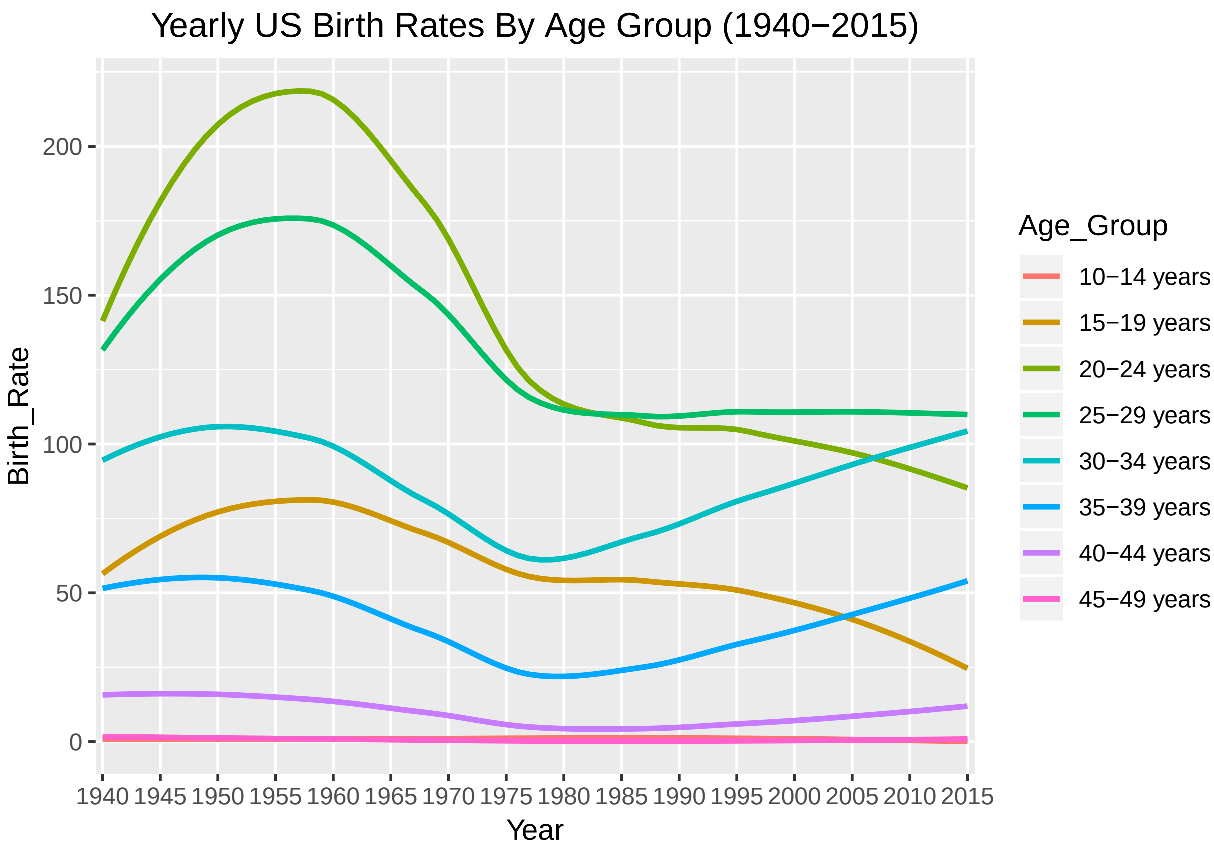
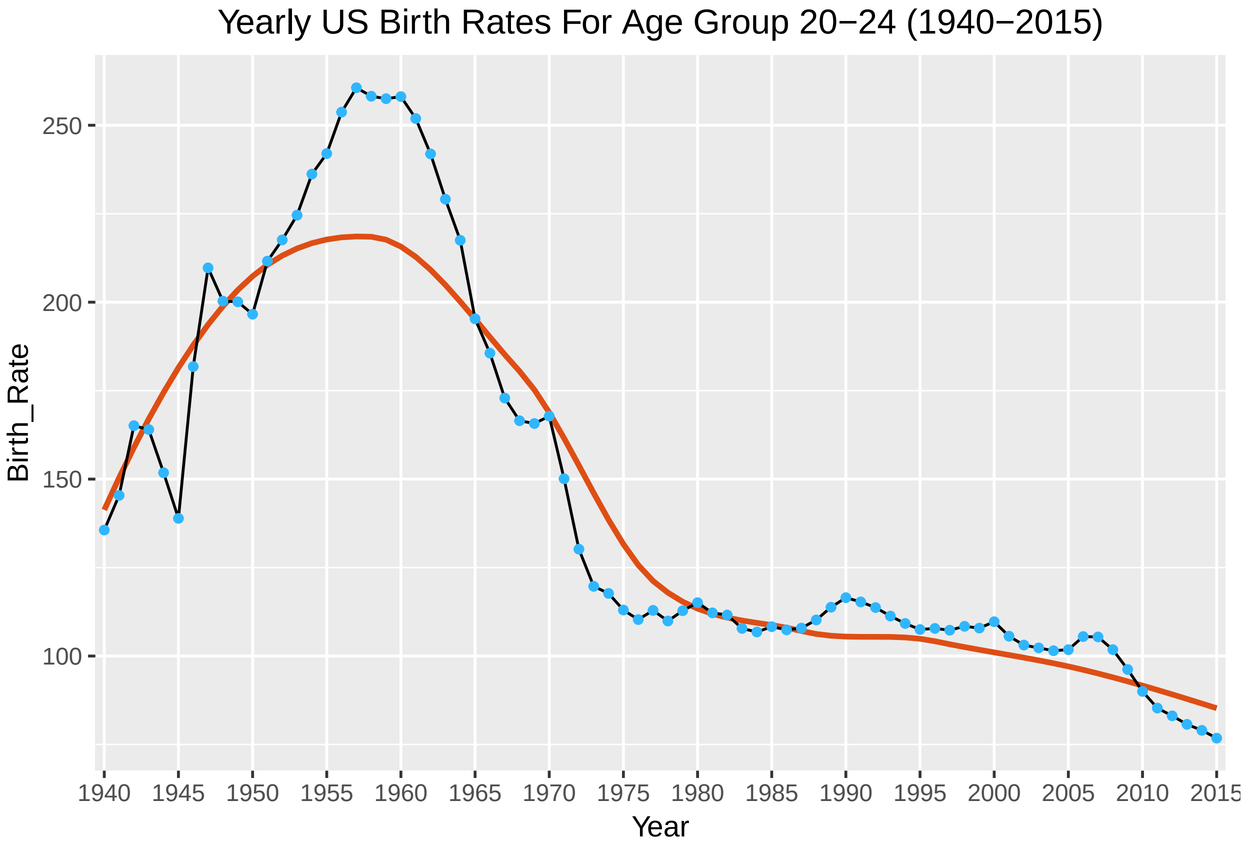
146 Project Code

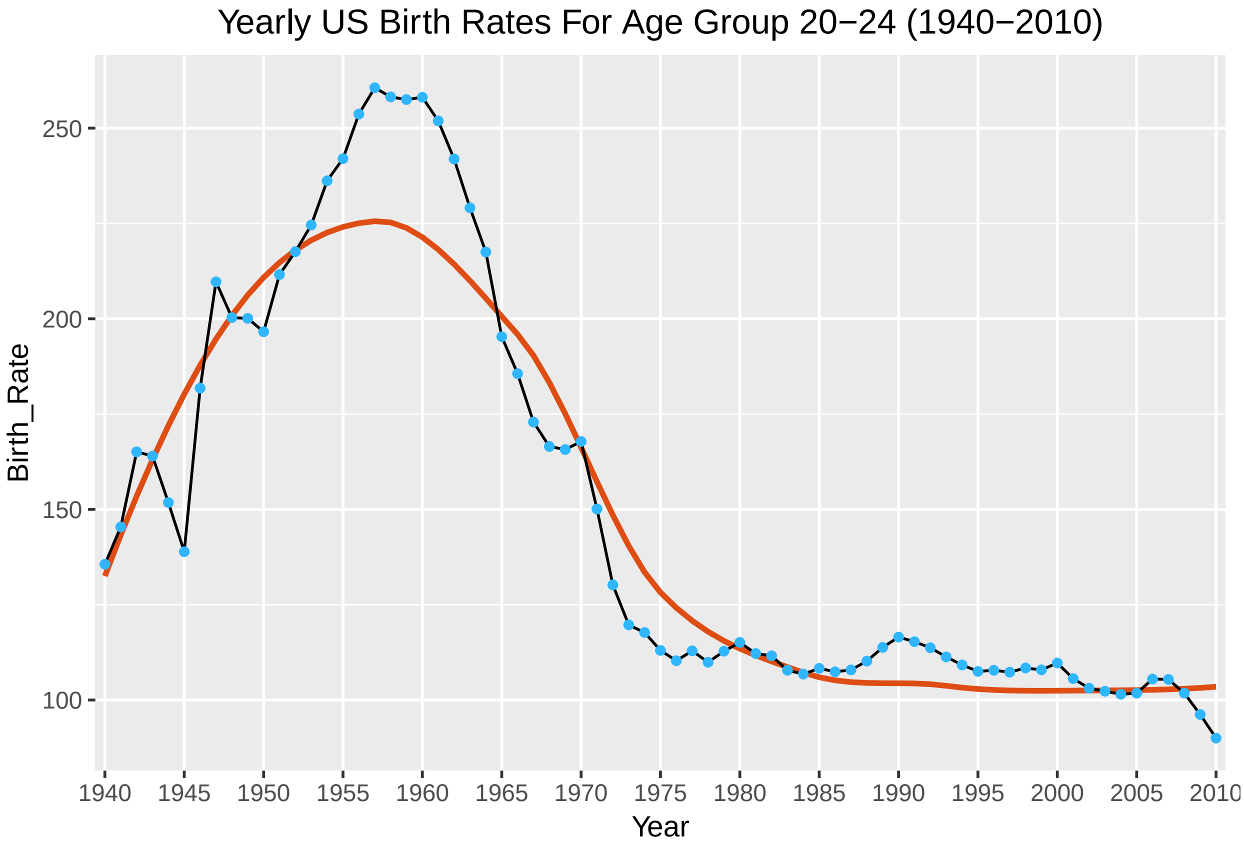
All Age Groups Plotted



Age Group of 20-24 Year Old Females

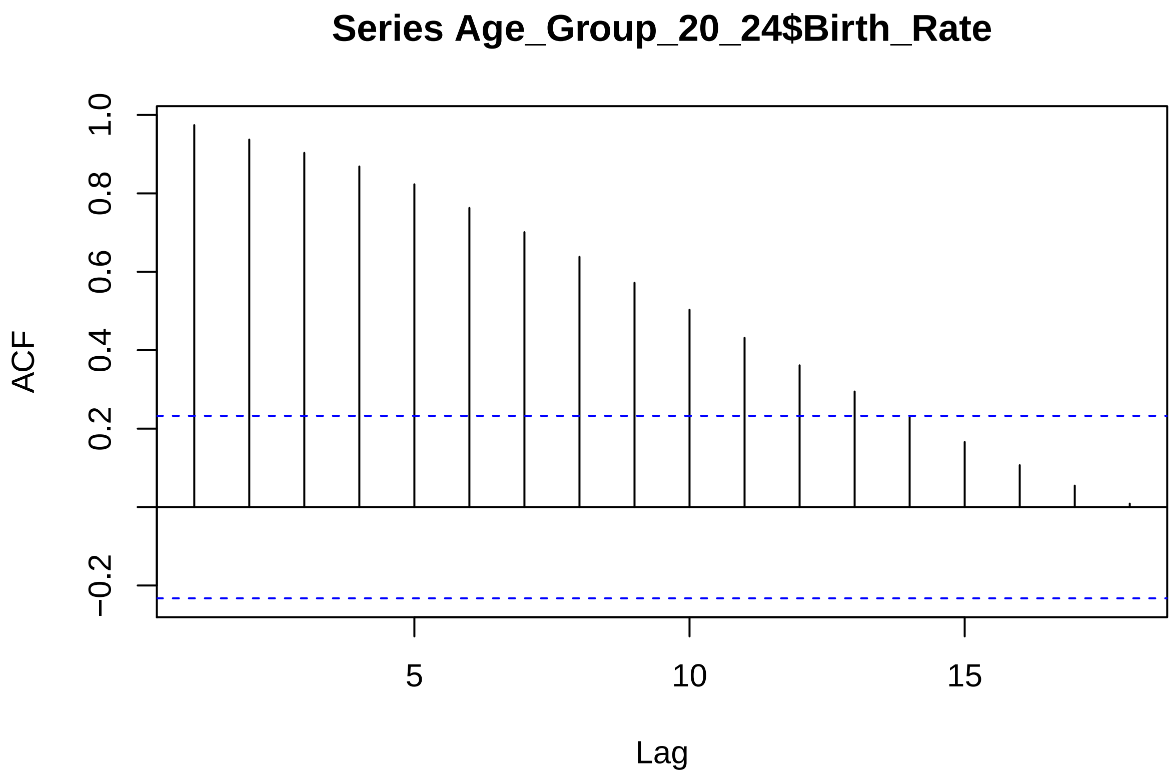


Removing 5 years for Prediction:

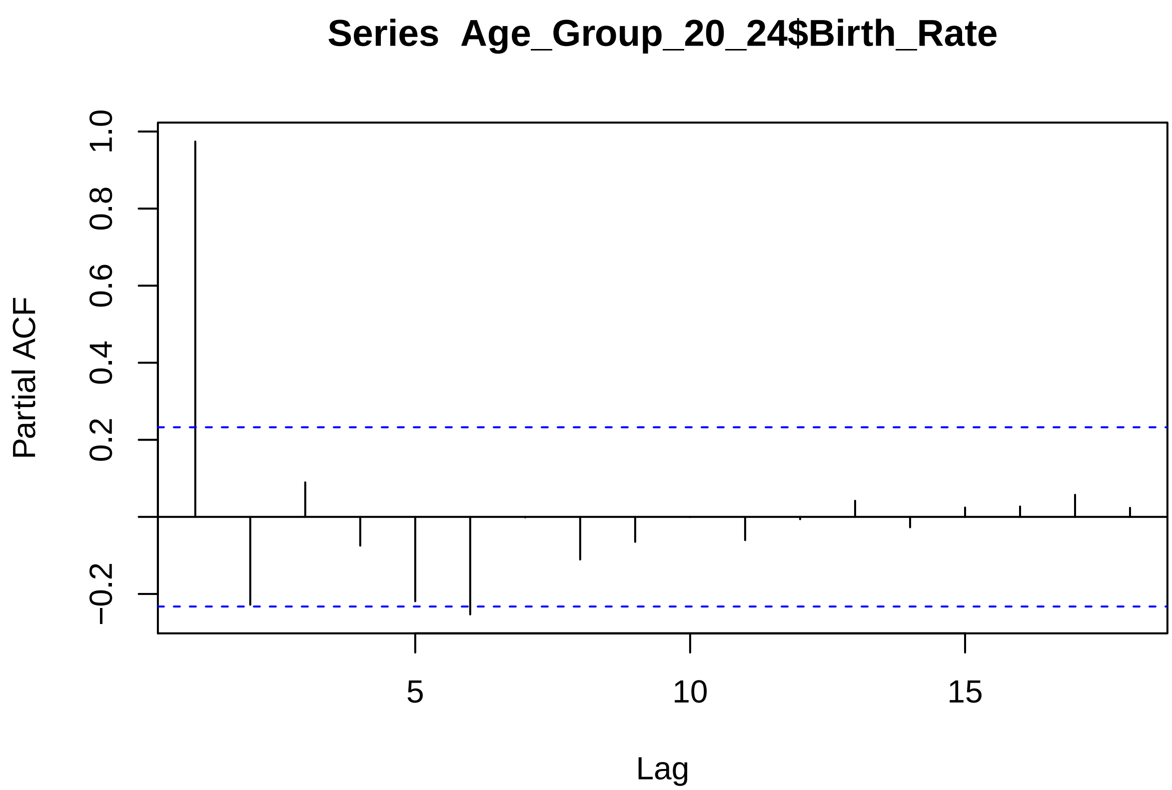


ACF , PACF , & EACF Plots:

```
acf( Age_Group_20_24$Birth_Rate )
```




```r
pacf( Age_Group_20_24$Birth_Rate )
```




```r
eacf( Age_Group_20_24$Birth_Rate )
```

```
## AR/MA
```

```
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x x x x x x x x x x x  x  x  o
## 1 x o o x x o o o o x o  o  o  o
## 2 x o o o x o o o o x o  o  o  o
## 3 x o o o x o o o o x o  o  o  o
## 4 x o o o x o o o o o o  o  o  o
## 5 x o x x o o o o o o o  o  o  o
## 6 x o x o o o o o o o o  o  o  o
## 7 x x x o o o o o o o o  o  o  o
```

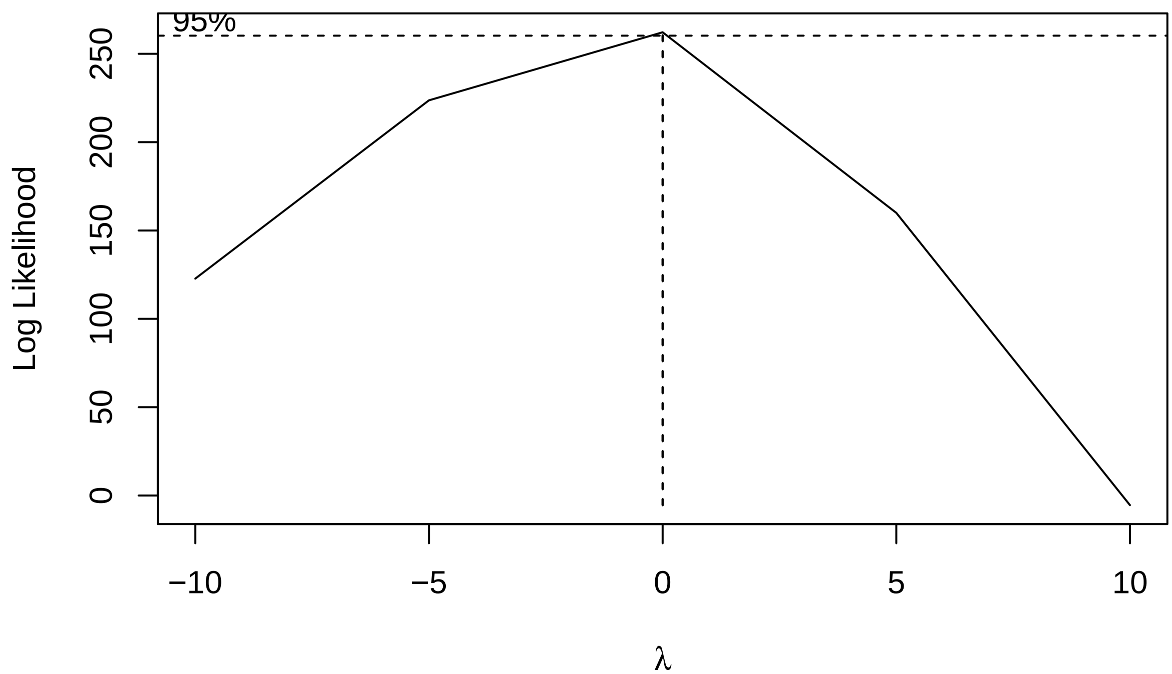
Stationarity Check:

Table 1: Augmented Dickey-Fuller Test: `Age_Group_20_24$Birth_Rate`

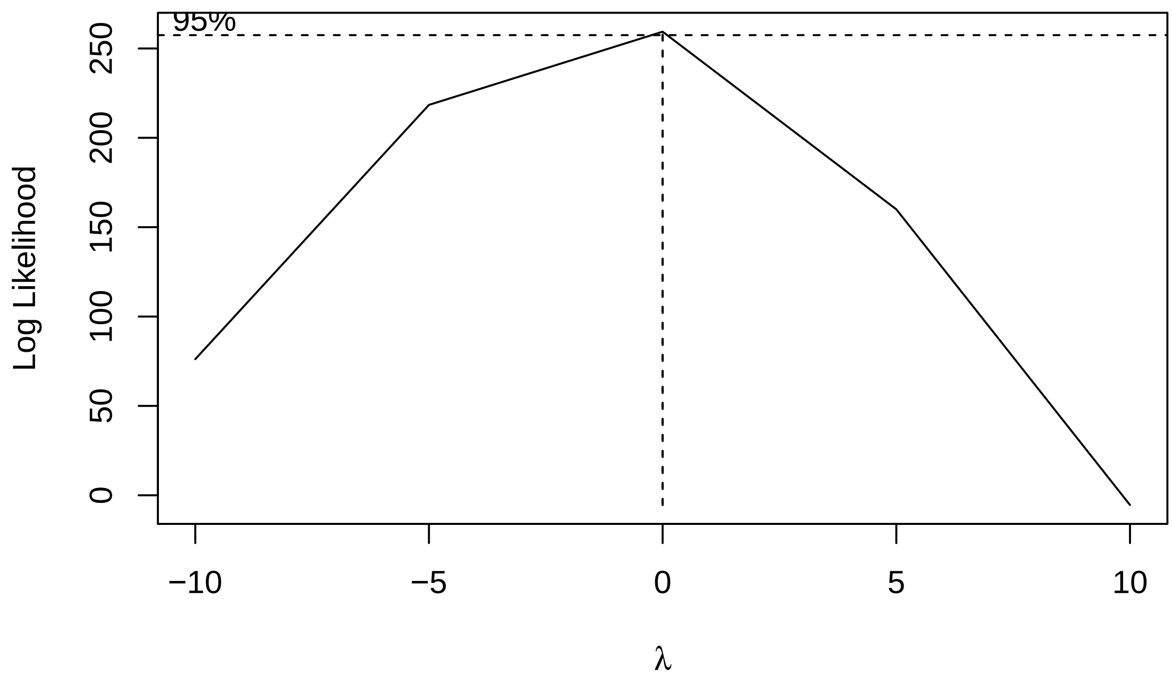
Test statistic	Lag order	P value	Alternative hypothesis
-1.583	4	0.745	stationary

Applying Transformations:

```r
BoxCox.ar( Age_Group_20_24$Birth_Rate , lambda = seq( from = -10 , to = 10 , by = 5 ))
```



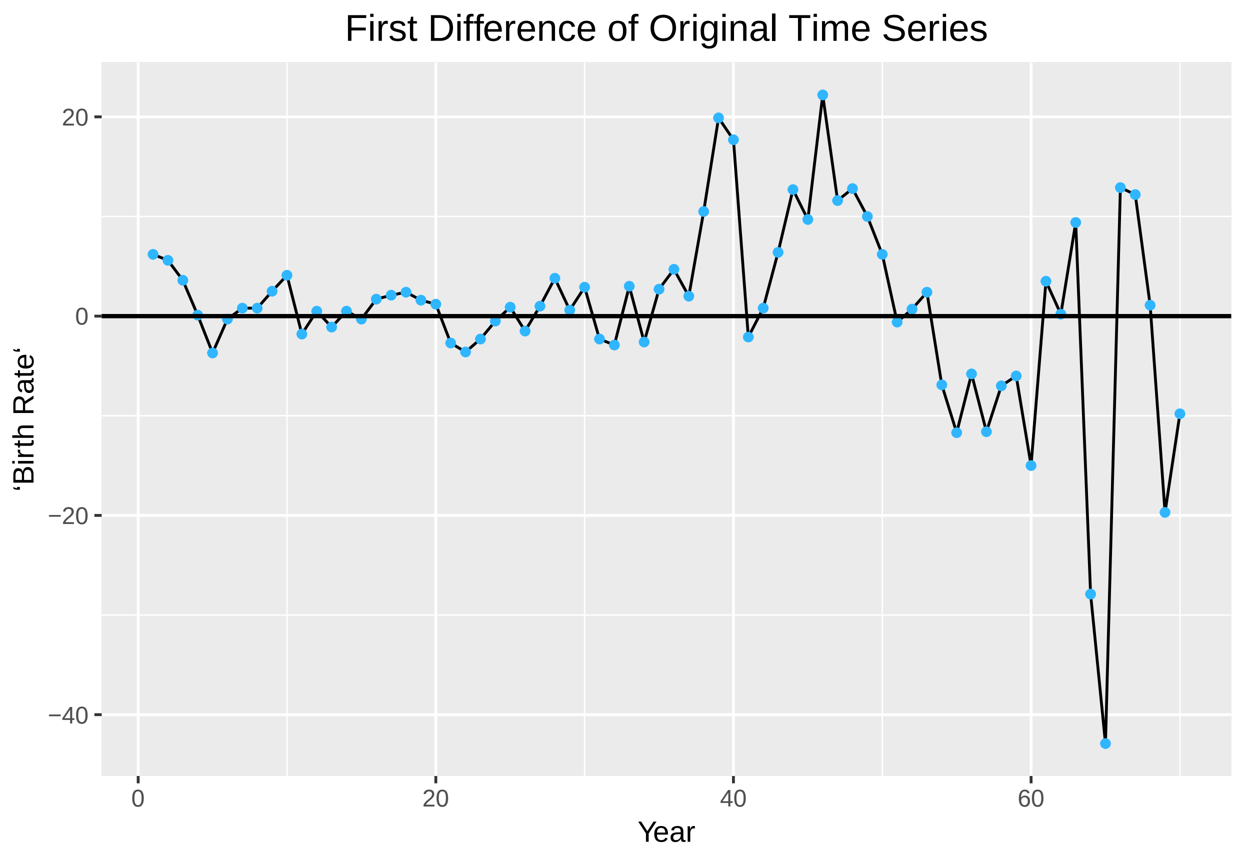

```r
# Second Box Cox Transformation is using burg instead of maximum likelihood estimation.

BoxCox.ar( Age_Group_20_24$Birth_Rate , lambda = seq( from = -10 , to = 10 , by = 5 ) , method = "burg")
```



Interpretation: Since the Box Cox plot gives such a wide confidence interval for lambda (λ) I will not apply any transformation for now.

Making the Time Series Stationary Via Differencing:

1st Difference:



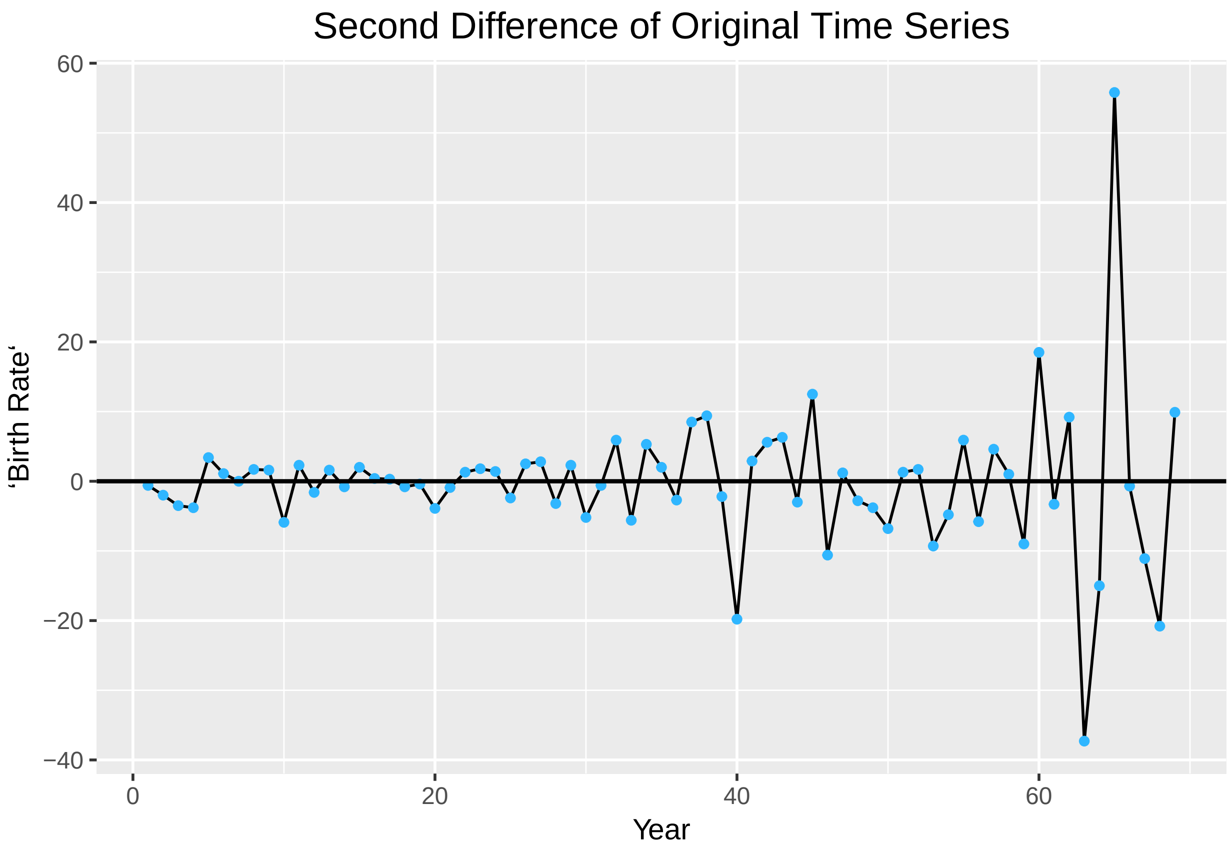
Stationarity Check:

Table 2: Augmented Dickey-Fuller Test: `Diff_Birth_Rate_20_24_Frame$`Birth Rate“

Test statistic	Lag order	P value	Alternative hypothesis
-1.877	4	0.6253	stationary

Interpretation: With $\alpha = 0.05$ we fail to reject H_0.

