Time Series Analysis

Assignment 3: Computer Based

Submit all Exercises

Submission date: 10pm 10 November 2019.

Read in the 3 time series in the file "assignment3.csv" and answer the following questions. Upload a **single file** containing all typed code and your answers to the questions below. It may be Word, PDF, or Rmd in format. Failure to submit a single file with all solutions will result in a penalty of 10/50 marks, but you may re-submit if you need to and in that case **the most recent file only will be graded**. Please **do not email solutions** to the tutor or the lecturer. Issues relating to upload to Brightspace must be identified clearly and well in advance of the deadline.

Late submissions will be penalised by 5/50 up to 1 hour late, 10/50 up to 24 hours, 25/50 up to 48 hours and no grades will be awarded for work that is more than 48 hours late.

Exercise 1.

(a) For Series X, identify the number of differences (if any) and order of the ARMA model for the stationary series based on the results of **tsdisplay()**. Give reasons for your answers.
[5]

(b) For Series Y, identify the number of differences (if any) and order of the ARMA model for the stationary series based on the results of **tsdisplay()**. Give reasons for your answers.
[5]

(c) Report the AR and / or MA parameter estimates for Series X and Y, with no drift by using **include.mean=FALSE**. Are they statistically significant?
[5]

(d) Look at Series Z. What type of model should you choose and why? Again, focus on models without drift (i.e. mean zero).
[5]

(e) Now try ARMA(1,1), ARMA(1,2) and ARMA(2,1) models fit to Series Z. Do the parameters come out to be statistically significant? Why?
[5]

Exercise 2. We will now examine model fit for Series Z in more detail.

(a) Examine the residuals of an ARMA(1,1) model fit to Series Z and describe what you find.
[7]

(b) Examine the residuals of an AR(2) model fit to Series Z and describe what you find.
[8]

Exercise 3. Now fit an AR(10) model to Series Z.

(a) Are the parameters significant? Examine the residuals and describe your findings. **[5]**

(b) Why might this model fit the Series well? **[5]**

Assignment 3 - 16374403

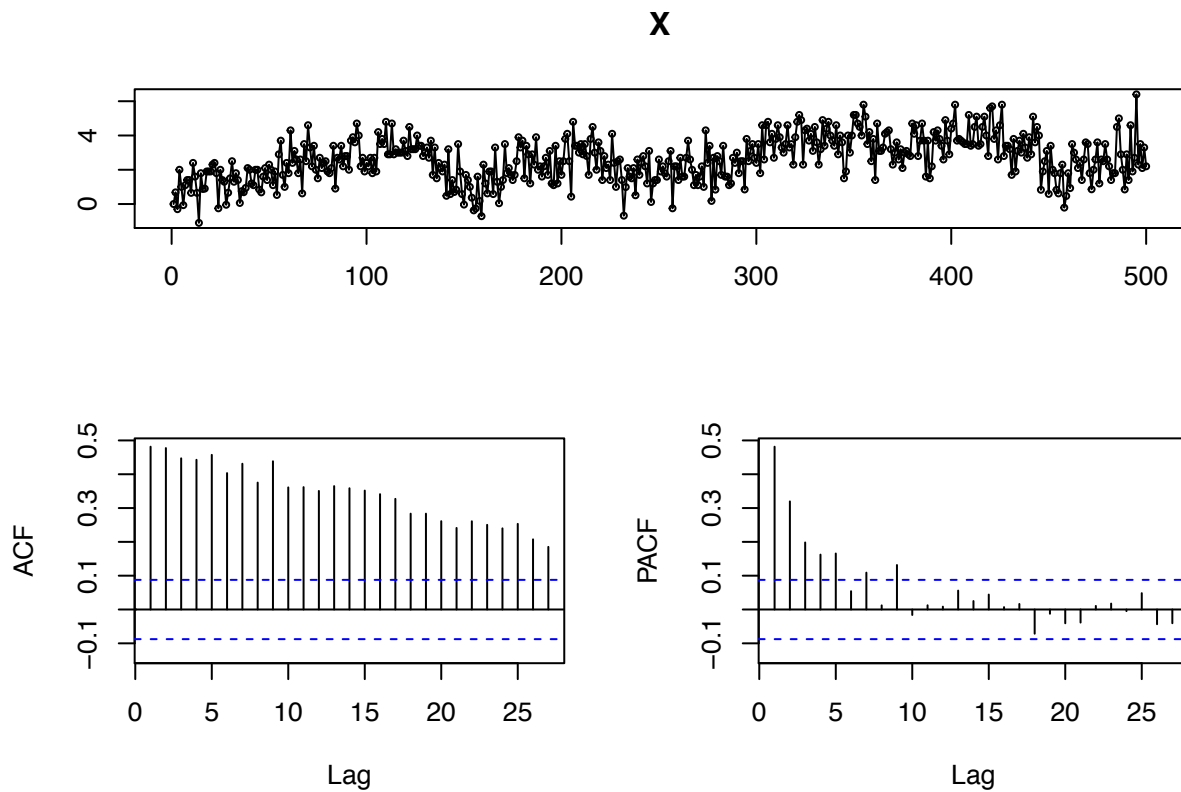## Exercise 1

Firstly, load in the data as follows

```
data =  read.csv('assignment3.csv')
X = data$X
Y = data$Y
Z = data$Z
```

(a)

To find the number of differences we first observe the results from the `tsdisplay()` command.
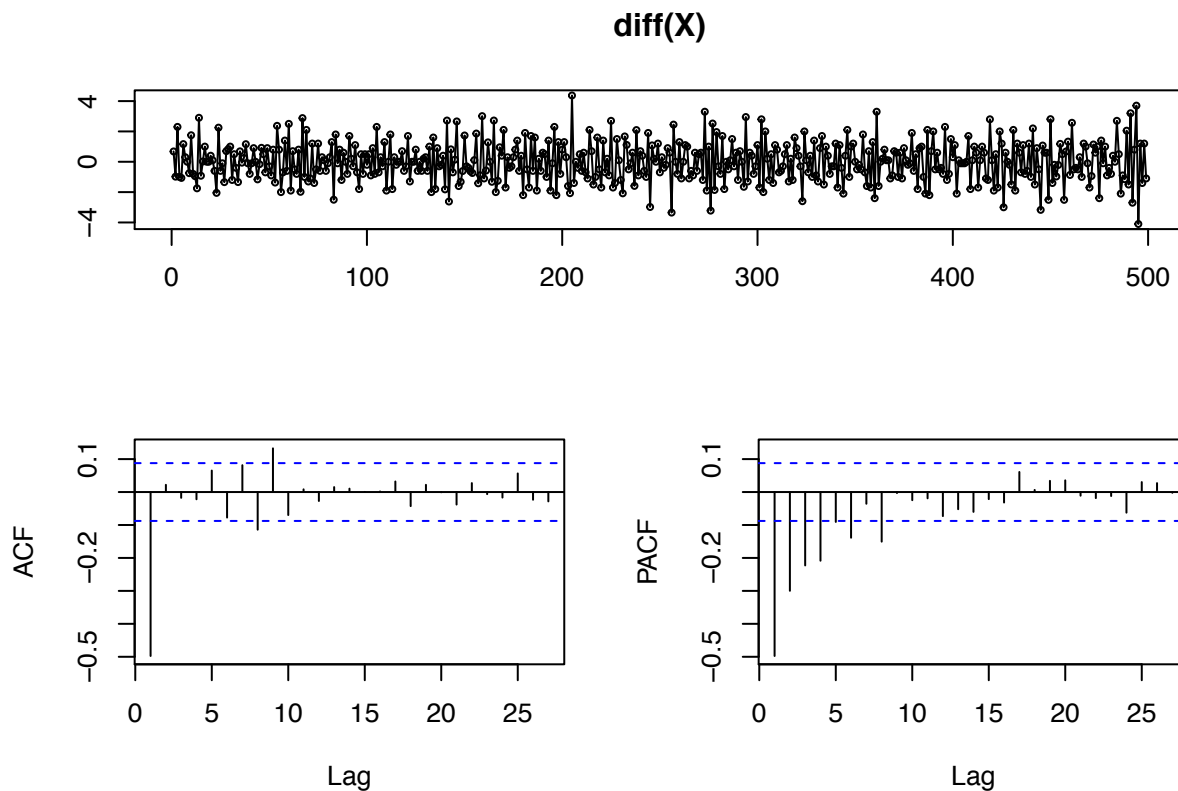
tsdisplay() for series X

```
library(forecast)
tsdisplay(X)
```


From the above results we can see a linear decrease of the ACF and so we know that we need to take differences of the model at least once. We will now take check the results for the `tsdisplay()` command applied to the series ΔX

**tsdisplay() for series ΔX**

diff(X)

We can see that the PACF plot exhibits exponential decay for the ΔX series. In addition, the ACF at lag 1 for ΔX is non-null but all lags greater than 1 are null. This indicates that we do not need to take any more differences and we can conclude that the series ΔX is MA(1).

