Project

Submit your project report and the video of your oral presentation via Canvas by

***4:00pm on Friday 15th December 2023**.*

Please note that:

- *The project comprises two parts: an applied part a theoretical part. The two parts of the project are disjoint.*
- *For the applied part, you must write a report which should not exceed seven pages, written in Word or Latex. Project reports exceeding this limit will be penalised. Note that you are advised to include an Appendix, which does not count towards the page limit, detailing enough R code to allow the reader to reproduce your analysis. You may also like to include supplementary tabular and graphical output.*
- *There is no page restriction for the theoretical part.*
- *You should submit the applied and theoretical parts of your project as a single electronic file in PDF format. For ease of processing, please save the file as* `f_surname_mas8382_project.pdf` *where* `f` *is the first letter of your first name and* `surname` *is your surname.*
- *The video of your oral presentation should not exceed three minutes. The format must be mp4 and the file must be zipped up as a Zip or Tar Archive file before uploading it to Canvas.*

1 Project brief

The project is worth 100% of the overall mark for the module and comprises two parts: an applied part (worth 75% of the marks) and a theoretical part (worth 25% of the marks). The two parts of the project are disjoint.

The oral presentation is pass / fail. This means it does not carry a mark, but must be passed in order to pass the module. Its main purpose is to encourage you to focus on explaining statistical ideas in your own words.

1.1 Applied Part

For this part of the project, you will be given a time series. In each case, the series is meant to represent monthly sales of a product over a ten-year period, from January 2011 to December 2020.

You are to identify and fit a suitable time series model to the data. This model is to be used for forecasting and you should produce a forecast for the period January to June 2021. You should also assess the reliability of your forecasts. Further details are given in Section 1.1.2 below.

Your work should be presented as a report. There will be marks for the quality and presentation of this report. See Section 1.1.3 below.

1.1.1 Data

The file `projectdata.txt` is available from Canvas.

There are 10 different datasets in the file. You can choose any of these 10 datasets. Your data is then in the column of the file corresponding to your chosen dataset. So, for example, if you choose dataset number 2, then assuming you have the file `projectdata.txt` saved in your current working directory, you can extract your data as follows:

```
## Read in project data file
filename = "projectdata.txt"
data = read.table(filename, header=FALSE)
## Extract the data according to your reference number
y = data[,2]
```

Your data should be a numeric vector of length 120:

```
str(y)


##  num [1:120] 168 106 178 219 281 ...
```

If you have any difficulty obtaining your data, please let me know. You should obtain your data straight away to avoid difficulties later.

1.1.2 Tasks

Modelling Using methods covered in the module, identify and fit a suitable model for the data. You should use the steps of the iterative approach to modelling, with identification, estimation and diagnostic checking, repeated as necessary. You should consider whether it would be better to fit the model to the data as they are or whether it would be better to apply a transformation to the data.

You should explain your analysis clearly, illustrate it with suitable tables and graphs and state your conclusions clearly.

The module covers three general approaches to modelling nonstationary time series:

- ARIMA models.
- Time-series regression models.
- Dynamic linear models.

You should try *at least* two of the three methods and compare the results. Different methods may have different advantages and disadvantages, for example one method may be better for short-term forecasting and another for longer-term forecasting.

Forecasting You are to produce forecasts of monthly sales for the period January to June 2021 inclusive. You should provide 95% forecast limits.

In order to assess the reliability of your forecasting method, you should also fit your model using only the first nine years of data and then forecast the next twelve months and compare the forecast with the actual values.

Comment on your findings.

Report You should present your work in the form of a report. This should be well structured and written. It should have sections, beginning with an introduction and ending with conclusions. You should comment on the behaviour of the time series. For example, do sales seem to be increasing or decreasing? Is there evidence of seasonal behaviour and, if so, what is its nature? You should explain your analysis clearly. Illustrate your work with suitable graphs and tables. Your conclusions should be clearly stated.

You are advised to include an Appendix, which does not count towards the page limit, detailing enough R code to allow the reader to reproduce your analysis. You should refer to the code in the appendix at appropriate points in the main body of the report. You may also like to include supplementary tabular and graphical output.

1.1.3 Marking

This applied part of the project counts for 75% of the overall project mark. It will be marked according to the following scheme.