2nd Difference:



Stationarity Check:

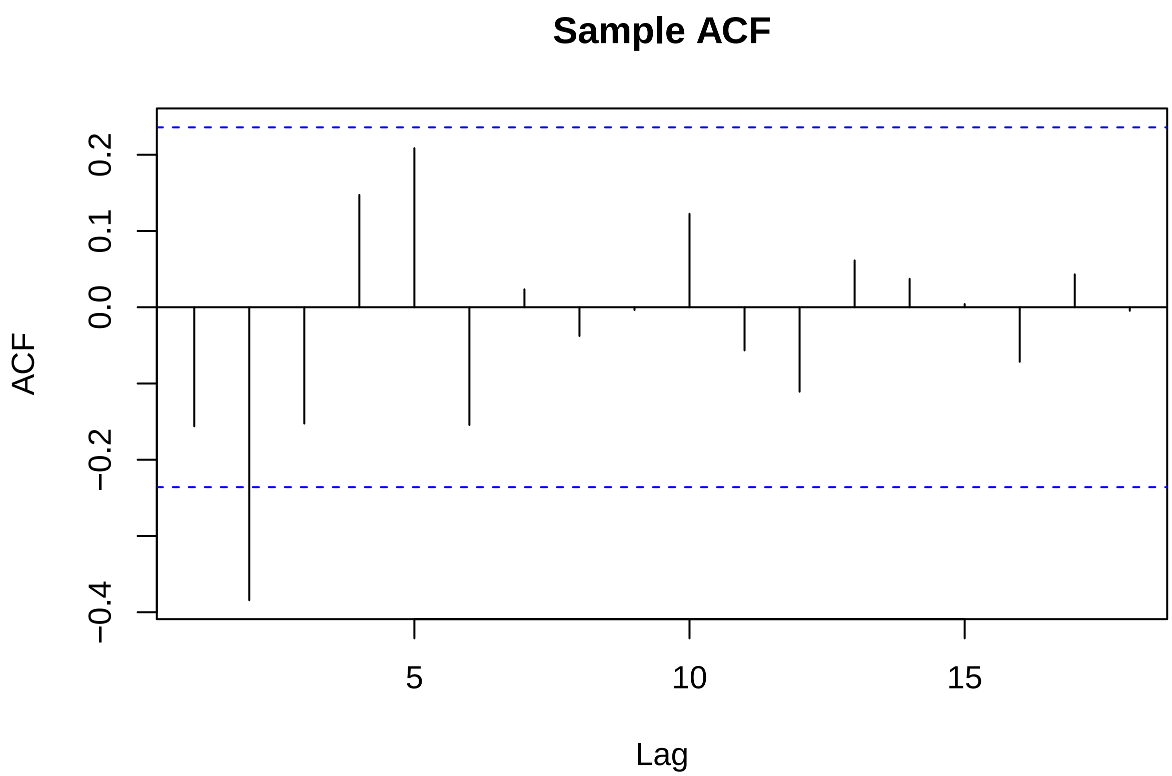
Table 3: Augmented Dickey-Fuller Test: `Diff_Diff_Birth_Rate_20_24_Frame$`Birth Rate"

Test statistic	Lag order	P value	Alternative hypothesis
-4.339	4	0.01 * *	stationary

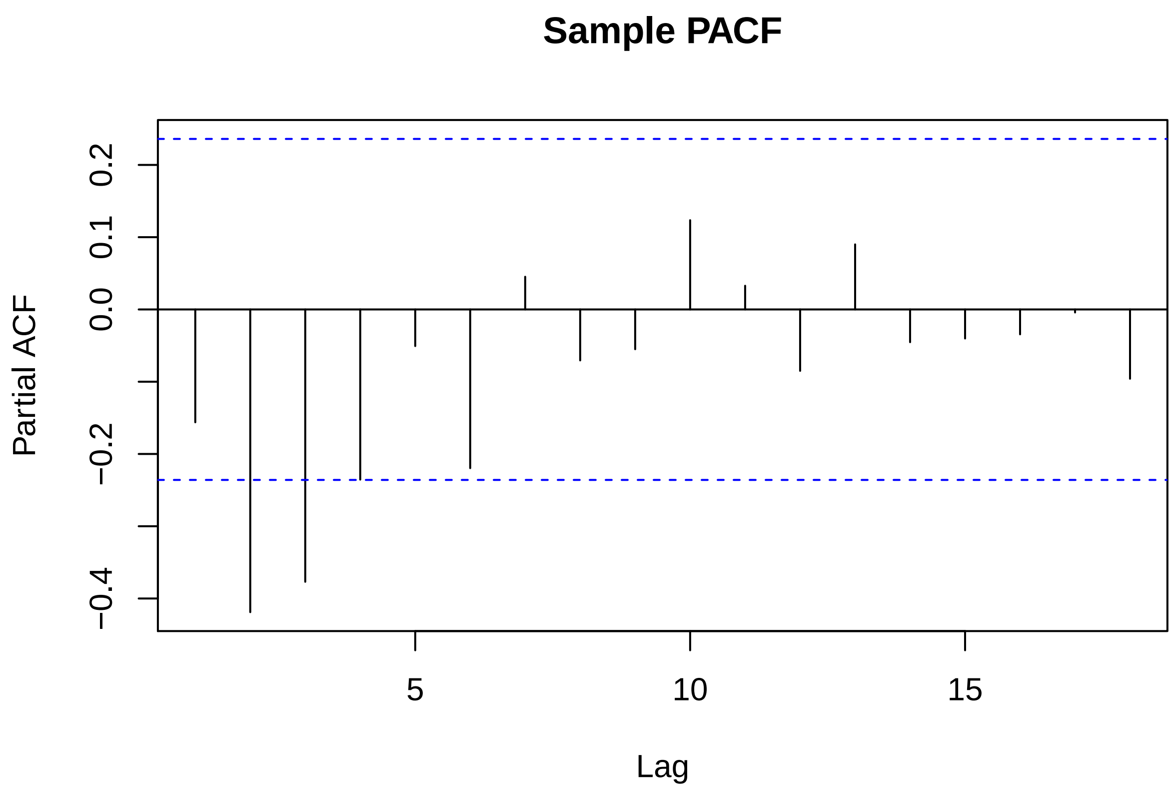
Interpretation: At $\alpha = 0.05$ we reject H_0.

ACF , PACF , & EACF Plots:

```
acf( Diff_Diff_Birth_Rate_20_24_Frame$`Birth Rate` , main = "Sample ACF")
```




```r
pacf( Diff_Diff_Birth_Rate_20_24_Frame$`Birth Rate` , main = "Sample PACF" )
```




```r
eacf( Diff_Diff_Birth_Rate_20_24_Frame$`Birth Rate` )
```

```
## AR/MA
```

```
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 o x o o o o o o o o o  o  o  o
## 1 x x x o o o o o o o o  o  o  o
## 2 x o o o x o o o o o o  o  o  o
## 3 x o o x o o o o o o o  o  o  o
## 4 x o x o o o o o o o o  o  o  o
## 5 x x x o o o o o o o o  o  o  o
## 6 o o x o o o o o o o o  o  o  o
## 7 x o o o o o o o o o o  o  o  o
```

Possible Models to fit:

- ARIMA (2 , 2 , 1)
- ARIMA (2 , 2 , 0)
- ARIMA (3 , 2 , 4)
- Other ARIMA models could also be suggested.

ARIMA (2 , 2 , 1) Model:

Call: arima(x = Age_Group_20_24$Birth_Rate, order = c(2, 2, 1), method = "ML")

Table 4: Coefficients

	ar1	ar2	ma1
	0.246	-0.4122	-0.7084
s.e.	0.1319	0.1207	0.1077

sigma^2 estimated as 68.38: log likelihood = -244.25, aic = 494.51

95 Percent CI for Parameter Significance:

- AR(1) $\longrightarrow$ 0.246 $\pm$ 0.2585 $\longrightarrow$ $\otimes$
- AR(2) $\longrightarrow$ -0.4122 $\pm$ 0.1207 $\longrightarrow$ ✓
- MA(1) $\longrightarrow$ -0.7084 $\pm$ 0.1077 $\longrightarrow$ ✓

ARIMA (2 , 2 , 0) Model:

Call: arima(x = Age_Group_20_24$Birth_Rate, order = c(2, 2, 0), method = "ML")

Table 5: Coefficients

	ar1	ar2
	-0.2118	-0.4302
s.e.	0.1088	0.1102

sigma^2 estimated as 85.23: log likelihood = -251.49, aic = 506.98

95 Percent CI for Parameter Significance:

- AR(1) $\longrightarrow$ -0.2118 $\pm$ 0.1088 $\longrightarrow$ $\otimes$ (very close to being significant)
- AR(2) $\longrightarrow$ -0.4302 $\pm$ 0.1102 $\longrightarrow$ ✓

ARIMA (3 , 2 , 4) Model:

Call: arima(x = Age_Group_20_24$Birth_Rate, order = c(3, 2, 4), method = "ML")

Table 6: Coefficients

	ar1	ar2	ar3	ma1	ma2	ma3	ma4
	-0.3266	-0.6274	-0.6567	-0.1399	-0.001863	0.2597	-0.4235
s.e.	0.2484	0.1858	0.1775	0.2638	0.2738	0.1758	0.1846

sigma^2 estimated as 63.69: log likelihood = -242.19, aic = 498.39

95 Percent CI for Parameter Significance:

- AR(1) $\longrightarrow$ -0.3266 $\pm$ 0.2484 $\longrightarrow$ $\otimes$
- AR(2) $\longrightarrow$ -0.6274 $\pm$ 0.1858 $\longrightarrow$ ✓
- AR(3) $\longrightarrow$ -0.6567 $\pm$ 0.1775 $\longrightarrow$ ✓
- MA(1) $\longrightarrow$ -0.1399 $\pm$ 0.2638 $\longrightarrow$ $\otimes$
- MA(2) $\longrightarrow$ -0.001863 $\pm$ 0.2738 $\longrightarrow$ $\otimes$
- MA(3) $\longrightarrow$ 0.2597 $\pm$ 0.1758 $\longrightarrow$ $\otimes$
- MA(4) $\longrightarrow$ -0.4235 $\pm$ 0.1846 $\longrightarrow$ ✓

ARIMA (0 , 1 , 1) Model:

Call: arima(x = Age_Group_20_24$Birth_Rate, order = c(0, 1, 1), method = "ML")

Table 7: Coefficients

	ma1
	0.522
s.e.	0.1076

sigma^2 estimated as 72.11: log likelihood = -249.22, aic = 500.44

95 Percent CI for Parameter Significance:

- MA(1) $\longrightarrow$ 0.522 $\pm$ 0.2108 $\longrightarrow$ ✓

ARIMA (1 , 2 , 1) Model:

Call: arima(x = Age_Group_20_24$Birth_Rate, order = c(1, 2, 1), method = "ML")

Table 8: Coefficients

	ar1	ma1
	0.3188	-0.8829
s.e.	0.1521	0.08722

sigma^2 estimated as 78.54: log likelihood = -248.93, aic = 501.86

95 Percent CI for Parameter Significance:

- AR(1) $\longrightarrow$ 0.3188 $\pm$ 0.2981 $\longrightarrow$ ✓
- MA(1) $\longrightarrow$ -0.8829 $\pm$ 0.1709 $\longrightarrow$ ✓

ARIMA (2 , 2 , 2) Model:

Call: arima(x = Age_Group_20_24$Birth_Rate, order = c(2, 2, 2), method = "ML")

Table 9: Coefficients

	ar1	ar2	ma1	ma2
	0.502	-0.5356	-1.001	0.3037
s.e.	0.235	0.13	0.2674	0.2279

sigma^2 estimated as 66.86: log likelihood = -243.52, aic = 495.04

95 Percent CI for Parameter Significance:

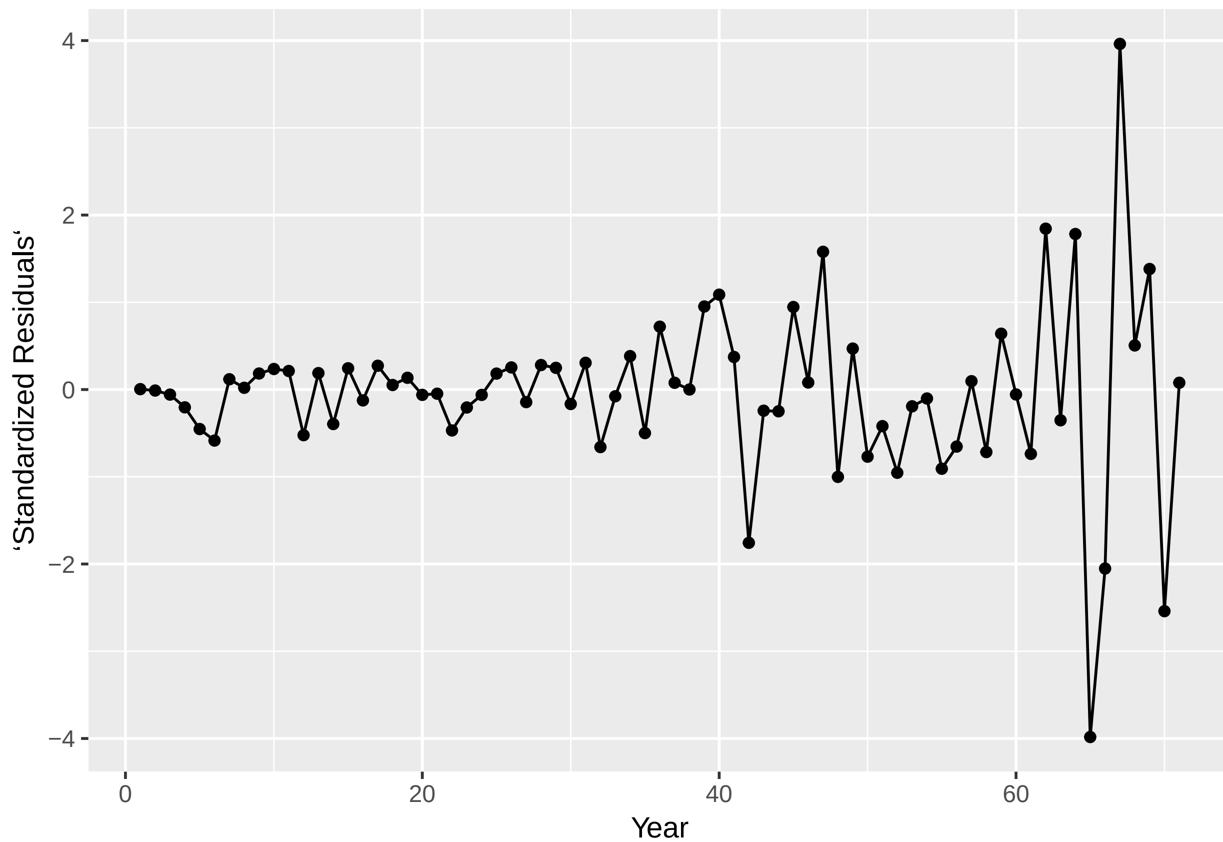
- AR(1) $\longrightarrow$ 0.502 $\pm$ 0.4606 $\longrightarrow$ ✓
- AR(2) $\longrightarrow$ -0.5356 $\pm$ 0.2548 $\longrightarrow$ ✓
- MA(1) $\longrightarrow$ -1.001 $\pm$ 0.5241 $\longrightarrow$ ✓
- MA(2) $\longrightarrow$ 0.3037 $\pm$ 0.4466 $\longrightarrow$ $\otimes$