In summary, we can now conclude that the original series X is ARIMA(0,1,1).

(b)

```
tsdisplay(Y)
```


Y
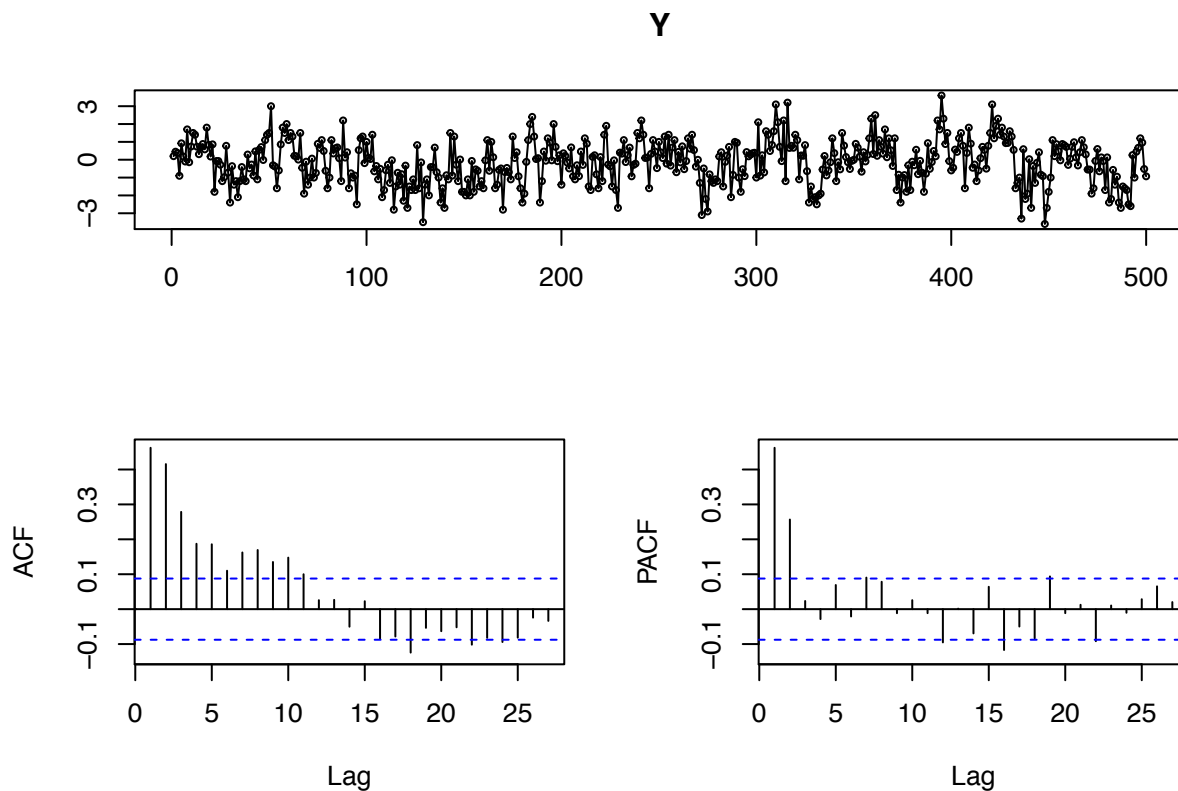
For the series Y we observe an exponential decay of the ACF and the PACF is null after lag 2. From this we can conclude that Y is ARIMA(2,0,0) and no differences are needed.

(c)

We must activate the *lmtest* library in order to access the `coeftest` function.

In (a) we found that the series X was ARIMA(0,1,1) and so we are looking to check the MA parameter estimate. We can get more information on this using the code below.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

3

```
modelX = arima(X, include.mean = FALSE, order = c(0,1,1))
coeftest(modelX)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value  Pr(>|z|)
## ma1 -0.824827   0.025328 -32.565 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this code we can see that we get a MA parameter estimate of $\theta = -0.824827$ with a p-value of 2.2e-16. This means that our parameter estimate is statistically significant at any reasonable level of significance. For example at $\alpha = 0.05$ we have that $2.2e - 16 < 0.05$ and so we reject the null hypothesis that $\theta = 0$ and conclude that our MA parameter estimate is significant.

From (b) we know that series Y is ARIMA(2,0,0). We can consider the parameter estimates using the code below:

```
modelY = arima(Y, include.mean = FALSE, order = c(2,0,0))
coeftest(modelY)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value  Pr(>|z|)
## ar1 0.347070   0.043090  8.0545 7.983e-16 ***
## ar2 0.260208   0.043091  6.0386 1.555e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This gives us our parameter estimates of $\phi = (0.347070, 0.260208)$. The p-values are 7.983e-16 and 1.555e-09 respectively. These are both much smaller than 0.05 and so we can conclude that the both parameter estimates are statistically significant at 5% level of significance (and again at any reasonable level of significance given how small the p-values are).

(d)

```
tsdisplay(Z)
```

Z
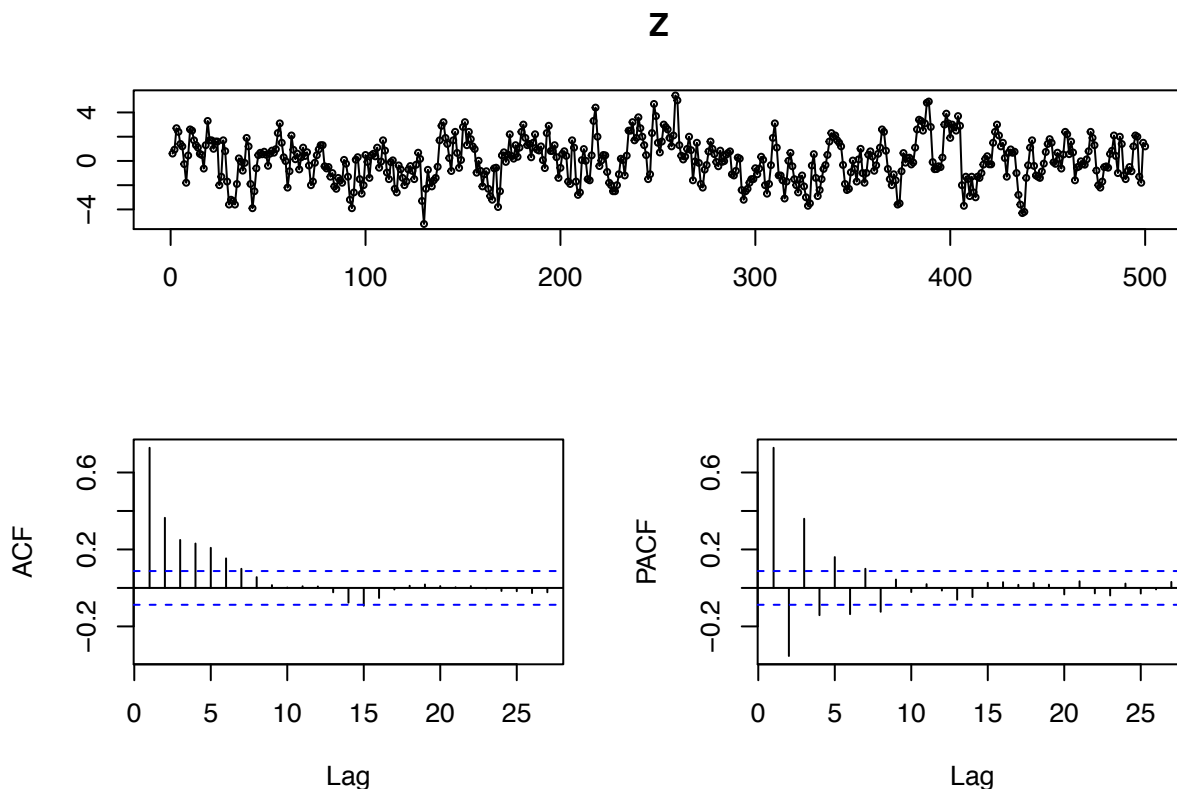
From the results above we can see that both the ACF and PACF are decaying exponentially. This tells us that the series Z is ARIMA$(p, 0, q)$ where both $p > 0$ and $q > 0$. We cannot simply find the values of p and q by reading them off the plots. Instead we must build up a model by adding parameters one by one and then assessing if each of the newly added parameters is significant or not after each added parameter. Since we know the model is of the form ARIMA$(p > 0, 0, q > 0)$ we can start in the simplest case of ARIMA$(1, 0, 1)$.

```
modelZ = arima(Z,include.mean = FALSE, order = c(1,0,1))
coeftest(modelZ)
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ar1 0.456554   0.042733  10.684 < 2.2e-16 ***
## ma1 0.804087   0.027625  29.108 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All parameters significant, continue to add next parameter.

