Time Series 413, Assignment 5

Nonstationary Time Series and Modeling Covariates (TS5)

Reed Ballesteros

Northwestern University SPS, Fall 2022
MSDS-413-DL
Instructor: Dr. Jamie D. Riggs, Ph.D
2022-10-24

The following list defines the data sets and their respective variables. The monthly market liquidity measures are from Professors Pastor and Stambaugh. The data are available from Wharton WRDS and are in the file **m-PastorStambaugh.txt**. See

https://breakingdownfinance.com/finance-topics/equity-valuation/pastor-stambaugh-model/

The following list defines the variables:

- DATE: is the month the data were collected
- PS_LEVEL: levels of aggregate liquidity
- PS_INNOV: innovations in aggregate liquidity
- PS_VWF: traded liquidity factor

The monthly Fama-Bliss bond yields have maturities of 1 and 3 years. The data are available from CRSP and are in the file **m-FamaBlissdbndyields.txt**.The following list defines the variables:

- qdate: month of the yields
- yield1: 1-year yields
- yield3: 3-year yields

1. Outlier Management (20 points)

Consider the monthly market liquidity measure of Professors Pastor and Stambaugh. The data are available from Wharton WRDS and are in the file **m-PastorStambaugh.txt**. Consider the variable PS level and denote the series by x_t.

1.1. Perform EDA.

Validate data as a time series:

```
## [1] 605
```

```
## [1] 605
```

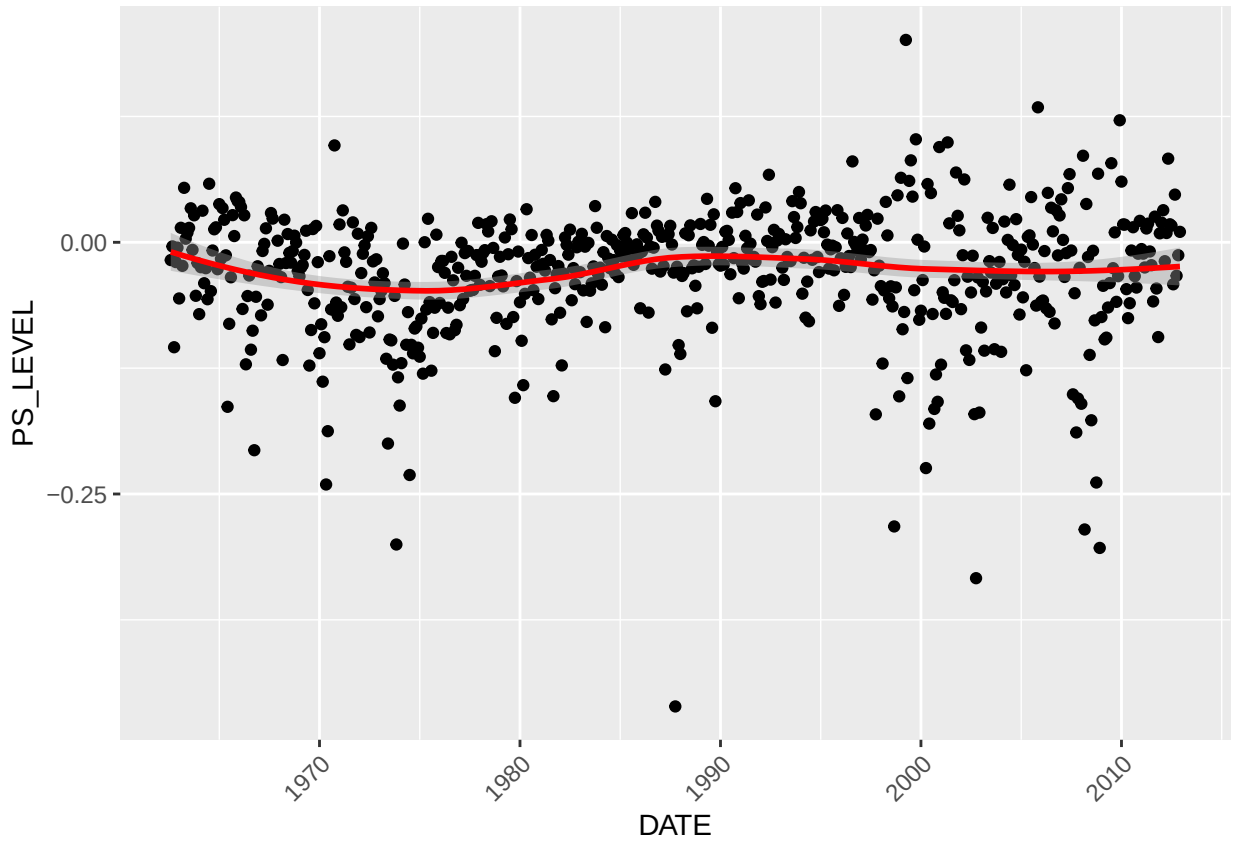
We have 140 unique years in 140 observations, which meets the $H_{10} : x_{it},\ i \in \{1, 2\},\ t \in \{1, 2, ..., n\}$ requirement for time series validation.

```
## [1] 605
```

```
## df
##   1
## 604
```

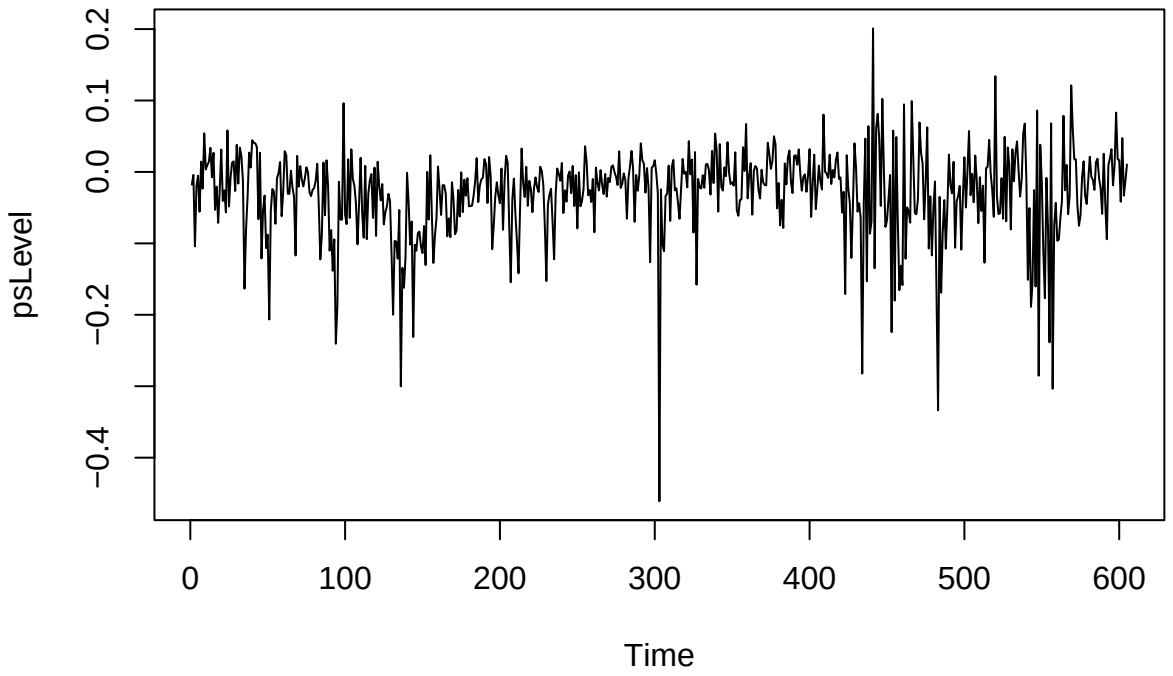
From the test above, we can verify that the constant time span between each date is only one month, denoted by the single value 1. This meets the $H_{20} : (t+1)-t = c, t \in \{1, 2..., n\}$ requirement for time series validation.

Let us create a general plot of the data:



We seem to observe a general flat trend in the data, but might have to specifically test via t-test for mean zero to confirm linear trend does not exist.

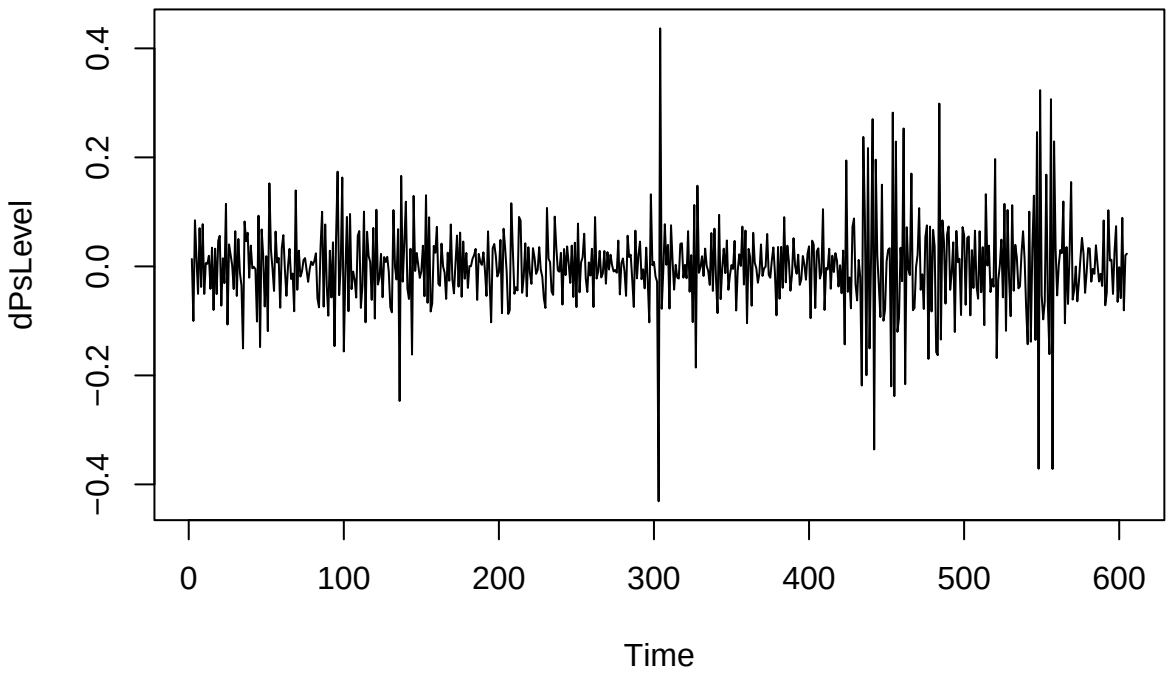
Plot: PS_LEVEL



Plotting as a PS_LEVEL time series data visually identifies some possible outliers, such as the PS_LEVEL drop to almost -0.45 at around time 300, and the mean is not zero and that some kind of trend exists.

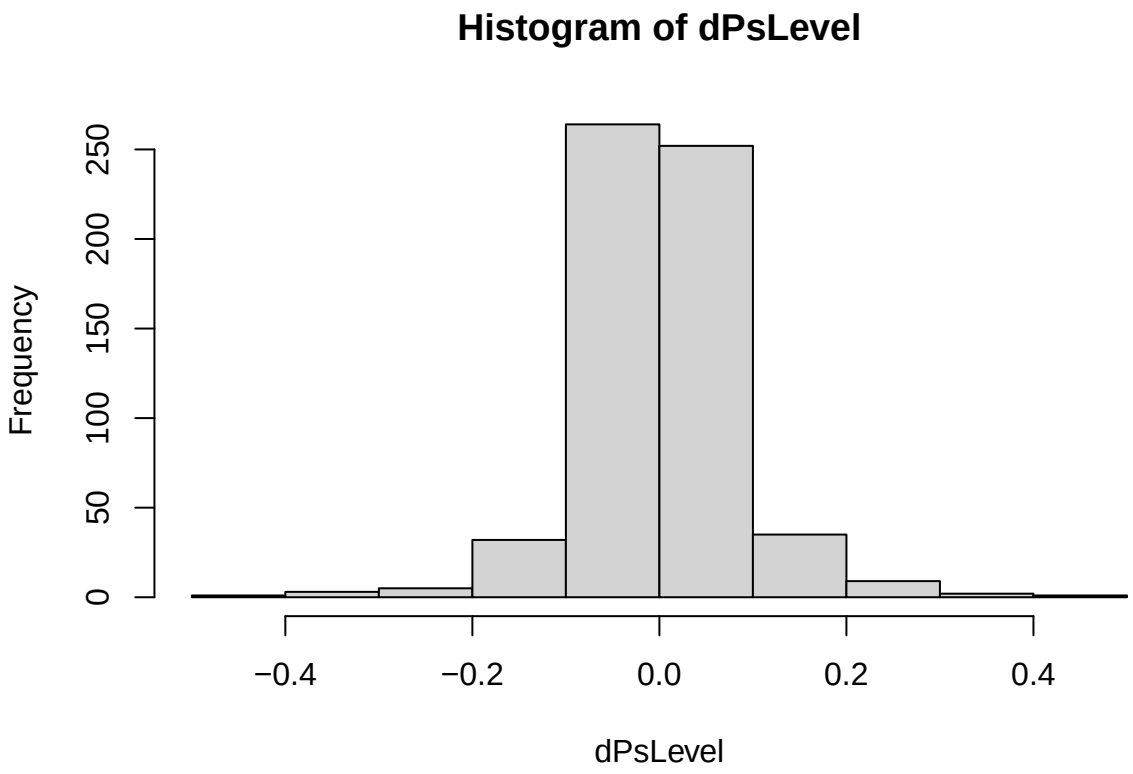
At this point instead of continuing with the EDA since we do not yet have mean zero we should transform the data, performing diff(PS_LEVEL).

Plot: diff(PS_LEVEL)



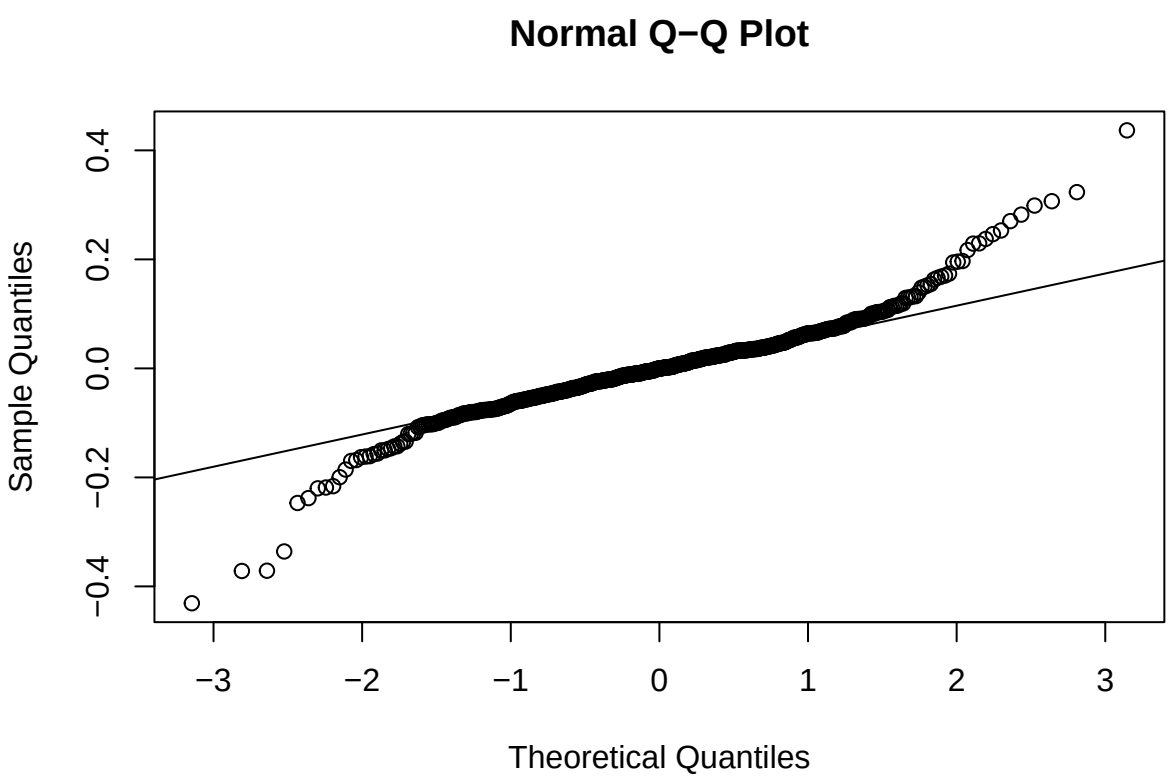
We now see that a linear trend is removed with mean 0 of diff(PS_LEVEL). We still notice the outliers at around the 300 mark.

Histogram: diff(PS_LEVEL)



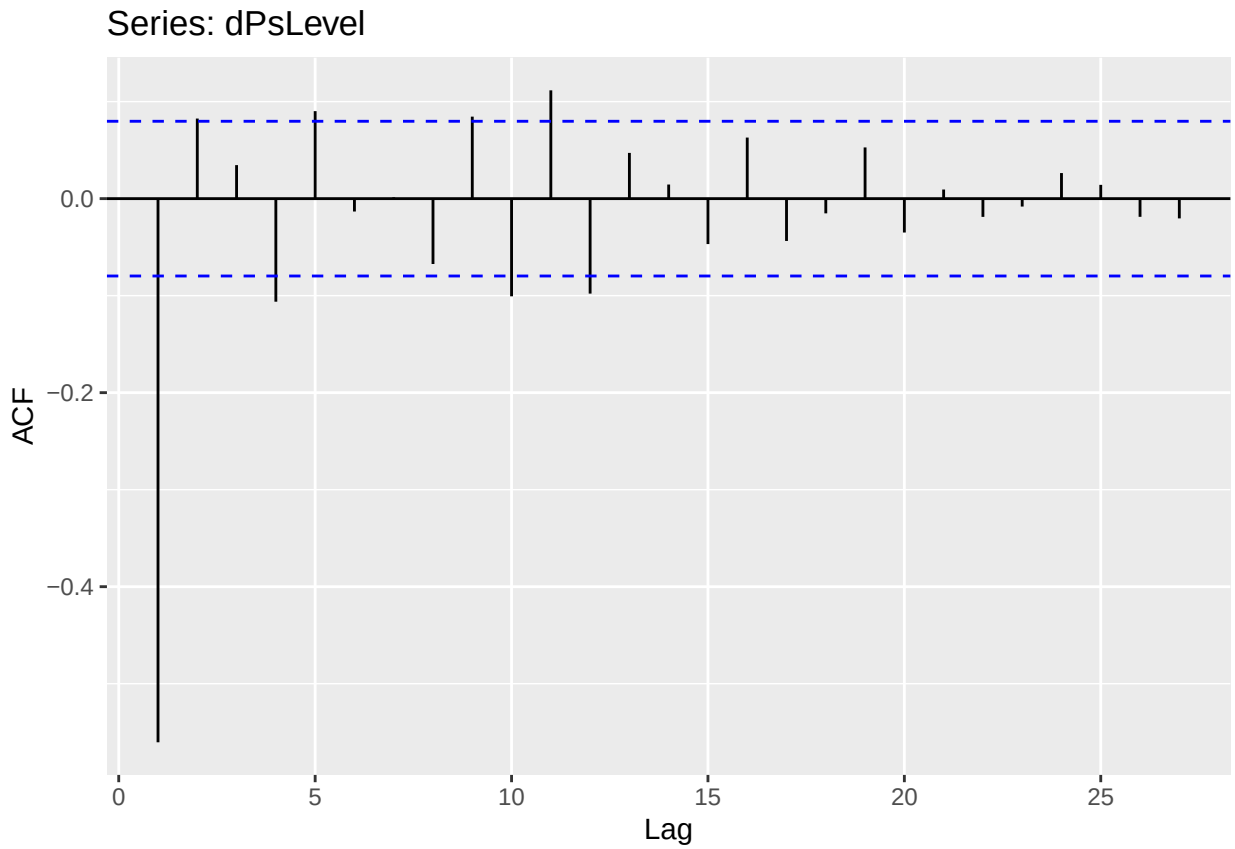
We can observe a left-skewed, tall distribution, thus showing non-normalcy in respect to a Gaussian PDF.

Q-Q Plot: diff(PS_LEVEL)



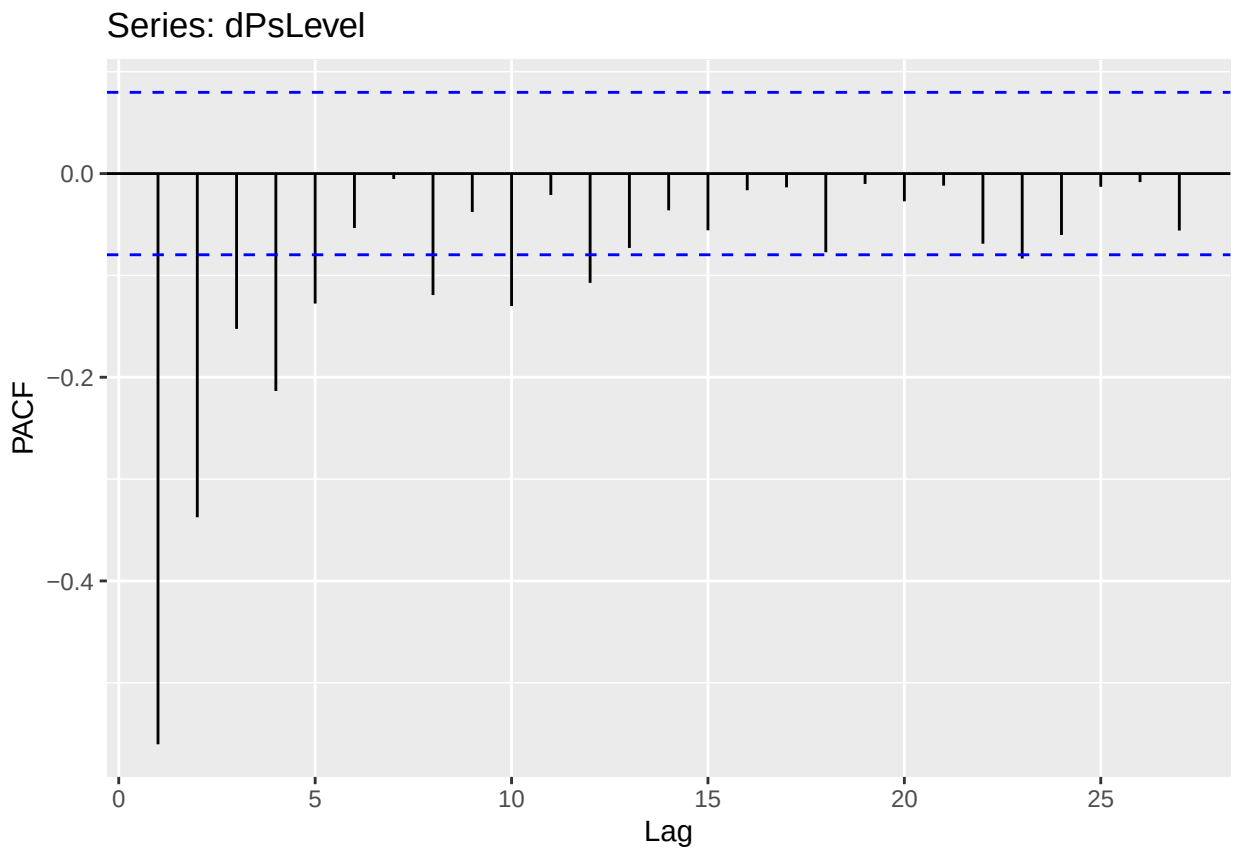
We can notice skewness and (excess) Kurtotis from the Q-Q plot as the ends of the plot veer off the idea normal line, thus showing non-normalcy in respect to a Gaussian PDF.

Stationarity: ACF diff(PS_LEVEL)



Based on the ACF plot above, we'll use an MA(1) for our ARIMA model.

Stationarity: PACF diff(PS_LEVEL)



Based on the PACF plot above, we'll use an AR(5) for our ARIMA model.

Mean 0: T-Test diff(PS_LEVEL)

```
##
##  One Sample t-test
##
## data:  data
## t = 0.01374, df = 603, p-value = 0.989
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.006583674  0.006676448
## sample estimates:
##    mean of x
## 4.638707e-05
##
## T-Test: mean is statistically zero, linear trend *REMOVED* ->
## *FAIL* to reject H0

## [1] TRUE
```

The 95% Confidence Interval (CI) of the diff(PS_LEVEL) data does contain 0, thus showing normalcy in respect to a Gaussian PDF, as well as the linear trend removed.

Normalcy: Skewness diff(PS_LEVEL)

```
##       skew     lwr.ci     upr.ci
```

```
## 0.03720920 0.01046075 0.05876747
## Skew: has *RIGHT* skewness,
## property does *NOT* conform to normality and Gaussian PDF
```

```
## [1] FALSE
```

The diff(PS_LEVEL) data has a distribution with right skewness, thus showing non-normalcy in respect to a Gaussian PDF.

Normalcy: (excess) Kurtosis diff(PS_LEVEL)

```
##     kurt   lwr.ci   upr.ci
## 6.265555 6.528789 6.795701
## Kurt: has *TALL thick-tailed* (excess) kurtosis,
## property does *NOT* conform to normality and Gaussian PDF
```

```
## [1] FALSE
```

The diff(PS_LEVEL) data has a distribution with tall (excess) Kurtosis, thus showing non-normalcy in respect to a Gaussian PDF.

Constant Variance: Breusch-Pagan Test diff(PS_LEVEL)

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(data ~ seq(1, length(data)))
## BP = 4.1297, df = 1, p-value = 0.04214
##
## Breusch-Pagan: *NON*-constant variance, possible clustering,
## heteroscedastic -> reject H0
```

```
##    BP
## FALSE
```

While we can see signs of non-constant variance in the time series plot, the Breusch-Pagen test confirms it.

Lag independence: Box-Ljung test diff(PS_LEVEL)

```
##
##  Box-Ljung test
##
## data:  data
## X-squared = 143.38, df = 30, p-value < 2.2e-16
##
## Box-Ljung: implies dependency present over 30 lags,
## autocorrelation present -> reject H0
```

```
## [1] FALSE
```

The Box-Ljung test shows the diff(PS_LEVEL) data contains lag dependency and thus autocorrelation.

1.2. Build a time series model for x_t (the expected value equation) using the model-building process. Write the equation of the model to be fitted (not the fitted model).

Based on the EDA above we will create an ARIMA(5,1,1) model, with d=1 since we've differenced the PS_LEVEL data.

