Stats326 Exam

Hasnain Cheena

01/07/2020

## Question 1

### 1a)

The plot of the daily sales of unleaded petrol shows a strong seasonal component with no clear trend or cyclical behaviour. Moreover, the seasonal component shows a trough on Sunday and crest on Wednesday in petrol sales.

### 1b)

Yes, the modelling assumptions are satisified for model . Firstly, the plot of the Residual Series shows a patternless band around 0. Secondly, the acf of the Residual Series shows no significant lags. Finally, the residual summary output shows that the residuals are approximately normally distributed. Therefore, all modelling assumptions are satisified.

### 1c)

Forecasts for the week starting 22nd September:

t79.pred = 5415 + 2225

t80.pred = 5415 + 3015

t81.pred = 5415 - 1060

t82.pred = 5415 - 907

t83.pred = 5415 - 107

t84.pred = 5415 - 1440

forecasts = data.frame("Date"=c("22-09-2008","23-09-2008","24-09-2008","25-09-2008",  
 "26-09-2008","27-09-2008","28-09-2008"),  
 "Predictions"= paste(c(t78.pred,t79.pred,t80.pred,t81.pred,  
 t82.pred,t83.pred,t84.pred), "Litres"))  
  
forecasts

## Date Predictions  
## 1 22-09-2008 5415 Litres  
## 2 23-09-2008 7640 Litres  
## 3 24-09-2008 8430 Litres  
## 4 25-09-2008 4355 Litres  
## 5 26-09-2008 4508 Litres  
## 6 27-09-2008 5308 Litres  
## 7 28-09-2008 3975 Litres

### 1d) Discuss autocorrelation in Non-stationary Time Series data. Use sketch diagrams to illustrate your answer.

A key aspect of time series data is that the future cannot influence the past, but the past can influence the future. This dependence on the past is called autocorrelation. This autocorrelation has to accounted for when building time series regression models. This is because in regression modelling a critical assumption is that the errors are , independent and identically distributed. Autocorrelation causes this independence assumption to be violated.

When we build a regression model of a Non-stationary time series first we model the trend and seasonal aspects of the series. After this we will have to check a plot of the Residual series and acf plot for any evidence of autocorrelation If the plot of the Residual series shows a white noise series (zero mean, constant variance and no autocorrelation structure) then our errors are independently distributed and model assumptions are satisified. If not then we need to extend our model to account for auto-correlation (assuming all other patterns such as seasonality and trend are modelled correctly).

For example see any of the following patterns in our plot of the Residual Series and acf plot then we have to extend the model to account for autocorrelation.

1. Clustering indicating positive autocorrelation
2. Oscillation indicating negative autocorrelation
3. Significant lags

* What does it look like and how can you seee it? Note that if the residual series shows a pattern clustering or oscillation it is a sign of positive autocorrelation, negative autocorrelation respectively (as shown below). This tells us there is a pattern in the data that has not been modelled yet.
* How do you model it? There are a few ways to deal with autocorrelation:

1. Use generalised difference equation
2. Add lagged response variable as an additional explanatory variable
3. Model data and then aim to model the residuals creating a two-part model.

* White noise?

## Question 2

### 2a)

Series of differences has a huge negative residual rest of it seems ok.ACF shows decay/persistence and PACF also shows decay/persistance. Therefore we have no indication on trend or seasonal terms to use in the model and we start with SARIMA(1,1,0)x(1,1,0).

### 2b)

Out of the 3 models is deeemed to be the best. This is because:

is a SARIMA(1,1,0)x(1,1,0) model. It has signficant estimates (at a 10% level) and is a simple model, but doesn’t have the lowest AIC score.

is a SARIMA(0,1,1)x(0,1,1) model. It was deemed to be the best model as all its estimates are significant (at a 10% level) and it has the lowest AIC score. Furthermore, it is a simpler model than .

is a SARIMA(1,1,1)x(1,1,1) model. It has some estimates that are not significant (at a 10% level) and does not have the lowest AIC score. Furthermore, it has 4 parameters versus the other models just have 2.

### 2c)

The plot of the Residual series of is a reasonably patternless band around 0. However, there is a large residual in the series, but because this is in the early part of the series it is not a worry. Furthermore, the acf plot of the Residual series shows no significant lags.

add RMSEP Predictions for july august and oct are slightly lower than actual and september is very simiar.

### 2d)

Derive equation using back-shift notation:

Prediction for November 2019:

nov.2019.predict = 332.1 + 331.5 - 331.3 + 0 - (0.9986 \* -0.005) + (0.1441 \* 0.054) - (0.14389826 \* 0.032)  
nov.2019.predict

## [1] 332.3082

The prediction for November 2019 is 332.3 ppb.

### 2e)

The best model to predict global N2O atmospheric concentration from November 2019 to October 2020 is the Reduced Harmonic Significant Harmonics model. This is because reduced harmonic (significant harmonics model) has the lowest RMSEP of all the models. Furthemore, the acf plot of the residual series of this model shows a weakly significant lag at lag 1. As this lag is weakly signficant it is not of concern.  
However, a key reservation I have that may make the RMSEP values slightly unreliable are fact that the calculation only occurred over the months of July to October (4 months). This doesn’t inform us about how the models predict over a whole year of observations.