```
modelZ = arima(Z,include.mean = FALSE, order = c(2,0,1))
coeftest(modelZ)
```

5

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ar1 0.441360   0.055377  7.9700 1.586e-15 ***
## ar2 0.023233   0.054056  0.4298    0.6673
## ma1 0.811964   0.032237 25.1876 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

AR2 parameter p-value is 0.6673. This is not significant at 5% level of significance and so we drop it from the model. We now add a MA2 parameter to our ARIMA(1,0,1) model to see if it is significant or not.

```
modelZ = arima(Z,include.mean = FALSE, order = c(1,0,2))
coeftest(modelZ)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value  Pr(>|z|)
## ar1  0.527975   0.123222  4.2847 1.830e-05 ***
## ma1  0.717124   0.148199  4.8389 1.305e-06 ***
## ma2 -0.082674   0.138975 -0.5949    0.5519
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

MA2 parameter p-value is 0.5519 and so we conclude it is not statistically significant at 5% level of significance and thus we drop it from the model.

Conclude ARMA(1,1) is the model with all significant parameters.

(e)

From the outputs above in part (d) we can see that the ARIMA(1,0,1) model is the only model out of the three which has all paramaters statistically significant at a 5% level of significance. The ARIMA(1,0,2) and ARIMA(2,0,1) models have parameters which are not statistically significant and so will be overfitting the data.
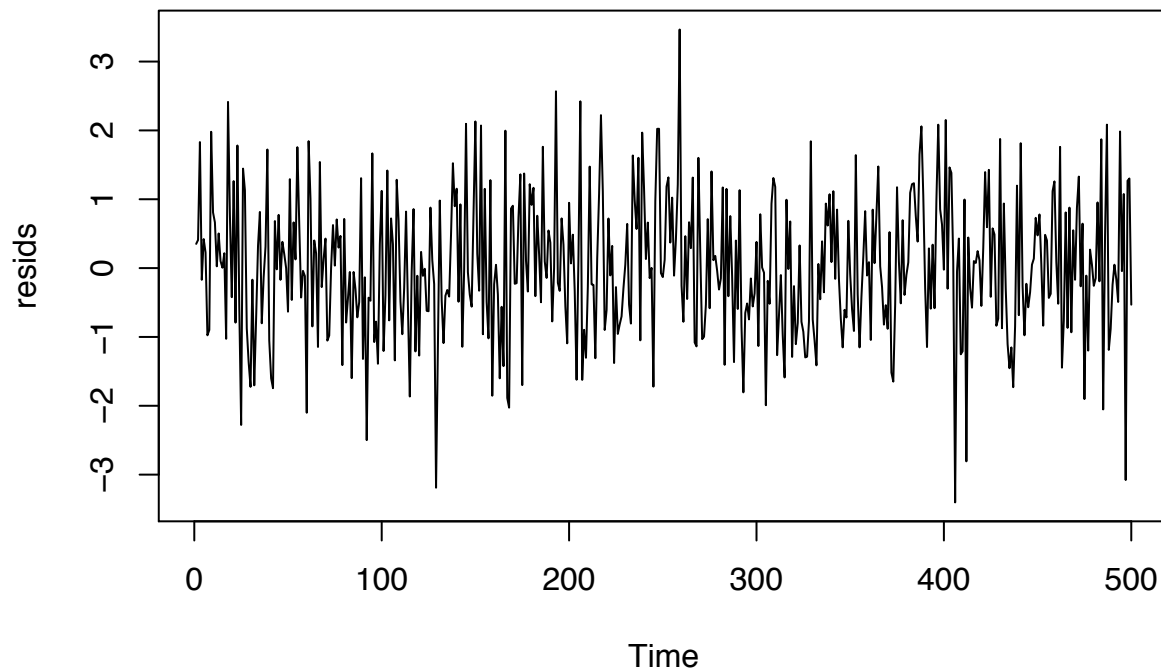
Conclude that the ARIMA(1,0,1) is the best fit of the three models since it is the only model with all significant parameters and is in the form ARMA($p > 0, q > 0$) which we found from the ACF and PACF plots.

Exercise 2

(a)

Step 1 - Plot of Residuals vs. Time

```
modelZ = arima(Z,include.mean = FALSE, order = c(1,0,1))
resids = residuals(modelZ)
# Step 1
# Plot of Residuals vs. Time
plot(resids)
```
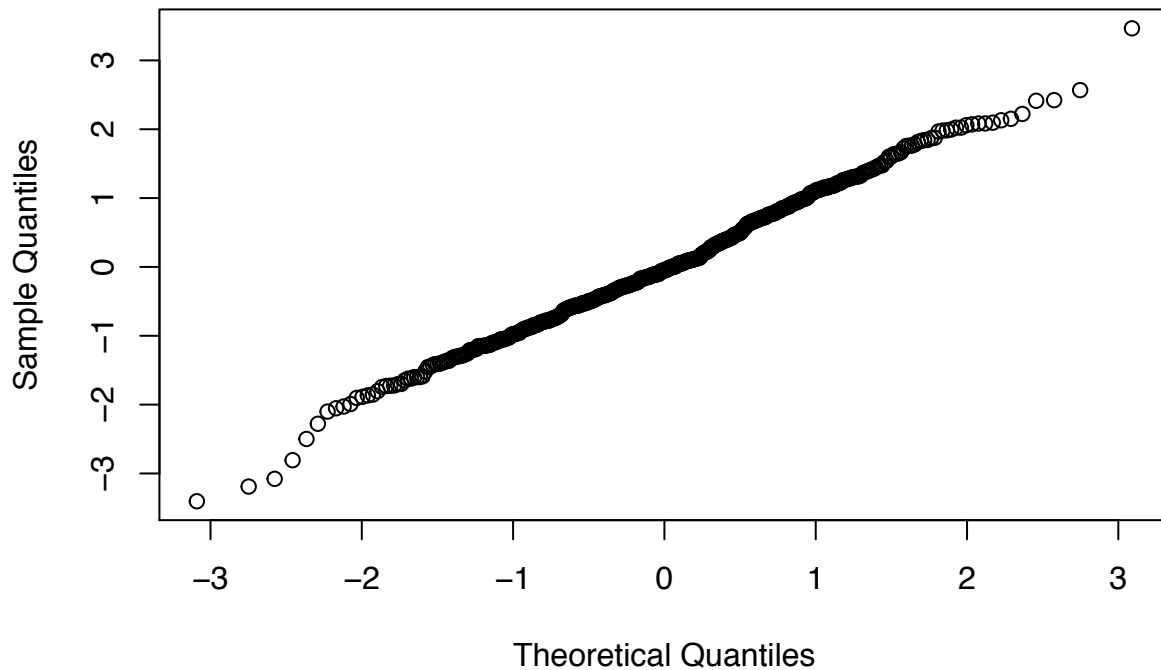

A plot of the residuals shows that they are centered around 0 and have no pattern as we would hope.