Best Possible Models Thus Far:

- ARIMA (2 , 2 , 0)

- ARIMA (0 , 1 , 1) (Model Suggested by auto.arima function)
- ARIMA (1 , 2 , 1)
- ARIMA (2 , 2 , 1) (Model Suggested by auto.arima function)

ARIMA (2 , 2 , 0) Assumptions Check:

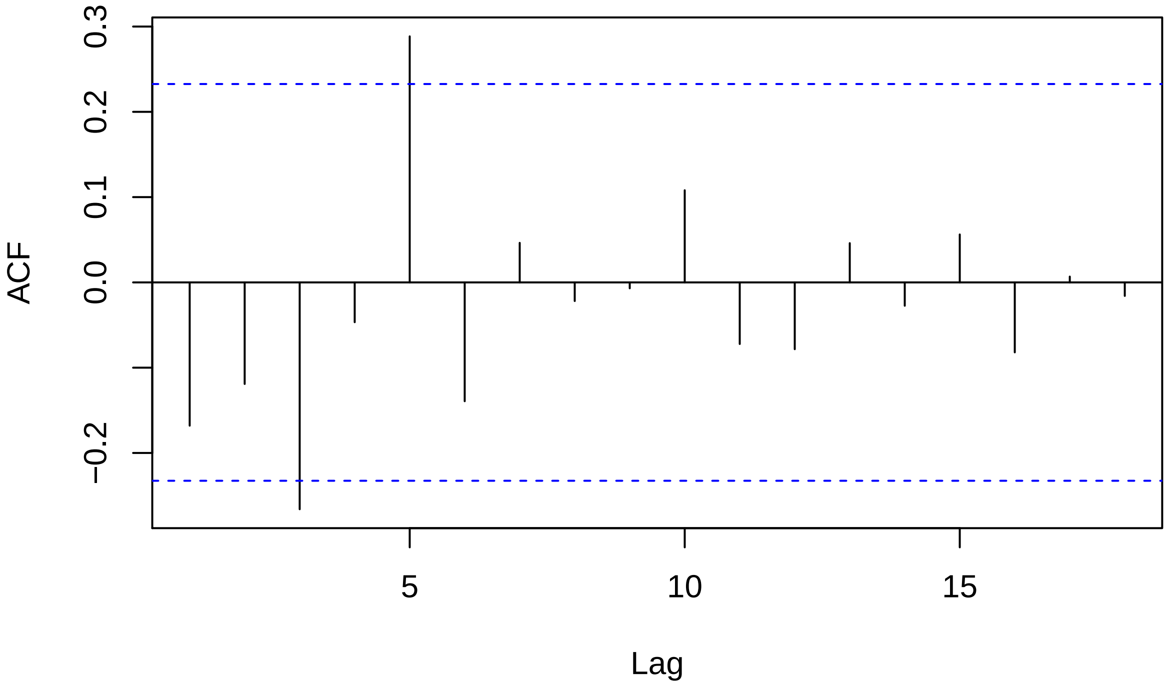


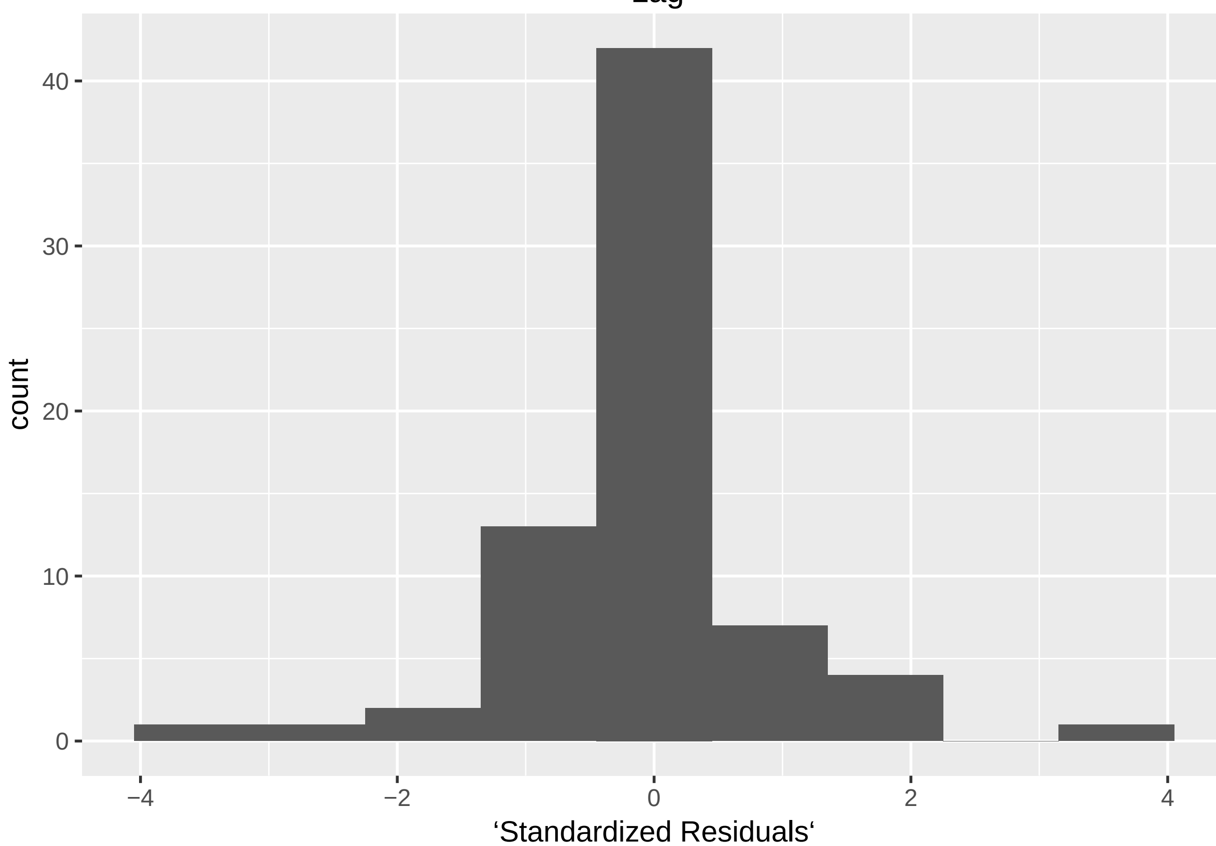
- **pvalue**: *0.811*
- **observed.runs**: *35*
- **expected.runs**: *36.49*
- **n1**: *36*
- **n2**: *35*
- **k**: *0*

Table 10: Augmented Dickey-Fuller Test: `ARIMA_2_2_0$residuals`

Test statistic	Lag order	P value	Alternative hypothesis
-3.648	4	0.03561 *	stationary

Sample ACF for ARIMA (2,2,0) Standardized Residuals





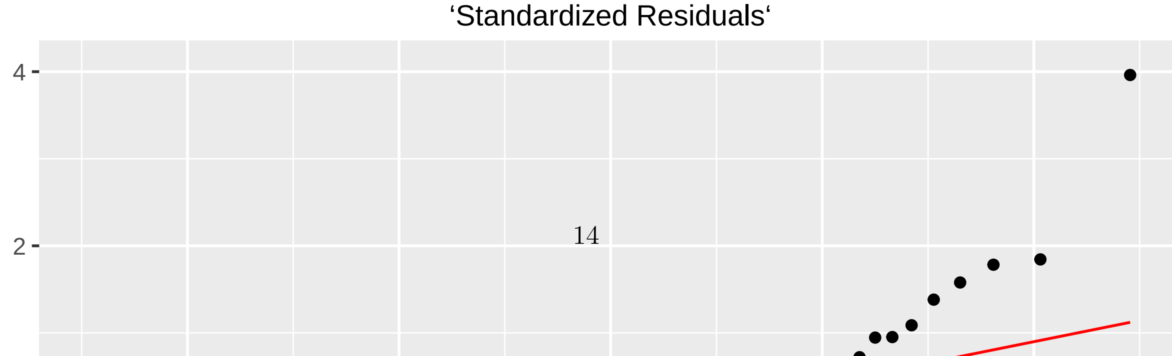


Table 11: Shapiro-Wilk normality test: `ARIMA_2_2_0_Frame$`Standardized Residuals“

Test statistic	P value
0.8571	9.845e-07 * * *

- Independence (Runs Test): ✓
- Normality Test: ⊗ (Maybe another Difference Idea from Lab 7 Pg.8)
- ADF Test: ✓

ARIMA (1 , 2 , 1) Assumptions Check:

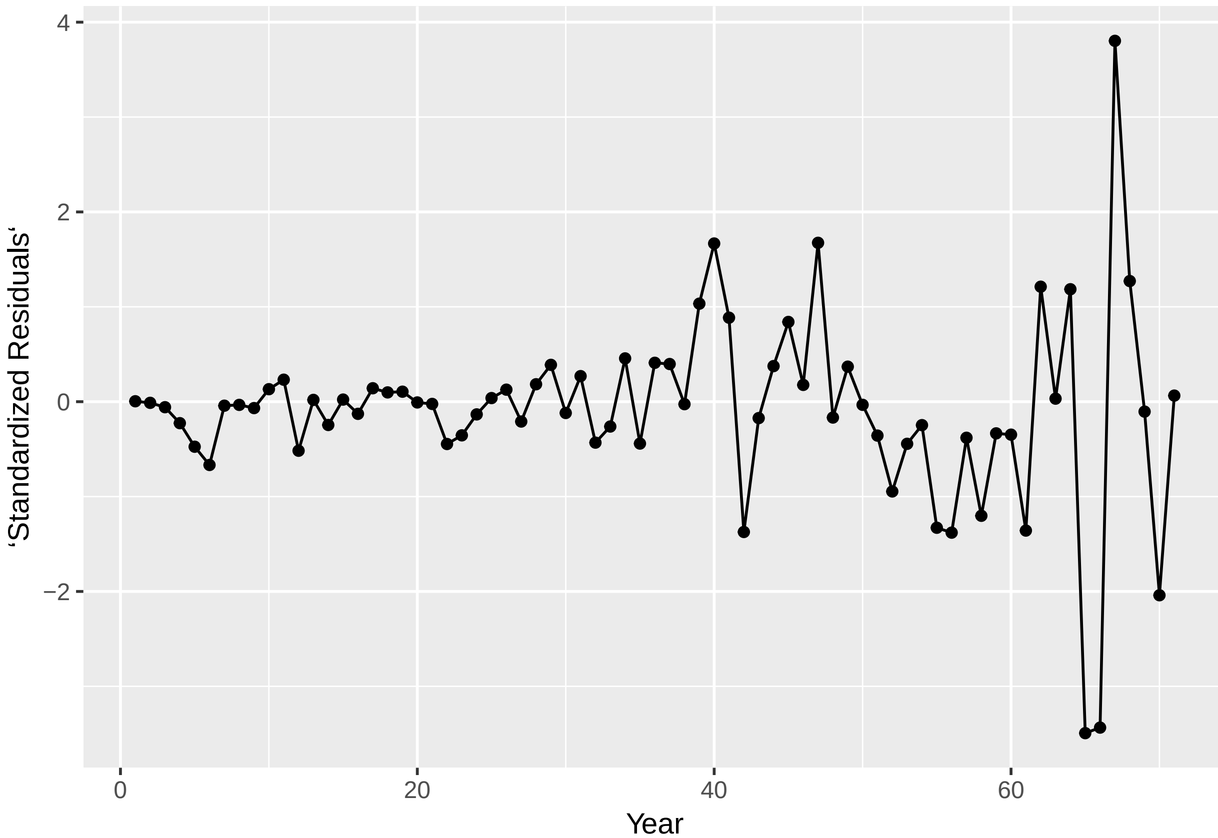
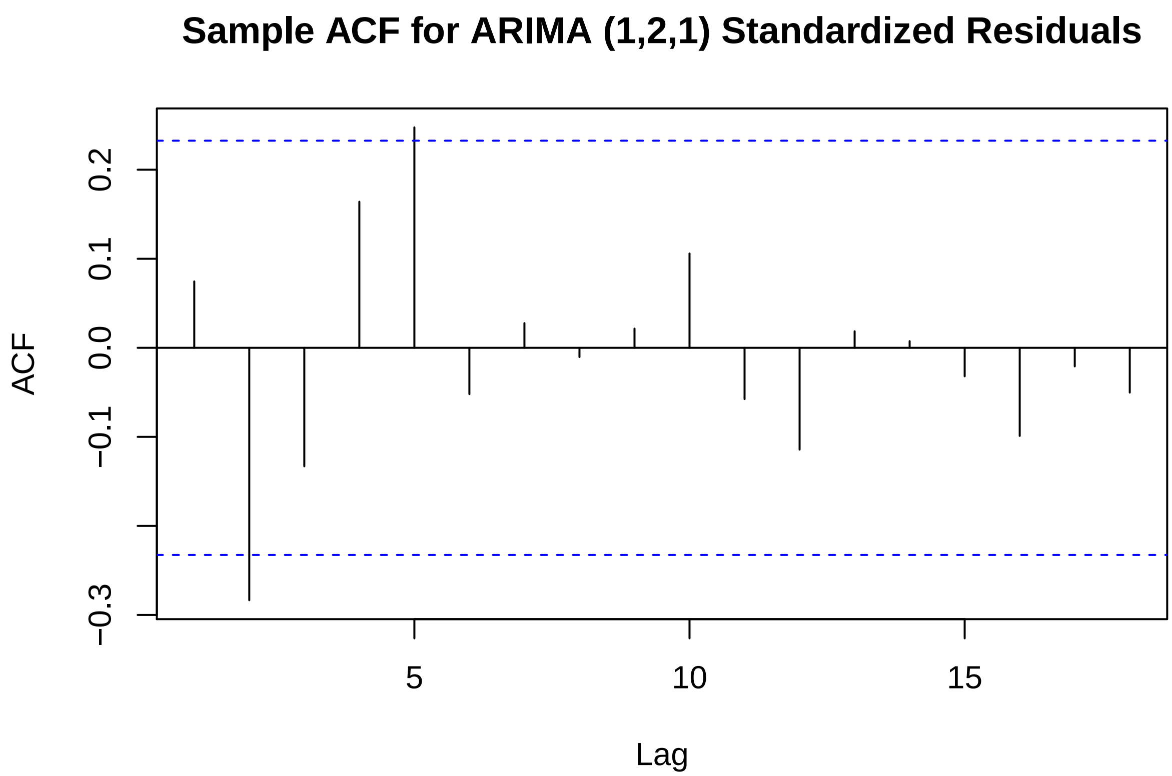
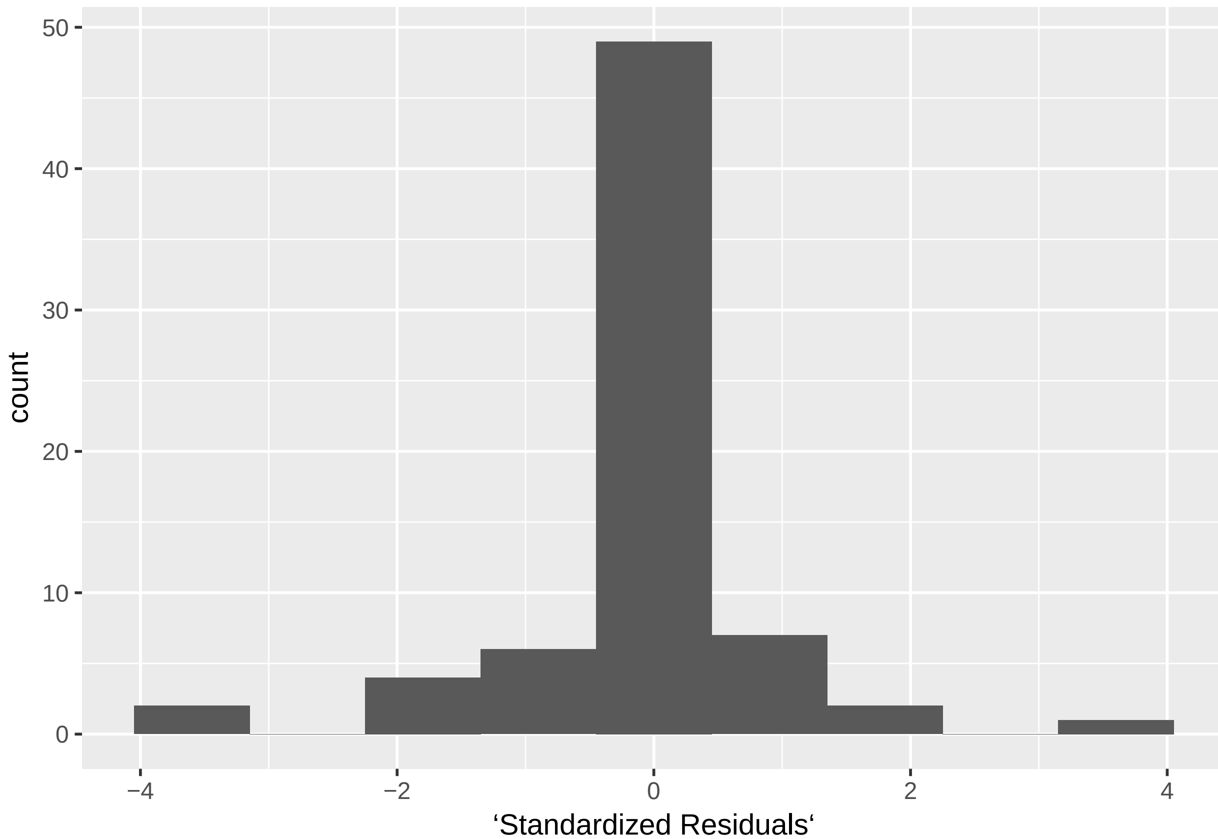