```
p<-5;d<-1;q<-1
m <- Arima(psLevel,order=c(p,d,q))
```

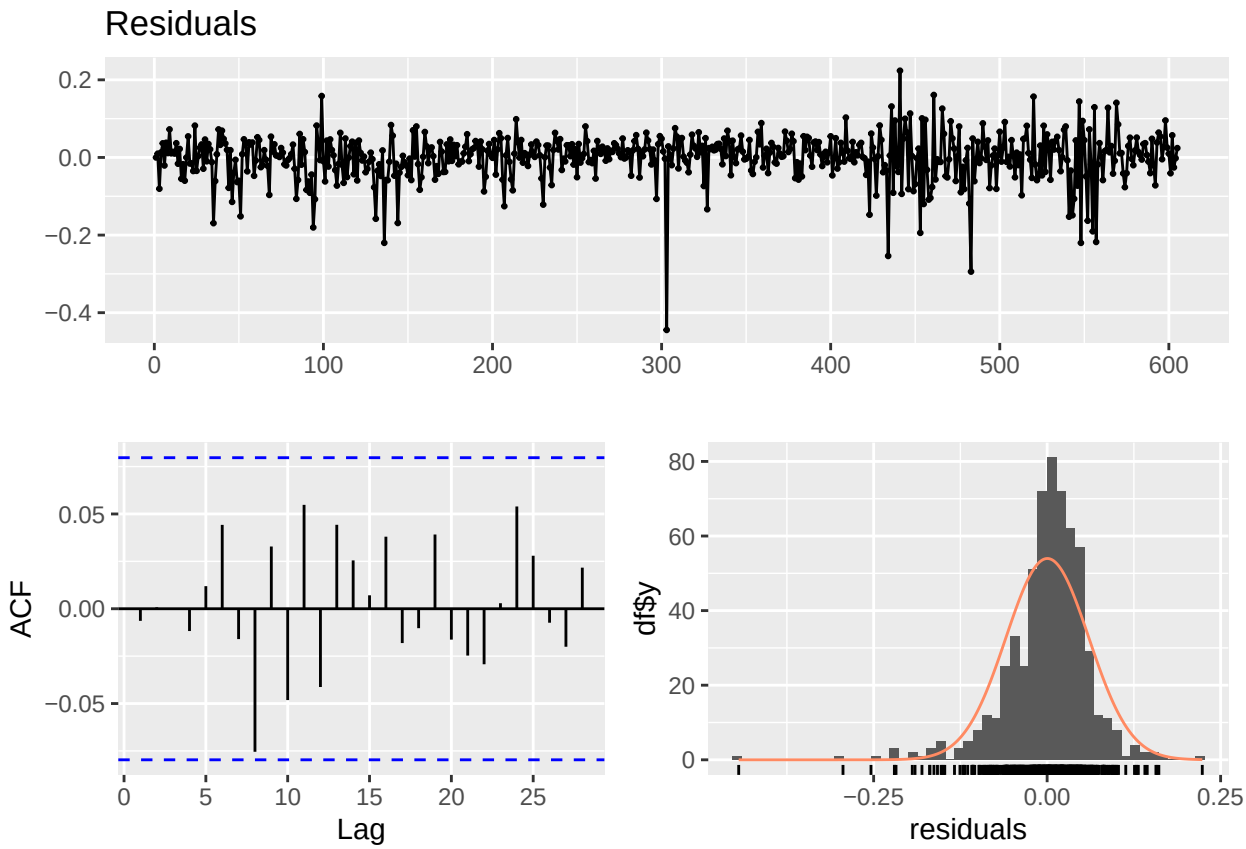
```
## Series: psLevel
## ARIMA(5,1,1)
##
## Coefficients:
##          ar1     ar2     ar3     ar4     ar5      ma1
##       0.0631  0.1863  0.1318  0.0111  0.1102  -0.9966
## s.e.  0.0414  0.0413  0.0416  0.0413  0.0413   0.0137
##
## sigma^2 = 0.003605:  log likelihood = 842.91
## AIC=-1671.83   AICc=-1671.64   BIC=-1641
##
## Training set error measures:
##                        ME       RMSE        MAE      MPE     MAPE      MASE
## Training set 0.0003298096 0.05969545 0.04142743 164.9897 421.9081 0.7242494
##                      ACF1
## Training set -0.006376128
```

Based on the summary above we create the following generalized equation:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + \phi_4 x_{t-4} + \phi_5 x_{t-5} + \theta_1 z_{t-1}$$

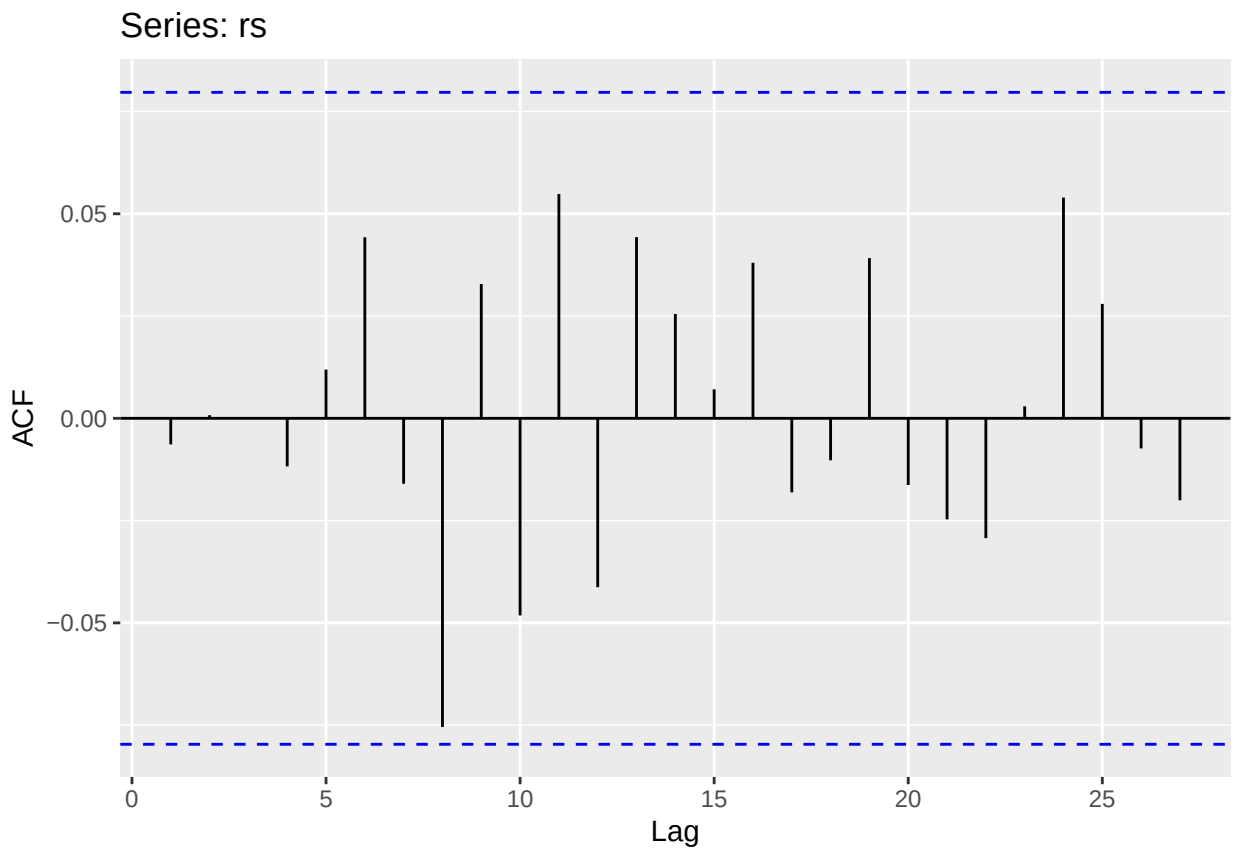
We will conduct model diagnostics.

Check residuals:



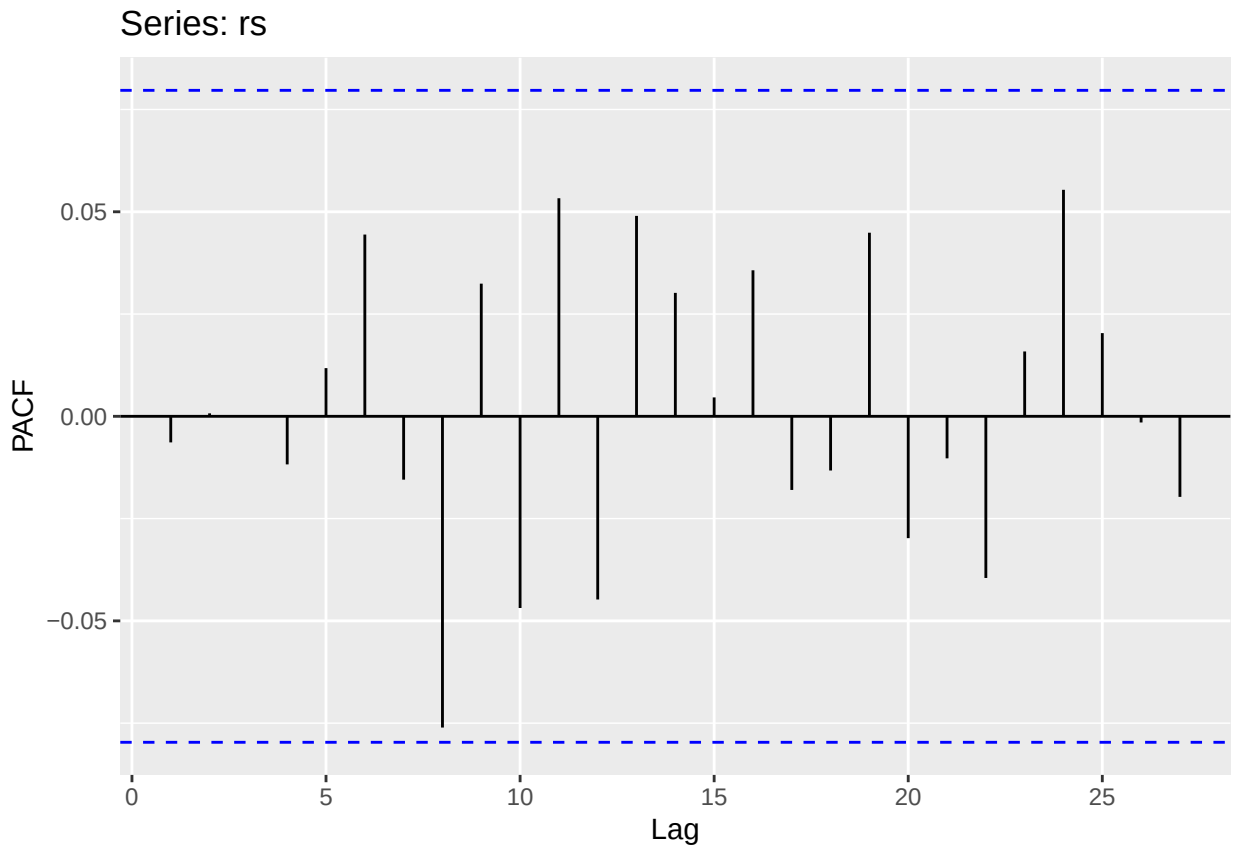
We see the residuals plot with maybe constant variance (we'll further test with McLeod-Li below) and mean 0, the ACF plot looks stationary, and the distribution is tall and left-skewed, therefore not normal in respect to a Gaussian PDF.

Stationarity: ACF Plot



The model looks stationary, will test with KPSS/ADF below.

Stationarity: PACF Plot



The model looks stationary, will test with KPSS/ADF below.

Mean Zero: T-Test

```
##
##  One Sample t-test
##
## data:  data
## t = 0.13578, df = 604, p-value = 0.892
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.004440377  0.005099996
## sample estimates:
##    mean of x
## 0.0003298096
##
## T-Test: mean is statistically zero, linear trend *REMOVED* ->
## *FAIL* to reject H0
```

```
## [1] TRUE
```

The 95% CI contains zero, therefore the mean is statistically 0 and the linear trend is removed.

Normalcy: Skewness

```
##      skew    lwr.ci    upr.ci
```

```
## -1.437539 -1.528881 -1.467790
## Skew: has *LEFT* skewness,
## property does *NOT* conform to normality and Gaussian PDF
```

```
## [1] FALSE
```

The model has exhibits left skewness, showing non-normalcy in respect to a Gaussian PDF.

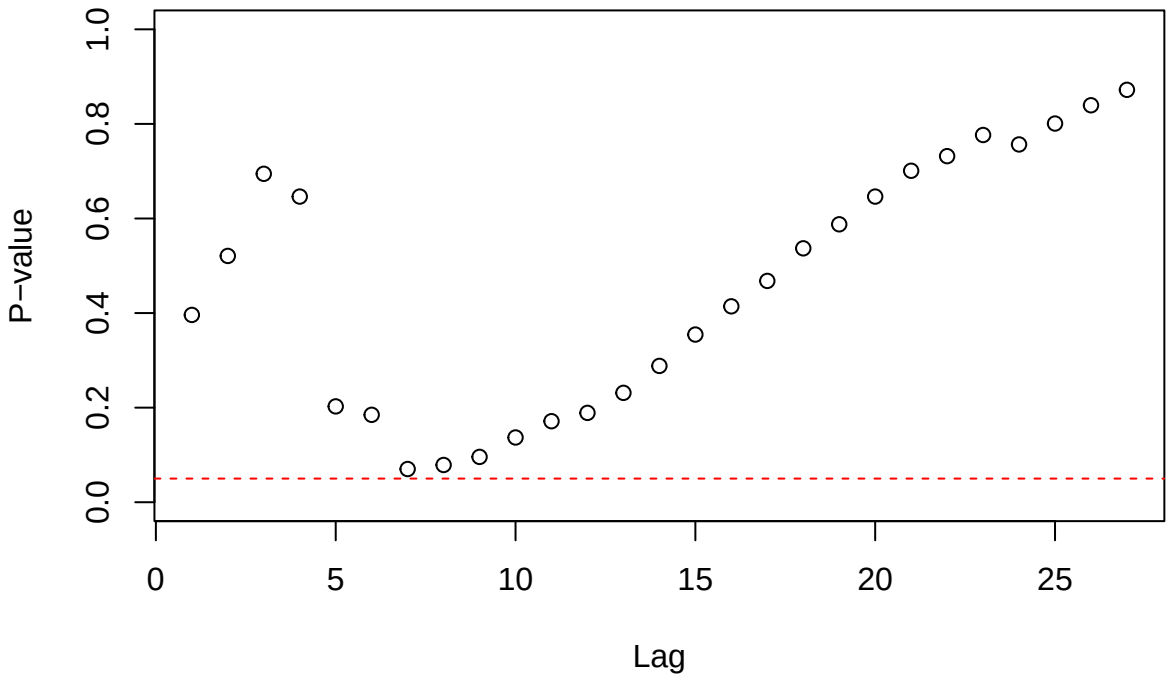
Normalcy: (excess) Kurtosis

```
##     kurt   lwr.ci   upr.ci
## 7.324127 7.480358 7.876927
## Kurt: has *TALL thick-tailed* (excess) kurtosis,
## property does *NOT* conform to normality and Gaussian PDF
```

```
## [1] FALSE
```

The model has exhibits tall (excess) Kurtosis, showing non-normalcy in respect to a Gaussian PDF.

Constant Variance: McLeod-Li Test




```
## McLeod-Li: constant variance, homoscedastic -> *FAIL* to reject H0
## McLeod-Li: Lags >= 0.05:
##   (none)
```

```
## [1] TRUE
```

The model exhibits constant variance.

Lag independence: Box-Ljung test

```
##
##  Box-Ljung test
##
## data:  data
## X-squared = 28.486, df = 30, p-value = 0.5447
##
## Box-Ljung: implies independence over 30 lags,
## *NO* autocorrelation -> *FAIL* to reject H0
```

```
## [1] TRUE
```

Box-Ljung test shows model has lag independence and thus no autocorrelation.

Stationarity: ADF test

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -3.4863
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:54:44 2023 by user: Reed
##
## ADF: contains *NO* unit roots over 30 lags, indicates *NO* mean drift,
## business cycles *NOT* present, series is stationary -> reject H0
```

```
##
## FALSE
```

The ADF test shows the model is stationary.

Stationarity: KPSS test

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 6 lags.
##
## Value of test-statistic is: 0.1281
##
## Critical value for a significance level of:
```

```
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216
##
## KPSS: *NO* unit roots, *NO* linear trend, slope zero,
## series is trend stationary -> *FAIL* to reject H0
```

```
## [1] TRUE
```

The KPSS test shows the model is stationary.

Checking for ARIMA(5,1,1) business cycles:

```
## [1] 13.854  2.406  4.018
```

We see there are three business cycles, 14-month, 2-month, and 4-month business cycles.

1.3. Identify the largest outlier in the series. Refine the fitted model by using an indicator for the outlier. Write the equation of the refined model (not the fitted model).

From the residuals plot in section 1.2, the largest outlier in the series lies below the mean, so we will look for the lowest value in the residuals.

```
which.min(m$residuals)
```

```
## [1] 303
```

We find the largest outlier is at index 303.

Let's create a model using the outlier:

```
m <- Arima(psLevel,order=c(p,d,q), xreg=i303)
```

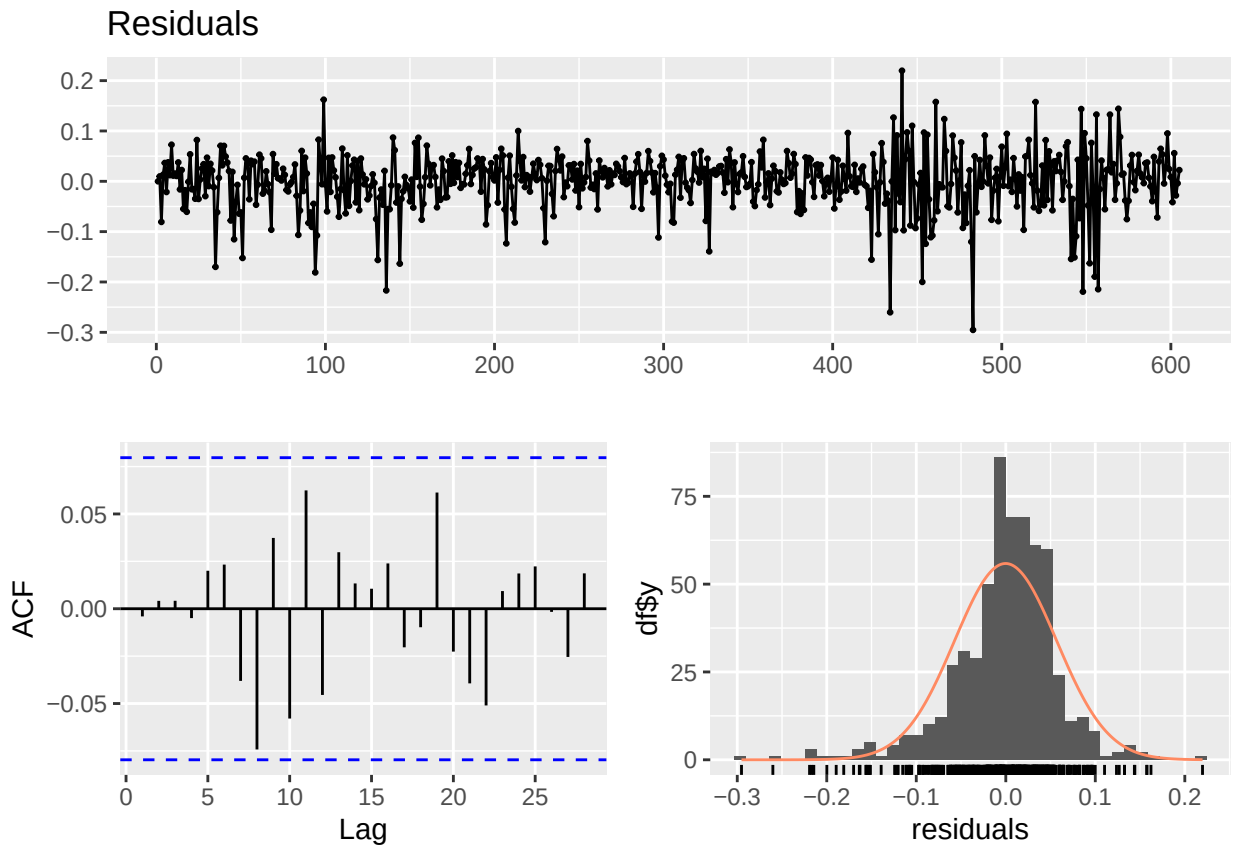
```
## Series: psLevel
## Regression with ARIMA(5,1,1) errors
##
## Coefficients:
##          ar1     ar2     ar3     ar4     ar5      ma1     xreg
##       0.0558  0.1789  0.1222  0.0074  0.1169  -0.9852  -0.4274
## s.e.  0.0447  0.0442  0.0441  0.0438  0.0438   0.0190   0.0551
##
## sigma^2 = 0.003292:  log likelihood = 871.5
## AIC=-1727   AICc=-1726.75   BIC=-1691.77
##
## Training set error measures:
##                         ME       RMSE        MAE      MPE     MAPE      MASE
## Training set -0.0002075432 0.05699791 0.04070132 172.7945 422.2474 0.7115554
##                      ACF1
## Training set -0.004040577
```

Based on the summary above we create the following generalized equation:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + \phi_5 y_{t-5} + \theta_1 z_{t-1} + \beta x_t + \epsilon_t,\ \epsilon_t \sim WN(0, \sigma_\epsilon^2)$$

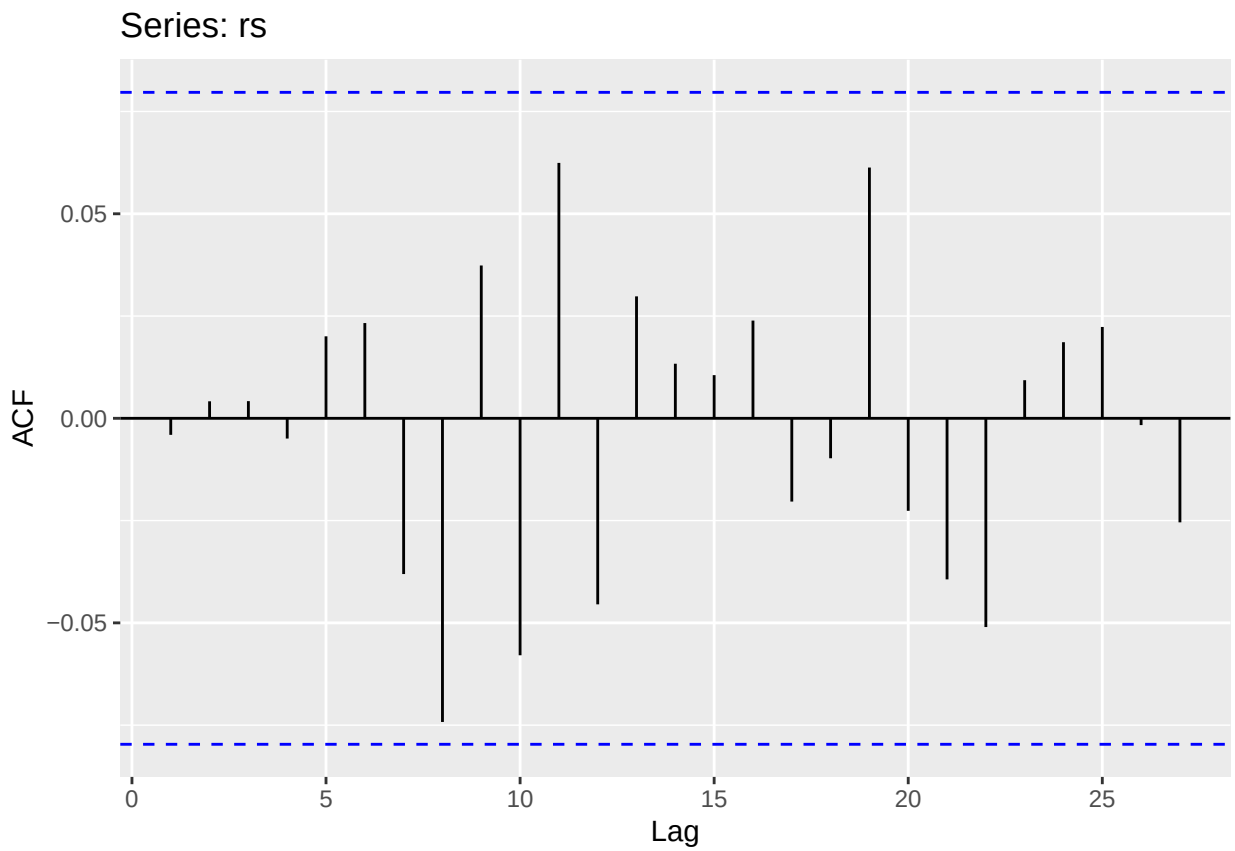
We will conduct model diagnostics.

Check residuals:



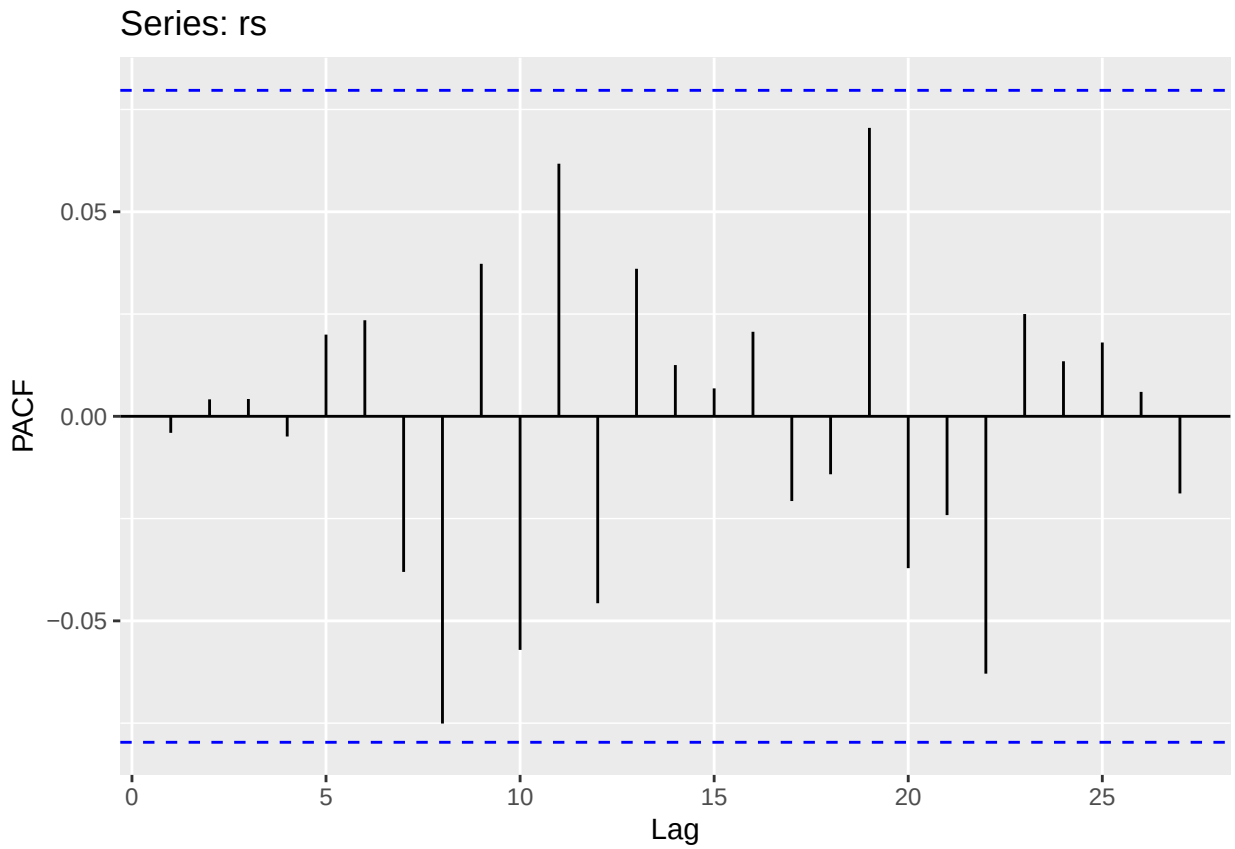
The plot most likely shows mean 0 with no linear trend and maybe eyeball non-constant variance (we'll test with McLeod-Li below). The ACF shows the model lags are stationary. The distribution shows left skewness and tall Kutosis, showing non-normalcy in respect to a Gaussian PDF.

Stationarity: ACF Plot



The model looks stationary with the ACF plot, will test with KPSS/ADF below.

Stationarity: PACF Plot



The model looks stationary with the PACF plot, will test with KPSS/ADF below.

Mean Zero: T-Test

