Stats326 Exam

Hasnain Cheena, 190411106, hche737

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 behavior. Moreover, the seasonal component shows a trough on Sunday and crest on Wednesday.

1b)

Yes, the modelling assumptions are satisfied for model $ULP.fit2$.
This is because, firstly, the plot of the Residual Series shows a patternless band around 0. Secondly, the ACF plot of the Residual Series shows no significant lags. Finally, the summary output shows that the residuals are approximately normally distributed. Therefore, all modelling assumptions are satisfied.

1c)

Forecasts for the week starting 22nd September:

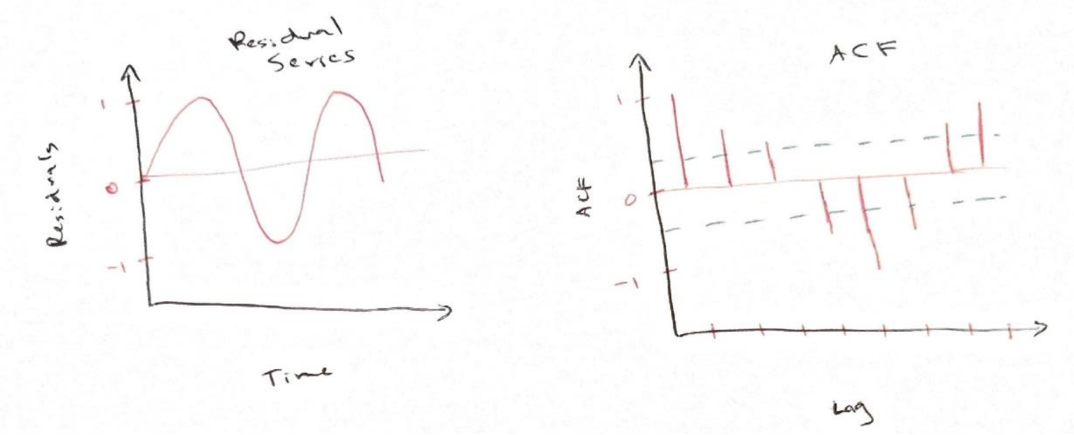
```
t79.pred = 5415 + 2225

t80.pred = 5415 + 3015

t81.pred = 5415 - 1060

t82.pred = 5415 - 907

t83.pred = 5415 - 107

t84.pred = 5415 - 1440

forecasts

##          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)

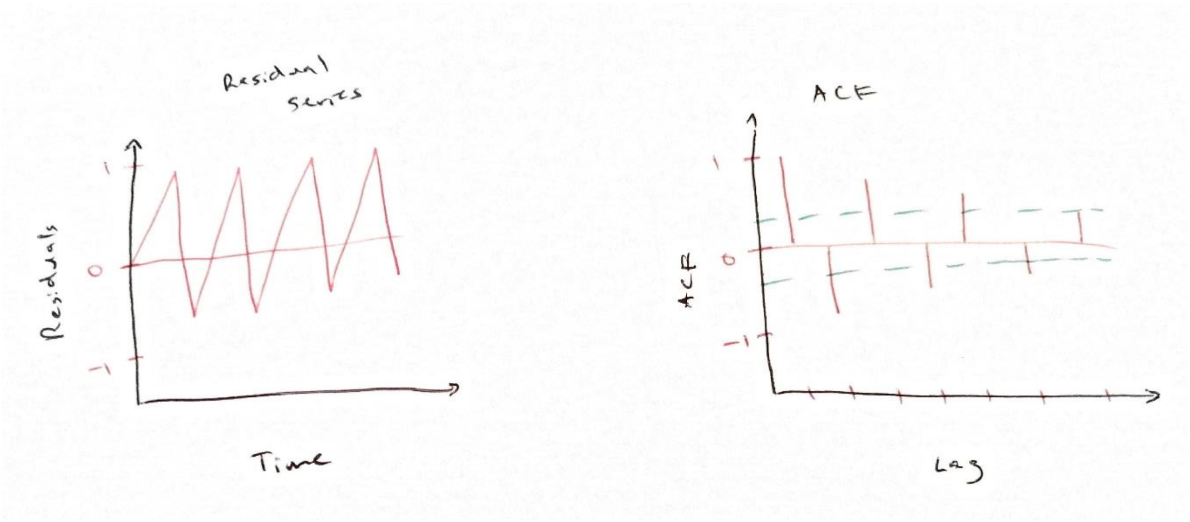
A key aspect of time series data is that the future cannot influence the past, but the past can influence the future. This dependence on the past is called autocorrelation. Autocorrelation must be accounted for when building time series regression models. This is because in regression modelling a critical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

When building a regression model of a Non-stationary time series, first we need to account for the trend and seasonal aspects of the series. After modelling these factors, we will check the Residual Series plot (Residual Series versus Time) and the plot of the ACF for any evidence of autocorrelation.
If the Residual Series plot shows a stationary series and the ACF shows no significant lags the Residual Series is White Noise. This means the errors are *iid* and the independence assumption is satisfied.
If the Residual Series is not White Noise, then we need to extend our model to account for autocorrelation. This assumes all other patterns such as seasonality and trend are accounted for and modelled correctly. For example, if we see any of the following patterns in the Residual Series plot and ACF plot then we must extend the model to account for autocorrelation.

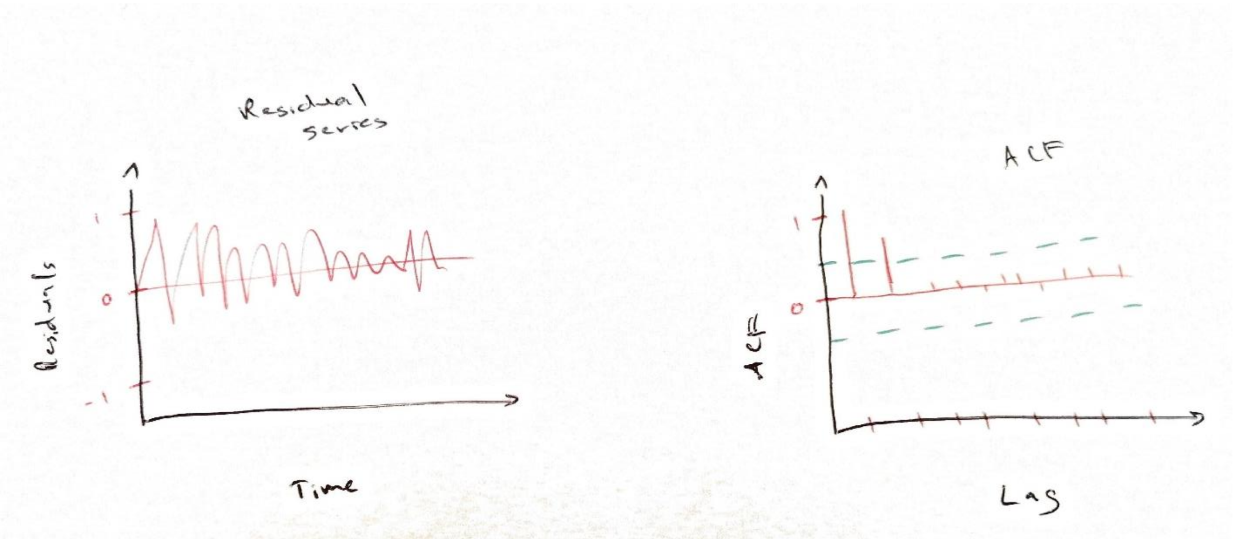
1. Clustering of the Residual Series and ACF lags indicating positive autocorrelation



2. Oscillation of the Residual Series and ACF lags indicating negative autocorrelation