Table 12: Augmented Dickey-Fuller Test: `ARIMA_1_2_1$residuals`

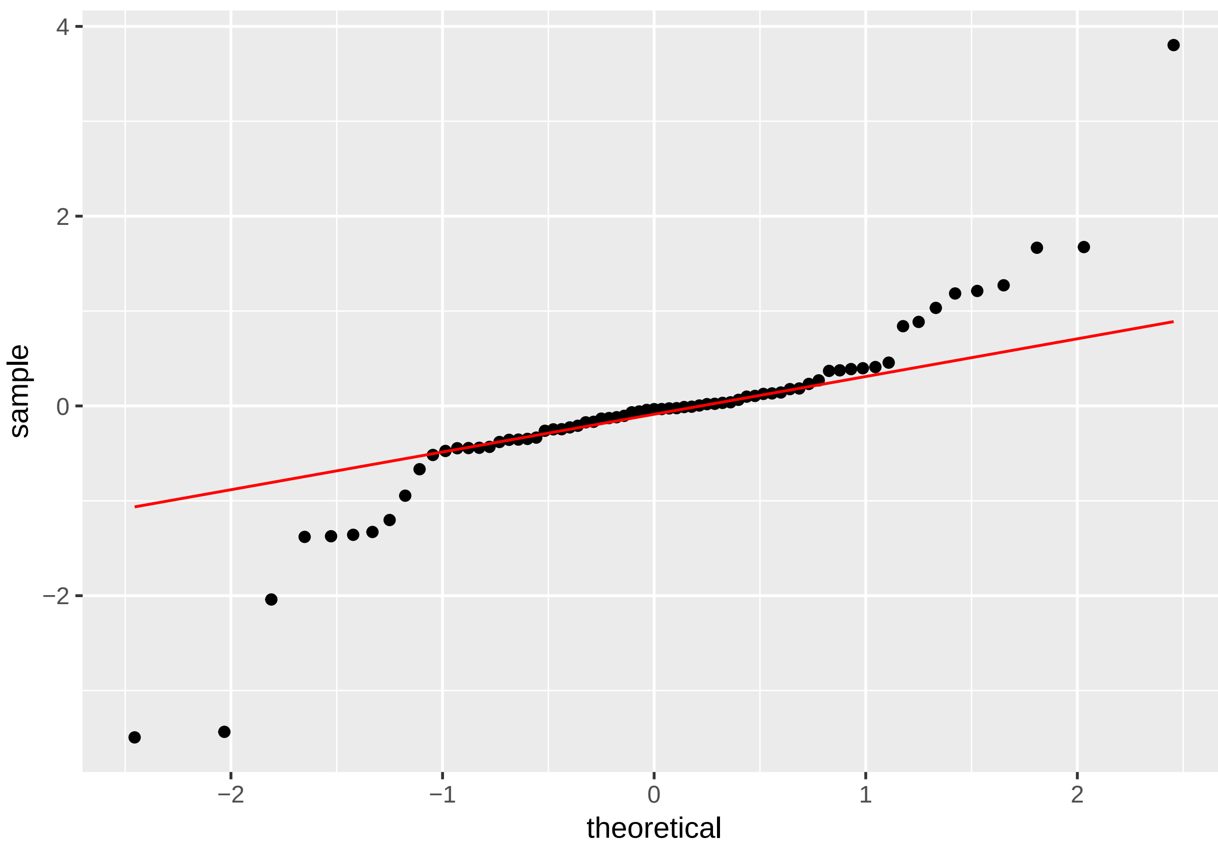
Test statistic	Lag order	P value	Alternative hypothesis
-2.488	4	0.3769	stationary



- **pvalue**: *0.31*
- **observed.runs**: *31*
- **expected.runs**: *35.65*
- **n1**: *41*
- **n2**: *30*
- **k**: *0*



50
40
30
20
10
0
count
−4
−2
0
2
4
'Standardized Residuals'



4
2
0
−2
sample
−2
−1
0
1
2
theoretical

Table 13: Shapiro-Wilk normality test: `ARIMA_1_2_1_Frame$`Standardized Residuals“

Test statistic	P value
0.8497	5.665e-07 * * *

- Independence (Runs Test): ✓
- Normality Test: ⊗
- ADF Test: ⊗

ARIMA (0 , 1 , 1) Assumptions Check:

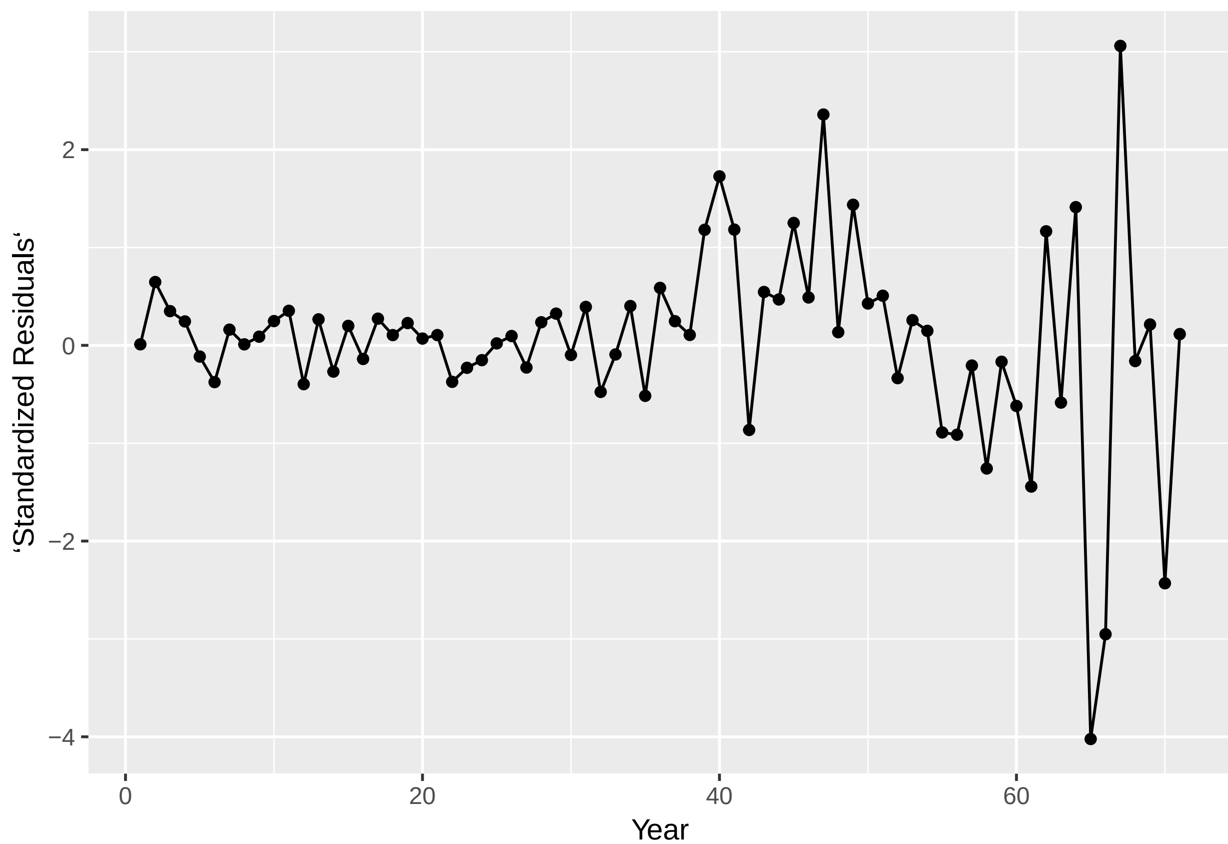
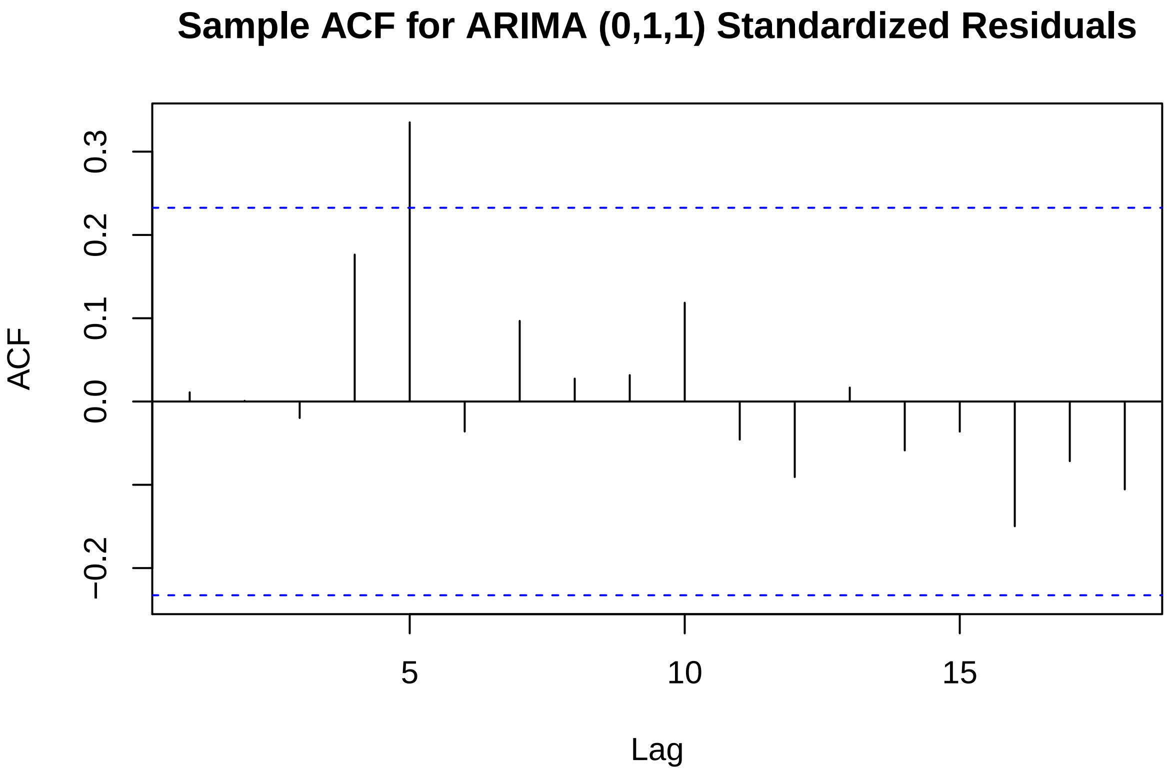
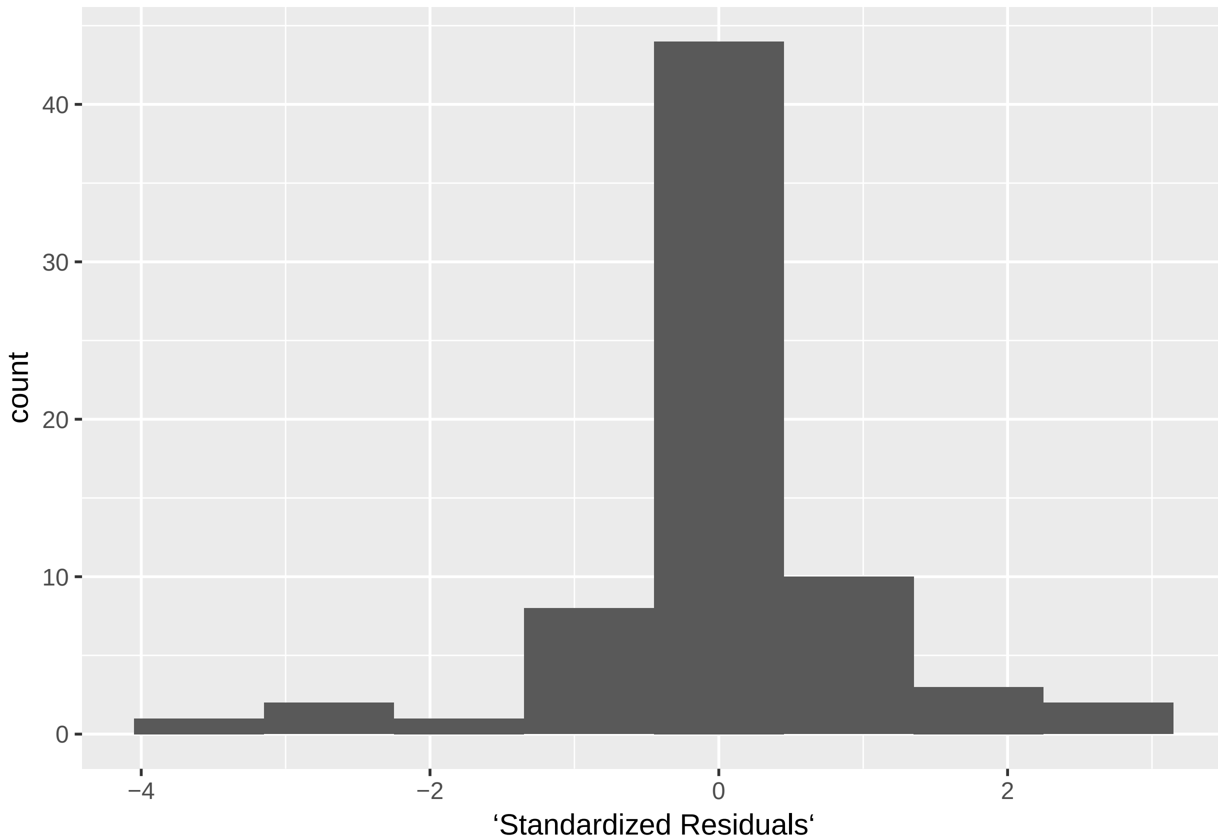