```
##
##  One Sample t-test
##
## data:  data
## t = -0.089489, df = 604, p-value = 0.9287
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.004762212  0.004347126
## sample estimates:
##     mean of x
## -0.0002075432
##
## T-Test: mean is statistically zero, linear trend *REMOVED* ->
## *FAIL* to reject H0
```

```
## [1] TRUE
```

The 95% CI contains zero, therefore the mean is statistically 0 and the linear trend is removed.

Normalcy: Skewness

```
##       skew     lwr.ci     upr.ci
```

```
## -0.8972678 -0.9219062 -0.8924053
## Skew: has *LEFT* skewness,
## property does *NOT* conform to normality and Gaussian PDF
```

```
## [1] FALSE
```

The model has exhibits left skewness, showing non-normalcy in respect to a Gaussian PDF.

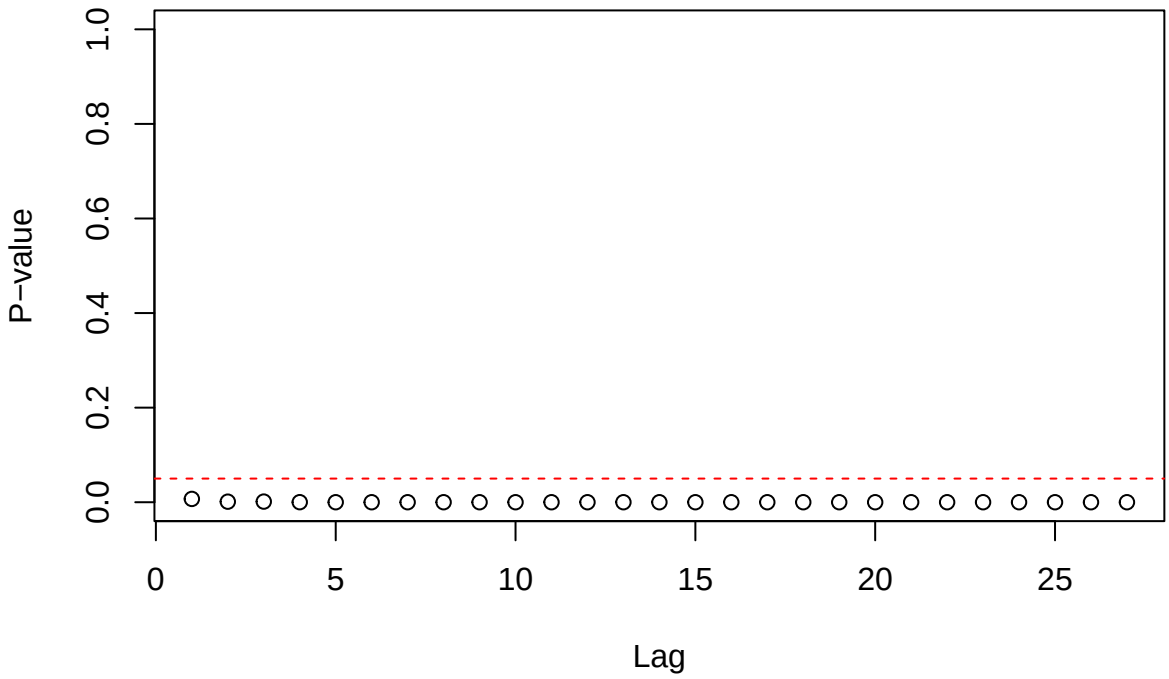
Normalcy: (excess) Kurtosis

```
##     kurt   lwr.ci   upr.ci
## 3.363427 3.412904 3.509188
## Kurt: has *TALL thick-tailed* (excess) kurtosis,
## property does *NOT* conform to normality and Gaussian PDF
```

```
## [1] FALSE
```

The model has exhibits tall (excess) Kurtosis, showing non-normalcy in respect to a Gaussian PDF.

Constant Variance: McLeod-Li Test




```
## McLeod-Li: *NON*-constant variance, heteroscedastic -> reject H0
## McLeod-Li: Lags < 0.05:
## 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27
```

```
## [1] FALSE
```

The model exhibits non-constant variance with many lags with a p-value of under 0.05.

Lag independence: Box-Ljung test

```
##
##  Box-Ljung test
##
## data:  data
## X-squared = 33.304, df = 30, p-value = 0.3095
##
## Box-Ljung: implies independence over 30 lags,
## *NO* autocorrelation -> *FAIL* to reject H0
```

```
## [1] TRUE
```

Box-Ljung test shows model has lag independence and thus no autocorrelation.

Stationarity: ADF test

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -3.6683
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:54:45 2023 by user: Reed
##
## ADF: contains *NO* unit roots over 30 lags, indicates *NO* mean drift,
## business cycles *NOT* present, series is stationary -> reject H0
```

```
##
## FALSE
```

The ADF test shows the model is stationary.

Stationarity: KPSS

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 6 lags.
##
## Value of test-statistic is: 0.0954
##
## Critical value for a significance level of:
```

```
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216
##
## KPSS: *NO* unit roots, *NO* linear trend, slope zero,
## series is trend stationary -> *FAIL* to reject H0

## [1] TRUE
```

The KPSS test shows the model is stationary.

Checking for ARIMA(5,1,1) with outlier business cycles:

```
## [1] 14.129  2.461  4.206
```

The model finds 3 business cycles, a 14-month, 2-month, and 4-month, same as the 'regular' ARIMA(5,1,1) model in section 1.2.

1.4. Further refine the model by setting the least significant parameters to zero. Write the equation of the revised model to be fitted (not the fitted model).

Let us find the least significant parameters from our model in section 1.3:

```
##                t       pval_t       pval_z Pr(>|t|)
## ar1    1.2497804 2.118687e-01 2.113798e-01
## ar2    4.0442261 5.937023e-05 5.249625e-05      ***
## ar3    2.7724796 5.736444e-03 5.563100e-03       **
## ar4    0.1698324 8.651994e-01 8.651419e-01
## ar5    2.6675014 7.848785e-03 7.641758e-03       **
## ma1  -51.9563236 0.000000e+00 0.000000e+00      ***
## xreg  -7.7542408 3.841372e-14 8.881784e-15      ***
```

We find that only ar1 and ar4 are the insignificant components in the model.

Let's create a reduced model:

```
#        A1,A2,A3,A4,A5,M1,XR
c11 <- c(00,NA,NA,00,NA,NA,NA)
m <- Arima(psLevel,order=c(p,d,q),xreg=i303,fixed=c11)   # AICc=-1729.18
```

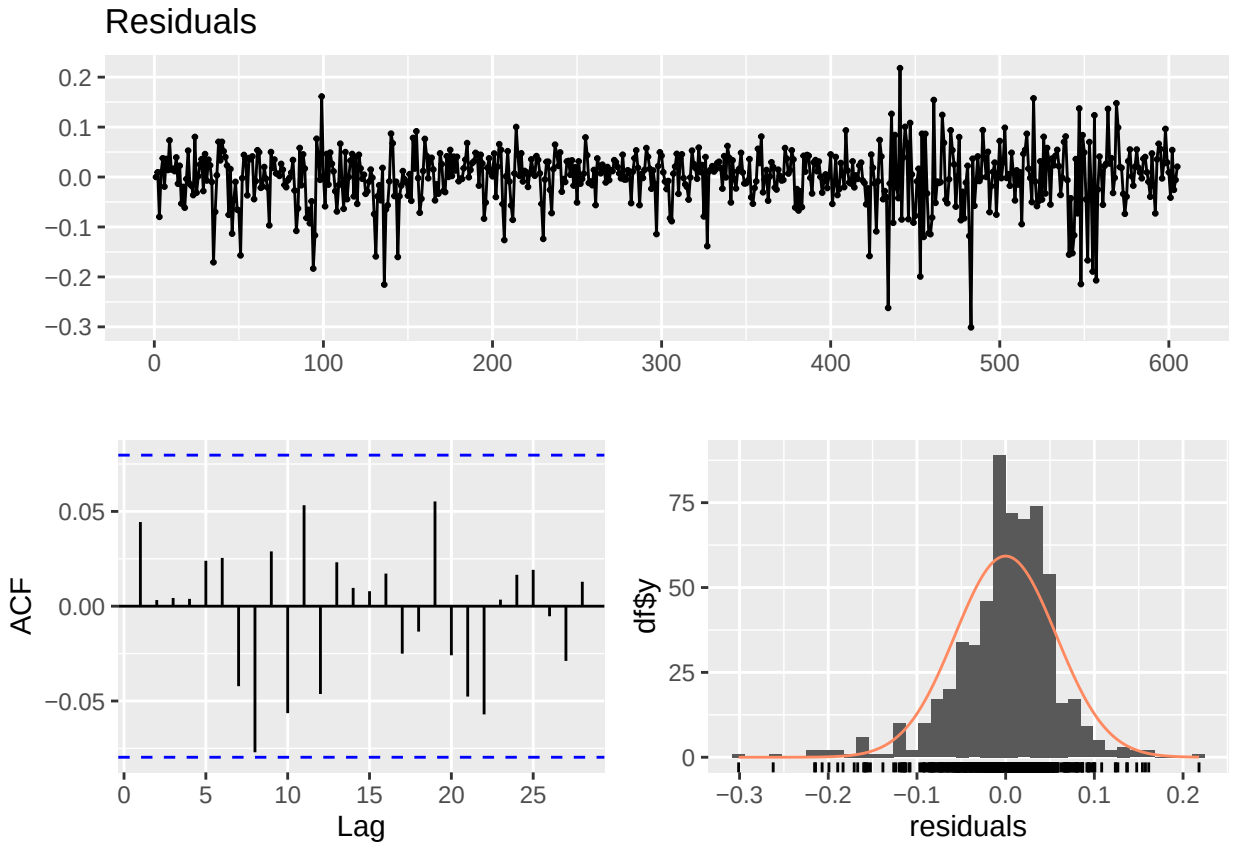
```
## Series: psLevel
## Regression with ARIMA(5,1,1) errors
##
## Coefficients:
##       ar1     ar2     ar3  ar4     ar5      ma1     xreg
##         0  0.1750  0.1247    0  0.1106  -0.9771  -0.4280
## s.e.    0  0.0426  0.0426    0  0.0421   0.0136   0.0551
##
## sigma^2 = 0.003291:  log likelihood = 870.66
## AIC=-1729.32   AICc=-1729.18   BIC=-1702.9
##
## Training set error measures:
##                        ME       RMSE        MAE      MPE     MAPE      MASE
## Training set -6.95523e-05 0.05708635 0.04077557 175.8037 429.6039 0.7128533
##                    ACF1
## Training set 0.04440602
```

Based on the summary above we create the following reduced generalized equation:

$$y_t = \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_5 y_{t-5} + \theta_1 z_{t-1} + \beta x_t + \epsilon_t,\ \epsilon_t \sim WN(0, \sigma_\epsilon^2)$$

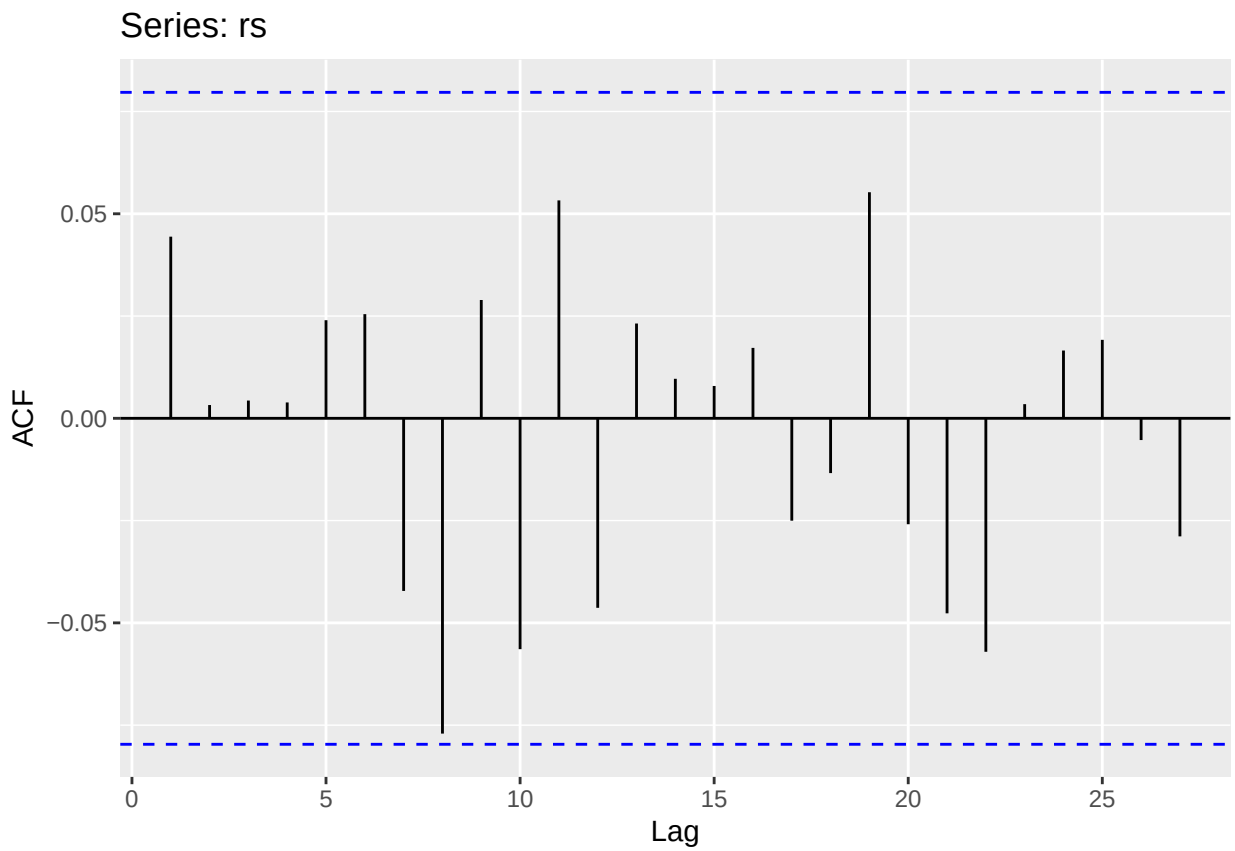
We will conduct model diagnostics.

Check residuals:



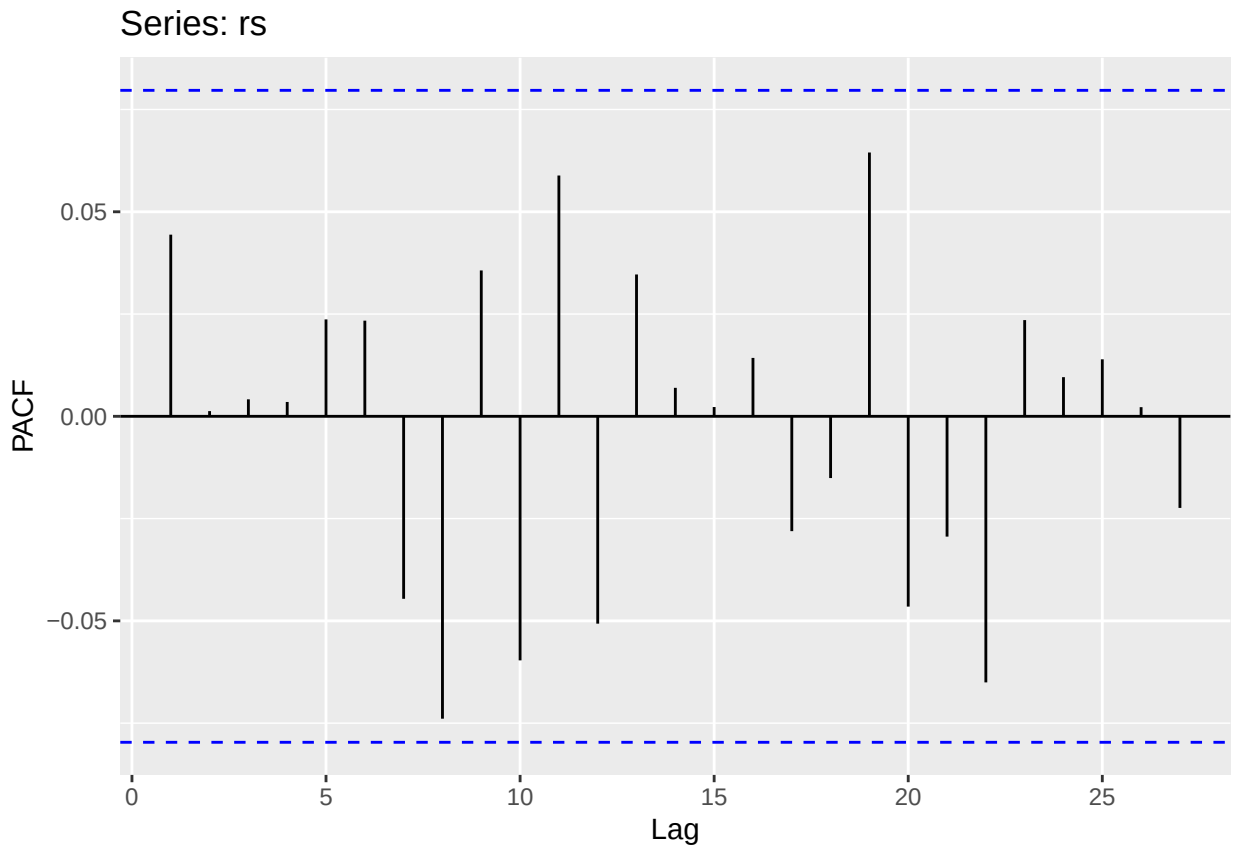
The plot most likely shows mean 0 with no linear trend and maybe eyeball non-constant variance (we will test with McLeod-Li below). The ACF shows the model lags are stationary. The distribution shows left skewness and tall Kutosis, showing non-normalcy in respect to a Gaussian PDF.

Stationarity: ACF Plot



The model looks stationary, will test with KPSS/ADF below.

Stationarity: PACF Plot



The model looks stationary, will test with KPSS/ADF below.

Mean Zero: T-Test

```
##
##  One Sample t-test
##
## data:  data
## t = -0.029943, df = 604, p-value = 0.9761
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.004631315  0.004492211
## sample estimates:
##    mean of x
## -6.95523e-05
##
## T-Test: mean is statistically zero, linear trend *REMOVED* ->
## *FAIL* to reject H0
```

```
## [1] TRUE
```

The 95% CI contains zero, therefore the mean is statistically 0 and the linear trend is removed.

Normalcy: Skewness

```
##       skew     lwr.ci     upr.ci
```

```
## -0.9209662 -0.9285384 -0.8987008
## Skew: has *LEFT* skewness,
## property does *NOT* conform to normality and Gaussian PDF
```

```
## [1] FALSE
```

The model has exhibits left skewness, showing non-normalcy in respect to a Gaussian PDF.

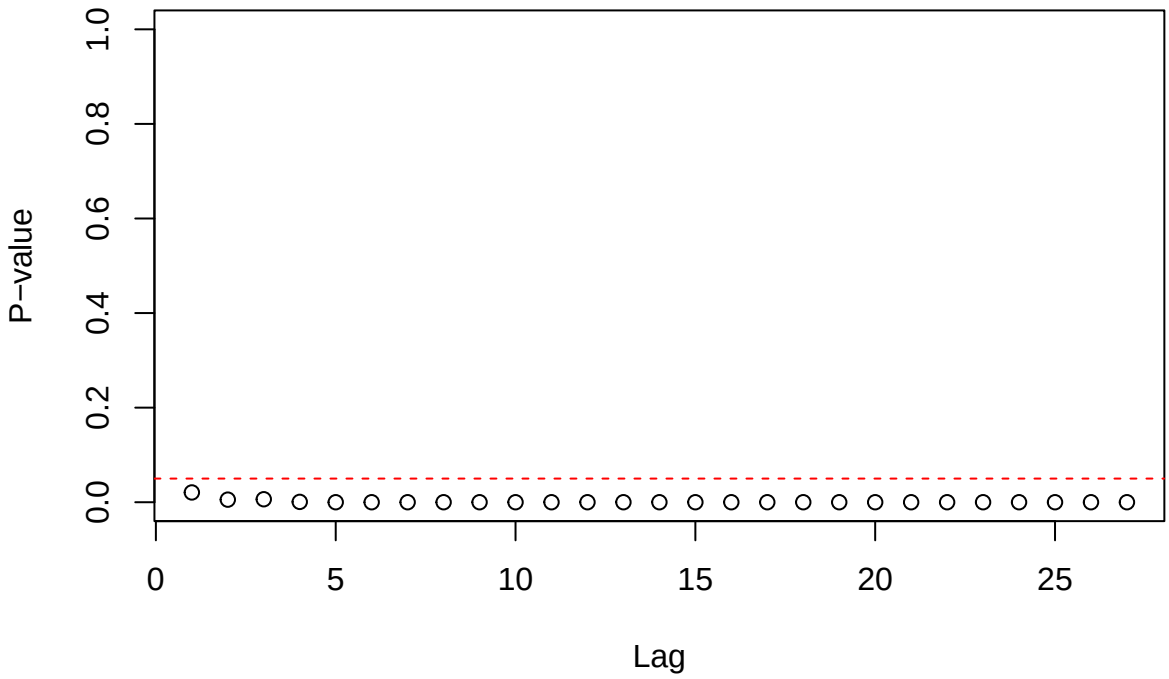
Normalcy: (excess) Kurtosis

```
##     kurt   lwr.ci   upr.ci
## 3.387039 3.419581 3.522317
## Kurt: has *TALL thick-tailed* (excess) kurtosis,
## property does *NOT* conform to normality and Gaussian PDF
```

```
## [1] FALSE
```

The model has exhibits tall (excess) Kurtosis, showing non-normalcy in respect to a Gaussian PDF.

Constant Variance: McLeod-Li Test




```
## McLeod-Li: *NON*-constant variance, heteroscedastic -> reject H0
## McLeod-Li: Lags < 0.05:
## 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27
```

```
## [1] FALSE
```

The model exhibits non-constant variance with many lags with a p-value of under 0.05.

Lag independence: Box-Ljung test

```
##
##  Box-Ljung test
##
## data:  data
## X-squared = 33.248, df = 30, p-value = 0.3119
##
## Box-Ljung: implies independence over 30 lags,
## *NO* autocorrelation -> *FAIL* to reject H0
```

```
## [1] TRUE
```

Box-Ljung test shows model has lag independence and thus no autocorrelation.

Stationarity: ADF test

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -3.9075
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:54:46 2023 by user: Reed
##
## ADF: contains *NO* unit roots over 30 lags, indicates *NO* mean drift,
## business cycles *NOT* present, series is stationary -> reject H0
```

```
##
## FALSE
```

The ADF test shows the model is stationary.