Step 2 - Normality of Residuals

```
# Check normality of residuals via QQ plot
qqnorm(resids)
```
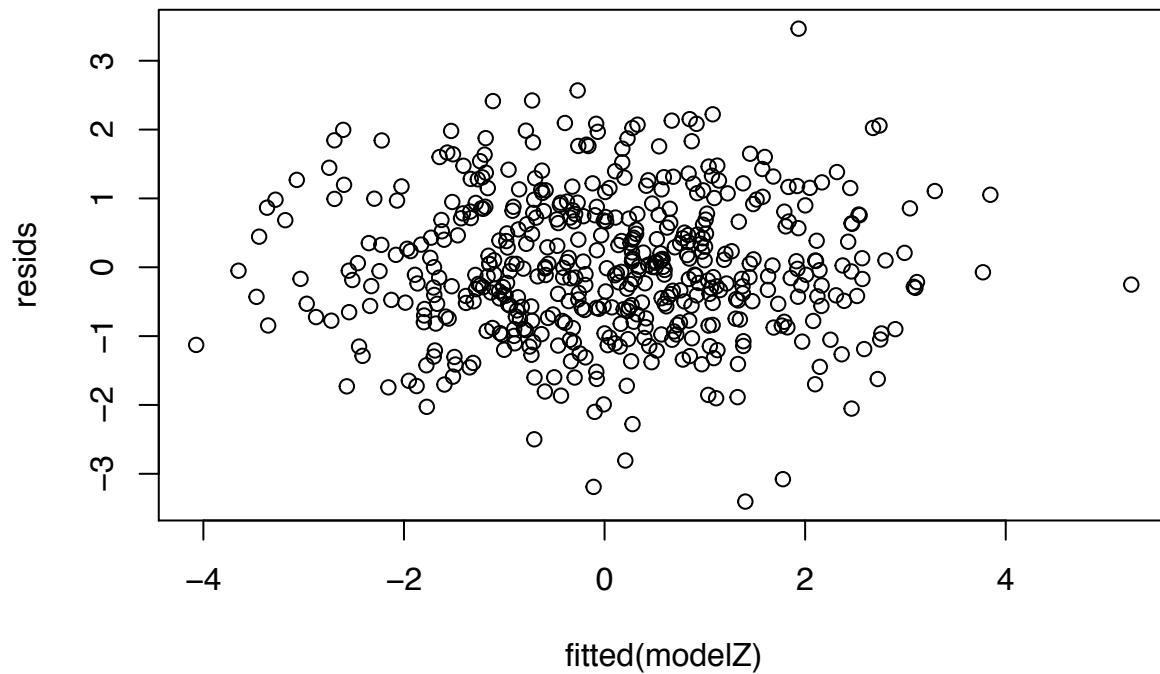
Normal Q–Q Plot


```r
# Shapiro-Wilks Test
shapiro.test(resids)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resids
## W = 0.9958, p-value = 0.2032
```

Firstly, a normal qq-plot of the residuals shows a near straight line (tracking $y = x$) with a little bit of deviation at the tails. The straight line is a good indication of normality. A Shapiro-Wilks test can be used to confirm this. The null hypothesis for a Shapiro-Wilks test is that the samples come from a normal distribution. The p-value for this Shapiro-Wilks test is 0.2032 and so we fail to reject the null hypothesis and conclude that there is *not* enough evidence at a 5% level of significance to conclude that the the residuals are *not* normally distributed.

Step 3 - Plot Residuals vs. Fitted Values

```r
# Plot residuals vs fitted values
plot(fitted(modelZ), resids)
```
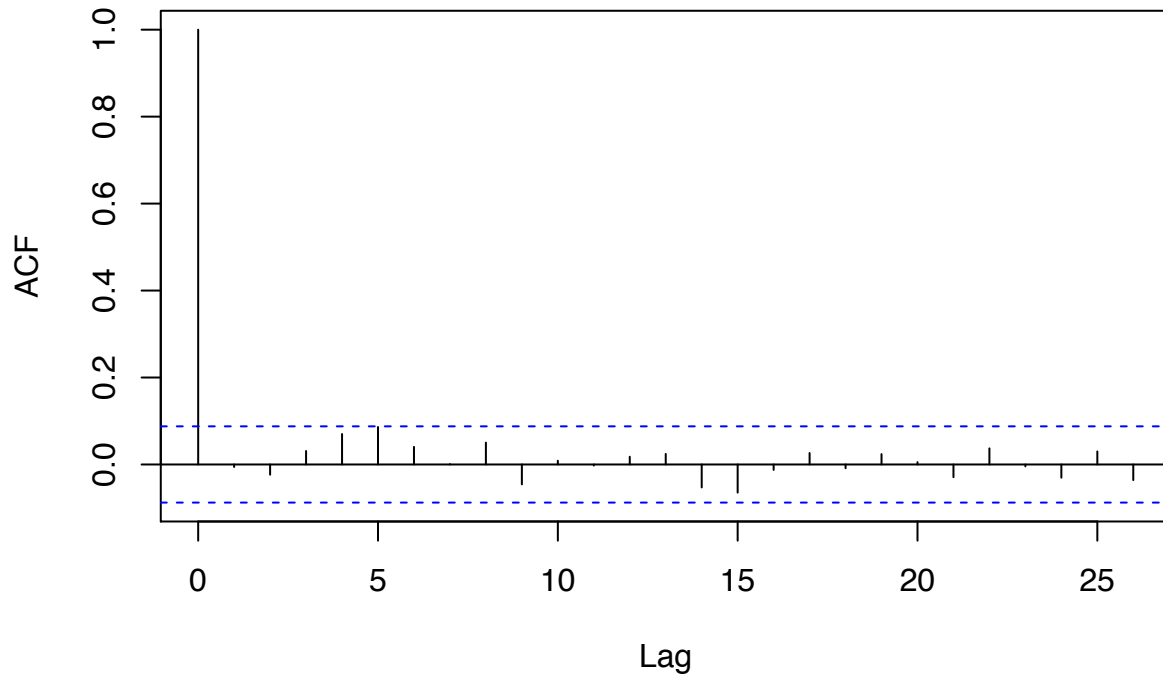
There is no pattern evident here. There seems to be no relation between residuals and fitted values. Also, residuals are centred at 0 as we would hope.

Step 4 - ACF Residuals and Ljung-Box-Pierce Tests

```
# ACF Residuals
acf(resids)
```

Series resids


```r
# Do a Ljung-Box-Pierce test to check these in more detail
Box.test(resids, lag = 5, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  resids
## X-squared = 7.0369, df = 5, p-value = 0.2179
```

The ACF plot above shows no statistically significant (Non Null) ACF at a significance level of 5%. We can conduct a more rigorous test of this using a Ljung-Box test with the lag set to 5 (a lag of 5 or 10 is the standard choice when conducting a Ljung-Box test). This tests the null hypothesis that there is no autocorrelation for any of the lags up to and including lag 5. The p-value comes out as 0.2179. Since the p-value is greater than 0.05 we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest any significant autocorrelation for the first 5 lags at $\alpha = 0.05$.
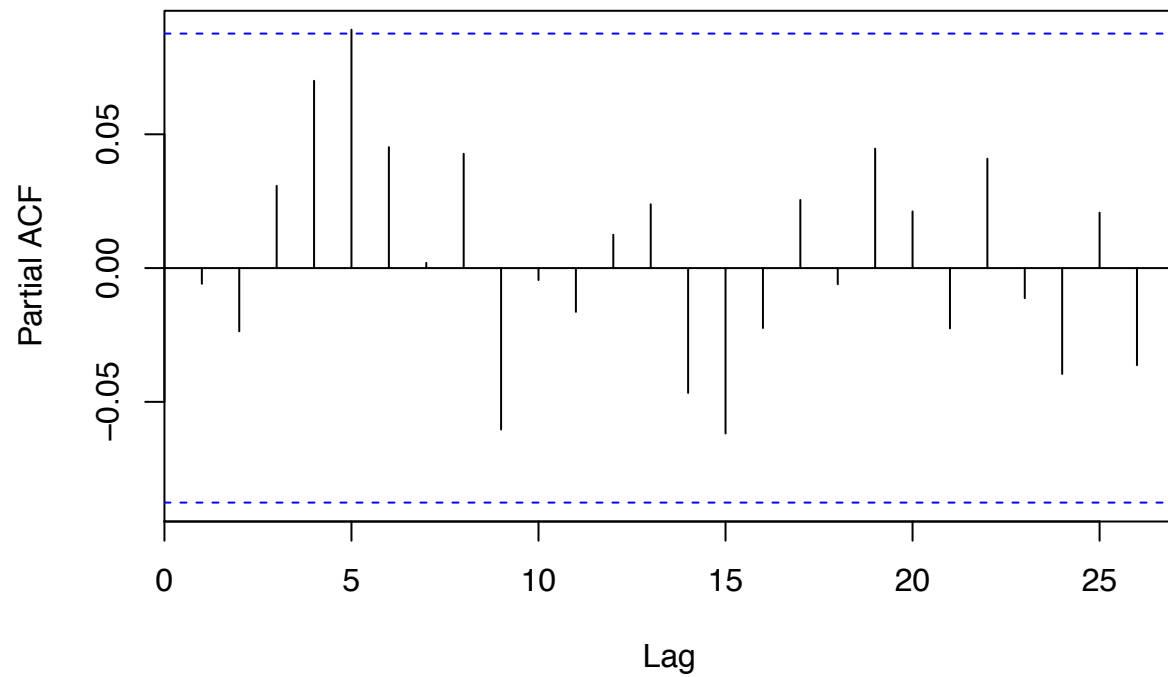
Step 5 - PACF Residuals

```r
pacf(resids)
```

