ST4064 Class Test 2024

March 20, 2024

Instructions

PLEASE READ THESE INSTRUCTIONS CAREFULLY and FOLLOW THEM!

1. Answer ALL 10 questions: all questions carry 10 marks.
2. Use the following template file for your R code, this is available on Canvas:

   template_ST4064.R
3. Please rename this file:

   [your_student_number]_ST4064.R

   e.g. if my student number was 123456789, then I'd rename this file

   123456789_ST4064.R
4. At the end of the exam, please upload your submissions onto Canvas at the submission point for this assignment.
5. Upload ONLY your R script.

SAVE YOUR WORK REGULARLY!!

Introduction:

A bike share company collects daily data on its assets in order to plan for servicing schedules and also future demand. Customers are classed as either "registered" if they have an account with company and pay a monthly fee, or "casual" if they just borrow bikes on an ad-hoc basis. The company have hired you as a consultant to develop a predictive model for the number of bike hires.

The important variables for this analysis are as follows:

- `casual` the number of casual bike hires on that day.
- `registered` the number of registered bike hires on that day.
- `dteday` the date
- `t` the date in numeric format

The data are contained in a file "sharedbikes.RData"; this file contains a `data.frame` object named `dat`.

Later in the assignment, you will also need the file "sharedbikes_forecast.RData"; this file contains a `data.frame` object named `newdat`.

Question 1

Load the the object `dat` into R.

[10 marks]

[ ]:

Question 2

Make a scatter plot of `casual` over time, ensuring the x-axis is correctely formatted as a date. Add a lowess smoother to the plot. Ensure your plot has a title and axis labels.

[10 marks]

[ ]:

A linear model is first fitted to the data to capture annual and weekly variation as well as trend over time. You will use time-series analysis to form a model for the residuals of this process, stored in an object `res` in the dataframe `dat`, as defined below:

[ ]:
```
# model annual and weekly periodicity and trend component
mod = lm(casual~t+s1+c1+s2+c2+s7_1+c7_1+s7_2+c7_2,data=dat)

# The linear model removes most seasonality at annual
# and weekly periodicity, but does not do a perfect job.
# => Use time-series models to imporve fit.
dat$res = ts(residuals(mod),start=2018,frequency=365)
```

Question 3

Plot the residual series, the acf and the pacf of the residuals.

[10 marks]

[ ]:

Question 4

Decompose the residual series series into trend, seasonal and random components.

[10 marks]

[ ]:

Question 5

Decompose the first order differences of the residual series into trend, seasonal and random components.

[10 marks]

[ ]:

Question 6

Produce cumulative periodograms of (a) the residual series and (b) the first order differences of the residual series.

[10 marks]

[ ]:

Question 7

Use the function `arima0` to fit candidate models, explore models of the form ARIMA$(p, 1, q)$ for the **residual series**. What is the best value of p and q?

[10 marks]

[ ]:

Question 8

Fit the best model using the `arima` function and plot the cumulative periodogram of the residuals of this model.

[10 marks]

[ ]:

Question 9

Re-plot the casual rentals series along with forecasts of the number of rentals for the next 60 days, including a confidence interval.

HINTS: For convenience the prediction dataset has been provided in the file "shared-bikes_forecast.RData", this can be used to forecast from the linear model. You will need to forecast the residual series too.

[10 marks]

[ ]:

Question 10

The company are also interested in the relationship in bike usage between casual and registered users. They ask you to investigate this relationship:

[ ]:
```
# remove annual periodicity and trend component
mod1 = lm(registered~t+s1+c1+s2+c2+s7_1+c7_1+s7_2+c7_2,data=dat)

dat$res1 = ts(residuals(mod1),start=2018,frequency=365)

p = plot(registered~dteday,data=dat)
lines(lowess(dat$dteday,dat$registered,f=0.1),col="red",lwd=2)
lines(dat$t,fitted(mod1),col="green",lwd=2)
```

Find the best vector autoregressive model that describes the relationship between the first differences of the residual series for casual bike usage and the first differences of the residual series for registered bike usage. Use the Schwartz information criterion (SC) to choose between models

[10 marks]

[ ]:
```

```