Stationarity: KPSS

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 6 lags.
##
## Value of test-statistic is: 0.0735
##
## Critical value for a significance level of:
```

```
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216
##
## KPSS: *NO* unit roots, *NO* linear trend, slope zero,
## series is trend stationary -> *FAIL* to reject H0

## [1] TRUE
```

The KPSS test shows the model is stationary.

Checking for reduced ARIMA(5,1,1) model with outlier business cycles:

```
## [1] 13.925  2.457  4.183
```

Similar to the full model, this model contains 3 business cycles, 14-month, 2-month, and 4-month cycles.

1.5. Compare your model from part 1.3. with your model from 1.4.. Which is preferred and why?

Full ARIMA(5,1,1) model with outlier summary:

```
## Series: psLevel
## Regression with ARIMA(5,1,1) errors
##
## Coefficients:
##          ar1     ar2     ar3     ar4     ar5      ma1     xreg
##       0.0558  0.1789  0.1222  0.0074  0.1169  -0.9852  -0.4274
## s.e.  0.0447  0.0442  0.0441  0.0438  0.0438   0.0190   0.0551
##
## sigma^2 = 0.003292:  log likelihood = 871.5
## AIC=-1727   AICc=-1726.75   BIC=-1691.77
##
## Training set error measures:
##                         ME       RMSE        MAE      MPE     MAPE      MASE
## Training set -0.0002075432 0.05699791 0.04070132 172.7945 422.2474 0.7115554
##                      ACF1
## Training set -0.004040577
```

Reduced ARIMA(5,1,1) model with outlier summary:

```
## Series: psLevel
## Regression with ARIMA(5,1,1) errors
##
## Coefficients:
##       ar1     ar2     ar3  ar4     ar5      ma1     xreg
##         0  0.1750  0.1247    0  0.1106  -0.9771  -0.4280
## s.e.    0  0.0426  0.0426    0  0.0421   0.0136   0.0551
##
## sigma^2 = 0.003291:  log likelihood = 870.66
## AIC=-1729.32   AICc=-1729.18   BIC=-1702.9
##
## Training set error measures:
##                        ME       RMSE        MAE      MPE     MAPE      MASE
## Training set -6.95523e-05 0.05708635 0.04077557 175.8037 429.6039 0.7128533
##                     ACF1
## Training set 0.04440602
```

The reduced model with the outlier has a lower AICc of -1729.18 compared to the full model with AICc -1726.75. While difference seems like a marginal gain, despite it parsimony is favored in this case. And since we are dealing with bond yields, the marginal difference can impact money markets (and especially peoples' money in general), and therefore would favor the model with its lower AICc score, and it happens to be the reduced model.

2. Box-Jenkins Methodology (20 points)

Consider the monthly Fama-Bliss bond yields with maturities of 1 and 3 years. The data are available from CRSP and are in the file **m-FamaBlissdbndyields.txt**. Denote the yields by y_{1t} and y_{3t}, respectively.

2.1. Perform EDA.

Validate data as a time series:

```
## [1] 636
```

```
## [1] 636
```

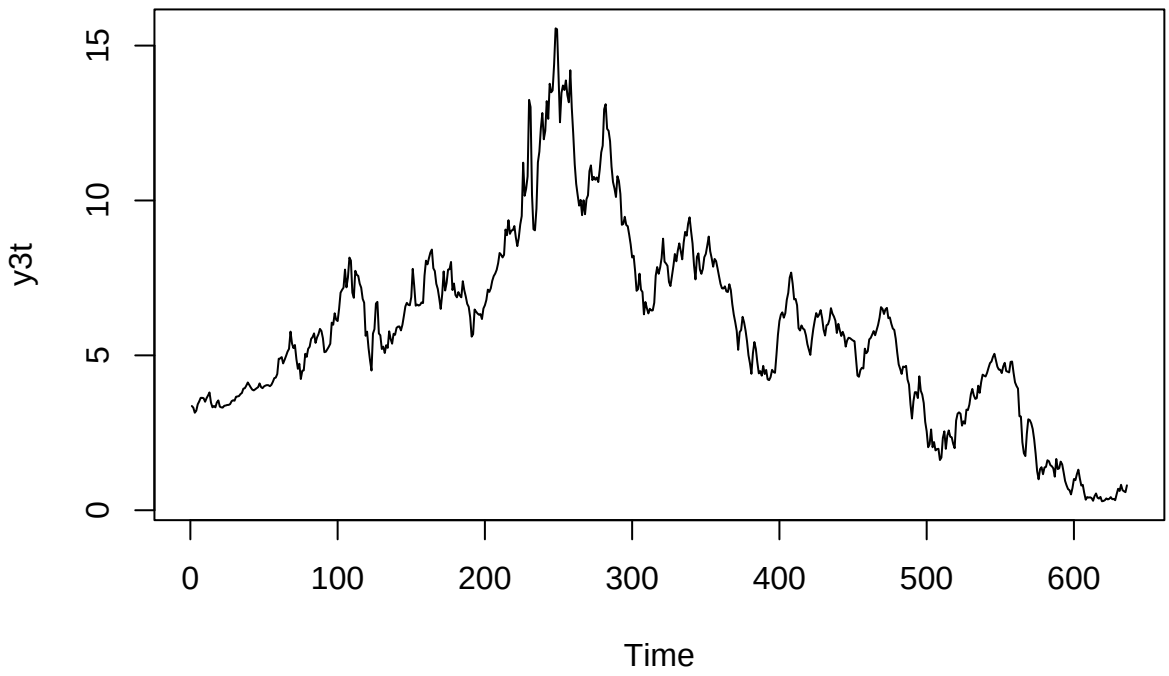
We have 636 unique years in 636 observations, which meets the $H_{10} : x_{it},\ i \in \{1, 2\},\ t \in \{1, 2, ..., n\}$ requirement for time series validation.

```
## [1] 636
```

```
## df
##   1
## 635
```

From the test above, we can verify that the constant time span between each date is only one month, denoted by the single value 1. This meets the $H_{20} : (t+1)-t = c, t \in \{1, 2..., n\}$ requirement for time series validation.

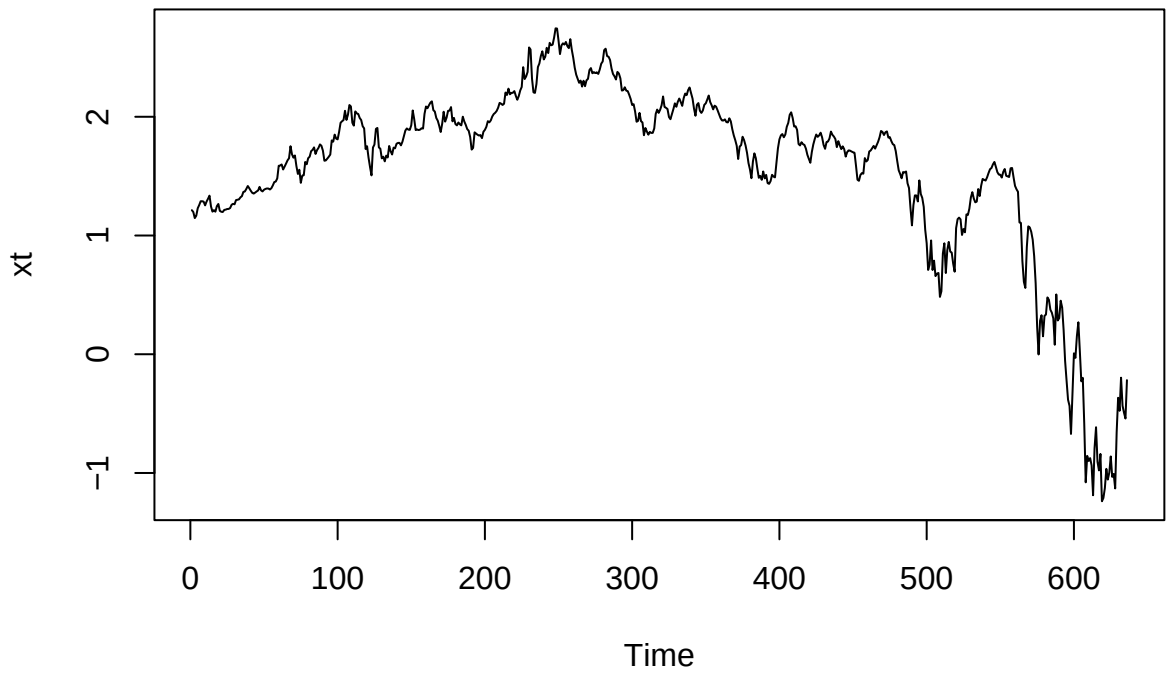
Let us create a time series plot of yield3:



Plotting the yield3 time series data visually shows the mean is not zero and not relatively flat with non-constant variance, with the data peaking at about 15 in 3-year bond yields at around month 200 to 300.

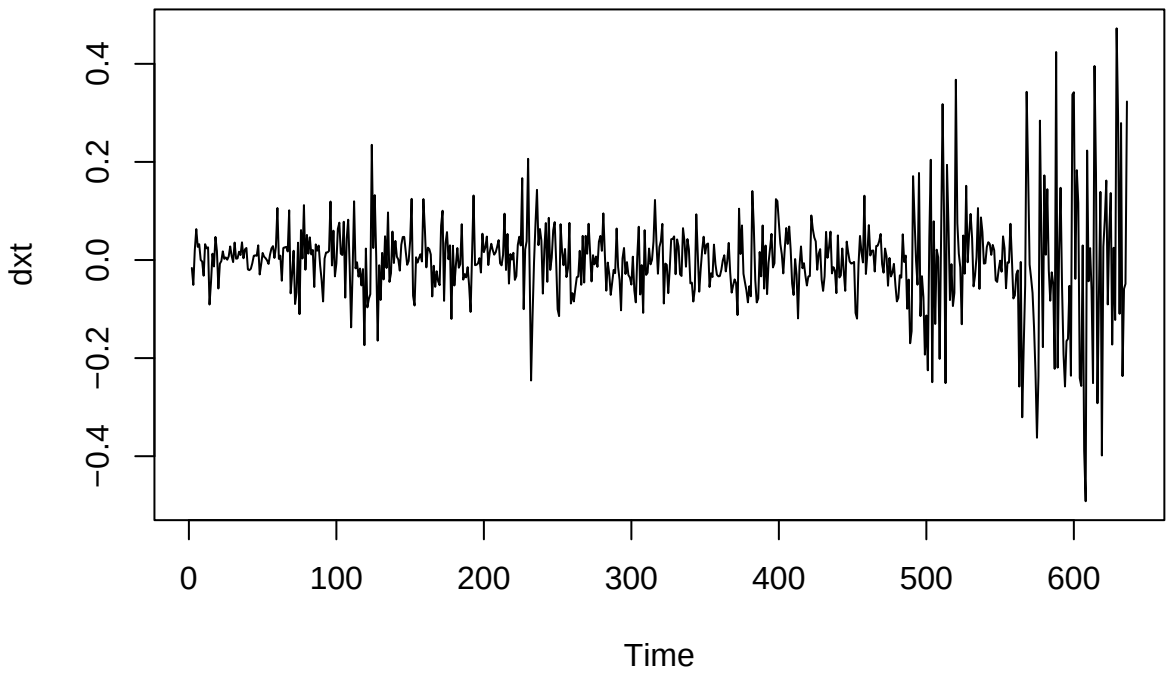
At this point instead of continuing with the EDA since we do not yet have mean zero we should transform the data, performing log(yield3).

Plot: log(yield3)



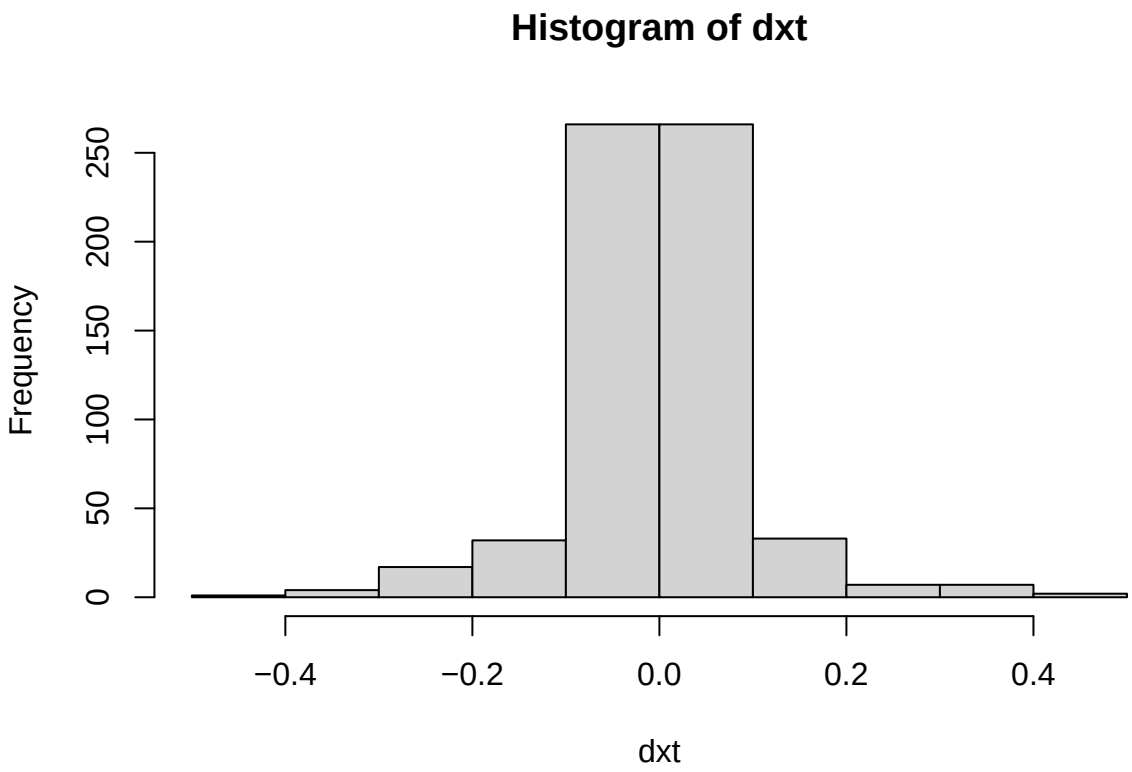
Transforming the data to log(yield3) still does not exhibit mean zero and constant variance. We will further transform the data to diff(log(yield3)).

Plot: diff(log(yield3))



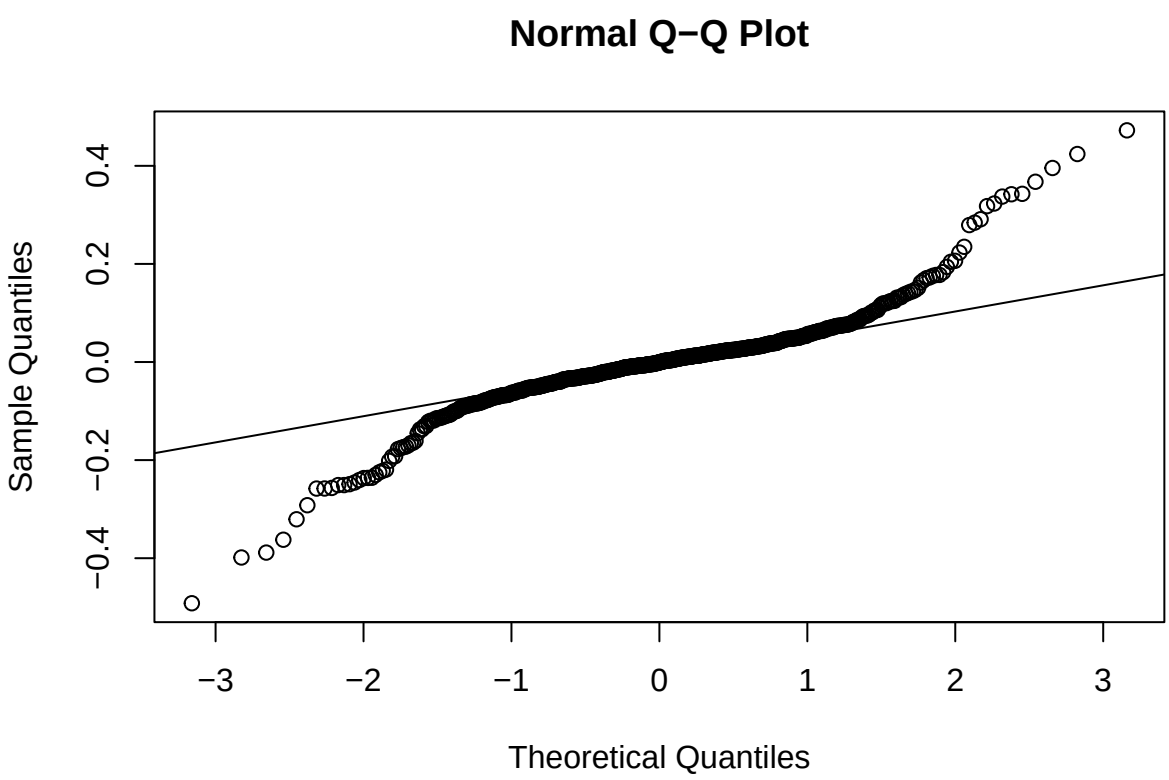
While transforming the data to diff(log(yield3)) still does not exhibit constant variance, we can eyeball mean 0 and will perform a full EDA moving forward.

Histogram: diff(log(yield3))



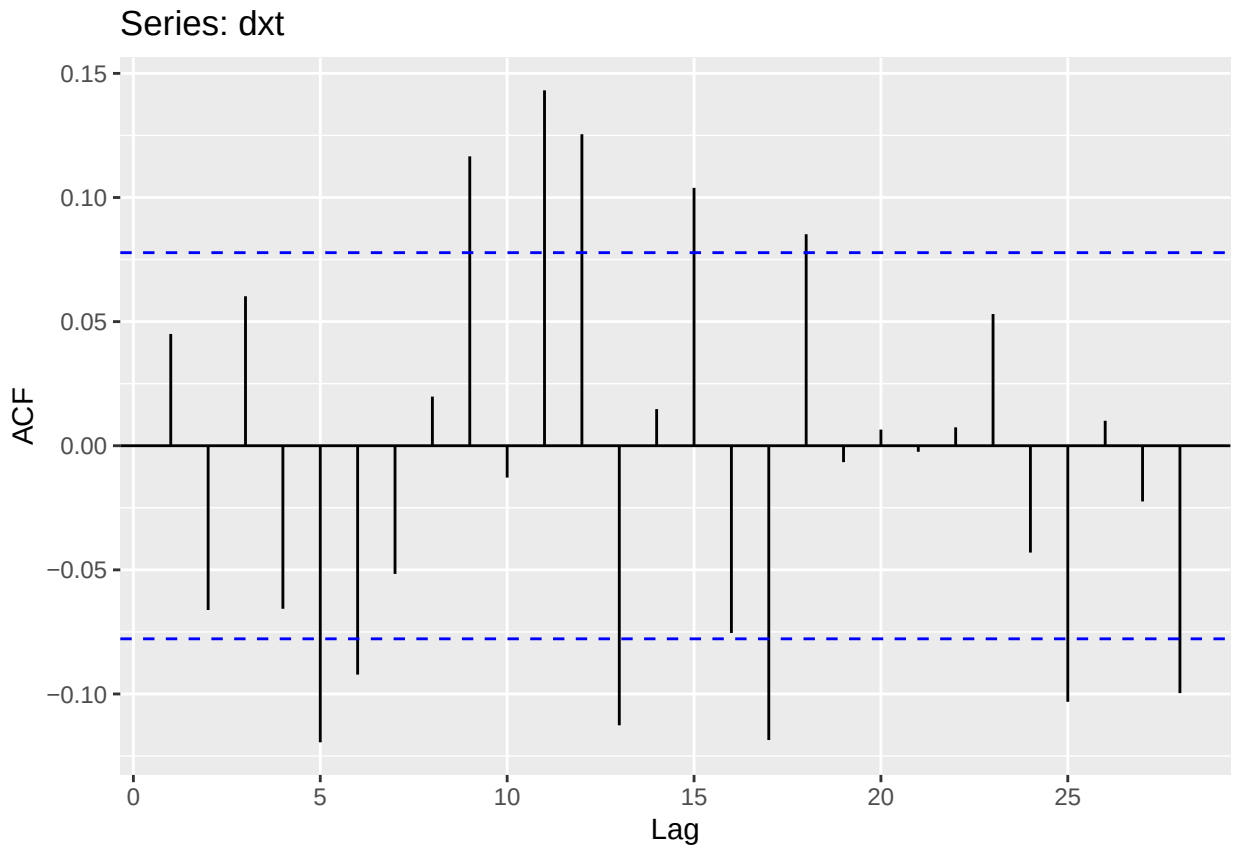
We can observe a fairly even-skewed but tall distribution, which might not fully qualify for normalcy in respect to a Gaussian PDF.

Q-Q Plot: diff(log(yield3))



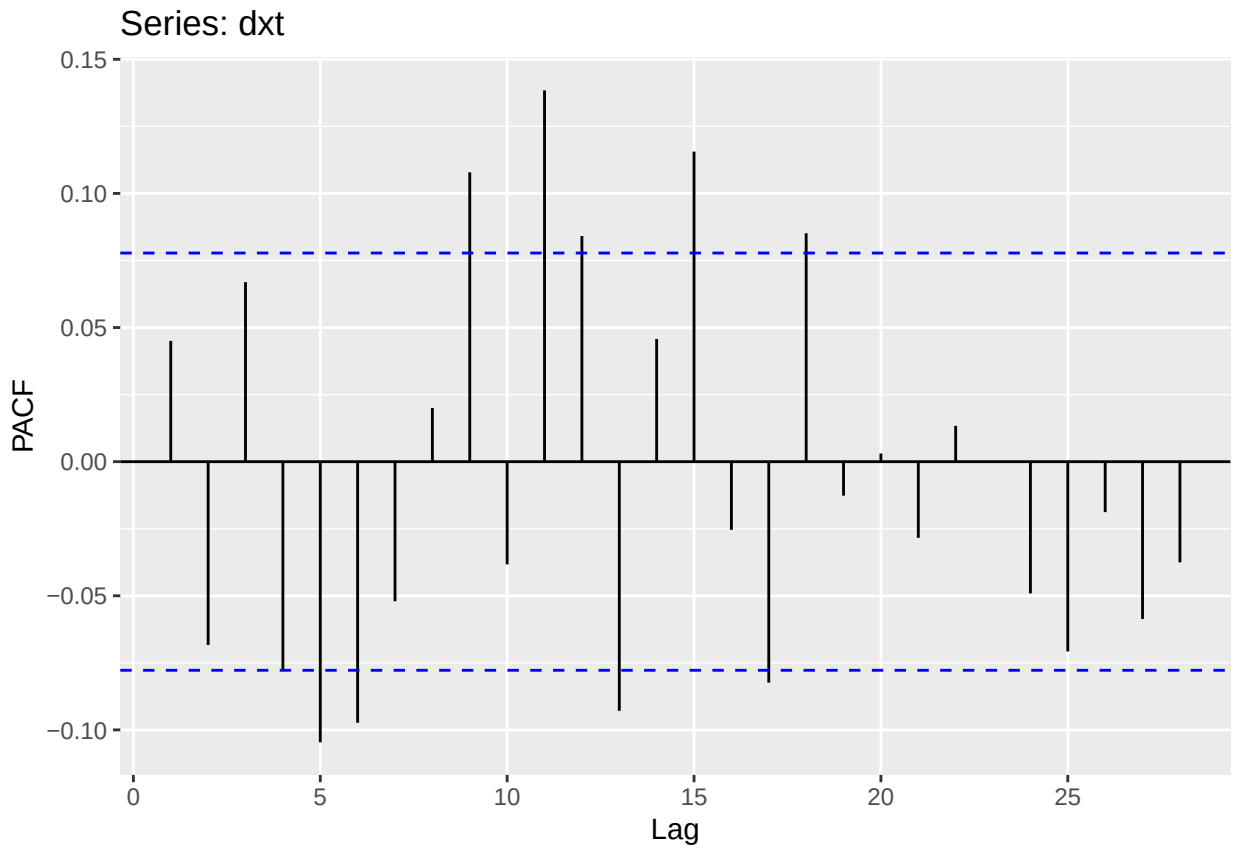
We can notice slight skewness and a good amount (excess) Kurtotis from the Q-Q plot as the ends of the plot veer off the idea normal line, thus showing non-normalcy in respect to a Gaussian PDF.

Stationarity: ACF diff(log(yield3))



Based on the ACF plot above, we'll use MA(6) for our ARIMA model.

Stationarity: PACF diff(PS_LEVEL)



Based on the PACF plot above, we'll use AR(6) for our ARIMA model.

Mean 0: T-Test diff(log(yield3))

```
##
##  One Sample t-test
##
## data:  data
## t = -0.59962, df = 634, p-value = 0.549
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.009637654  0.005128728
## sample estimates:
##    mean of x
## -0.002254463
##
## T-Test: mean is statistically zero, linear trend *REMOVED* ->
## *FAIL* to reject H0

## [1] TRUE
```

The 95% Confidence Interval (CI) of diff(log(yield3)) does contain 0, thus the mean is statistically zero, showing normalcy in respect to a Gaussian PDF, as well as the linear trend removed.

Normalcy: Skewness diff(log(yield3))

```
##      skew    lwr.ci    upr.ci
```

```
## 0.1161441 0.1132442 0.1615245
## Skew: has *RIGHT* skewness,
## property does *NOT* conform to normality and Gaussian PDF
```

```
## [1] FALSE
```

The PS_LEVEL data has a distribution with slight right skewness, thus showing non-normalcy in respect to a Gaussian PDF.

Normalcy: (excess) Kurtosis diff(log(yield3))

```
##     kurt   lwr.ci   upr.ci
## 5.668950 5.745001 5.850422
## Kurt: has *TALL thick-tailed* (excess) kurtosis,
## property does *NOT* conform to normality and Gaussian PDF
```

```
## [1] FALSE
```

The PS_LEVEL data has a distribution with tall (excess) Kurtosis, thus showing non-normalcy in respect to a Gaussian PDF.

Constant Variance: Breush-Pagan Test diff(log(yield3))

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(data ~ seq(1, length(data)))
## BP = 77.17, df = 1, p-value < 2.2e-16
##
## Breusch-Pagan: *NON*-constant variance, possible clustering,
## heteroscedastic -> reject H0
```

```
##    BP
## FALSE
```

While we can see signs of non-constant variance in the time series plot, the Breusch-Pagen test confirms it.

Lag independence: Box-Ljung test diff(log(yield3))

```
##
##  Box-Ljung test
##
## data:  data
## X-squared = 113.28, df = 30, p-value = 1.334e-11
##
## Box-Ljung: implies dependency present over 30 lags,
## autocorrelation present -> reject H0
```

```
## [1] FALSE
```

The Box-Ljung test shows the diff(log(yield3)) transformation contains lag dependency and thus autocorrelation.