Model identification and fitting: (20 marks) Use (at least) the methods covered in the course to identify a suitable model, fit it and check that it appears to be satisfactory. You should check carefully that the model is appropriate, as described in the module. Explain and present your analysis clearly and state your conclusions.

Forecasting: (10 marks) This covers both the forecast (with limits) for January to June 2021 and the testing of the forecast. Again, explain your analysis and state your conclusions clearly.

Model comparison: (10 marks) These marks are available for a comparison of the different approaches to modelling (*i.e.* regression, ARIMA, dynamic linear models).

Report and presentation: (15 marks) Your report should be well structured and written. See Section 1.1.2 above.

Bonus marks: (20 marks) These marks are included to give extra marks for especially good features, use of initiative, imagination and so on. For example, consideration of all *three* classes of model covered in the module, using an appropriate method which we have not specifically covered or particularly good features of the explanation. They will be used only sparingly.

1.2 Theoretical Part

This theoretical part of the project counts for 25% of the overall project mark. There are two questions and they are marked as indicated below. If you would prefer not to type your solutions, you can scan or photograph handwritten answers. The resulting file(s) can then be embedded in, or appended to, your report for the applied part of the project.

1.2.1 Question 1 (15 marks)

Consider a seasonal ARIMA model

$$X_t = \varepsilon_t + \theta^* \varepsilon_{t-2}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

(a) Identify the model using the notation ARIMA$\{(p,d,q)\times(p^*,d^*,q^*)_s\}$.

(b) A seasonal ARIMA model is invertible if and only if the roots of the seasonal and non-seasonal moving average characteristic equations, $\Theta(x)=0$ and $\Theta^*(x^s)=0$, lie outside the unit circle. It is stationary if and only if the roots of the (extended) seasonal and non-seasonal autoregressive characteristic equations, $\Phi(x)(1-x)^d=0$ and $\Phi^*(x^s)(1-x^s)^{d^*}=0$, lie outside the unit circle. Explain why the series is always stationary and invertible for $|\theta^*|<1$.

(c) Assuming $|\theta^*|<1$, find the coefficients π_k in the representation of the series as an infinite order autoregression:

$$X_t=\varepsilon_t+\sum_{k=1}^{\infty}\pi_k X_{t-k}$$

(d) Given data $x_1,\ldots,x_n$, find equations for the k-step ahead forecast $\hat{x}_n(k)$ and the forecast error variance for $k=1,2,3,\ldots$. Your expressions for the forecasts may be written in terms of previous one-step ahead forecasts.

1.2.2 Question 2 (10 marks)

Regression-based methods for time-series analysis often account for seasonal variability by incorporating season as a fixed effect. Suppose the period of the seasonal behaviour is p. Then, in a very simple model for a time series we might have

$$X_t=m+s_t+R_t$$

where m is the intercept, R_t is an error term and s_t is the seasonal effect, such that

$$s_i=s_{i+p}=s_{i+2p}=\ldots \tag{1}$$

for each $i=1,2,\ldots,p$. To give the correct number of degrees of freedom we normally impose a sum-to-zero constraint

$$\sum_{i=1}^{p}s_i=0. \tag{2}$$

We can represent seasonal effects using a Fourier series

$$s_t=\sum_{k=1}^{K}\left\{a_k\cos\left(\frac{2\pi kt}{p}\right)+b_k\sin\left(\frac{2\pi kt}{p}\right)\right\}$$

where $K\leq\lfloor p/2\rfloor$ in which $\lfloor\cdot\rfloor$ is the floor function. Note that if p is even and $K=p/2$ then $b_{p/2}\equiv 0$. The parameters to be estimated are the coefficients, $a_1,b_1,a_2,b_2,\ldots$ which, unlike the distinct s_i, are unconstrained.[1]

Show that seasonal effects, constructed in this way, satisfy the conditions (1) and (2).

1.3 Oral presentation

Present a summary of the main findings from the applied part of your project report. You can make your slides using whatever presentation software you like, for example Latex Beamer, PowerPoint or Keynote.

[1] Taking $K<\lfloor p/2\rfloor$ obviously gives a more parsimonious model, with fewer than $p-1$ degrees of freedom, but this is at the cost of a loss of flexibility in the model.