Series resids

The PACF plot of the residual values displayed above indicates null PACF values for all lags.
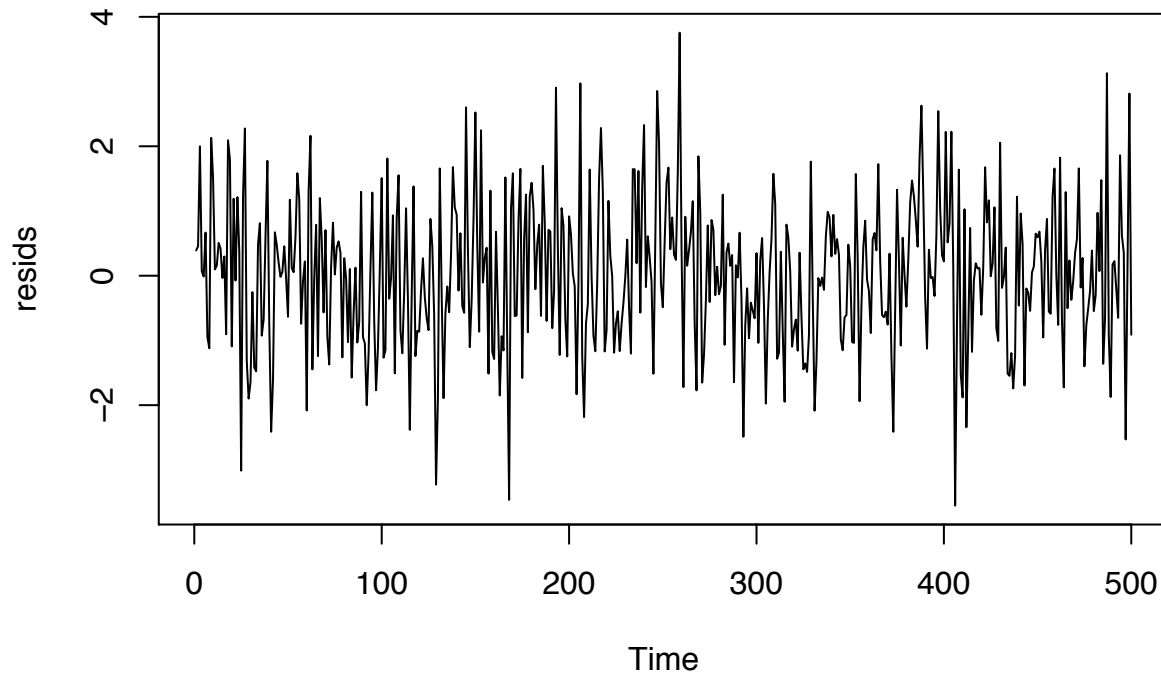
Summary

All steps passed and so we fail to reject the ARMA(1,1) model.

(b)

Step 1 - Plot of Residuals vs. Time

```
modelZ = arima(Z,include.mean = FALSE, order = c(2,0,0))
resids = residuals(modelZ)
# Step 1
# Plot of Residuals vs. Time
plot(resids)
```
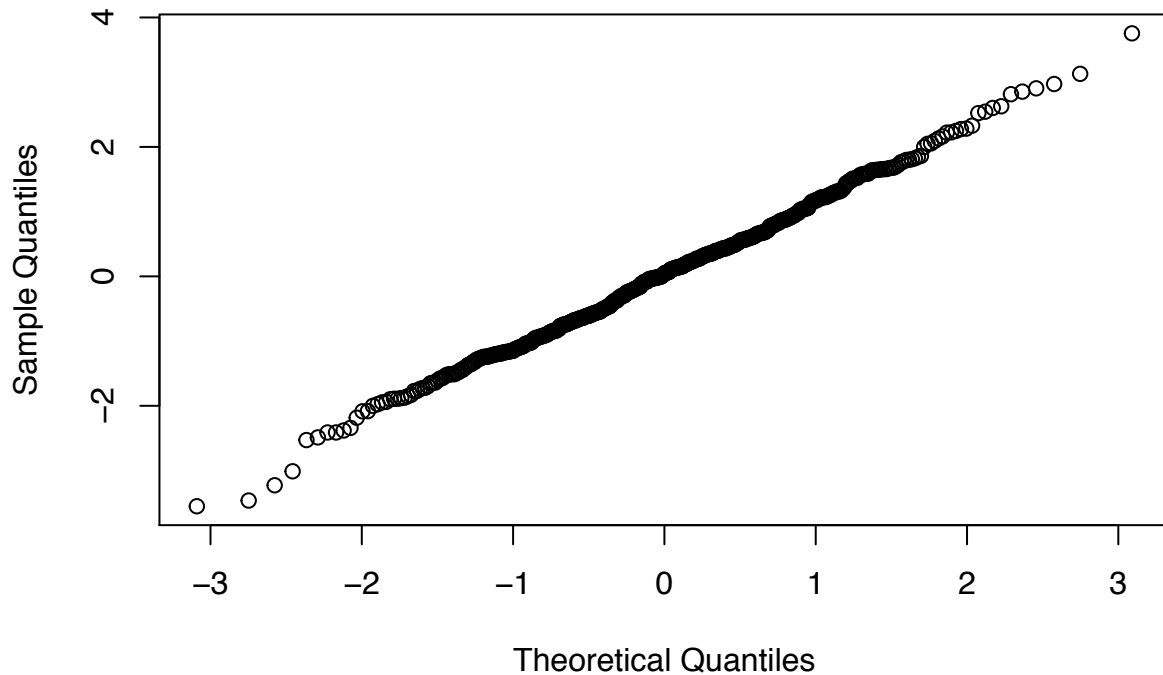

A plot of the residuals shows that they are centered around 0 and have no pattern as we would hope.

Step 2 - Normality of Residuals

```
# Check normality of residuals via QQ plot
qqnorm(resids)
```

Normal Q–Q Plot


```r
# Shapiro-Wilks Test
shapiro.test(resids)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resids
## W = 0.99757, p-value = 0.6861
```

Firstly, a normal qq-plot of the residuals shows a near straight line along $y = x$ with a little bit of deviation at the tails, similar to what we saw in (a). The straight line along $y = x$ is a good indication of normality. A Shapiro-Wilks test can be used to confirm this. The null hypothesis for a Shapiro-Wilks test is that the samples come from a normal distribution. The p-value for this Shapiro-Wilks test is 0.6861 and so we fail to reject the null hypothesis and conclude that there is not enough evidence at a 5% level of significance to conclude that the the residuals are not normally distributed.