2.2. Build a time series model using the Box-Jenkins method for the log of the year three (y_{3t}) data: $x_t = log(y_{3t})$. For simplicity, you may ignore possible outliers, but describe how you would treat outliers if they were not to be ignored.

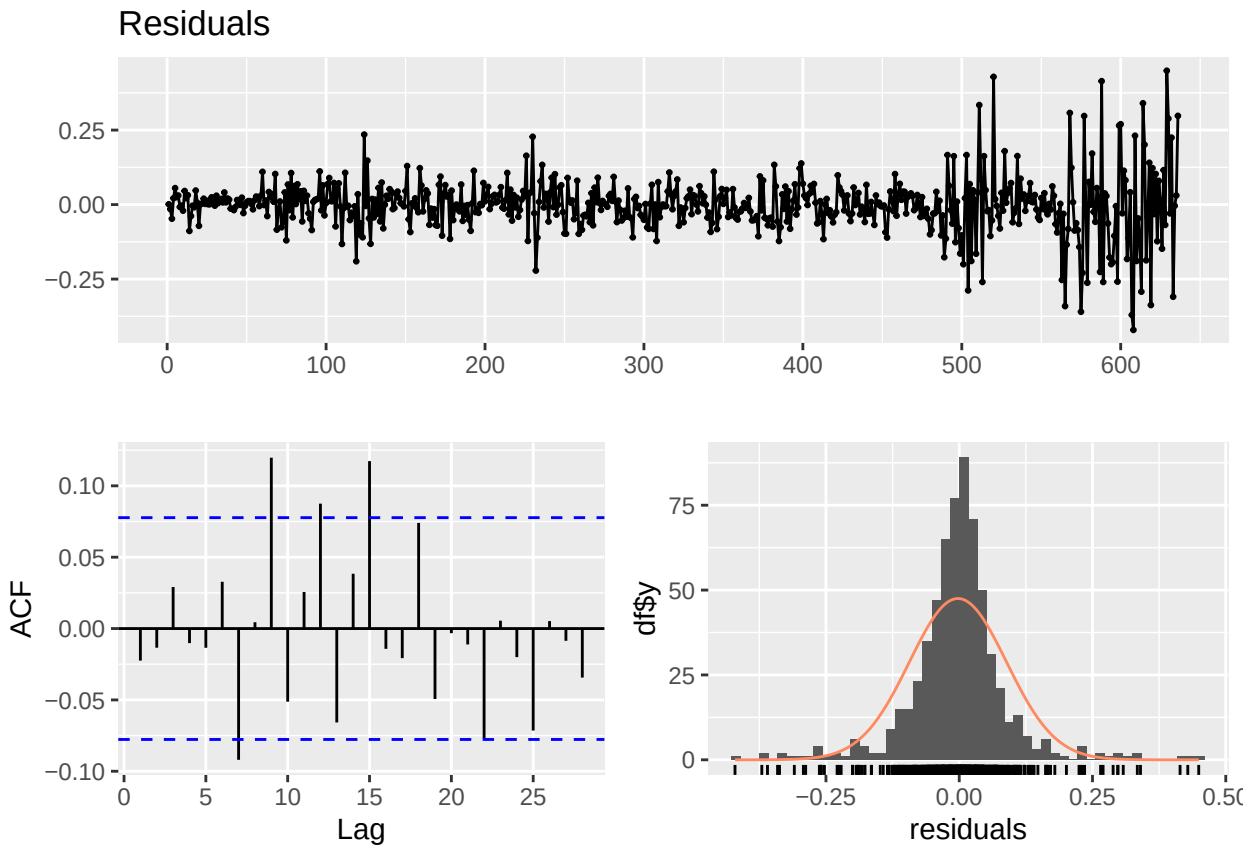
Based on our EDA from section 2.1 above, we will set our ARIMA model with AR(6) and MA(6) components, and since we performed the EDA on differenced log(yield3) data, we will also set d = 1 to denote the differencing.

```
# set d=1 for diff(log(yield3))
p<-6;d<-1;q<-6 # based on EDA in 2.1
m <- Arima(xt,order=c(p,d,q),method="ML")
```

```
## Series: xt
## ARIMA(6,1,6)
##
## Coefficients:
##           ar1     ar2      ar3     ar4      ar5      ar6     ma1      ma2
##       -0.3804  0.1168  -0.3046  0.2864  -0.4008  -0.7909  0.4417  -0.1459
## s.e.      NaN  0.0682   0.0519     NaN      NaN      NaN  0.0395   0.0714
##          ma3      ma4     ma5     ma6
##       0.3164  -0.3244  0.2698  0.7361
## s.e.  0.0704   0.0371     NaN  0.0687
##
## sigma^2 = 0.008465:  log likelihood = 619.27
## AIC=-1212.54   AICc=-1211.95   BIC=-1154.64
##
## Training set error measures:
##                        ME       RMSE       MAE      MPE     MAPE      MASE
## Training set -0.002413176 0.09105988 0.0588405 21.85956 31.77556 0.9736578
##                     ACF1
## Training set -0.02245569
```

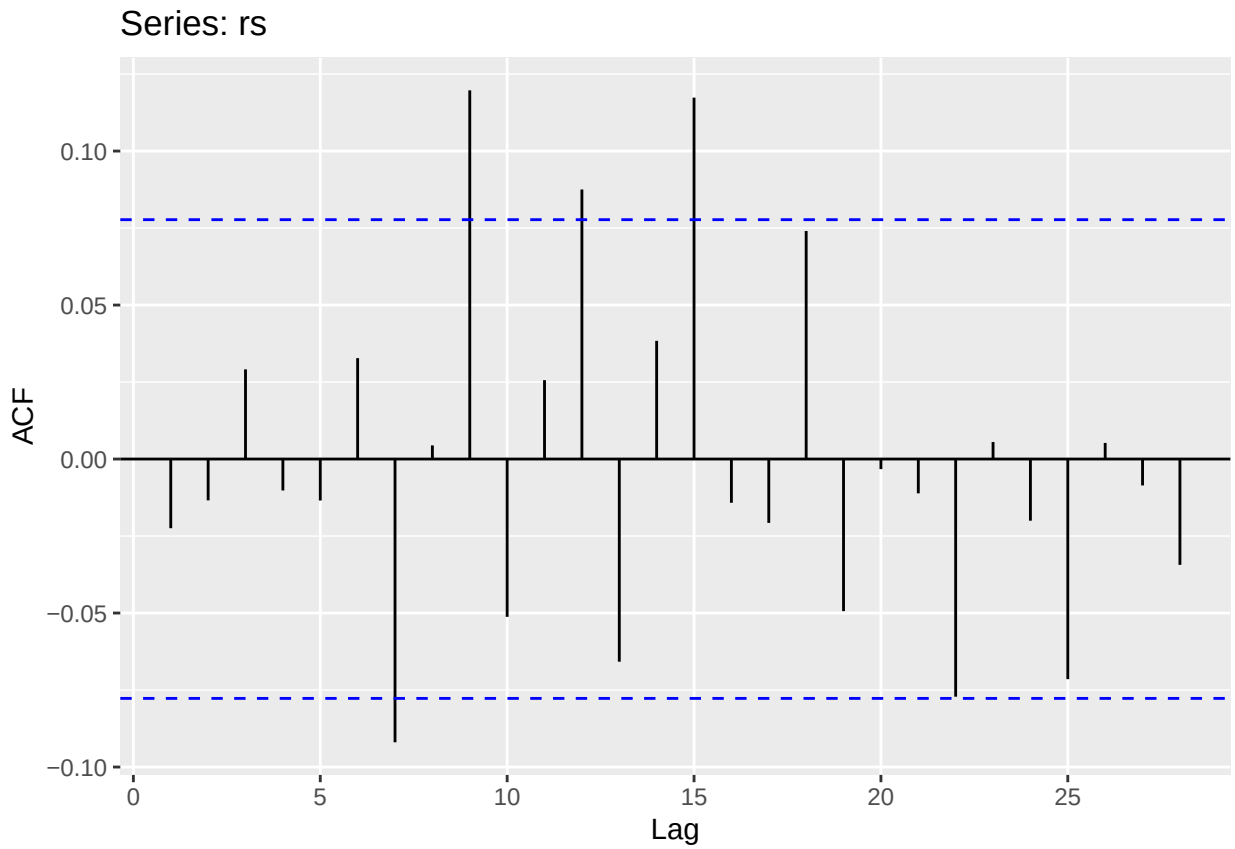
We will conduct model diagnostics for ARIMA(6,1,6).

Check residuals:



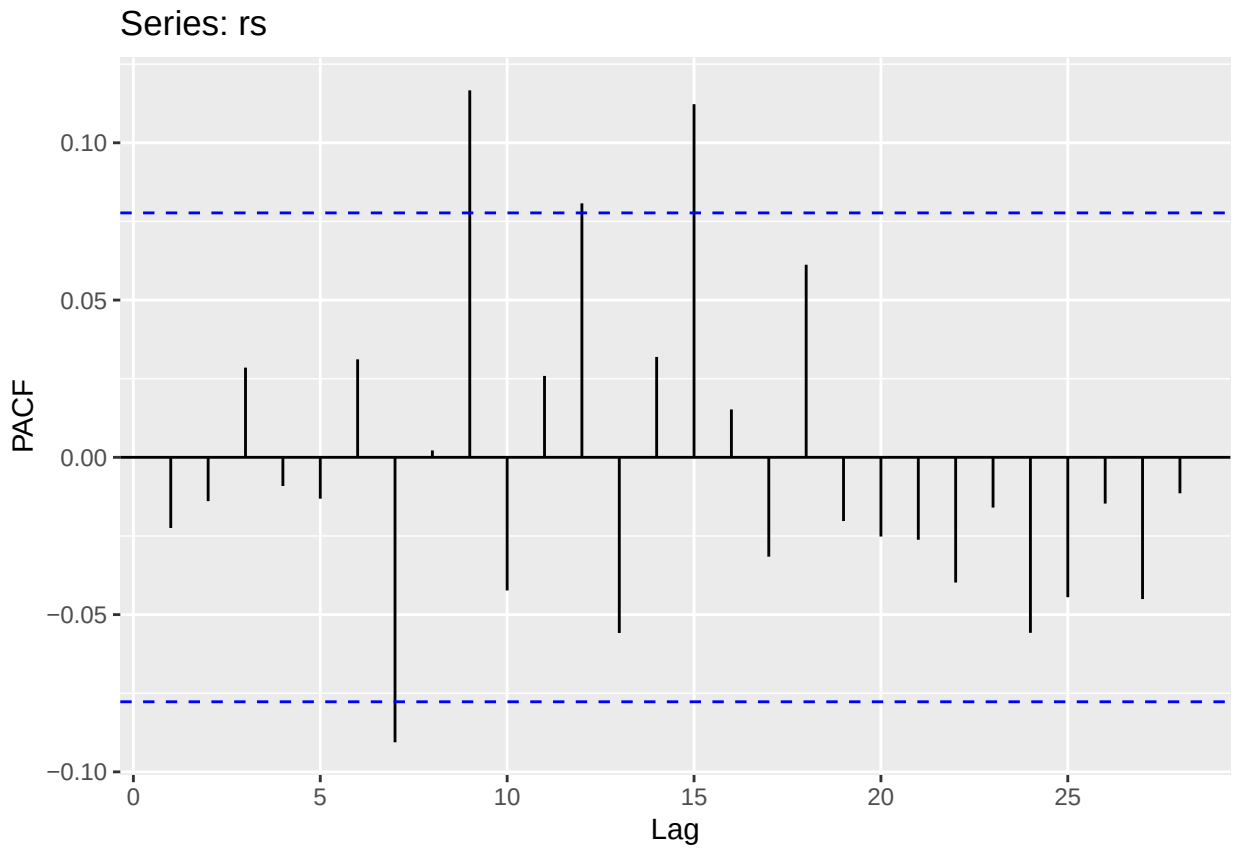
From the residuals plot we can eyeball a mean zero and non-constant variance. The ACF plot might look to be stationary (will test with KPSS/ADF), while the distribution looks tall but with about even skewness at mean 0.

Stationarity: ACF Plot



The model might look to be stationary, will test with KPSS/ADF below.

Stationarity: PACF Plot



The model might look to be stationary, will test with KPSS/ADF below.

Mean Zero: T-Test

```
##
##  One Sample t-test
##
## data:  data
## t = -0.66804, df = 635, p-value = 0.5044
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.009506737  0.004680385
## sample estimates:
##    mean of x
## -0.002413176
##
## T-Test: mean is statistically zero, linear trend *REMOVED* ->
## *FAIL* to reject H0

## [1] TRUE
```

The 95% Confidence Interval (CI) contains zero, therefore the mean is statistically 0 and the linear trend is removed.

Normalcy: Skewness

```
##      skew    lwr.ci    upr.ci
```

```
## 0.1212137 0.0969263 0.1405698
## Skew: has *RIGHT* skewness,
## property does *NOT* conform to normality and Gaussian PDF
```

```
## [1] FALSE
```

The model has exhibits slight right skewness, showing non-normalcy in respect to a Gaussian PDF.

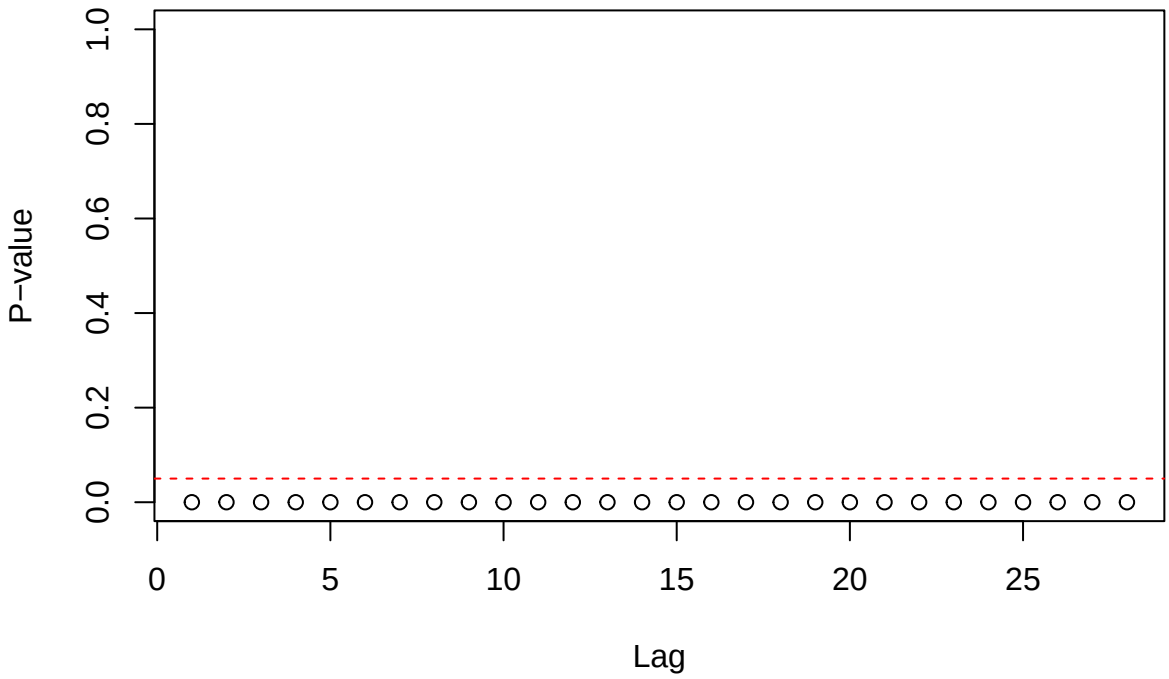
Normalcy: (excess) Kurtosis

```
##     kurt   lwr.ci   upr.ci
## 5.379094 5.457812 5.554380
## Kurt: has *TALL thick-tailed* (excess) kurtosis,
## property does *NOT* conform to normality and Gaussian PDF
```

```
## [1] FALSE
```

The model has exhibits tall (excess) Kurtosis, showing non-normalcy in respect to a Gaussian PDF.

Constant Variance: McLeod-Li Test




```
## McLeod-Li: *NON*-constant variance, heteroscedastic -> reject H0
## McLeod-Li: Lags < 0.05:
## 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28
```

```
## [1] FALSE
```

The model exhibits non-constant variance.

Lag independence:

```
##
##  Box-Ljung test
##
## data:  data
## X-squared = 52.798, df = 30, p-value = 0.00624
##
## Box-Ljung: implies dependency present over 30 lags,
## autocorrelation present -> reject H0
```

```
## [1] FALSE
```

Box-Ljung test shows the model displays lag dependence and thus has autocorrelation.

Stationarity: ADF test

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -4.5795
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:54:48 2023 by user: Reed
##
## ADF: contains *NO* unit roots over 30 lags, indicates *NO* mean drift,
## business cycles *NOT* present, series is stationary -> reject H0
```

```
##
## FALSE
```

The ADF test shows the model is stationary.

Stationarity: KPSS

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 6 lags.
##
## Value of test-statistic is: 0.0246
##
## Critical value for a significance level of:
```

```
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216
##
## KPSS: *NO* unit roots, *NO* linear trend, slope zero,
## series is trend stationary -> *FAIL* to reject H0
```

```
## [1] TRUE
```

The KPSS test shows the model is stationary.

Checking for business cycles:

```
## [1] 10.943  2.217  3.822  3.007  6.127
```

The model contains 5 business cycles, 11-month, 2-month, 4-month, 3-month, and 6-month cycles.

2.3. Fit the following model to the log earnings series: $m < -arima(xt, order = c(0,1,1), seasonal = list(order = c(0,0,1), period = 4))$ **where** xt **denotes the log of the earnings. Write the equation of the fitted model. Compare this model with the model in part 2.2. Which model is preferred? Why?**

```
m <- Arima(xt, order=c(0,1,1), seasonal=list(order=c(0,0,1), period=4))
```

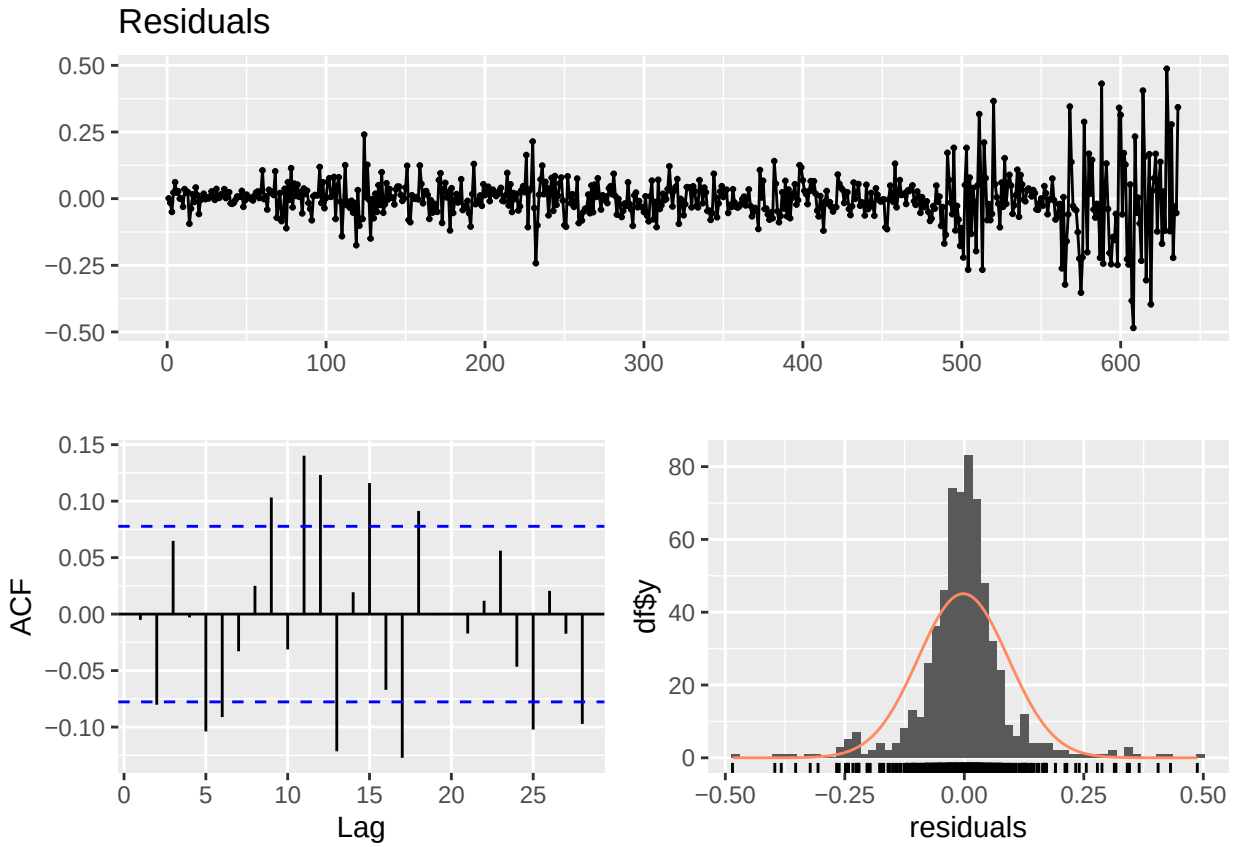
```
## Series: xt
## ARIMA(0,1,1)(0,0,1)[4]
##
## Coefficients:
##          ma1     sma1
##       0.0533  -0.0617
## s.e.  0.0441   0.0385
##
## sigma^2 = 0.008938:  log likelihood = 597.76
## AIC=-1189.52   AICc=-1189.48   BIC=-1176.16
##
## Training set error measures:
##                       ME       RMSE        MAE      MPE    MAPE      MASE
## Training set -0.002251931 0.09431684 0.06016146 22.23155 32.8514 0.9955162
##                     ACF1
## Training set -0.00506217
```

Based on the summary above, we have the following equation:

$$x_t = \theta_1 z_{t-1} + \Theta_1 z_{t-1}$$

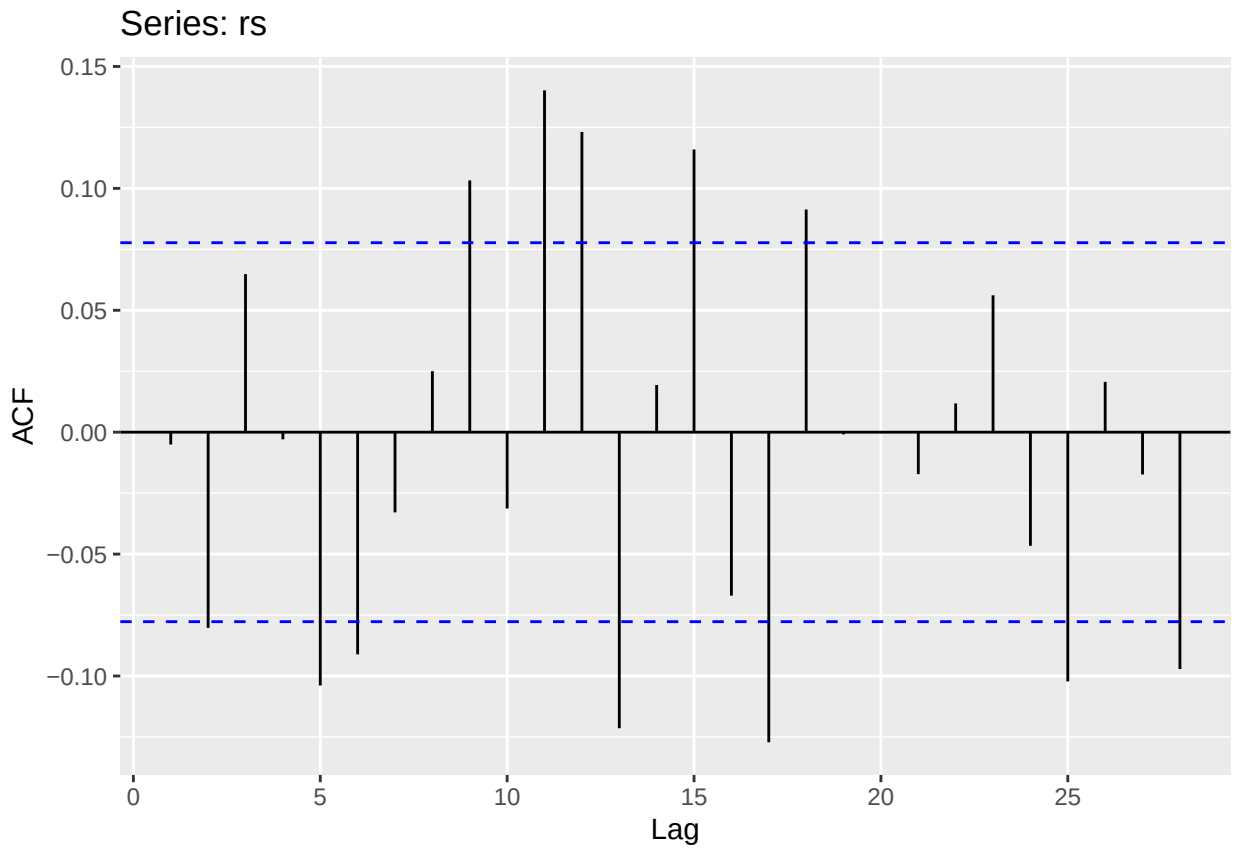
We will conduct model diagnostics for ARIMA(0,1,1) x SARIMA(0,0,1).