3. The Residual Series seems stationary but ACF plot shows significant lags



After examining the plots and if there is evidence for autocorrelation, we must deal with it. There are a few ways to deal with autocorrelation:

1. Use a generalized difference equation

2. Add a lagged response variable as an additional explanatory variable

3. Model the data and then aim to model the residuals separately creating a two-part model.

Question 2

2a)

The series of differences has a huge negative residual in the early part of the series. However, the remaining section of the series shows a reasonably patternless band around 0. Both the ACF and PACF plots show decay or persistence. Therefore, we have no indication of the order of trend or seasonal terms to use in the model and we start with SARIMA(1,1,0)x(1,1,0)$_{12}$.

2b)

$sarima.fit1$ is a SARIMA(1,1,0)x(1,1,0)$_{12}$ model. All estimates are significant at a 10% level. Moreover, it is a relatively simple model, but doesn't have the lowest AIC score.

$sarima.fit2$ is a SARIMA(0,1,1)x(0,1,1)$_{12}$ 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 relatively simpler model than $sarima.fit3$.

$sarima.fit3$ is a SARIMA(1,1,1)x(1,1,1)$_{12}$ model. It has some estimates that are not significant (at a 10% level) and does not have the lowest AIC score. Furthermore, it is a more complex model because it has four parameters versus the other models just have two.

2c)

The plot of the Residual series of $sarima.fit2$ 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.

Further, the output shows that the predictions for July, August and October 2019 are slightly lower than observed values and the prediction for September 2019 is very similar to the actual value. Moreover, the output shows that, on average, the prediction error of $sarima.fit2$ is 0.1 ppb.

2d)

Derive equation using back-shift notation:

$$(1-B)(1-B^{12})y_t = (1+\alpha_1 B)(1+A_1 B^{12})\varepsilon_t$$

$$(1-B)(1-B^{12})y_t = (1+0.1441B)(1-0.9986B^{12})\varepsilon_t$$

$$y_t - B^{12}y_t - By_t + B^{13}y_t = \varepsilon_t - 0.9986B^{12}\varepsilon_t + 0.1441B\varepsilon_t - 0.14389826B^{13}\varepsilon_t$$

$$y_t = y_{t-1} + y_{t-12} - y_{t-13} + \varepsilon_t - 0.9986\varepsilon_{t-12} + 0.1441\varepsilon_{t-1} - 0.14389826\varepsilon_{t-13}$$

Prediction for November 2019:

$$y_{t+1} = y_t + y_{t-11} - y_{t-12} + \varepsilon_{t+1} - 0.9986\varepsilon_{t-11} + 0.1441\varepsilon_t - 0.14389826\varepsilon_{t-12}$$