Step 3 - Plot Residuals vs. Fitted Values

```r
# Plot residuals vs fitted values
plot(fitted(modelZ), resids)
```
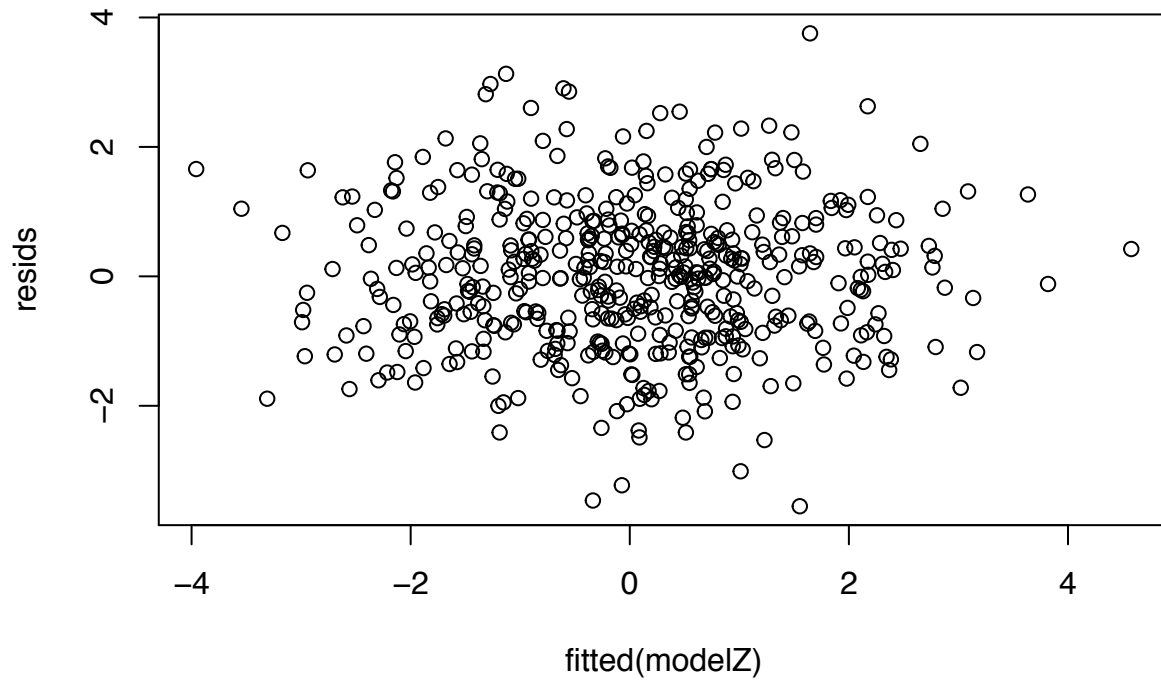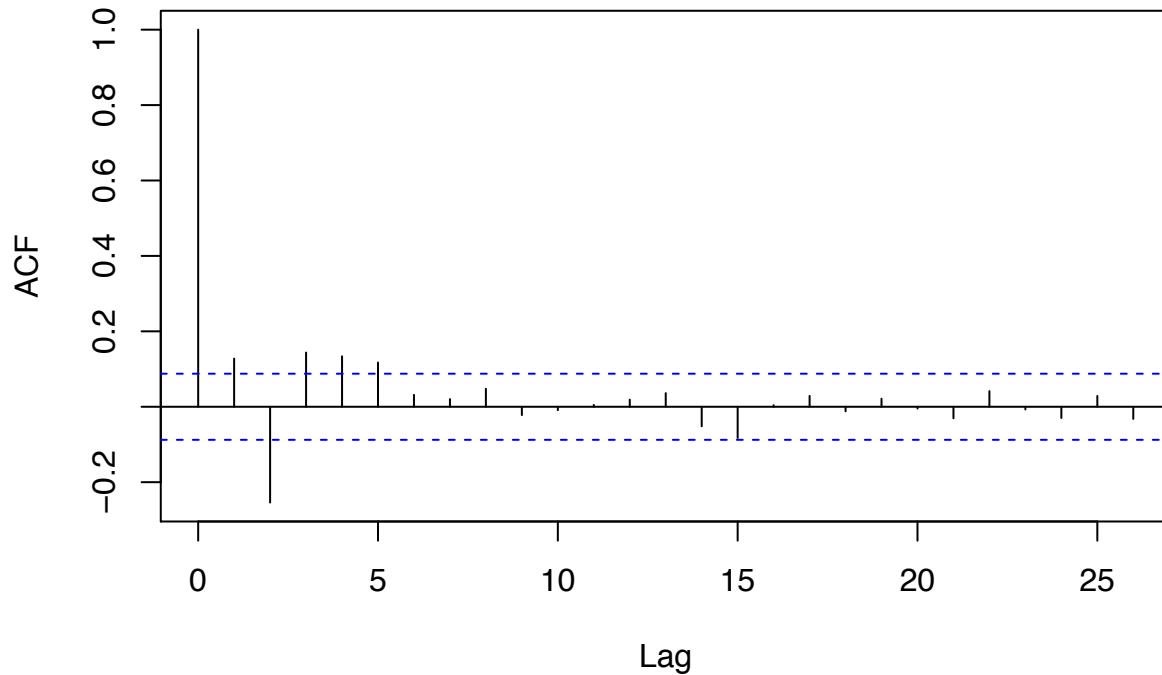
There is no pattern evident here, the residuls vs fitted values plot seems to show a random scatter with residuals centered around 0. There seems to be no relation between residuals and fitted values.

Step 4 - ACF Residuals and Ljung-Box-Pierce Tests

```r
# ACF Residuals
acf(resids)
```

Series resids


```r
# Do a Ljung-Box-Pierce test to check these in more detail
Box.test(resids, lag = 5, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  resids
## X-squared = 67.243, df = 5, p-value = 3.837e-13
```
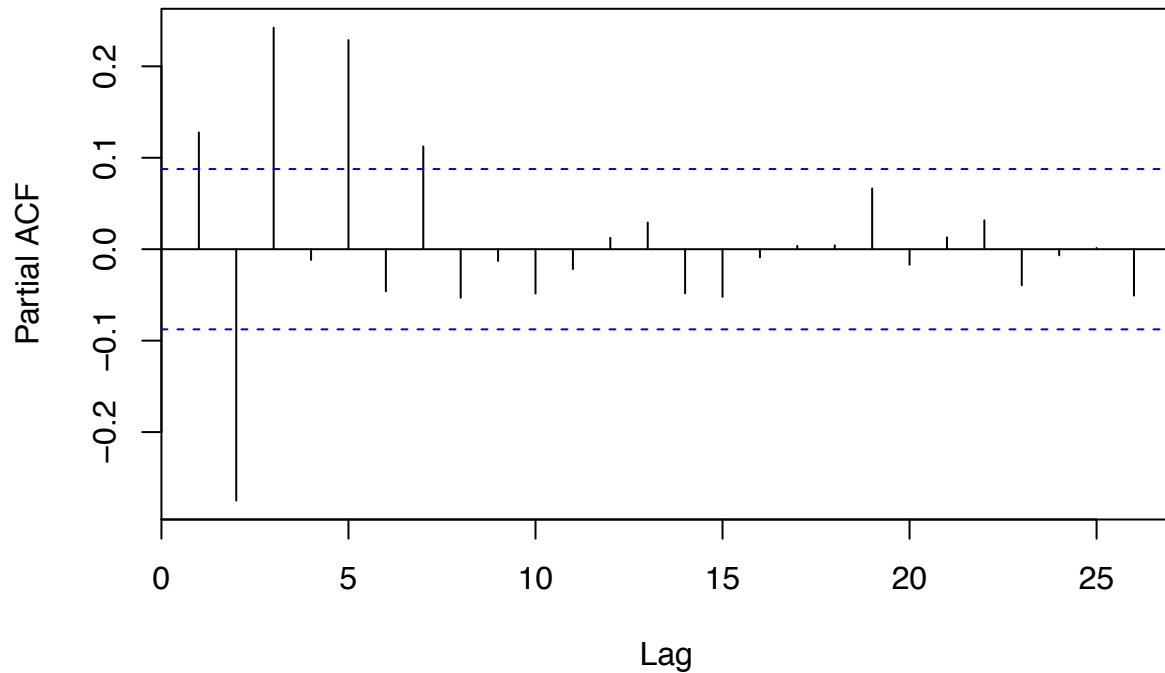
The ACF plot above shows statistically significant (non null) ACF values for low lags at a significance level of 5%. We can say this because the ACD values for lag 5 and below extend beyond the blue dashed lines.

We can conduct a more rigorous test of this using a Ljung-Box test with the lag set to 5. This tests the null hypothesis that there is no autocorrelation for any lags up to and including lag 5. The p-value comes out as 3.837e-13 with test statistic of 67.243. Since the p-value is smaller than 0.05 we reject the null hypothesis and conclude that there is enough evidence to suggest significant autocorrelation for at least one of first 5 lags at $\alpha = 0.05$.

Step 5 - PACF Residuals

```r
pacf(resids)
```

Series resids

The PACF plot of the residual values displayed above indicates statistically significant PACF values for low lags.

Summary

Based on steps 4 and 5 we reject the AR(2) model as an acceptable fit to Series Z.

Q3

(a)

Are the parameters significant?