Check residuals:



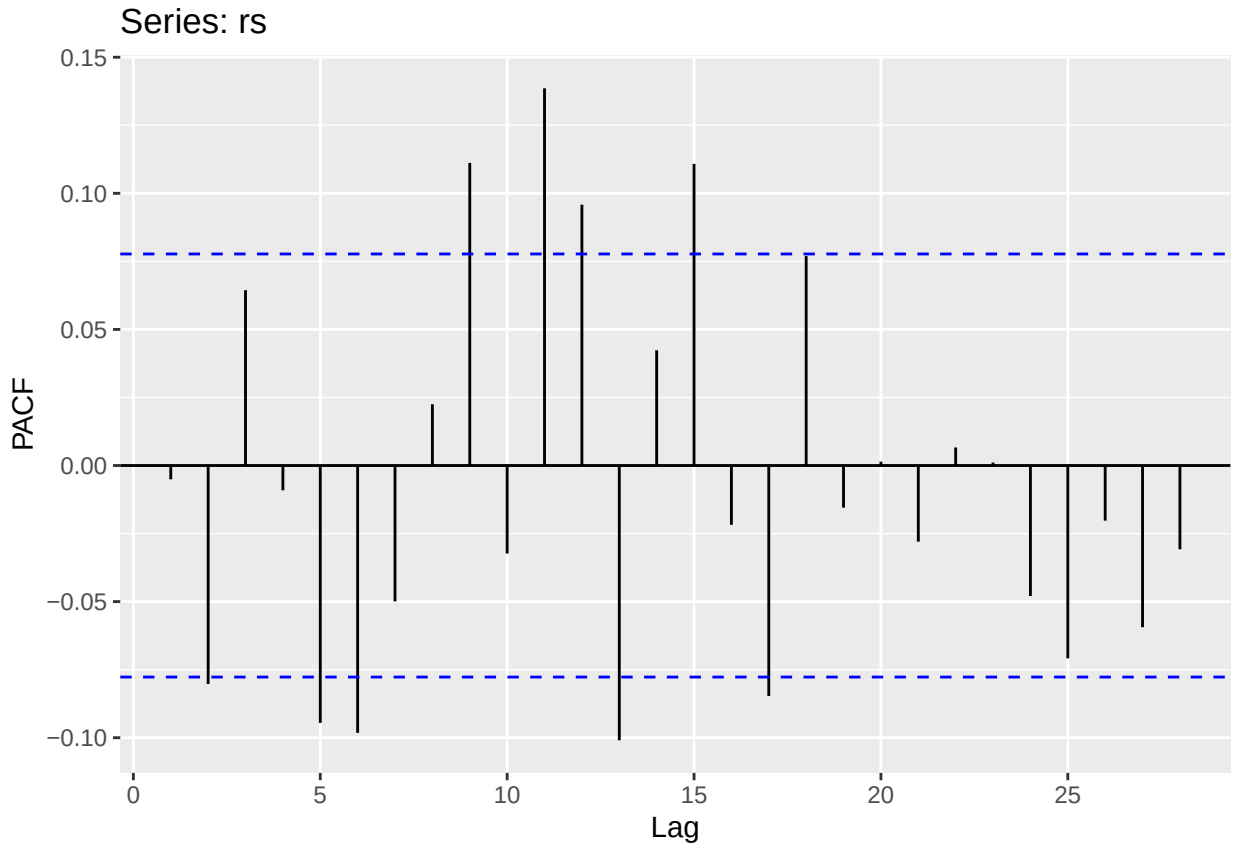
From the residuals plot we can eyeball a mean zero and non-constant variance. The ACF plot might not be stationary (will test with KPSS/ADF), while the distribution looks tall but with about even skewness at mean 0.

Stationarity: ACF Plot



The model doesn't look very stationary, will test with KPSS/ADF below.

Stationarity: PACF Plot



The model doesn't look very stationary, will test with KPSS/ADF below.

Mean Zero: T-Test

```
##
##  One Sample t-test
##
## data:  data
## t = -0.60183, df = 635, p-value = 0.5475
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.009599695  0.005095833
## sample estimates:
##    mean of x
## -0.002251931
##
## T-Test: mean is statistically zero, linear trend *REMOVED* ->
## *FAIL* to reject H0
```

```
## [1] TRUE
```

The 95% Confidence Interval (CI) contains zero, therefore the mean is statistically 0 and the linear trend is removed.

Normalcy: Skewness

```
##      skew    lwr.ci    upr.ci
```

```
## 0.1632234 0.1587314 0.2057411
## Skew: has *RIGHT* skewness,
## property does *NOT* conform to normality and Gaussian PDF
```

```
## [1] FALSE
```

The model has exhibits slight right skewness, showing non-normalcy in respect to a Gaussian PDF.

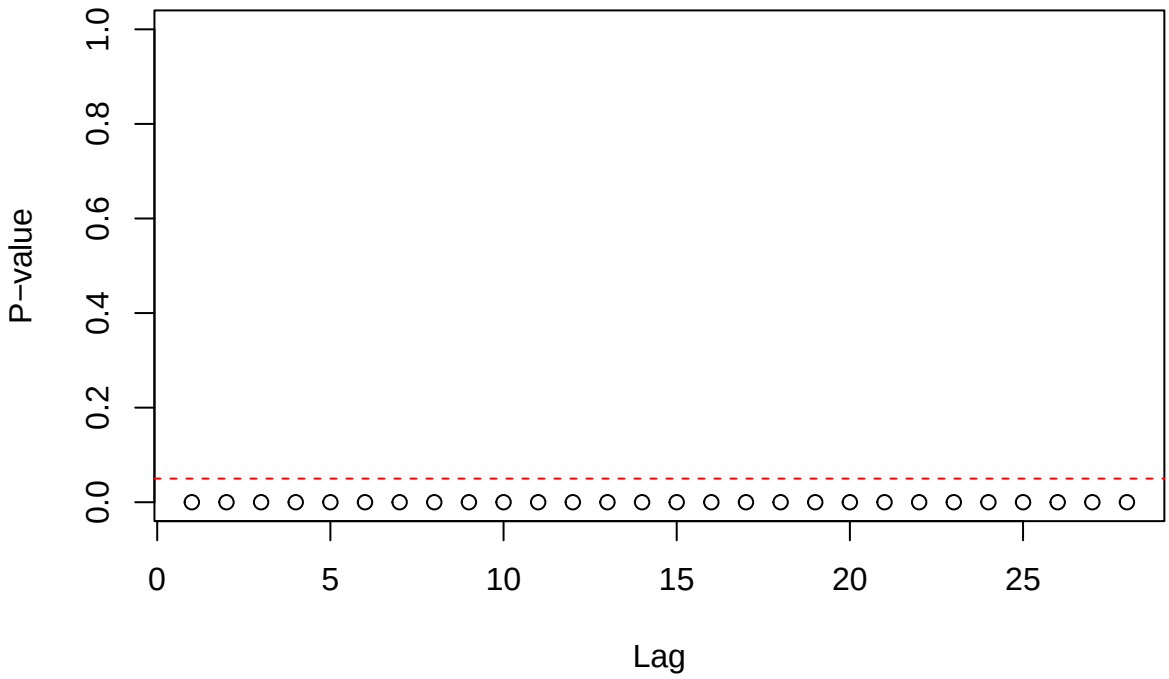
Normalcy: (excess) Kurtosis

```
##     kurt   lwr.ci   upr.ci
## 5.845872 5.903651 6.013982
## Kurt: has *TALL thick-tailed* (excess) kurtosis,
## property does *NOT* conform to normality and Gaussian PDF
```

```
## [1] FALSE
```

The model has exhibits tall (excess) Kurtosis, showing non-normalcy in respect to a Gaussian PDF.

Constant Variance: McLeod-Li Test




```
## McLeod-Li: *NON*-constant variance, heteroscedastic -> reject H0
## McLeod-Li: Lags < 0.05:
## 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28
```

```
## [1] FALSE
```

The model exhibits non-constant variance as all of the model lags are under the p-value of 0.05.

Lag independence:

```
##
##  Box-Ljung test
##
## data:  data
## X-squared = 110.37, df = 30, p-value = 4.002e-11
##
## Box-Ljung: implies dependency present over 30 lags,
## autocorrelation present -> reject H0
```

```
## [1] FALSE
```

Box-Ljung test shows the model displays lag dependence and thus has autocorrelation.

Stationarity: ADF test

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -4.6645
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:54:49 2023 by user: Reed
##
## ADF: contains *NO* unit roots over 30 lags, indicates *NO* mean drift,
## business cycles *NOT* present, series is stationary -> reject H0
```

```
##
## FALSE
```

The ADF test shows the model is stationary.

Stationarity: KPSS

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 6 lags.
##
## Value of test-statistic is: 0.023
##
## Critical value for a significance level of:
```

```
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216
##
## KPSS: *NO* unit roots, *NO* linear trend, slope zero,
## series is trend stationary -> *FAIL* to reject H0
```

```
## [1] TRUE
```

The KPSS test shows the model is stationary.

Checking for business cycles:

```
## [1] 4.294
```

The model shows only a 4-month business cycle.

Comparing the ARIMA(6,1,6) vs. the ARIMA(0,1,1)xSARIMA(0,0,1) model:

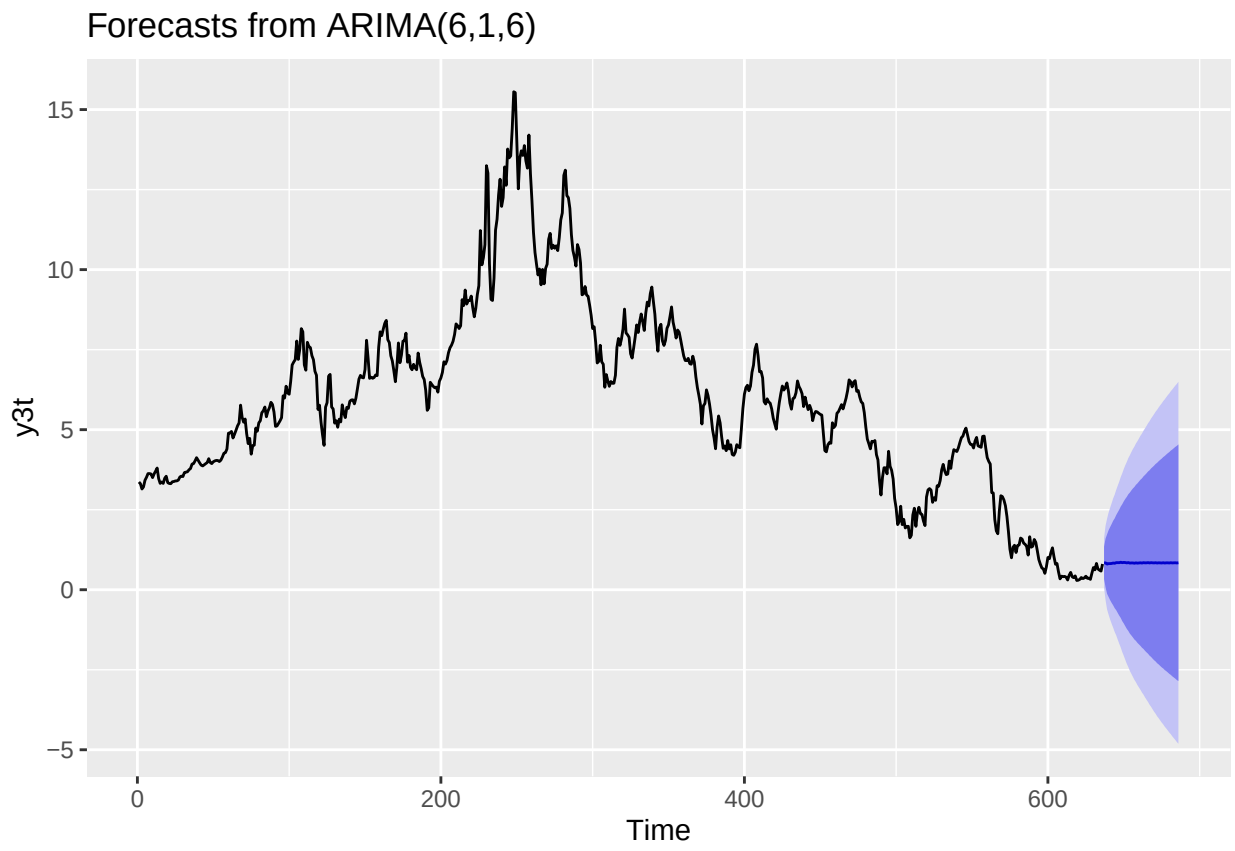
```
## Series: xt
## ARIMA(6,1,6)
##
## Coefficients:
##           ar1     ar2      ar3     ar4      ar5      ar6     ma1      ma2
##       -0.3804  0.1168  -0.3046  0.2864  -0.4008  -0.7909  0.4417  -0.1459
## s.e.      NaN  0.0682   0.0519     NaN      NaN      NaN  0.0395   0.0714
##          ma3      ma4     ma5     ma6
##       0.3164  -0.3244  0.2698  0.7361
## s.e.  0.0704   0.0371     NaN  0.0687
##
## sigma^2 = 0.008465:  log likelihood = 619.27
## AIC=-1212.54   AICc=-1211.95   BIC=-1154.64
##
## Training set error measures:
##                         ME       RMSE       MAE      MPE     MAPE      MASE
## Training set -0.002413176 0.09105988 0.0588405 21.85956 31.77556 0.9736578
##                      ACF1
## Training set -0.02245569
```

```
## Series: xt
## ARIMA(0,1,1)(0,0,1)[4]
##
## Coefficients:
##          ma1     sma1
##       0.0533  -0.0617
## s.e.  0.0441   0.0385
##
## sigma^2 = 0.008938:  log likelihood = 597.76
## AIC=-1189.52   AICc=-1189.48   BIC=-1176.16
##
## Training set error measures:
##                         ME       RMSE        MAE      MPE    MAPE      MASE
## Training set -0.002251931 0.09431684 0.06016146 22.23155 32.8514 0.9955162
##                      ACF1
## Training set -0.00506217
```

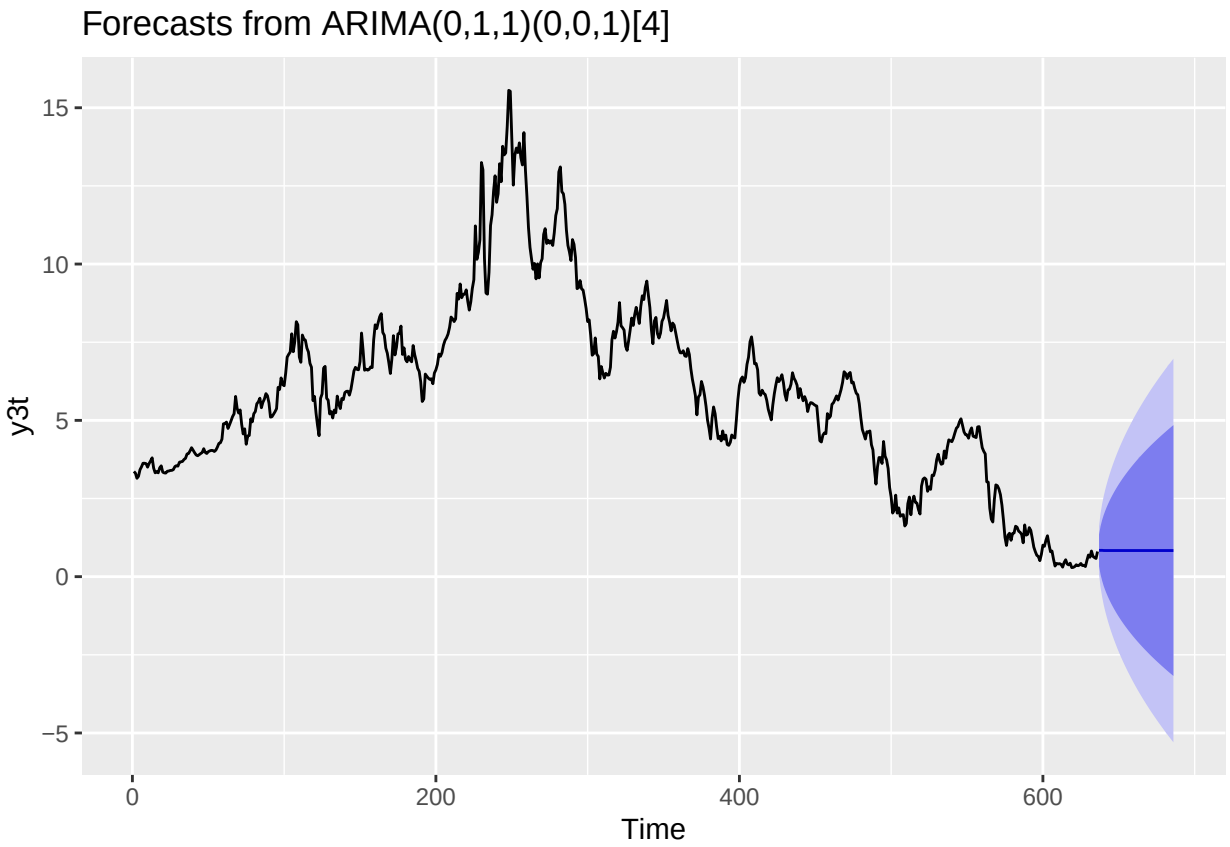
From the summary statistics above, the ARIMA(6,1,6) model has lower AICc (-1211.95), RMSE(0.09106), and MAE (0.05884) compared to the ARIMA(0,1,1) x SARIMA(0,0,1) model's AICc (-1189.48), RMSE(0.09432), and MAE (0.06016). Because of the lower statistics, we would choose the ARIMA(6,1,6) model over the ARIMA(0,1,1) x SARIMA(0,0,1) model.

2.4. You may use $t = 600$ as the starting forecast origin. Which model is preferred? Why?

Forecast: ARIMA(6,1,6)



Forecast: ARIMA(0,1,1) x SARIMA(0,0,1)



Based on the sample forecasts above, while both models are admittedly not great, the ARIMA(6,1,6) performs better with a smaller 80% and 95% CI range compared to the ARIMA(0,1,1) x SARIMA(0,0,1) model.

3. ARIMA and Regression Errors (20 points)

Consider the monthly Fama-Bliss bond yields with maturities of 1 and 3 years. The data are available from CRSP and are in the file **m-FamaBlissdbndyields.txt**. Denote the yields by y1t and y3t, respectively. The goal is to explore the dependence of the 3-year yield on the 1 year yield.

3.1. Fit the linear regression model $y3t = \beta_0 + \beta_1 y_{1t} + e_t$ using the model-building process. Write the equation of the model to be fitted.

We create the following linear regression model:

```
m <- lm(y3t~y1t)
```

```
##
## Call:
## lm(formula = y3t ~ y1t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.69599 -0.42038 -0.03045  0.37993  1.41445
##
```

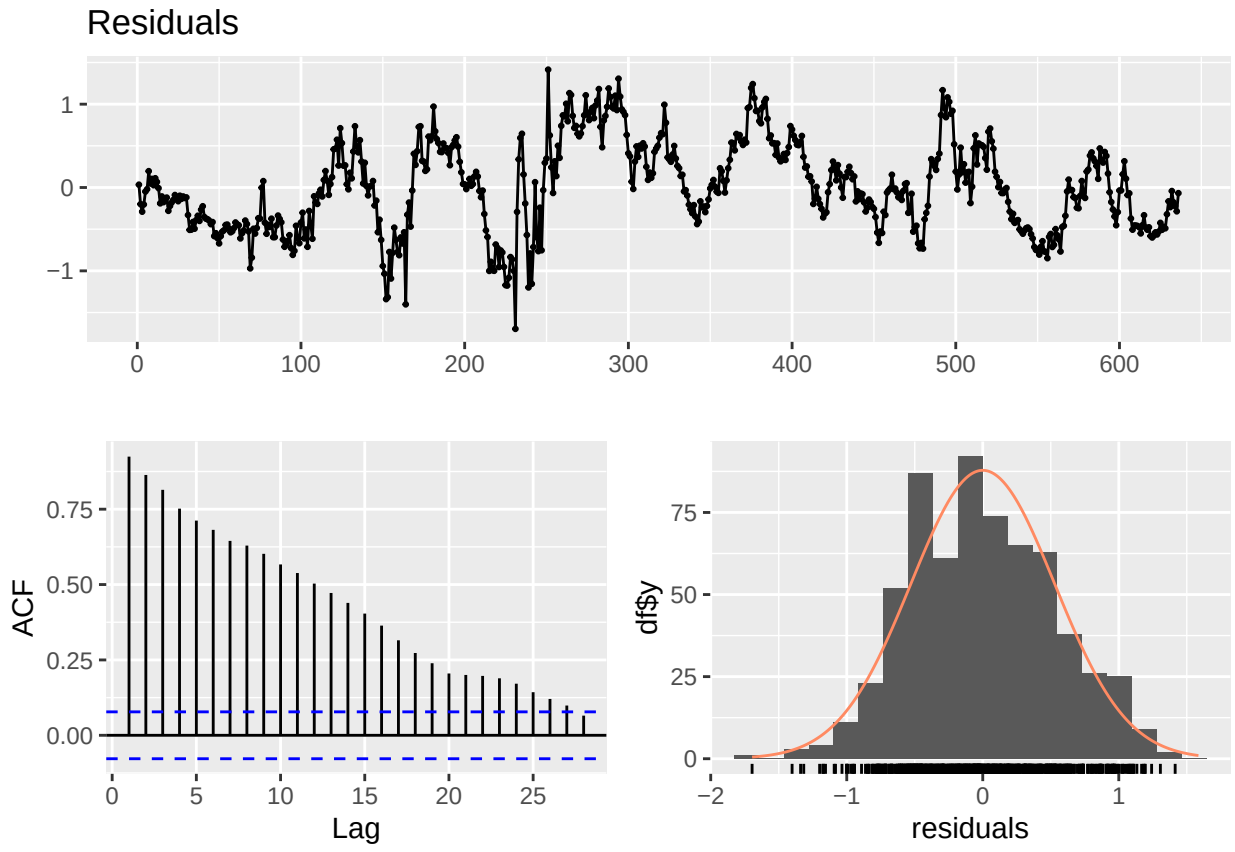
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.712645   0.041909    17.0   <2e-16 ***
## y1t         0.940816   0.006676   140.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.529 on 634 degrees of freedom
## Multiple R-squared:  0.9691, Adjusted R-squared:  0.969
## F-statistic: 1.986e+04 on 1 and 634 DF,  p-value: < 2.2e-16
```

From the summary above, we have the following equation:

$$y_{3t} = \beta_0 + \beta_1 y_{1t} + \eta_t$$

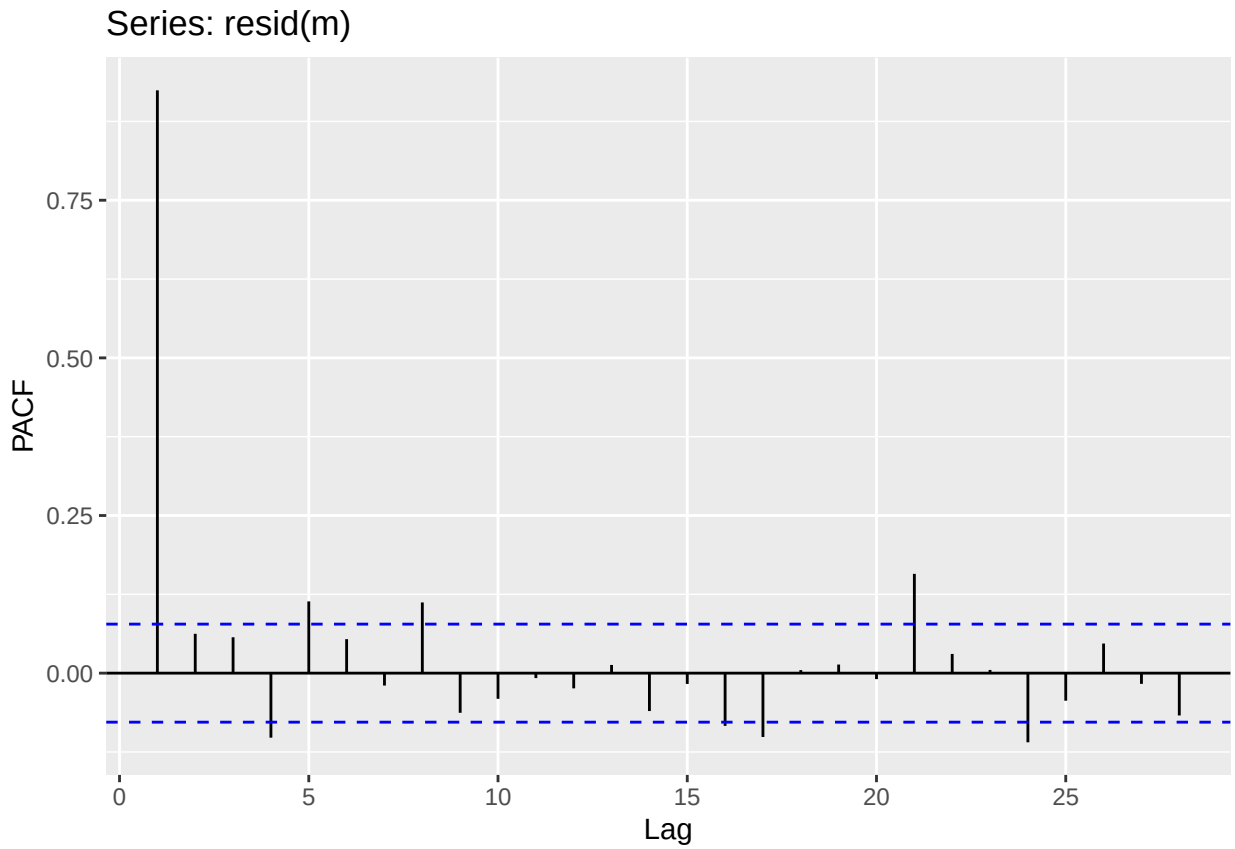
We will perform model diagnostics on the $y_{3t} = \beta_0 + \beta_1 y_{1t} + \eta_t$ model.

Check residuals plot:



The residuals plot might eyeball mean 0 but we'll have to run a t-test for mean 0 to find out. The ACF plot shows the lags of the model residuals is not stationary. The distribution looks tall and might be slightly skewed to the left, showing non-normalcy in respect to a Gaussian PDF.

PACF plot:



We see autocorrelation in the 1st lag, suggesting AR(1) for the regression model.

T-Test for Mean 0:

```
##
##  One Sample t-test
##
## data:  data
## t = 1.2327e-15, df = 635, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.04115617  0.04115617
## sample estimates:
##    mean of x
## 2.583521e-17
##
## T-Test: mean is statistically zero, linear trend *REMOVED* ->
## *FAIL* to reject H0

## [1] TRUE
```

The 95% CI of the linear regression model residuals contain 0 there the mean of the residuals is statistically 0, and that the linear trend is removed.

ADF Test:

```
##
```

```
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -4.0569
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:54:50 2023 by user: Reed
##
## ADF: contains *NO* unit roots over 30 lags, indicates *NO* mean drift,
## business cycles *NOT* present, series is stationary -> reject H0
```

```
##
## FALSE
```

The ADF test shows that there are no unit roots in regards to drifting.

KPSS Test:

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 6 lags.
##
## Value of test-statistic is: 0.6068
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216
##
## KPSS: contains unit roots, linear trend present, slope *NOT* zero,
## series *NOT* trend stationary -> reject H0
```

```
## [1] FALSE
```

The KPSS test shows that there are unit roots in regards to random walk and trending.