```
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 N_2O atmospheric concentration from November 2019 to October 2020 is the reduced harmonic (significant harmonics) model. This is mainly because the reduced harmonic (significant harmonics) model has the lowest RMSEP of all the tried models. Comparing the diagnostic plots of the models, the ACF plot of the residual series of the reduced harmonic (significant harmonics) model shows a weakly significant lag at lag 1. However, as this lag is only weakly significant it is not of concern.

A critical reservation I have is that the RMSEP values are slightly unreliable. This is because the RMSEP calculation only accounted for 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 shows that the series is not stationary. This is because it has non-constant variation.

Furthermore, the ACF plot of the NZ50 returns shows no significant lags and thus no evidence for autocorrelation. However, the ACF plot of the mean-centered squared values has significant lags at lag 4, 5 and 8. This is evidence for the series being conditionally heteroscedastic (volatile).

3b)

From the plots discussed above, we have found evidence for volatility in the series. As a result of this the time series modelling technique that should be used is ARCH/GARCH modelling. This is because ARCH/GARCH modelling will be able to account for the volatility in the series.

To employ this technique first we examine the ACF plot of the mean-centered squared series for exponential decay. If exponential decay exists, then we try ARCH(1).
However, if the ACF plot does not show exponential decay then we try an ARCH(p) model, increasing p incrementally while checking the model diagnostics within every iteration.
Once we find a good value for p and if it is large, we can try fit a GARCH(1,1) model as it may be more parameter efficient.

3c)

Once we have found the most appropriate ARCH/GARCH model, we are able to model the volatility in a series. This model can be used for simulation purposes.
A critical reason they are not useful for prediction is because the underlying data is affected by market conditions and government policy.

3d)

The modelling of the NZ50 monthly returns data started with modelling it as an ARCH(1) creating the $arch.1$ model. This model had non-significant estimates (at a 10% level). Moreover, this model still showed significant lags in the squared residuals. Therefore, it was not appropriate.

As ARCH(1) was found to not be appropriate, an ARCH(5) was tried creating the $arch.5$ model. This model had many insignificant estimates. Furthermore, 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) and the ACF plot of the squared residuals showed one weakly signficant lag. However, as it is a weakly significant lag, it is not a worry. Moreover, GARCH(1,1) is a simpler model than ARCH(5). Therefore, it was deemed to be best.

Note that all the models tried showed evidence against normality and no evidence against independence.

We estimate that for the model deemed best:

- α_0 will be between -0.04 and 2.03
- α_1 will be between 0.03 and 0.26
- β_1 will be between 0.64 and 0.94

Question 4

Note: I chose to answer 4a and 4c.

4a)

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

First, we start with modelling the Pooled model. The equation of this model is shown below:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

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

After fitting a Pooled model, we go ahead and 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 separated out. The equation of this model is shown below:

$$y_{it} = \beta_1 x_{it} + \alpha_i + \eta_{it}$$

After fitting the Fixed effects model, we can then test whether the fixed effects are necessary. To do this we perform an F-Test to compare the Fixed effects model to the Pooled model. If the p-value of the test is less than 0.05 then the fixed effects are needed. Thus, we have evidence that the Fixed Effects model is better than the Pooled model.

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:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + \eta_{it}$$

After fitting the Random effects model, we perform a test to determine whether the random effects are necessary. This test is called the Lagrange multiplier test. If the p-value of this test is less than 0.05 we have evidence that the random effects are necessary. Therefore, the Random effects model is better than the Pooled model.

Finally, after finding that both the fixed and random effects models are better than the Pooled model, we need to determine which out of those two is best. To do this we can perform a test for exogeneity using the Hausman test. If the test produces a p-value greater than 0.05 we have no evidence against exogeneity, and we accept the Random effects model. However, if the result of the test shows the p-value is less than 0.05, we accept the Fixed effects model.

Using the method described above, we can select an appropriate model to describe the panel data.

4c)

To perform co-integrated time series modelling the steps described below are necessary.

Firstly, you need to assess whether there exists a relationship between the two time series that are to be regressed. 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 with the modelling. If it does not, appropriate transformations must be implemented to make this relationship linear (e.g. log transformations).

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.

To assess condition 1 an augmented Dickey-Fuller test is performed on each series. If the result of the test is that a unit root exists in both of our series, then we need to reapply the test again to the differenced series to assess if a second unit root exists. If the output of the second test is that no unit root exists in both differenced series, then the time series are of the same order.

After confirming that condition 1 is met, perform the regression and use the augmented Dickey-Fuller test on the resulting 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.