Use the `coeftest()` command on the newly defined model:

```
modelZ = arima(Z, order = c(10,0,0), include.mean = FALSE)
resids = residuals(modelZ)

# test coefficients
coeftest(modelZ)
```

```
##
## z test of coefficients:
##
##        Estimate Std. Error  z value  Pr(>|z|)
## ar1    1.243151   0.044721  27.7977 < 2.2e-16 ***
## ar2   -1.021776   0.071533 -14.2840 < 2.2e-16 ***
## ar3    0.864821   0.084612  10.2210 < 2.2e-16 ***
```

16

```
## ar4  -0.629263    0.092152   -6.8285 8.578e-12 ***
## ar5   0.559293    0.094632    5.9102 3.417e-09 ***
## ar6  -0.435401    0.094739   -4.5958 4.311e-06 ***
## ar7   0.321994    0.092804    3.4696 0.0005212 ***
## ar8  -0.208950    0.085816   -2.4348 0.0148983 *
## ar9   0.075427    0.072839    1.0355 0.3004238
## ar10 -0.021601    0.045484   -0.4749 0.6348375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to this test all of the model parameters are significant except for ar9 and ar10 (ie. ϕ_9 and ϕ_{10}).

Examine residuals

Step 1 - Plot of Residuals vs. Time

```
# Plot of Residuals vs. Time
plot(resids)
```
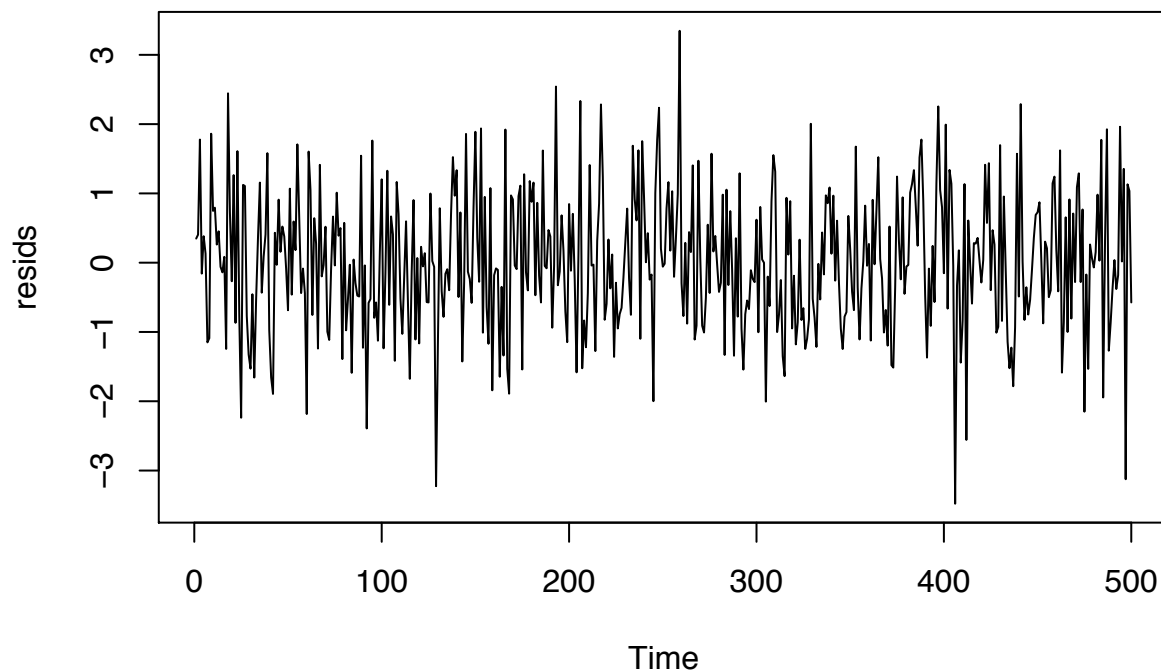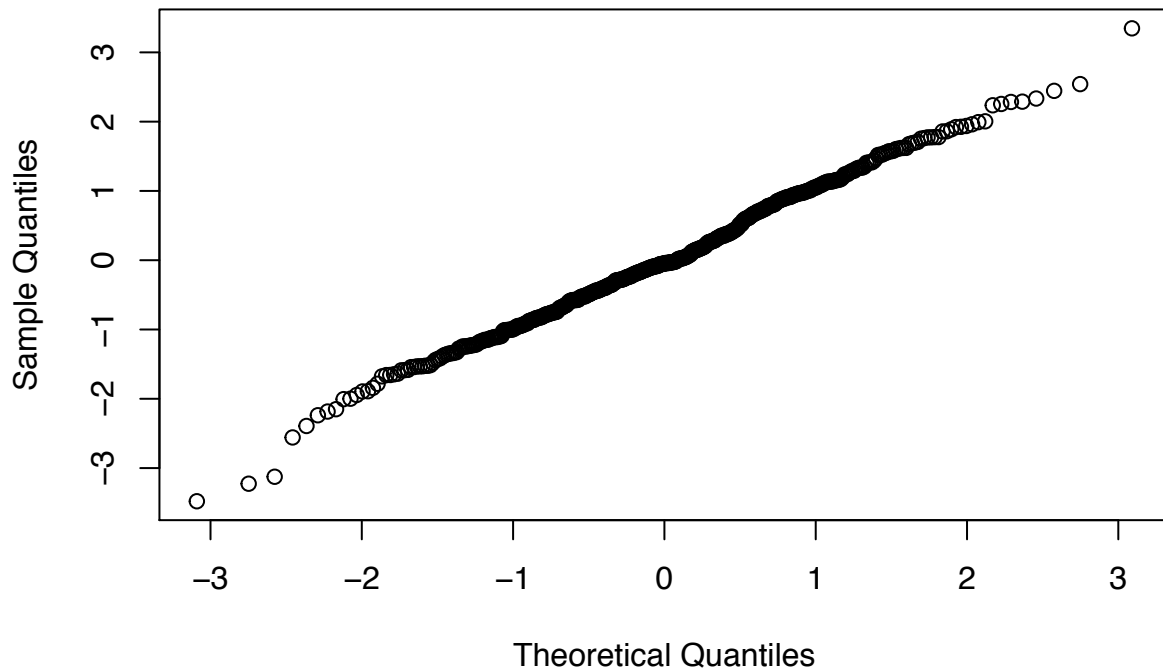

A plot of the residuals shows that they are centered around 0 and have no pattern as we would hope.

Step 2 - Normality of Residuals

```
# Check normality of residuals via QQ plot
qqnorm(resids)
```
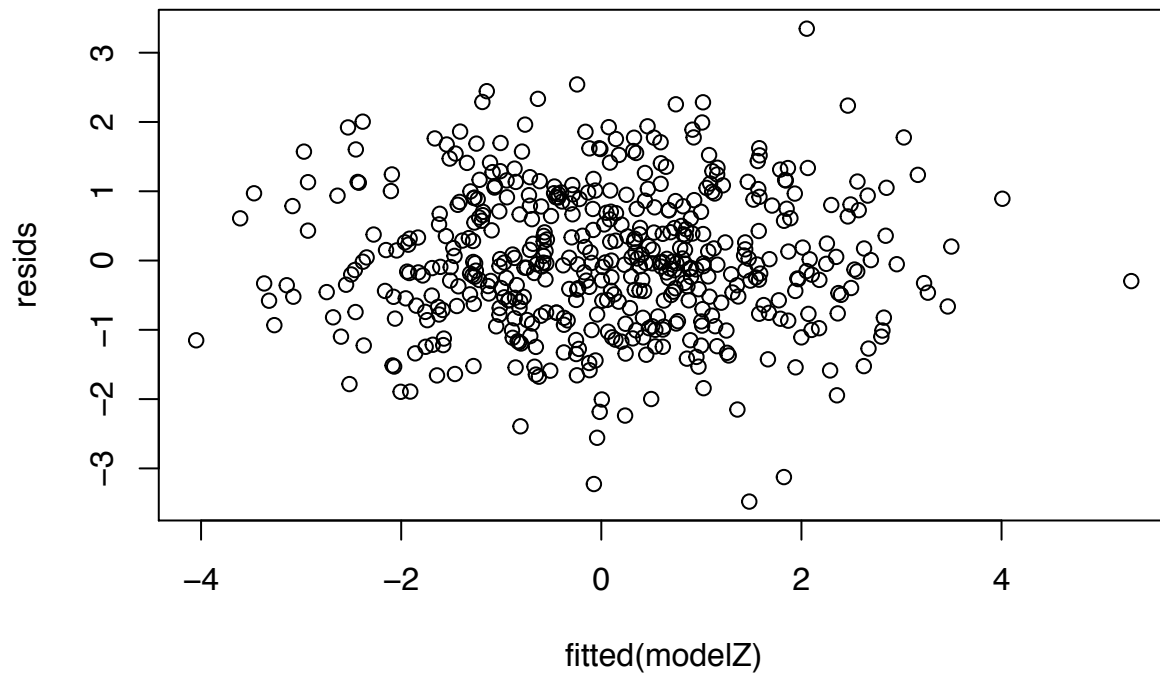
17

Normal Q–Q Plot