3.2. Fit a linear regression model letting $d_{1t} = \triangle y_{1t}$ and $d_{2t} = \triangle y_{3t}$, where $\triangle$ is the differencing operator. Here $d_{it}, i = 1, 2, 3$ denotes the change in monthly bond yields. Consider the linear regression $d_{3t} = \beta d_{1t} + e_t$. Write the equation of the model to be fitted. Is the model an adequate model? Why?

We create the following linear regression model:

```r
m <- lm(d3t ~ -1 + d1t)
```

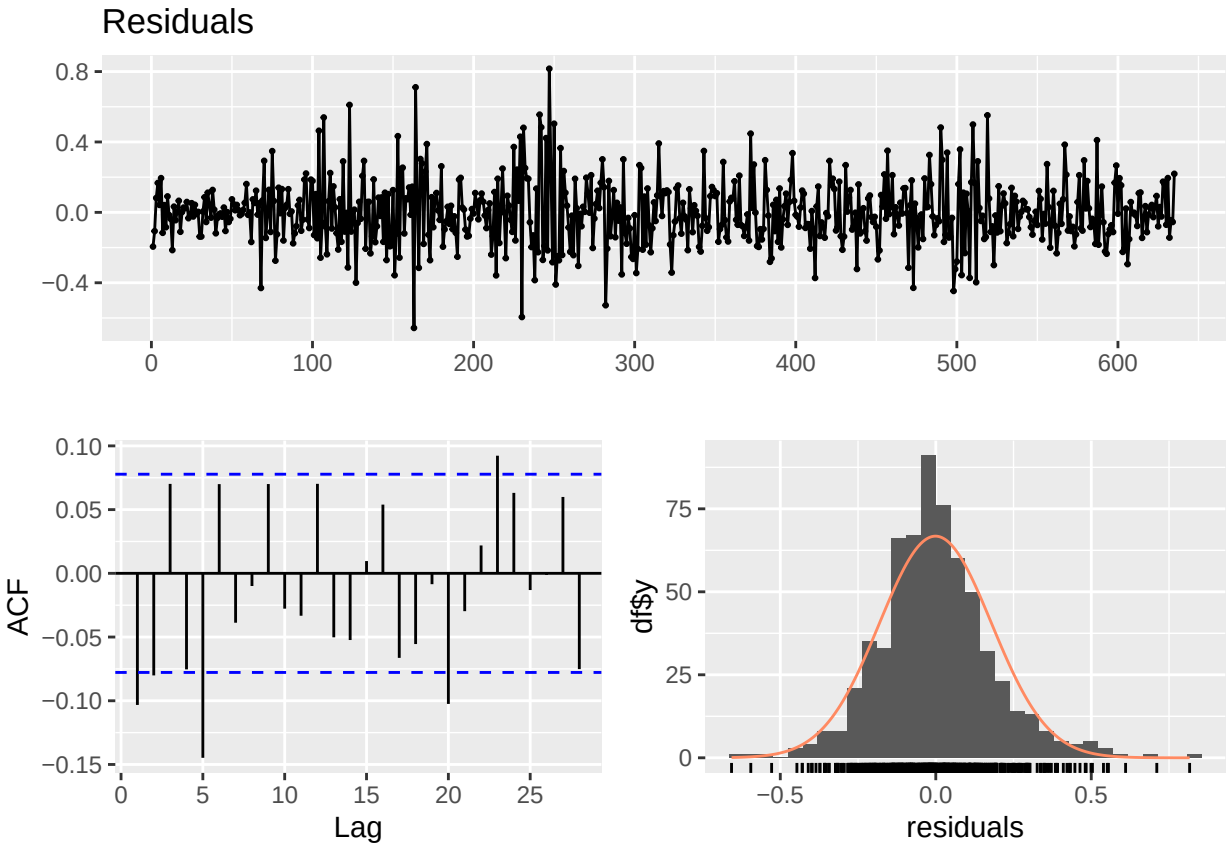
```
##
## Call:
## lm(formula = d3t ~ -1 + d1t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65669 -0.11132 -0.01054  0.09621  0.81624
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## d1t  0.73598    0.01478   49.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1803 on 634 degrees of freedom
## Multiple R-squared:  0.7963, Adjusted R-squared:  0.796
## F-statistic:  2478 on 1 and 634 DF,  p-value: < 2.2e-16
```

From the summary above, we have the following equation:

$$d_{3t} = \beta_0 + \beta_1 d_{1t} + \eta_t$$

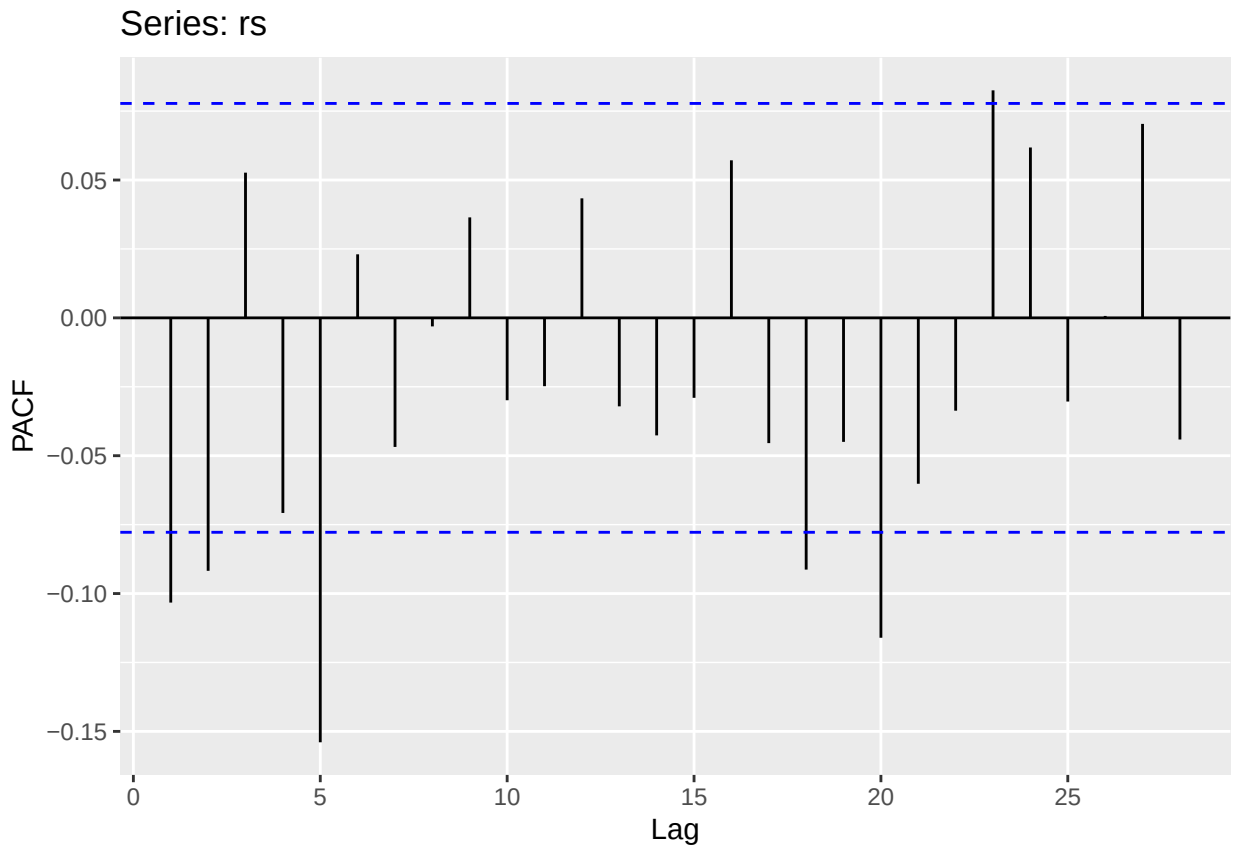
We will perform model diagnostics on the $d_{3t} = \beta_0 + \beta_1 d_{1t} + \eta_t$ model.

Check residuals plot:



The residuals plot might eyeball mean 0 but we'll have to run a t-test for mean 0 to find out. The ACF plot shows the lags of the model residuals is not stationary. The distirbution looks tall and might be slighly skewed to the left, showing non-normalcy in respect to a Gaussian PDF.

PACF Plot:



The PACF plot suggests an AR(5) model for the linear regression model.

T-Test for mean 0:

```
##
##  One Sample t-test
##
## data:  data
## t = -0.25445, df = 634, p-value = 0.7992
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.03515814  0.02709200
## sample estimates:
##    mean of x
## -0.004033071
##
## T-Test: mean is statistically zero, linear trend *REMOVED* ->
## *FAIL* to reject H0

## [1] TRUE
```

The 95% CI of the linear regression model residuals contain 0 there the mean of the residuals is statistically 0, and that the linear trend is removed.

ADF Test:

```
##
```

```
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -4.9154
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:54:51 2023 by user: Reed
##
## ADF: contains *NO* unit roots over 30 lags, indicates *NO* mean drift,
## business cycles *NOT* present, series is stationary -> reject H0

##
## FALSE
```

The ADF test shows no unit roots and that it is stationary.

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 6 lags.
##
## Value of test-statistic is: 0.0219
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216
##
## KPSS: *NO* unit roots, *NO* linear trend, slope zero,
## series is trend stationary -> *FAIL* to reject H0

## [1] TRUE
```

The KPSS test also shows no unit roots and that it is stationary.

3.3. Based on the model refinements, describe and compare the linear dependence between the bond yields of the two linear regression models.

```
##
## Call:
## lm(formula = y3t ~ y1t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.69599 -0.42038 -0.03045  0.37993  1.41445
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.712645   0.041909    17.0   <2e-16 ***
## y1t         0.940816   0.006676   140.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.529 on 634 degrees of freedom
## Multiple R-squared:  0.9691, Adjusted R-squared:  0.969
## F-statistic: 1.986e+04 on 1 and 634 DF,  p-value: < 2.2e-16
```

```
## Series: y3t
## Regression with ARIMA(0,0,0) errors
##
## Coefficients:
##       intercept    xreg
##          0.7126  0.9408
## s.e.     0.0418  0.0067
##
## sigma^2 = 0.2798:  log likelihood = -496.42
## AIC=998.84   AICc=998.88   BIC=1012.21
##
## Training set error measures:
##                        ME      RMSE       MAE       MPE     MAPE     MASE
## Training set 4.194351e-14 0.5281353 0.4314126 -5.011104 12.94632 1.582768
##                    ACF1
## Training set 0.9243193
```

```
##
## Call:
## lm(formula = d3t ~ -1 + d1t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65669 -0.11132 -0.01054  0.09621  0.81624
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## d1t  0.73598    0.01478   49.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1803 on 634 degrees of freedom
## Multiple R-squared:  0.7963, Adjusted R-squared:  0.796
## F-statistic:  2478 on 1 and 634 DF,  p-value: < 2.2e-16
```

```
## Series: d3t
## Regression with ARIMA(0,0,0) errors
##
## Coefficients:
##         xreg
##       0.7360
```

```
## s.e.  0.0148
##
## sigma^2 = 0.0325:  log likelihood = 187.38
## AIC=-370.76   AICc=-370.74   BIC=-361.85
##
## Training set error measures:
##                        ME      RMSE       MAE  MPE MAPE      MASE      ACF1
## Training set -0.001003403 0.1801395 0.1346783 -Inf  Inf 0.3719585 -0.103278
```

Based on the model summaries above, the $y_{3t} = \beta_0 + \beta_1 y_{1t} + \eta_t$ regression model has a higher R-squared score of 0.9691 over the $d_{3t} = \beta_0 + \beta_1 d_{1t} + \eta_t$ regression model R-squared score of 0.7963. That being said, as these regression models are integrated into an ARIMA model, the $d_{3t} = \beta_0 + \beta_1 d_{1t} + \eta_t$ model has a much lower AICc of -370.74 compared to the $y_{3t} = \beta_0 + \beta_1 y_{1t} + \eta_t$ model's AICc of 998.88.

Because of the much lower AICc score, the $d_{3t} = \beta_0 + \beta_1 d_{1t} + \eta_t$ regression model would be more preferred to integrate into an ARIMA time series model.

3.4. Fit an AR($m33\$order$) model to d_{3t} using d_{1t} as an explanatory variable using the model-building process. Write the equation of the model to be fitted.

The order of $m33\$order$ to d_{3t} using d_{1t} is:

```
m33$order
```

```
## [1] 5
```

We create the following model using $m33\$order$ 5:

```
m <- Arima(ts(d3t), order=c(m33$order,0,0), xreg=d1t, include.mean=F)
```

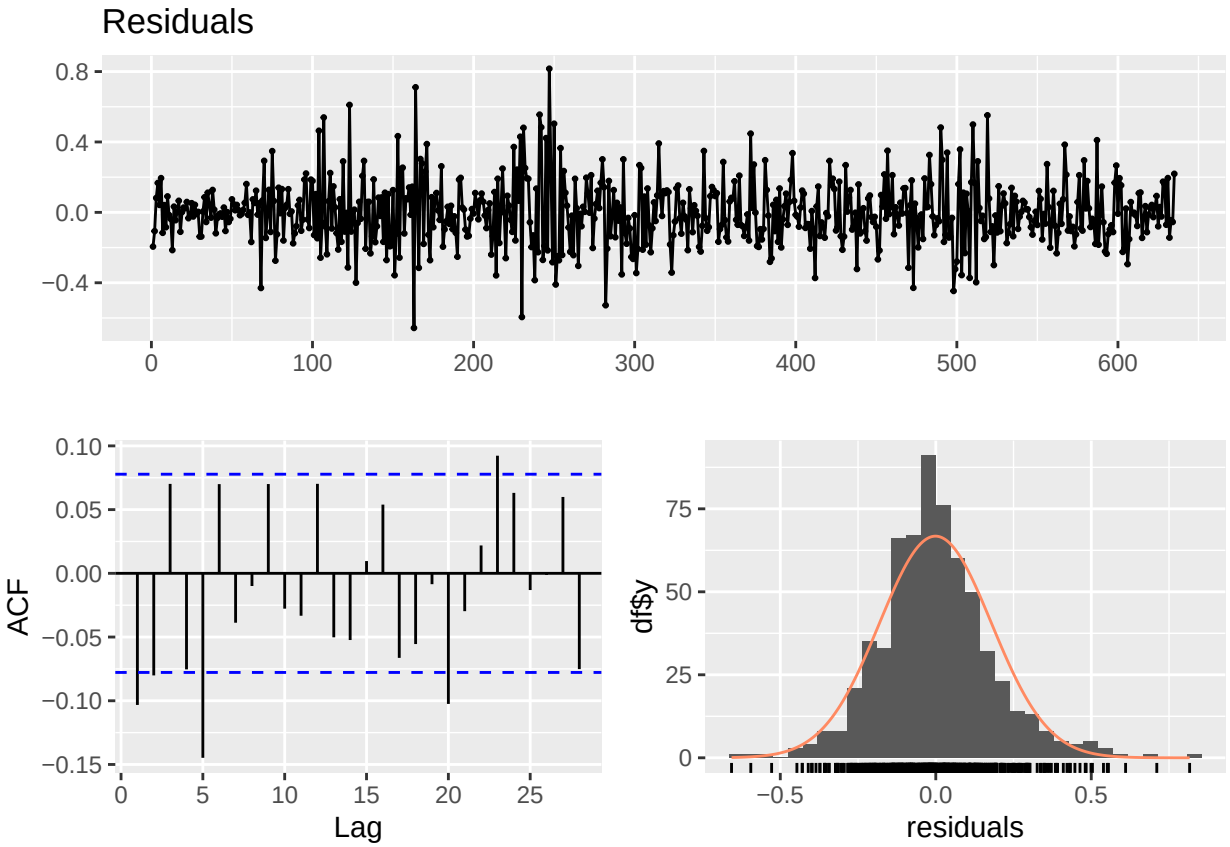
```
## Series: ts(d3t)
## Regression with ARIMA(5,0,0) errors
##
## Coefficients:
##           ar1      ar2     ar3      ar4      ar5    xreg
##       -0.1178  -0.0845  0.0326  -0.0893  -0.1533  0.7436
## s.e.   0.0395   0.0394  0.0396   0.0397   0.0393  0.0142
##
## sigma^2 = 0.03111:  log likelihood = 203.7
## AIC=-393.4   AICc=-393.23   BIC=-362.23
##
## Training set error measures:
##                        ME      RMSE       MAE  MPE MAPE      MASE        ACF1
## Training set -0.001448499 0.1755434 0.1295376 -Inf  Inf 0.3577608 0.003178031
```

The model above and the linear regression model of d_{3t} and d_{1t} give us the following equation:

$$y_i = \beta_0 + \beta_1 d_{1t} + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + \phi_4 x_{t-4} + \phi_5 x_{t-5} + e_i$$

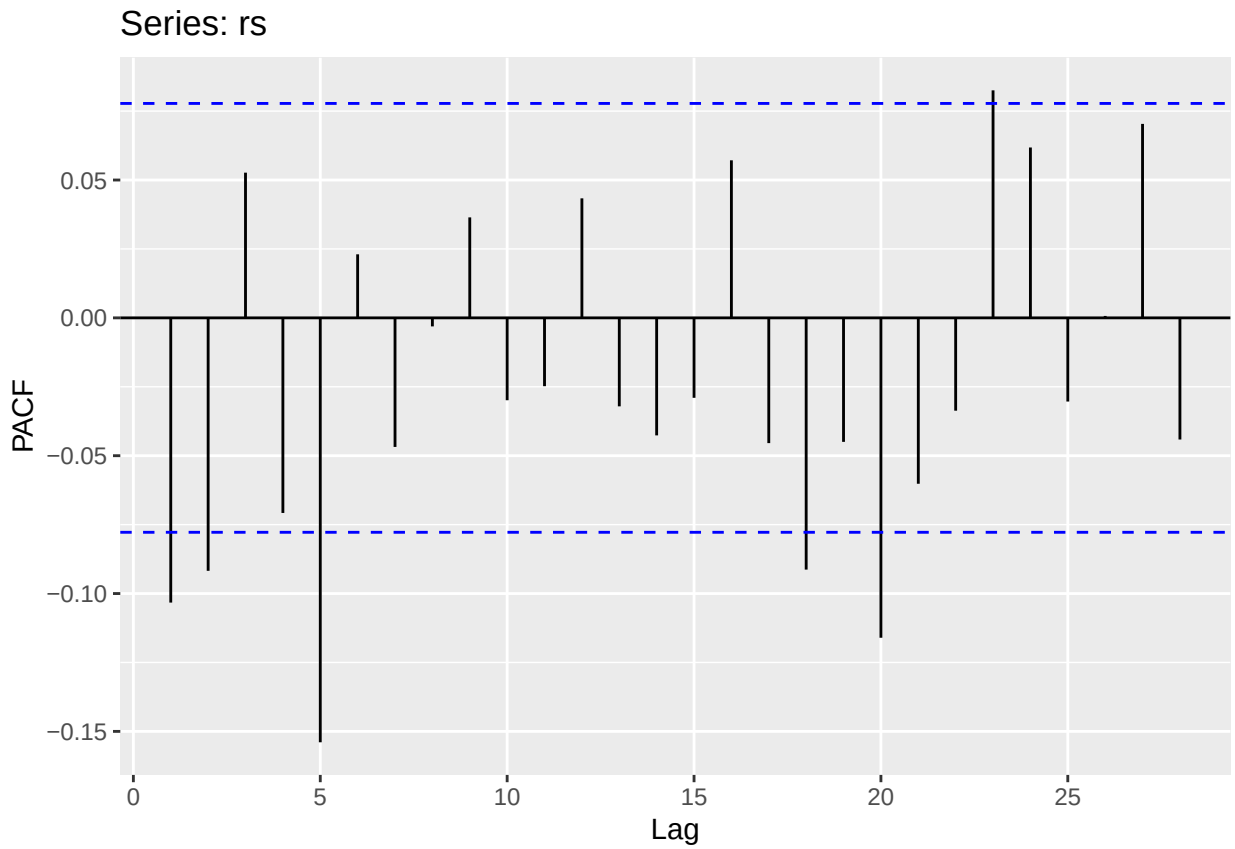
We will perform model diagnostics.

Check residuals plot:



The residuals plot might eyeball mean 0 but we'll have to run a t-test for mean 0 to confirm. The ACF plot shows the lags of the model residuals to be fairly stationary. The distribution looks tall and might be slightly skewed to the right, showing non-normalcy in respect to a Gaussian PDF.

PACF Plot:



The PACF plot suggests the model to be fairly stationary.

T-Test for mean 0:

```
##
##  One Sample t-test
##
## data:  data
## t = -0.14025, df = 634, p-value = 0.8885
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.01505208  0.01304527
## sample estimates:
##    mean of x
## -0.001003403
##
## T-Test: mean is statistically zero, linear trend *REMOVED* ->
## *FAIL* to reject H0

## [1] TRUE
```

The 95% CI of the linear regression model residuals contain 0 there the mean of the residuals is statistically 0, and that the linear trend is removed.

ADF Test:

```
##
```

```
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -4.9154
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:54:52 2023 by user: Reed
##
## ADF: contains *NO* unit roots over 30 lags, indicates *NO* mean drift,
## business cycles *NOT* present, series is stationary -> reject H0

##
## FALSE
```

The ADF test shows no unit roots and that the model is stationary.

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 6 lags.
##
## Value of test-statistic is: 0.0219
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216
##
## KPSS: *NO* unit roots, *NO* linear trend, slope zero,
## series is trend stationary -> *FAIL* to reject H0

## [1] TRUE
```

The KPSS test also shows no unit roots and that the model is stationary.

Check for business cycles:

```
## [1] 6.085 3.007
```

We find two business cycles in the model: 6-month and a 3-month cycles.

3.5. Refine the model in 3.4. by setting the insignificant coefficients to zero. Write the equation of the fitted model. Compare this model with your best linear regression model. Which is better? Why?

We test the model in section 3.4 for significant components:

```
##               t       pval_t       pval_z Pr(>|t|)
## ar1  -2.9821949 0.0029724542 2.861898e-03       **
## ar2  -2.1444754 0.0323767079 3.199481e-02        *
## ar3   0.8251224 0.4096144424 4.093021e-01
## ar4  -2.2479348 0.0249253839 2.458034e-02        *
## ar5  -3.9004771 0.0001063199 9.600331e-05      ***
## xreg 52.5257511 0.0000000000 0.000000e+00      ***
```

We find that the ar3 component of the model $\phi_3 x_{t-3}$ is statistically insignificant and will be removed.