## Question 3

### 3a)

The plot of the NZ50 monthly returns percentage shows that the series is not stationary. This is because it has non-constant variation. We can see

The acf plot of the NZ50 returns shows no significant lags and thus no evidence for autocorrelation. However, the acf of the mean-centred squared values has significant lags at lag 4, 5 and 8. Thus indicating the fact that this series is conditionally heteroscedastic (volatile).

### 3b)

* have evidence for volatility (conditionally heteroscedastic series) thus try ARCH/GARCH modelling
* First we examine if their is exponential decay in the acf of mean-centred squared values. If the acf shows exponential decay then ARCH(1) will work.
* However if the acf does not show exponential decay then we try an ARCH(p) model increasing p incrementally and checking the model diagnostics.
* Once we find a good p value and if it is large, we can try fit a GARCH(1,1) model as it may be more parameter efficient.

ARCH and GARCH - used for finance modelling, for high volatility models

### 3c)

Once we have found the most appropriate ARCH/GARCH model, they can be used to model volatility in the series which is often used for simulation purposes. A key reason they are not useful for prediction (like most other time series modelling techniques) is because the underlying data they are modelling is effected by market conditions and government policy.

### 3d)

The modelling of the NZ50 monthly returns percentage data started with modelling it as a ARCH(1) resulting in . This model had non-significant estimates (at a 10% level), had evidence against normality and no evidence against independence. However, this model still showed significant lags in the squared residuals. ARCH(1) did not work as the initial acf on page 14 did not show exponential decay. Therefore it was not appropriate.

As ARCH(1) was not appropriate an ARCH(5) was tried, resulting in This model had many insignificant estimates, had evidence against normality and no evidence against independence. The squared residuals showed no significant lags. This model would work however, it is quite complex. Therefore a GARCH(1,1) was also tried.

The GARCH(1,1) model is the most appropriate model to model the NZ50 Sharemarket data. This is because all of its estimates are significant (at a 10% level), had evidence against normality and no evidence against independence. The acf of the squared residuals showed one weakly signficant lag. However, as it is weakly significant lag it is not a worry. Therefore, as GARCH(1,1) is simpler than ARCH(5) it was deemed to be appropriate.

We estimate that will be between -0.04 and 2.03. We estimate that will be between 0.03 and 0.26. We esitmate that will be between 0.64 and 0.94.

## Question 4

### 4a) Discuss the modelling of panel data

Panel data modelling occurs in a sequential fashion using 3 types of models: Pooled model, Fixed effects model and the Random effects model.

We first start the modelling with applying the Pooled model. The equation of this model is shown below:

where is the overall intercept and is the error term.

The Pooled model ignores the panel structure of the data and thus is ideal to use a baseline for the Fixed effects and Random effects models - this is why it is performed first. The Pooled model assumes there is a common effect across all cross-sectional units (common intercept).

After fitting a Pooled model we then fit a Fixed effects model. A Fixed effects model is fitted to determine whether there are specific effects for each cross-sectional unit. In this model individual effects for each cross-sectional unit are seperated out. The equation of this model is shown below: Effects are correlated with the explanatory variables.

After we fit the Fixed effects model we can get estimates of the fixed effects.

= Ftext for pool vs fixed and what happens after that We can then test whether the fixed effects are needed. To do this we can do an F-Test to compare the Fixed effects model to the Pooled model. The null hypothesis of this test is:

* Random effects After fitting the Fixed effects model we can also fit a Random effects model. This model assumes the individual effects are uncorrelated with the explanatory variables. The equation for this model is given below:
* Larange Multipler After fitting the Random effects model we can test whether the random effects are necessary by using a Lagrange multipler test. If the p-value < 0.05 we cna reject the null hypothesis that the random effects are not needed.
* Test for exogenity After performin the Larange Multipler test (and if we see that both fixed effects and random effects are needed) we can perform a test for exogeneity using the Hausman test. If the test produces a p-value > 0.05 we have no evidence against exogeneity and we accept the Random Effects model. Vice versa, if the test produces a p-value < 0.05 we have evidence against exogeneity and we accept the Fixed Effects model.

Using the method described above, we can select a suitable/appropriate model to model the panel data.

### 4c) Discuss Cointegrated time series modelling

* What are co-integrated series If you want to perform co-integrated time series modelling there is a sequential method to follow. This is listed below.
* Assess whether a linear trend exists between the two time series and that it is valid Firstly, you need to assess whether there exists a relationship between the two time series you want to design the co-integrated model with. The easiest way to do this is to plot one series against the other on a scatterplot. If the scatterplot shows a linear relationship then it is possible to continue on witht the modelling. If it does not then you may need to perform appropriate transformations to the data to make the linear (e.g. log transformations).
* Assess conditions for co-integrated series are met The second aspect to co-integrated time series modelling is to fit the model and assess whether the conditions are met. The two conditions are:

1. The two series are of the same integrated order
2. The linear combination of the two series is stationary.

* Each model is of the same order
* How is that done To assess whether each series is of the same integrated order performed the augmented dickey-fuller test on each series. If a unit root exists in both of our series then applying the test again to the differenced series to assess if a second unit root exists. If there is no second root in either differenced series then our series are of the same order.

Then fit the model - perfom regression.

* Assess if the model fitted is a linear combination of the 2 series is stationary
* How is that done

Once you perform regression, use the augmented dickey-fuller test on the residual series. If the residual series if found to be stationary (p-value < 0.05) then the second condition is met.

If both these conditions are met then we have a valid co-integrated time series model and vice versa.

This is cointegrated modelling