Table 14: Augmented Dickey-Fuller Test: `ARIMA_0_1_1$residuals`

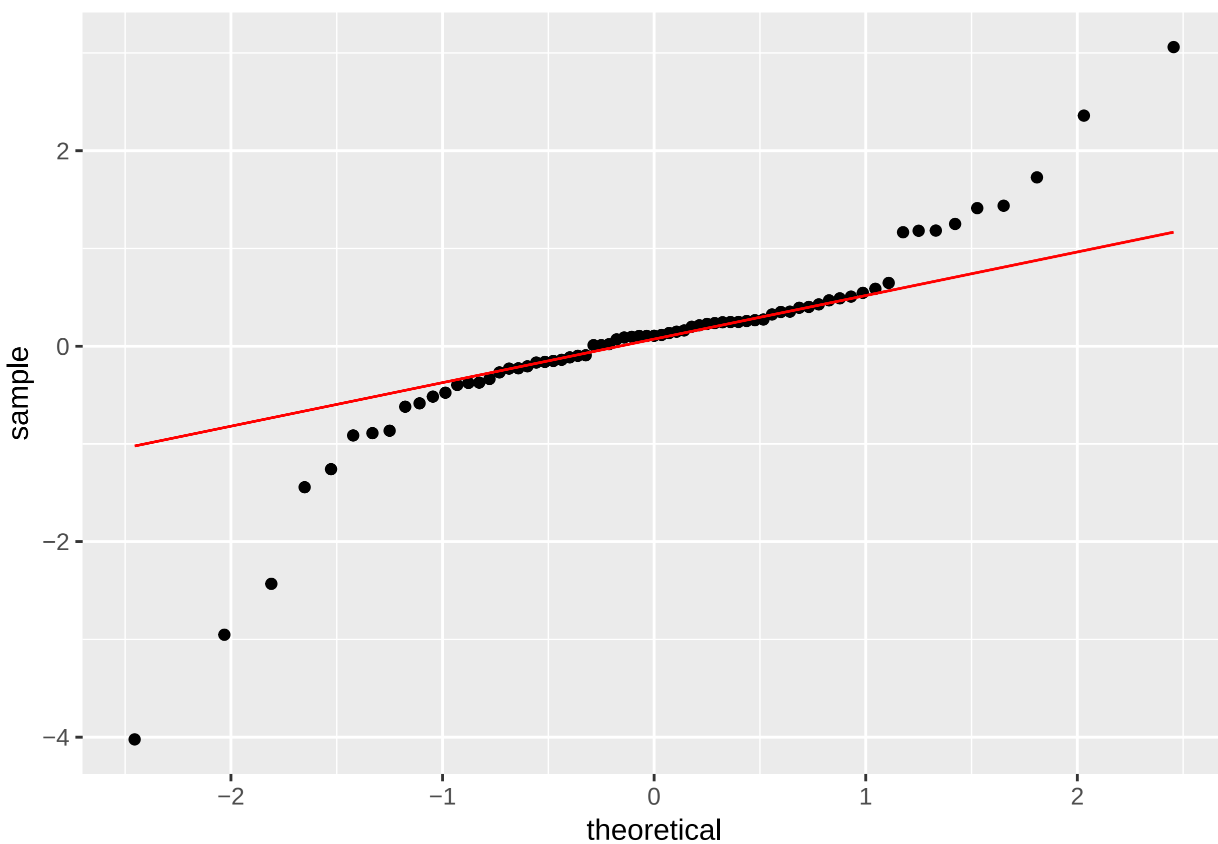
Test statistic	Lag order	P value	Alternative hypothesis
-1.985	4	0.5813	stationary



- **pvalue**: *0.807*
- **observed.runs**: *33*
- **expected.runs**: *34.46*
- **n1**: *27*
- **n2**: *44*
- **k**: *0*



count
40
30
20
10
0
−4
−2
0
2
‘Standardized Residuals‘



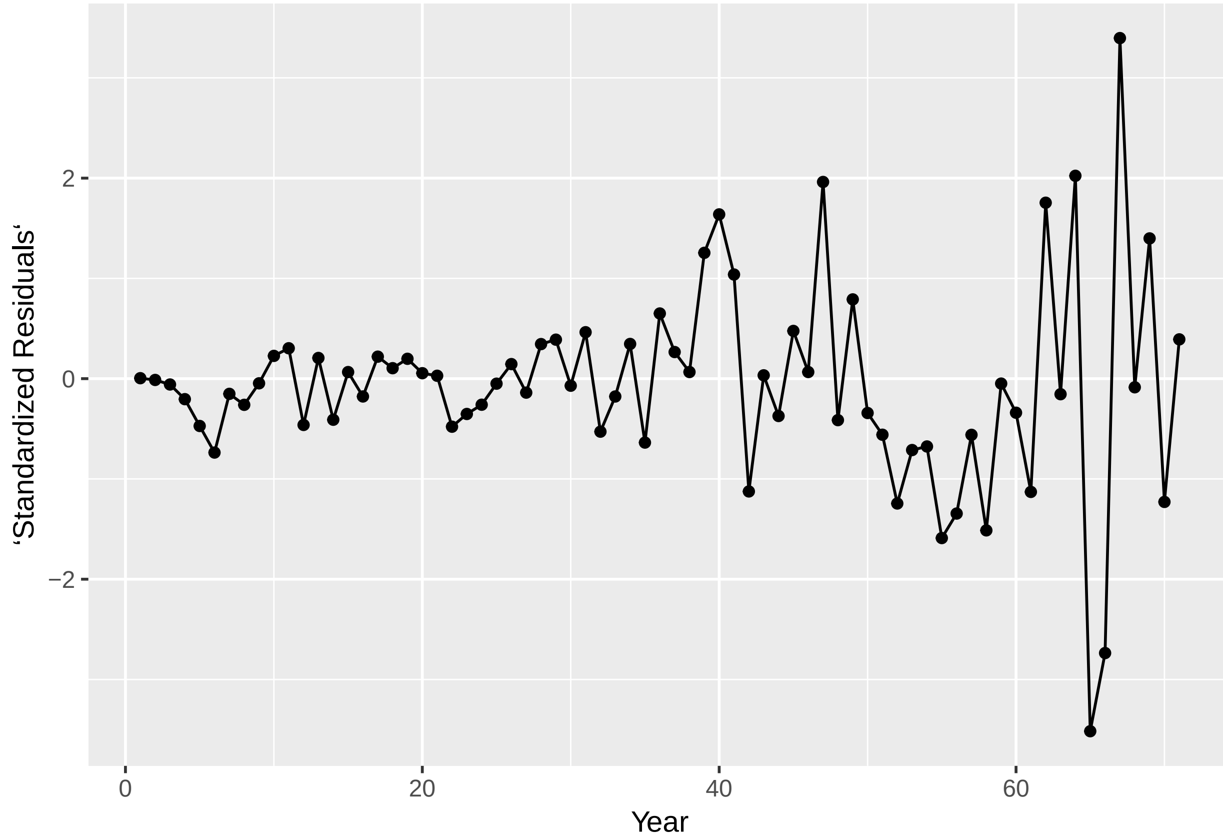
sample
2
0
−2
−4
−2
−1
0
1
2
theoretical

Table 15: Shapiro-Wilk normality test: `ARIMA_0_1_1_Frame$`Standardized Residuals“

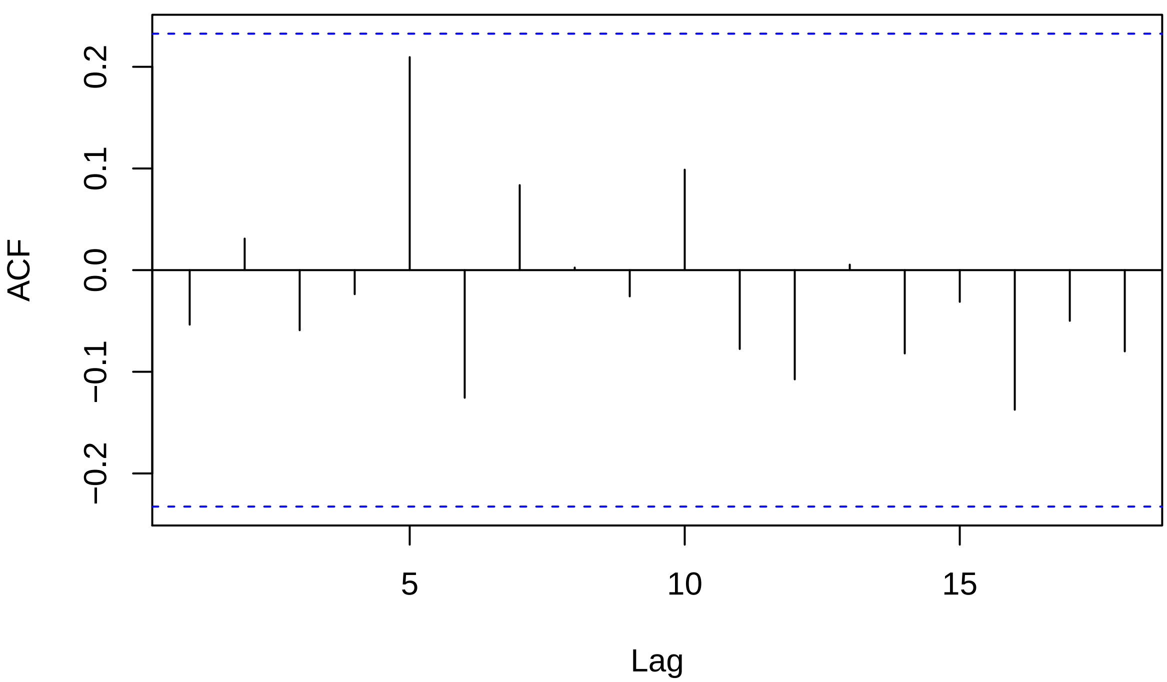
Test statistic	P value
0.8647	1.758e-06 * * *

- Independence (Runs Test): ✓
- Normality Test: ⊗
- ADF Test: ⊗

ARIMA (2 , 2 , 1) Assumptions Check:



Sample ACF for ARIMA (2,2,1) Standardized Residuals

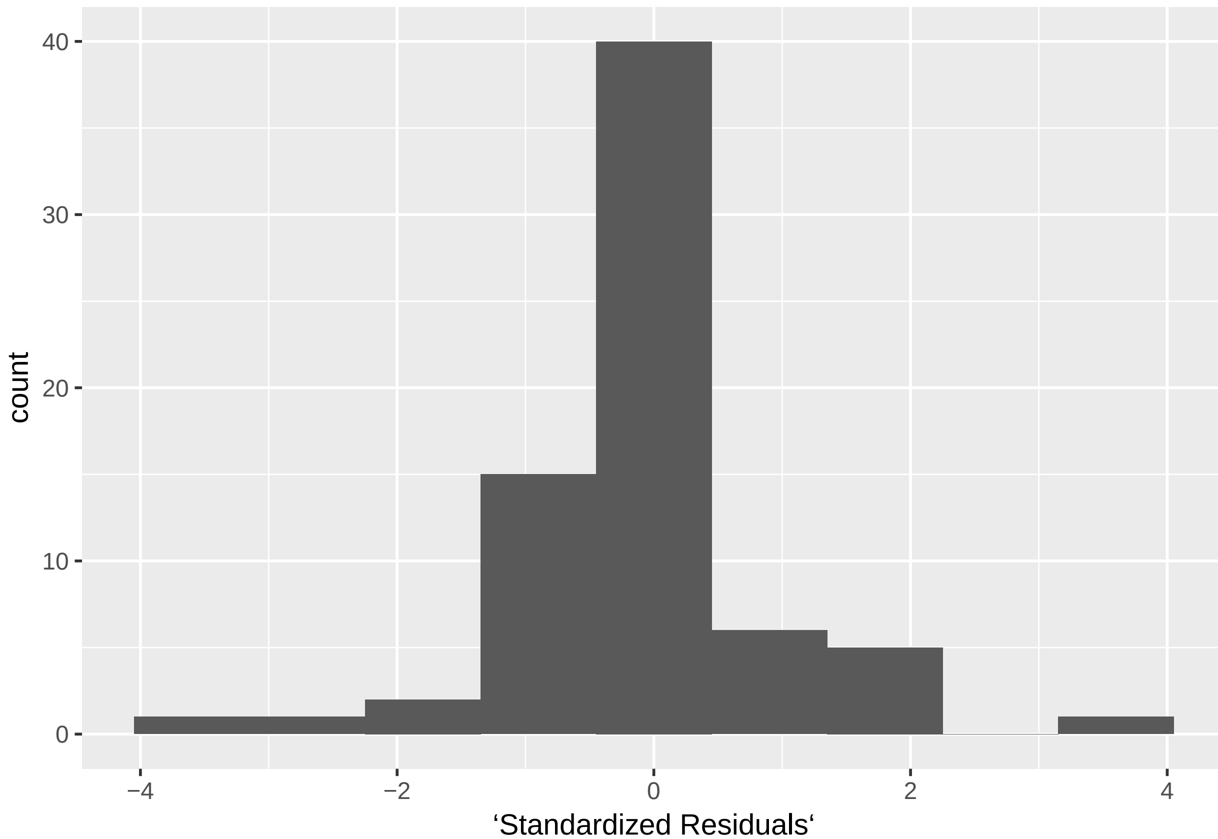


- **pvalue**: *0.916*

- **observed.runs**: *35*
- **expected.runs**: *35.93*
- **n1**: *40*
- **n2**: *31*
- **k**: *0*