We create the following reduced model:

```
c35 <- c(NA,NA, 0,NA,NA,NA)
m <- Arima(d3t, order=c(5,0,0), xreg=d1t, include.mean=F, fixed=c35)
```

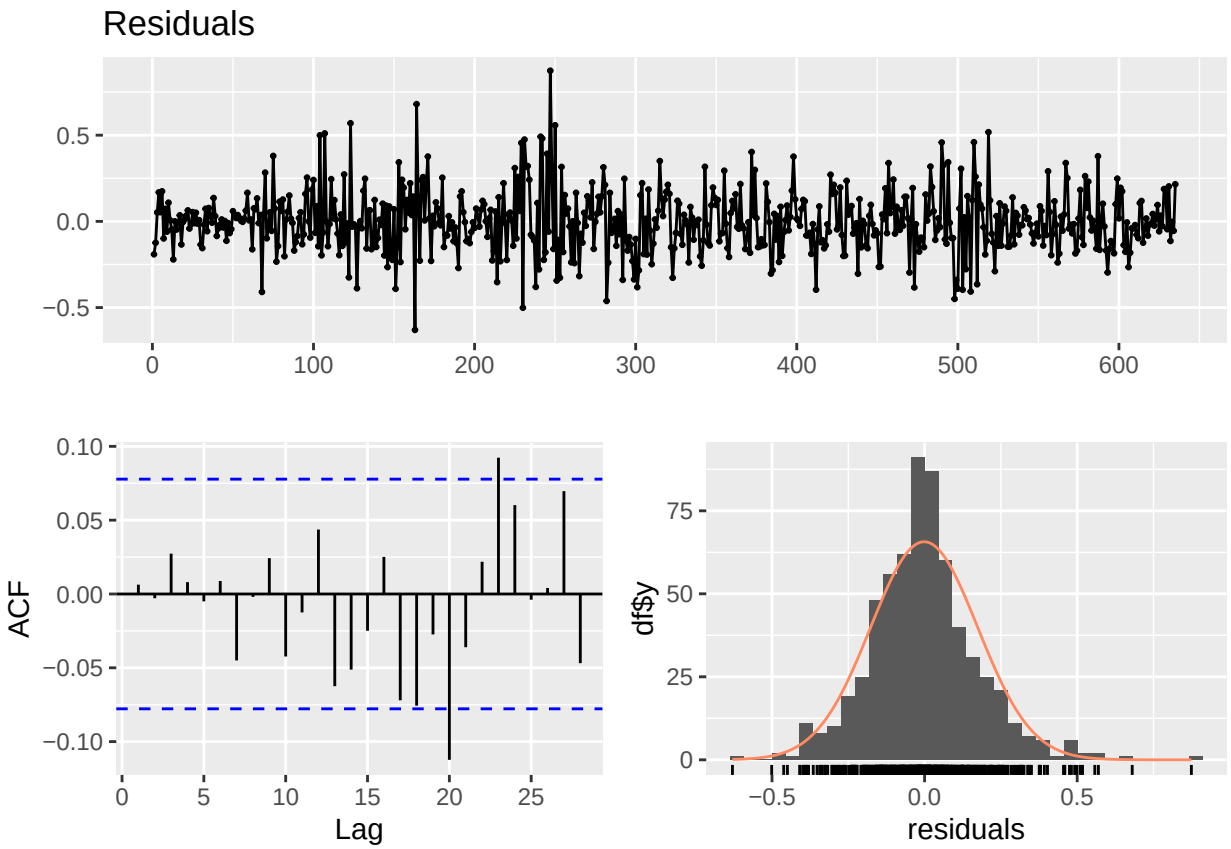
```
## Series: d3t
## Regression with ARIMA(5,0,0) errors
##
## Coefficients:
##           ar1      ar2  ar3      ar4      ar5    xreg
##       -0.1205  -0.0883    0  -0.0929  -0.1563  0.7429
## s.e.   0.0394   0.0391    0   0.0395   0.0391  0.0142
##
## sigma^2 = 0.03109:  log likelihood = 203.36
## AIC=-394.72   AICc=-394.59   BIC=-368
##
## Training set error measures:
##                         ME     RMSE       MAE  MPE MAPE      MASE        ACF1
## Training set -0.001501416 0.175638 0.1296148 -Inf  Inf 0.3579741 0.006322553
```

The reduced model above and the linear regression model of d_{3t} and d_{1t} give us the following equation:

$$y_i = \beta_0 + \beta_1 d_{1t} + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_4 x_{t-4} + \phi_5 x_{t-5} + e_i$$

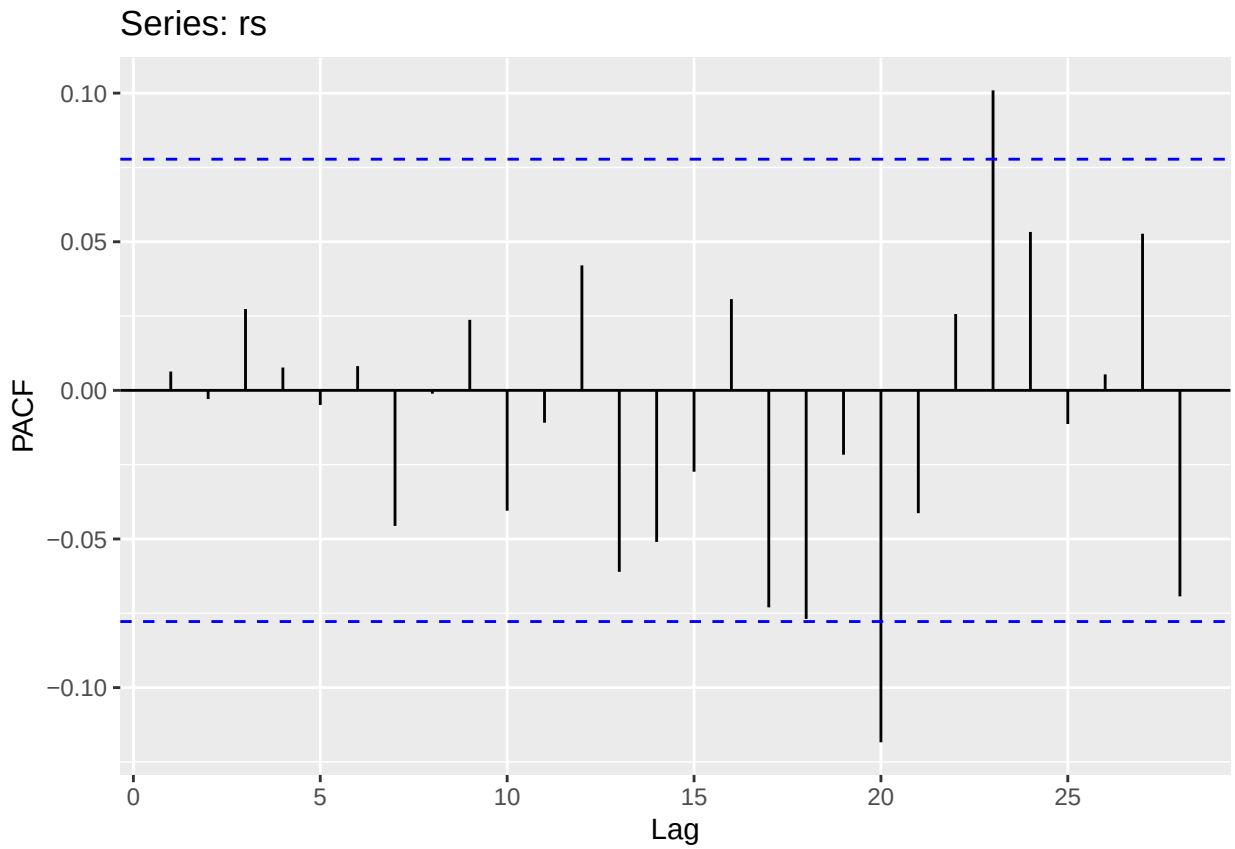
We will perform model diagnostics.

Check residuals plot:



The residuals plot might eyeball mean 0 but we'll have to run a t-test for mean 0 to find out. The ACF plot shows the lags of the model residuals to be fairly stationary. The distribution looks tall and might be slightly skewed to the right, showing non-normalcy in respect to a Gaussian PDF.

PACF Plot:



The PACF plot suggests the model to be fairly stationary.

T-Test for mean 0:

```
##
##  One Sample t-test
##
## data:  data
## t = -0.21525, df = 634, p-value = 0.8296
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.01519875  0.01219591
## sample estimates:
##    mean of x
## -0.001501416
##
## T-Test: mean is statistically zero, linear trend *REMOVED* ->
## *FAIL* to reject H0

## [1] TRUE
```

The 95% CI of the linear regression model residuals contain 0 there the mean of the residuals is statistically 0, and that the linear trend is removed.

ADF Test:

```
##
```

```
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -4.7489
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:54:52 2023 by user: Reed
##
## ADF: contains *NO* unit roots over 30 lags, indicates *NO* mean drift,
## business cycles *NOT* present, series is stationary -> reject H0

##
## FALSE
```

The ADF test shows no unit roots and that the model is stationary.

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 6 lags.
##
## Value of test-statistic is: 0.0275
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216
##
## KPSS: *NO* unit roots, *NO* linear trend, slope zero,
## series is trend stationary -> *FAIL* to reject H0

## [1] TRUE
```

The KPSS test also shows no unit roots and that the model is stationary.

Check for business cycles:

```
## [1] 6.082 3.008
```

We still find the same two business cycles in the reduced model: 6-month and a 3-month cycles.

3.6. Use the command polyroot in R to find the solutions of the characteristic equation of the refined AR($m33$\$$order$) model. How many real solutions are there?

```
## [1]  1.064082+0.930665i -0.560747+1.298226i -0.560747-1.298226i
## [4]  1.064082-0.930665i -1.600835+0.000000i
```

The real number is the solution without an imaginary number, which is only one value at -1.600835+0.000000i.

3.7. Compute the inverse of the absolute values of the solutions of the characteristic equation. Show the maximum value of the inverses. Does the maximum value imply that the AR($m33$order$) model likely contains a unit root? Why?

Solutions of the characteristic equation:

```
## [1] 1.413651 1.414153 1.414153 1.413651 1.600835
```

Inverse absolute solutions of the characteristic equation:

```
## [1] 0.7073883 0.7071370 0.7071370 0.7073883 0.6246740
```

Maximum value of the inverses:

```
## [1] 0.7073883
```

While the KPSS and ADF tests indicate no unit roots, we also find two business cycles in the model. But since we are able to inverse the solutions to the characteristic equation, these inverses are characteristic roots to the model. That said, the max inverse is 0.7074, which is less than one. All characteristic roots are less than one, which makes the model stationary, such as what the KPSS and ADF tests indicate as well.

While characteristic roots exist in the stationary model, they are less than 1. Therefore the characteristic roots are not unit roots, and in theory unit roots should not exist in a stationary model.

3.8. Compare the fit of your best linear regression model and your best AR model. Which is preferred and why?

Regression Model summary:

```
##
## Call:
## lm(formula = y3t ~ y1t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.69599 -0.42038 -0.03045  0.37993  1.41445
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.712645   0.041909    17.0   <2e-16 ***
## y1t         0.940816   0.006676   140.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.529 on 634 degrees of freedom
## Multiple R-squared:  0.9691, Adjusted R-squared:  0.969
## F-statistic: 1.986e+04 on 1 and 634 DF,  p-value: < 2.2e-16
```

Full ARIMA(5,0,0)/Regression Model summary:

```
## Series: ts(d3t)
## Regression with ARIMA(5,0,0) errors
##
## Coefficients:
##           ar1      ar2     ar3      ar4      ar5    xreg
##       -0.1178  -0.0845  0.0326  -0.0893  -0.1533  0.7436
## s.e.   0.0395   0.0394  0.0396   0.0397   0.0393  0.0142
##
## sigma^2 = 0.03111:  log likelihood = 203.7
## AIC=-393.4   AICc=-393.23   BIC=-362.23
##
## Training set error measures:
##                         ME      RMSE       MAE  MPE MAPE      MASE        ACF1
## Training set -0.001448499 0.1755434 0.1295376 -Inf  Inf 0.3577608 0.003178031
```

Reduced ARIMA(5,0,0)/Regression Model summary:

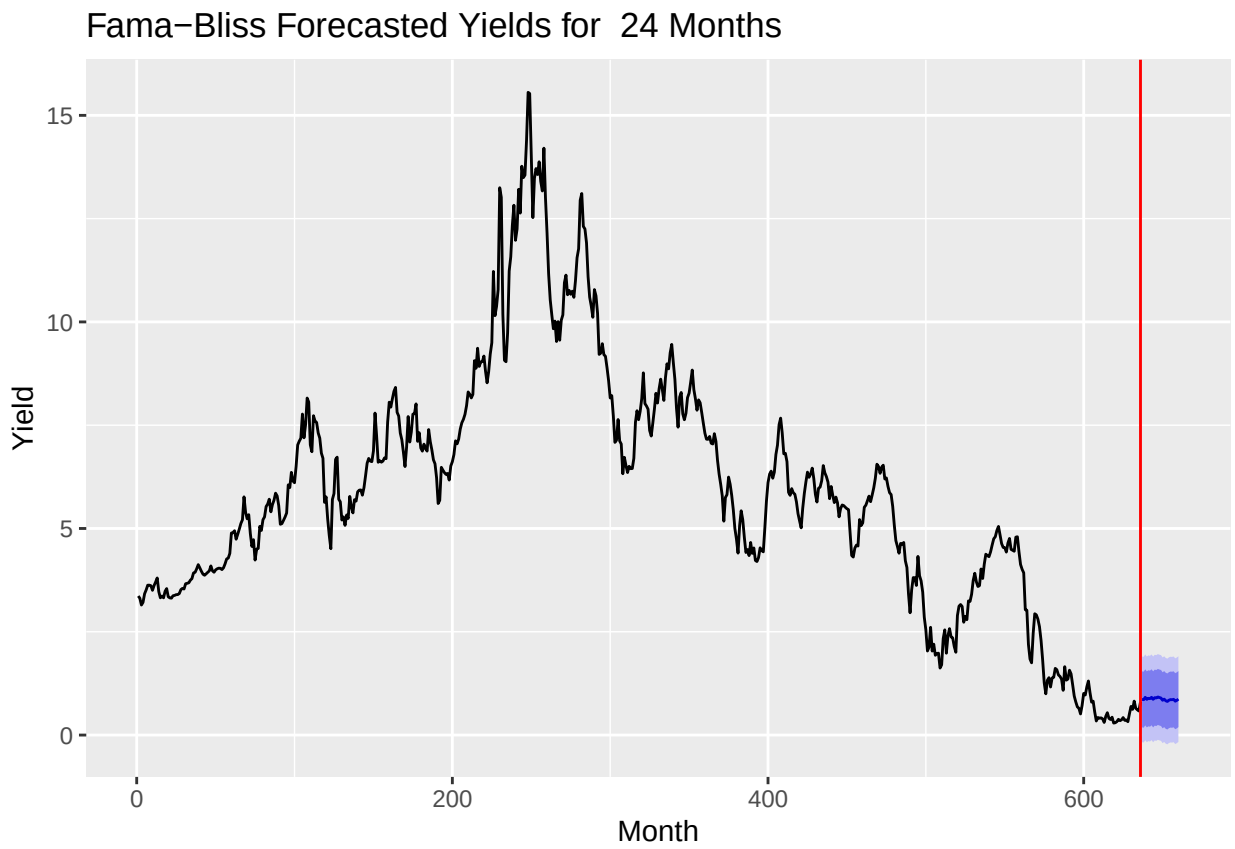
```
## Series: d3t
## Regression with ARIMA(5,0,0) errors
##
## Coefficients:
##           ar1      ar2  ar3      ar4      ar5    xreg
##       -0.1205  -0.0883    0  -0.0929  -0.1563  0.7429
## s.e.   0.0394   0.0391    0   0.0395   0.0391  0.0142
##
## sigma^2 = 0.03109:  log likelihood = 203.36
## AIC=-394.72   AICc=-394.59   BIC=-368
##
## Training set error measures:
##                         ME     RMSE       MAE  MPE MAPE      MASE        ACF1
## Training set -0.001501416 0.175638 0.1296148 -Inf  Inf 0.3579741 0.006322553
```

Comparing the Full and Reduced ARIMA(5,0,0)/Regression Models, the Reduced model has a lower AICc of -394.59 compared to the Full model AICc of 393.23. The Full model has lower RMSE and MAE. Comparing these statistics seem to be marginal. That being said, AIC tends to favor more complex models, but the reduced model has a more favorable AIC than the full model. Because of this, based on these statistics, we would choose the reduced model of the full model.

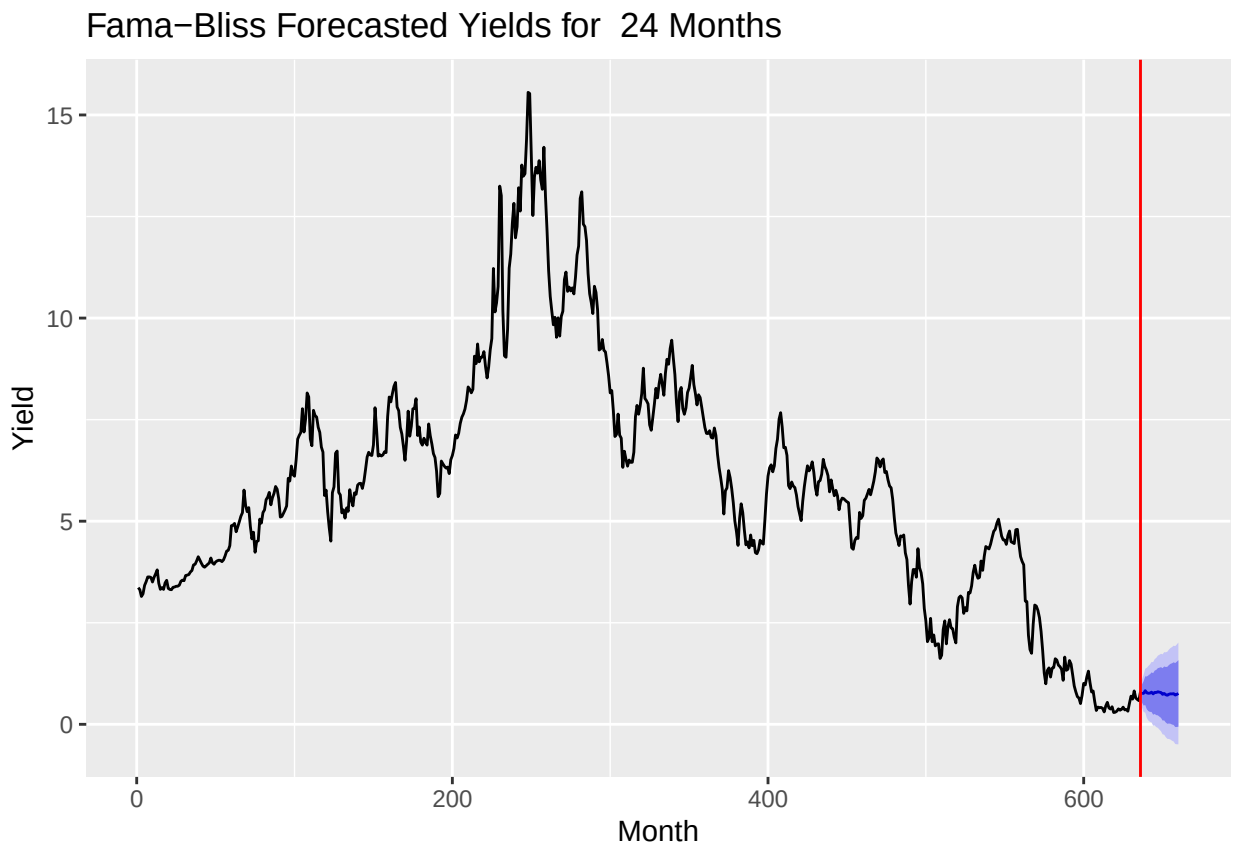
Fit regression model into ARIMA(0,0,0) white noise time series model

```
m311 <- Arima(y3t, order=c(0,0,0), xreg=y1t, include.mean=T)
```

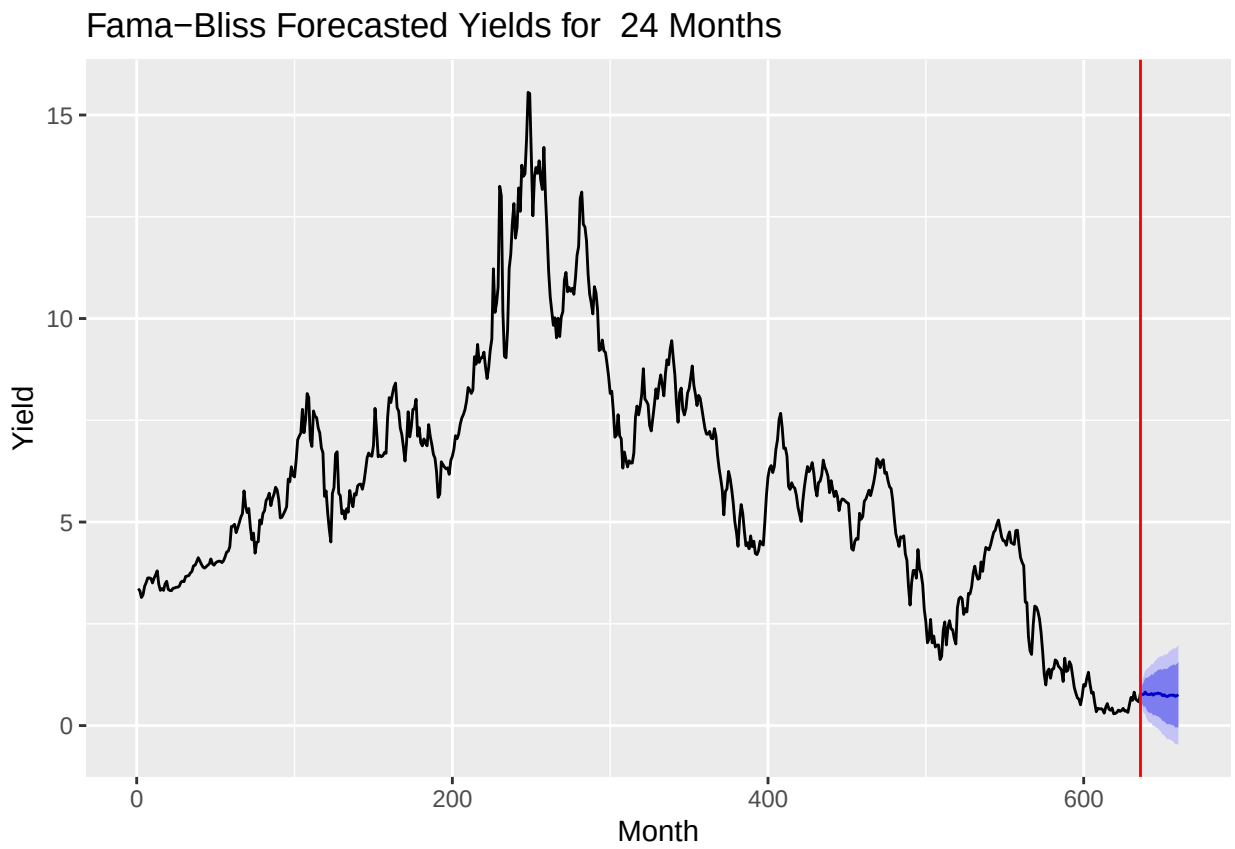
Forecasting using the Regression Model:



Forecasting using the full ARIMA(5,0,0)/Regression Model:



Forecasting using the reduced ARIMA(5,0,0)/Regression Model:



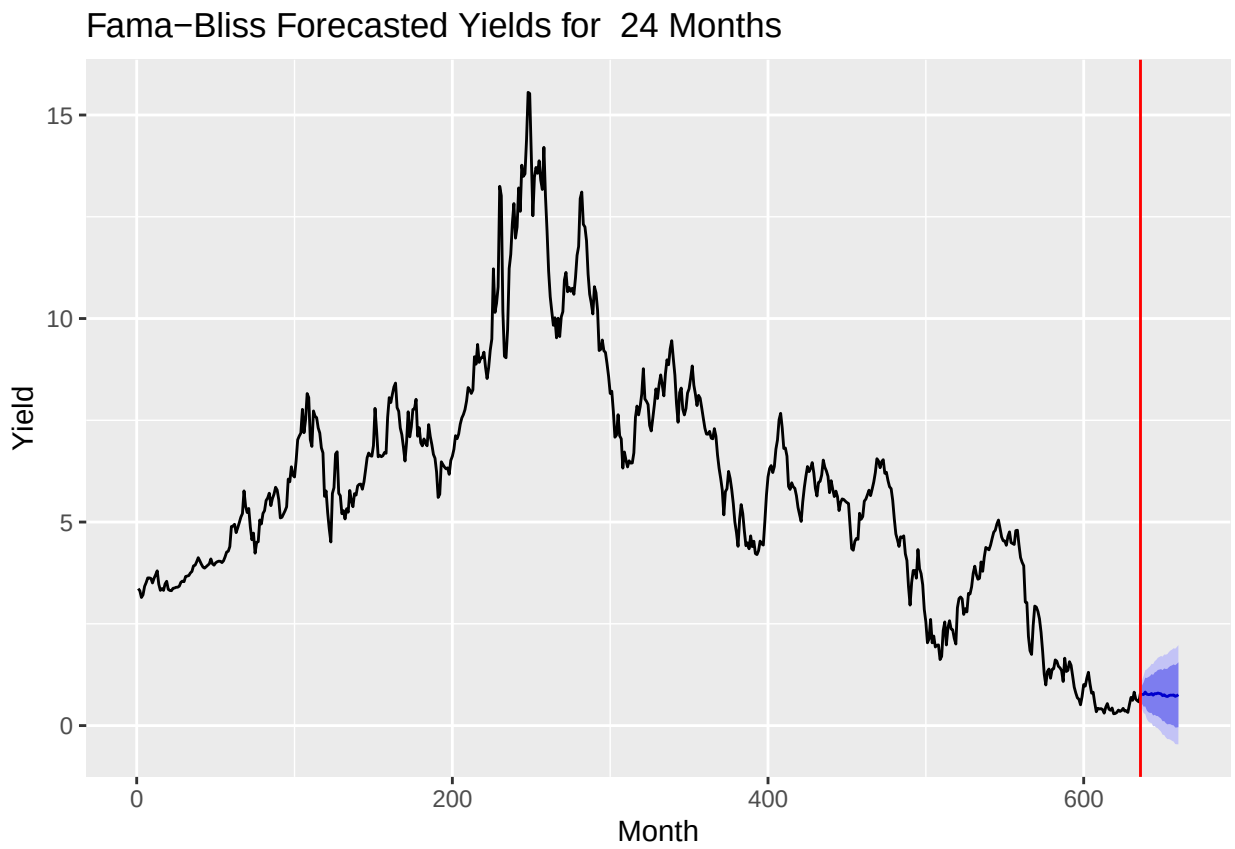
Comparing the forecast results above, while the overall forecasts of both the full and reduced ARIMA(5,0,0)/regression models are not that great, the reduced ARIMA(5,0,0)/regression model has a slightly smaller 80/95% CI range than the full model, implying better accuracy. Therefore we would prefer the reduced ARIMA(5,0,0)/regression model forecast.

The regression model forecast does not 'funnel outwards' like what the other two forecasts do, therefore short-term predictions for the regression model have a larger CI range. Because of that we find the regression model to perform worse than the other two models at forecasting.

4. Report (20 points)

For the Fama-Bliss bond yields analyses, choose what you think is the best model's outcomes and write an executive summary that allows stakeholders to make decisions or take action.

(Based on the analysis, modelling, testing, and forecasting performed in sections 2 and 3, we will use the reduced ARIMA(5,0,0)/regression model as noted in section 3.8 for our forecasting executive report.)



This forecasting model attempts to predict 3-year bond yields for the next 24 months. The data is sourced from the Fama-Bliss Discount Bonds report provided by the Center for Research in Security Prices, LLC (CRSP) website which contains 53 years of monthly 1-year and 3-year bond yield data from January of 1961 (indexed as month 1) to December of 2013 (indexed as month 636). The forecast presented here is based on the prototyping and evaluation of several financial models, and selected the best-performing model developed from the time series modelling tools we currently have available.

The forecast model shows the dark blue point forecast line slightly sloping down from the last 3-year yield data point of 0.804 to the predicted 3-year yield 24-months later to 0.758. In addition to the point forecast line we also have an 80% confidence interval (CI) shown by the blue area of possible values that could occur, as well as a 95% CI expanded in the lighter blue area. The CI ranges also include the possibility of negative yields by 21 months in the 80% CI and as early as 8 months for the 95% CI.

Given our current time series modelling tools we would recommend to limit the use of this model to fairly short-term recommendations. Though we can see an overall general drop of 3-year bond yields in which the forecast point line shows, and have been at the lowest point the prior three years, despite the CI ranges and barring an economic crisis we don't think we'll realistically reach negative yields anytime soon. But from the model we should also understand yields will also not be dramatically increasing. We can generalize from the model short-term that 3-year bond yields will be similar as to what is currently happening, but can also expect that it could slightly drop within the next few months as well.

We will continue to improve this model by iteration as our toolsets and knowledge base increases to hopefully provide longer term forecasts.