```r
# Shapiro-Wilks Test
shapiro.test(resids)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resids
## W = 0.99586, p-value = 0.2136
```

Firstly, a normal qq-plot of the residuals shows a near straight line with a little bit of deviation at the tails. The straight line is a good indication of normality. A Shapiro-Wilks test can be used to confirm this. The null hypothesis for a Shapiro-Wilks test is that the samples come from a normal distribution. The p-value for this Shapiro-Wilks test is 0.2136 and so we fail to reject the null hypothesis and conclude that there is not enough evidence at a 5% level of significance to conclude that the the residuals are not normally distributed.

Step 3 - Plot Residuals vs. Fitted Values

```r
# Plot residuals vs fitted values
plot(fitted(modelZ), resids)
```
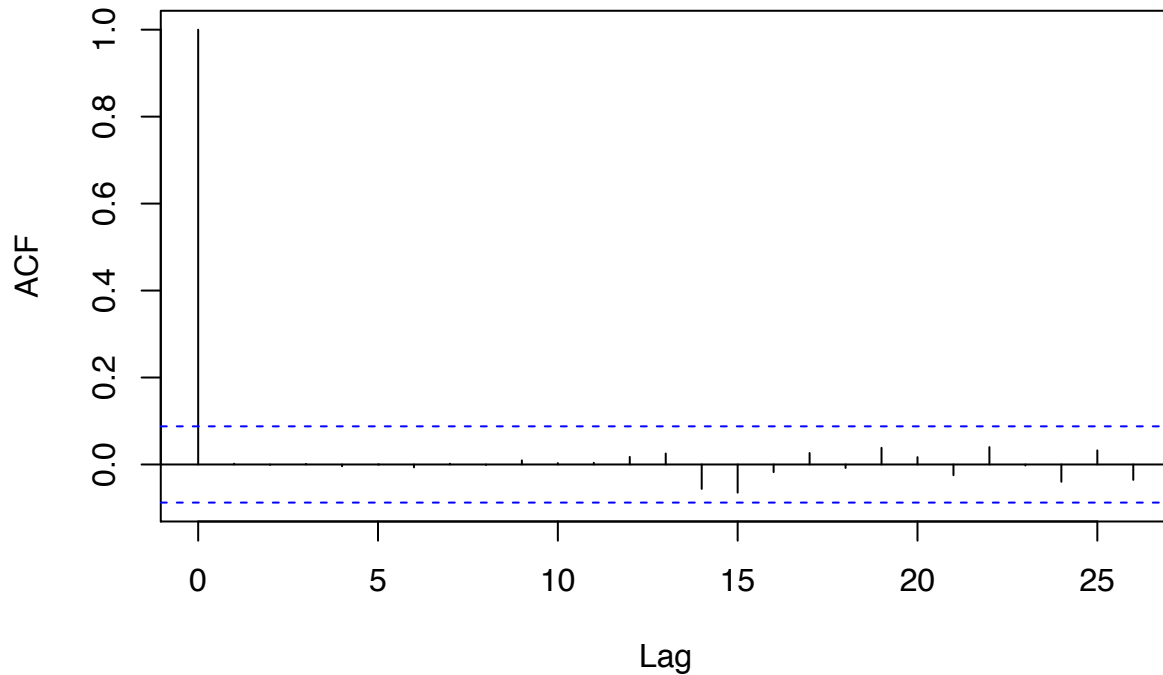
There is no pattern evident here. There seems to be no relation between residuals and fitted values. Also, residuals are centred at 0 as we would hope.

Step 4 - ACF Residuals and Ljung-Box-Pierce Tests

```
# ACF Residuals
acf(resids)
```

Series resids


```r
# ie. for lag = 1,2,3,4,5 the ACF crosses the dashed blue line
# Do a Ljung-Box-Pierce test to check these in more detail
Box.test(resids, lag = 5, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  resids
## X-squared = 0.011461, df = 5, p-value = 1
```

The ACF plot above shows no statistically significant (non null) ACF values at any lag (execpt the trivial case at lag 0 which will always have autocorrelation of 1) at a significance level of 5%.

We can conduct a more rigorous test of this using a Ljung-Box test with the lag set to 5. This tests the null hypothesis that there is no autocorrelation for lags up to and including 5. The p-value comes out as 1. Since the p-value is greater than 0.05 we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest any significant autocorrelation for any of the first 5 lags at $\alpha = 0.05$.
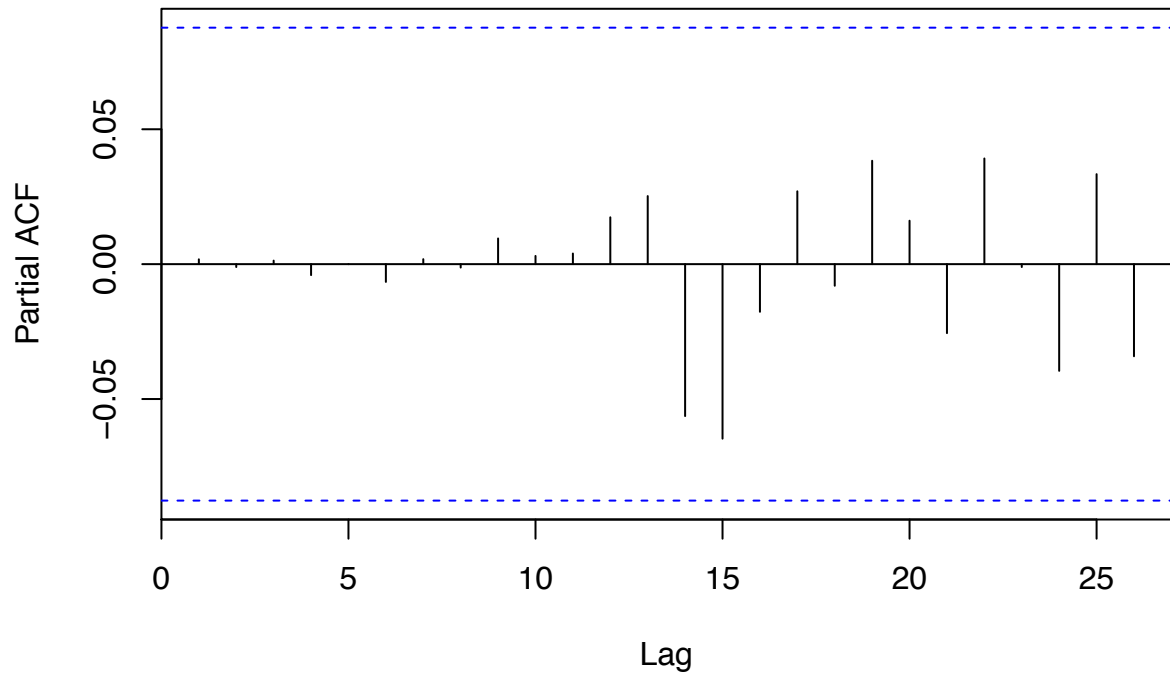
Step 5 - PACF Residuals

```r
pacf(resids)
```

Series resids

The PACF plot of the residual values displayed above indicates null PACF values for all lags.

(b)

An ARMA(p, q) time series can be approximated by a AR(p) time series for large values of p if the roots of the MA polynomial are strictly less than 1 in magnitude (ie. if $|\theta_i| < 1$ for all $i = 1, 2, ...q$). As we have seen in previous parts, this condition holds for our time series.

Note: The condition described above is equivalent to invertability and so we can say that this AR(10) fits the ARMA model well due to the invertibility of the MA portion of Series Z.