Table 16: Augmented Dickey-Fuller Test: `ARIMA_2_2_1_Frame$`Standardized Residuals“

Test statistic	Lag order	P value	Alternative hypothesis
-2.602	4	0.3301	stationary


40
30
20
10
0
count
−4
−2
0
2
4
‘Standardized Residuals‘

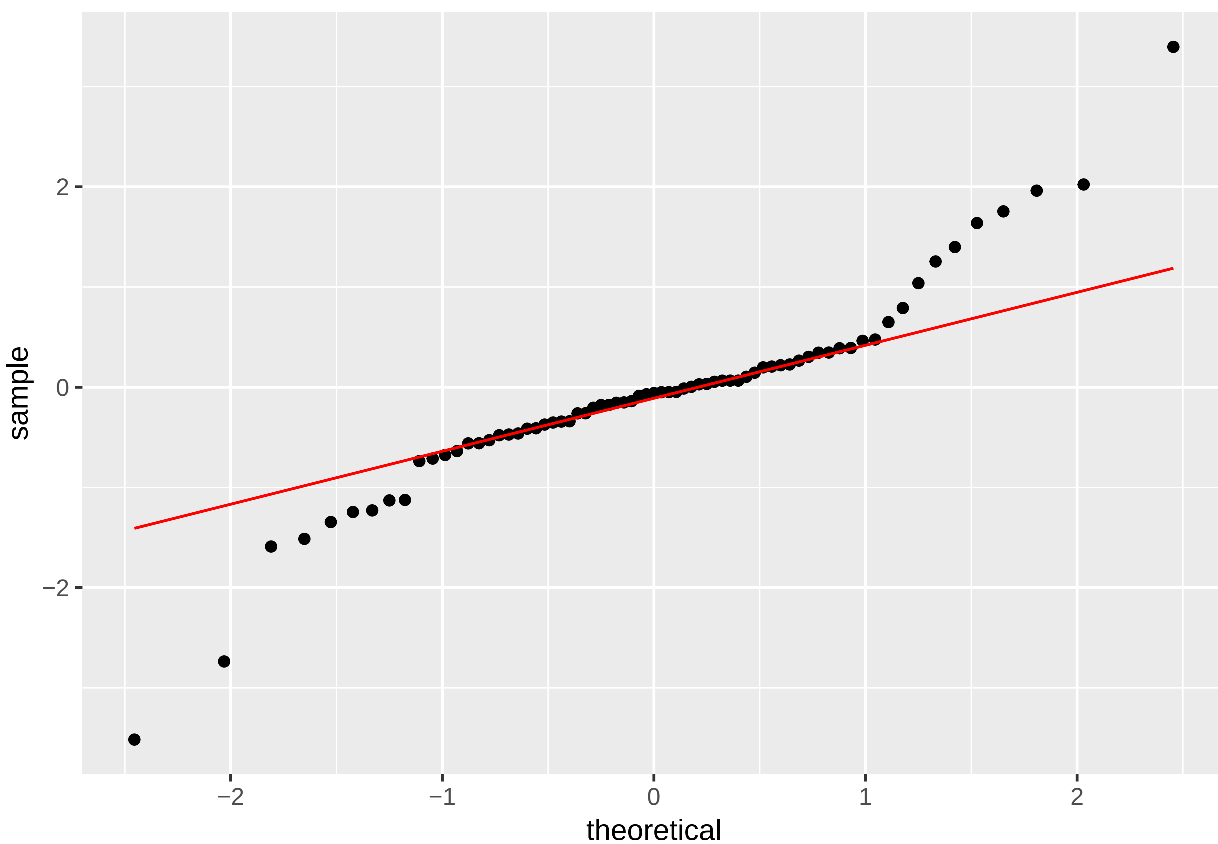

2
0
−2
sample
−2
−1
0
1
2
theoretical

Table 17: Shapiro-Wilk normality test: `ARIMA_2_2_1_Frame$`Standardized Residuals“

Test statistic	P value
0.9124	0.0001109 * * *

- Independence (Runs Test): ✓
- Normality Test: ⊗
- ADF Test: ⊗

Using ARIMA(2 , 2 , 0) OR Removing Points from the tails in attempt to fix Normality:

Just for the sake of it I am going to go ahead and use the ARIMA (2 , 2 , 0) model to predict the 5 years we removed. Down the road after removing some points from the tails we can compare the two models and how well they are at predicting.

Prediction with ARIMA(2 , 2 , 0):

```
Age_Group_20_24_Rev_1 = rev(Age_Group_20_24$Birth_Rate )

Year = 1940:2010

Age_Group_20_24_Rev_1 = cbind.data.frame( Age_Group_20_24_Rev_1 , Year)

colnames(Age_Group_20_24_Rev_1)[1] = "Birth_Rate"

ARIMA_2_2_0 =  arima( Age_Group_20_24_Rev_1$Birth_Rate , order = c( 2 , 2 , 0) , method = "ML")

pander( ARIMA_2_2_0)
```

Call: arima(x = Age_Group_20_24_Rev_1$Birth_Rate, order = c(2, 2, 0), method = “ML”)

Table 18: Coefficients

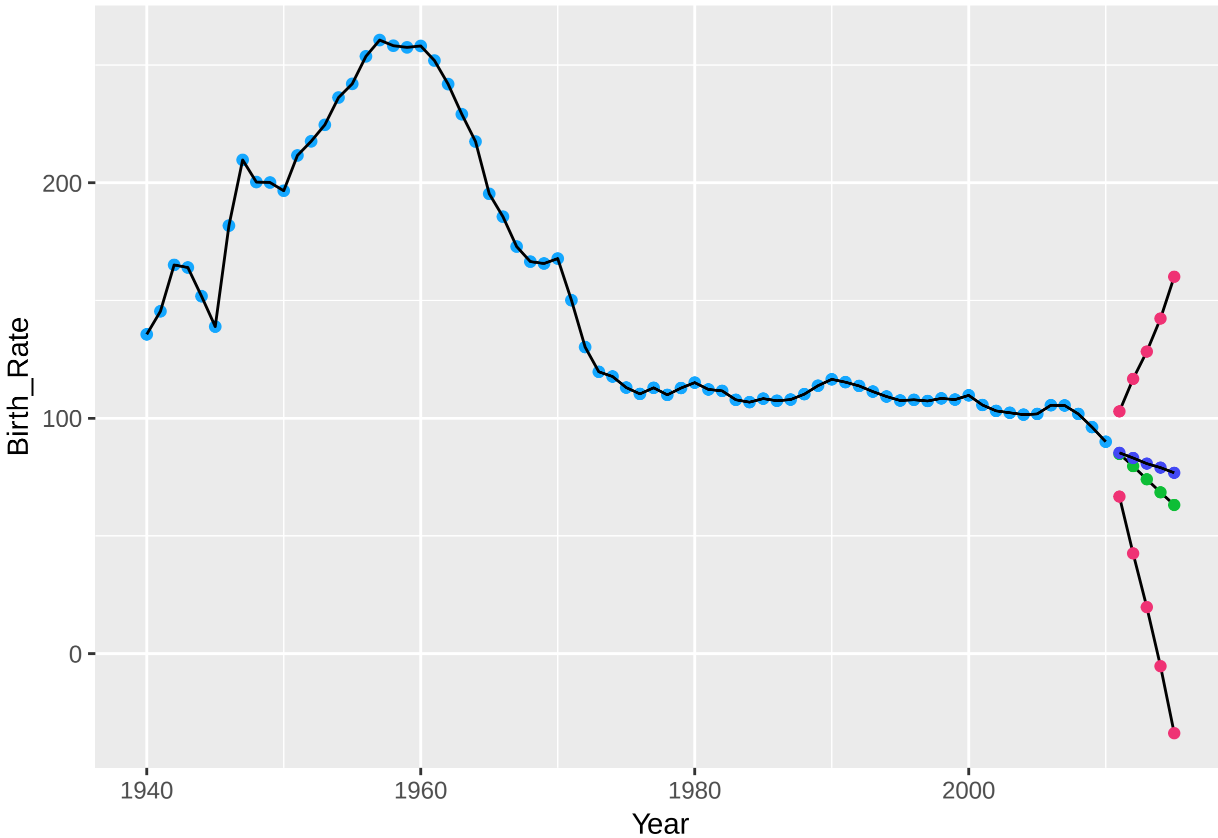
	ar1	ar2
	-0.2118	-0.4302
s.e.	0.1088	0.1102

sigma^2 estimated as 85.23: log likelihood = -251.49, aic = 506.98

```
## Time Series:
## Start = 72
## End = 76
## Frequency = 1
## [1] 84.787 79.624 74.025 68.498 63.142
```

```
## Time Series:
## Start = 72
```

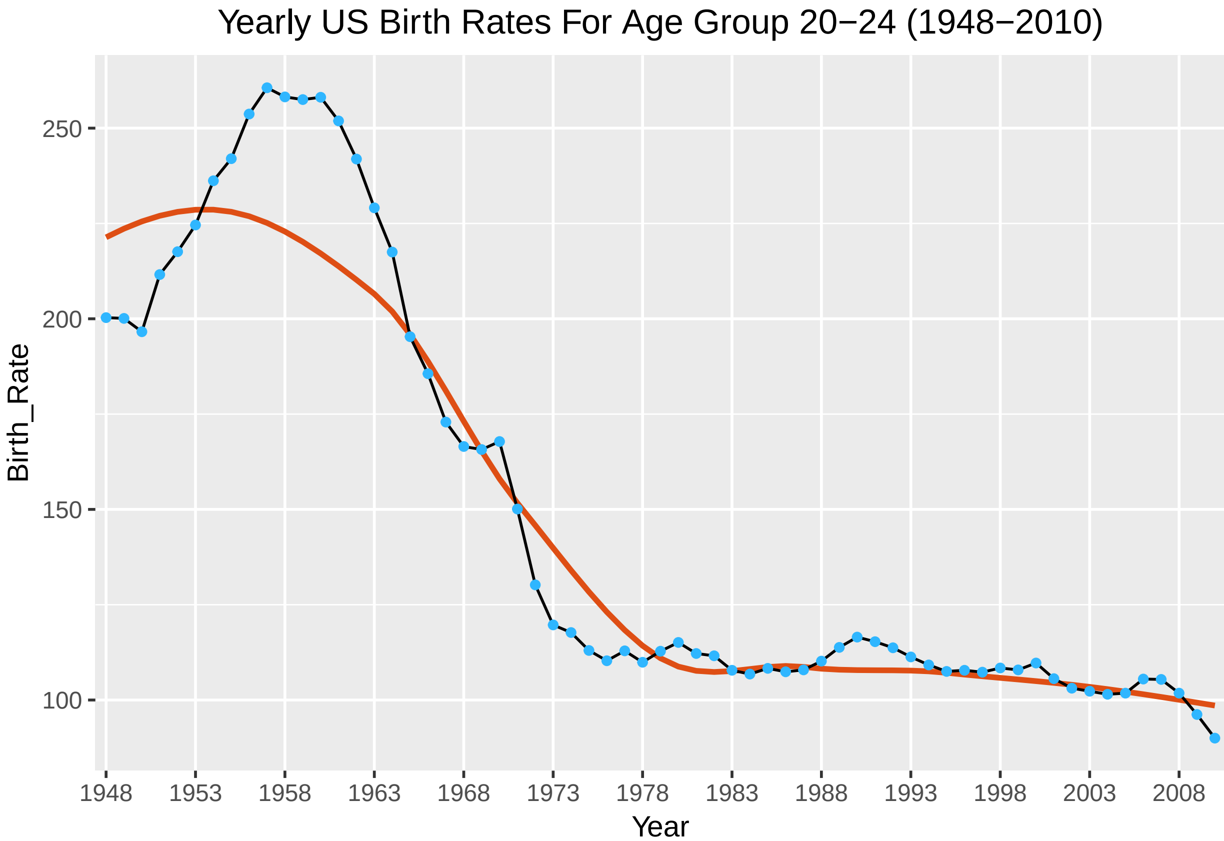
```
## End = 76
## Frequency = 1
## [1]  9.232 18.915 27.694 37.674 49.462
```



