric array named **Q2c**, where **Q2c[p,q]** contains the value of $BIC(p, 1, q)$ for $p, q \in \{1, 2, 3, 4\}$.

[5 marks]

- (d) Based on the BICs computed in (b) and (c), decide which model fits best the given time series data.

[2 marks]

[14 marks]

[END OF PROJECT QUESTIONS]