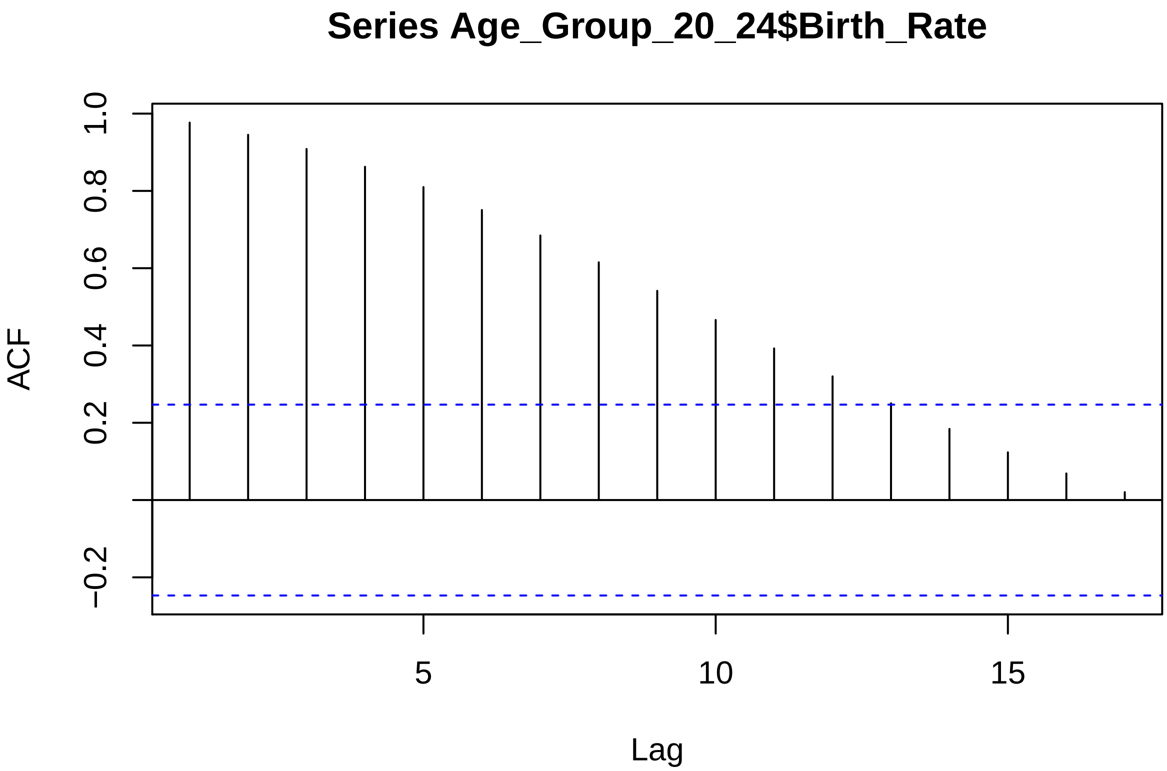
6.965

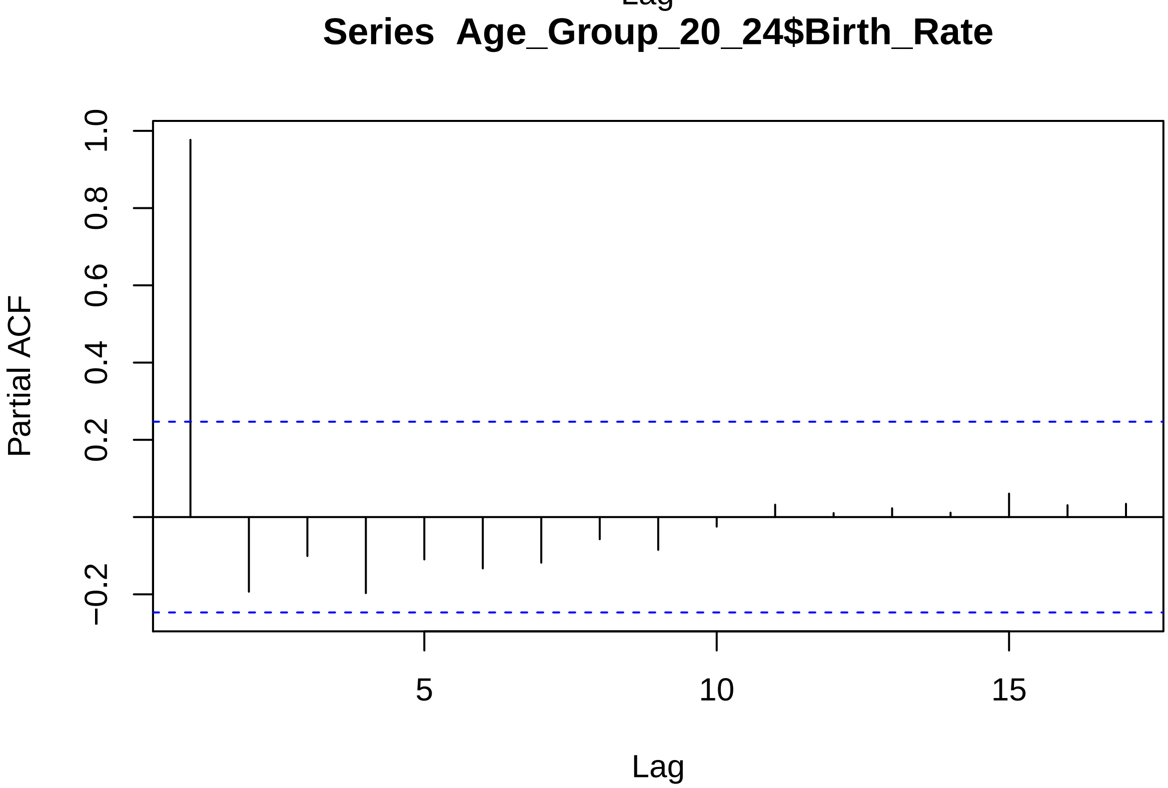
Maybe the next best thing to do is remove 8 values from the start of the data to remove the very large spike of in data attributed to the baby boom after the end of the second world war. After removing these points then we can start the model selection process from the beginning.

Plotting the Data:



ACF , PACF , & EACF Plots:






```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x x x x x x x x x x x  x  o  o
```

```
## 1 x x x x x x x o o o o  o  o  o
## 2 o o o o o o o o o o o  o  o  o
## 3 x o o o o o o o o o o  o  o  o
## 4 o o o o o o o o o o o  o  o  o
## 5 o o x o o o o o o o o  o  o  o
## 6 x o x o o o o o o o o  o  o  o
## 7 o o o o o o o o o o o  o  o  o
```

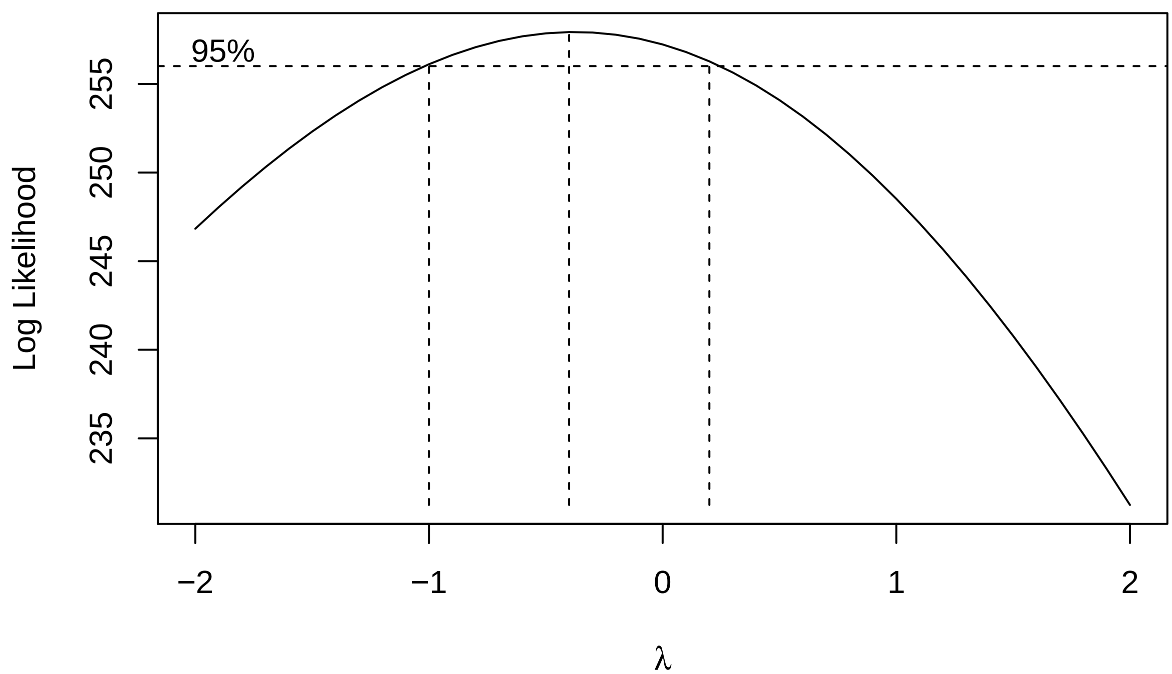
Stationarity Check:

Table 19: Augmented Dickey-Fuller Test: `Age_Group_20_24$Birth_Rate`

Test statistic	Lag order	P value	Alternative hypothesis
-2.612	3	0.3273	stationary

Interpretation: The data is not stationary so we should proceed with applying a Box Cox transformation & taking differences.

```
## Warning in arima0(x, order = c(i, 0L, 0L), include.mean = demean): possible
## convergence problem: optim gave code = 1
```



Applying Transformations:

Possible Transformations:

- Log
- Inverse Square Root
- Reciprocal

Quick check to see if Statonary:

Table 20: Augmented Dickey-Fuller Test: `Log_Age_Group_20_24$Birth_Rate`

Test statistic	Lag order	P value	Alternative hypothesis
-2.244	3	0.4762	stationary

Table 21: Augmented Dickey-Fuller Test: `Inverse_Root_Age_Group_20_24$Birth_Rate`

Test statistic	Lag order	P value	Alternative hypothesis
-2.121	3	0.526	stationary

Table 22: Augmented Dickey-Fuller Test: `Reciprocal_Age_Group_20_24$Birth_Rate`

Test statistic	Lag order	P value	Alternative hypothesis
-2.033	3	0.5616	stationary

Interpretation: None of the transformations make the model stationary so lets apply differencing and jump ahead to applying second differencing.

Applying Differences:

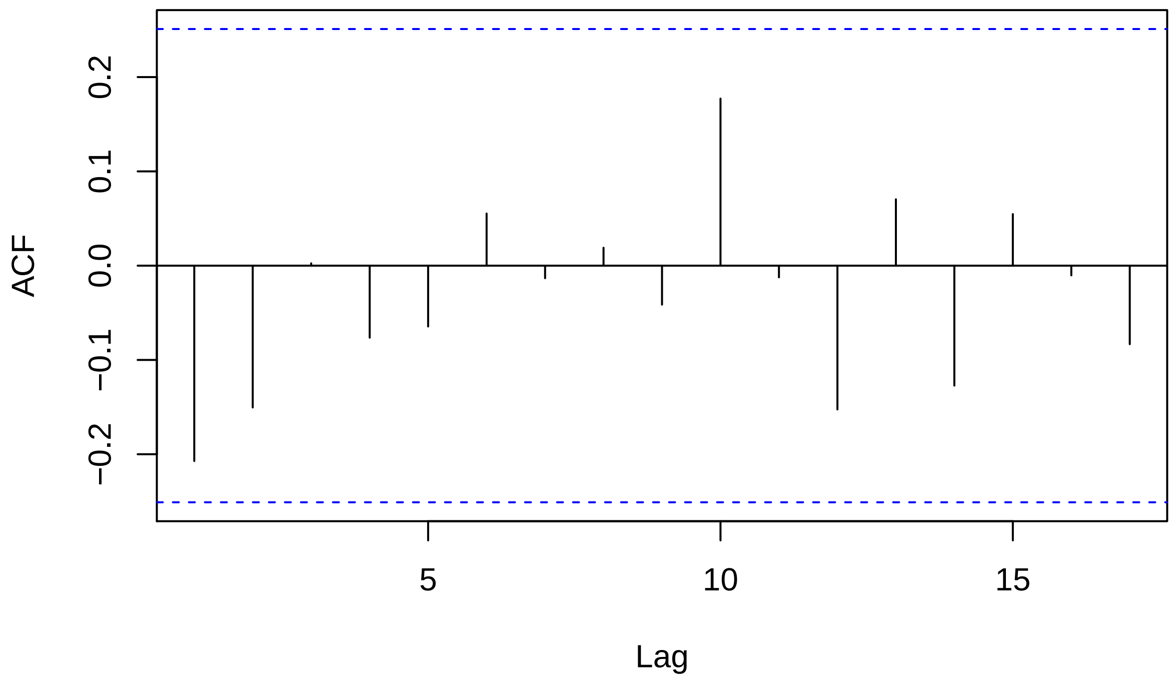
Log Transformation 2nd Difference:

```
## Warning in adf.test(Log_Age_Group_20_24_diff_Frame$Birth_Rate): p-value
## smaller than printed p-value
```

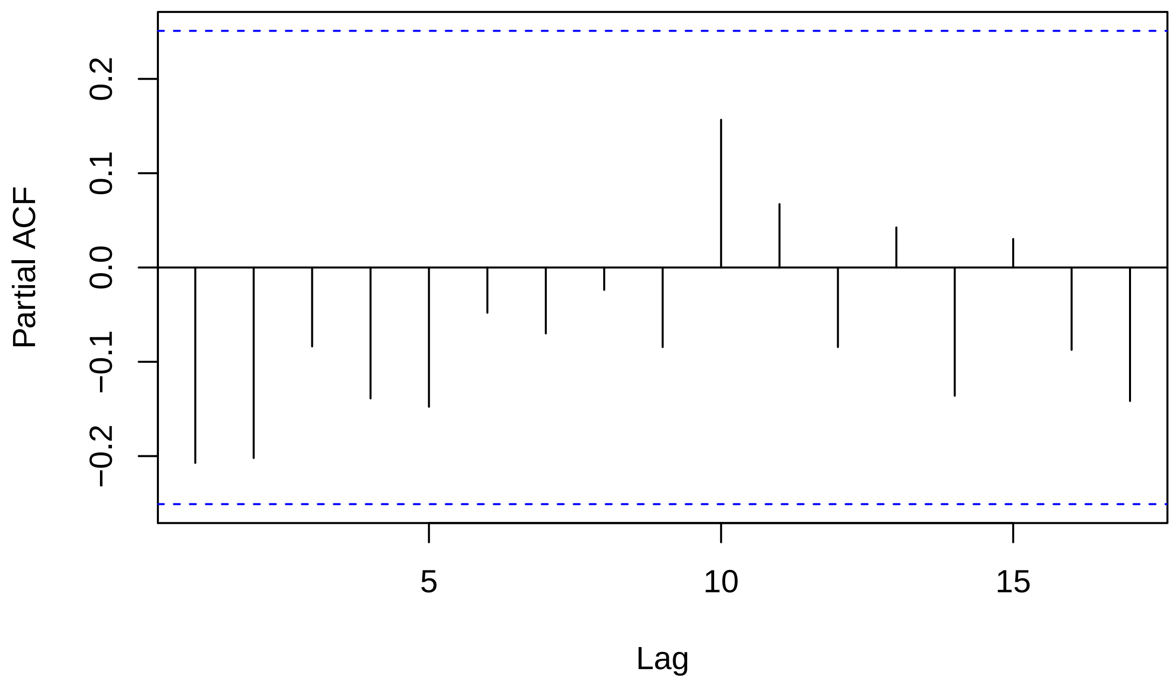
Table 23: Augmented Dickey-Fuller Test: `Log_Age_Group_20_24_diff_Frame$Birth_Rate`

Test statistic	Lag order	P value	Alternative hypothesis
-5.428	3	0.01 * *	stationary

Series Log_Age_Group_20_24_diff_Frame$Birth_Rate



Series Log_Age_Group_20_24_diff_Frame$Birth_Rate




```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 o o o o o o o o o o o  o  o  o
## 1 x o o o o o o o o o o  o  o  o
## 2 x x o o o o o o o o o  o  o  o
```

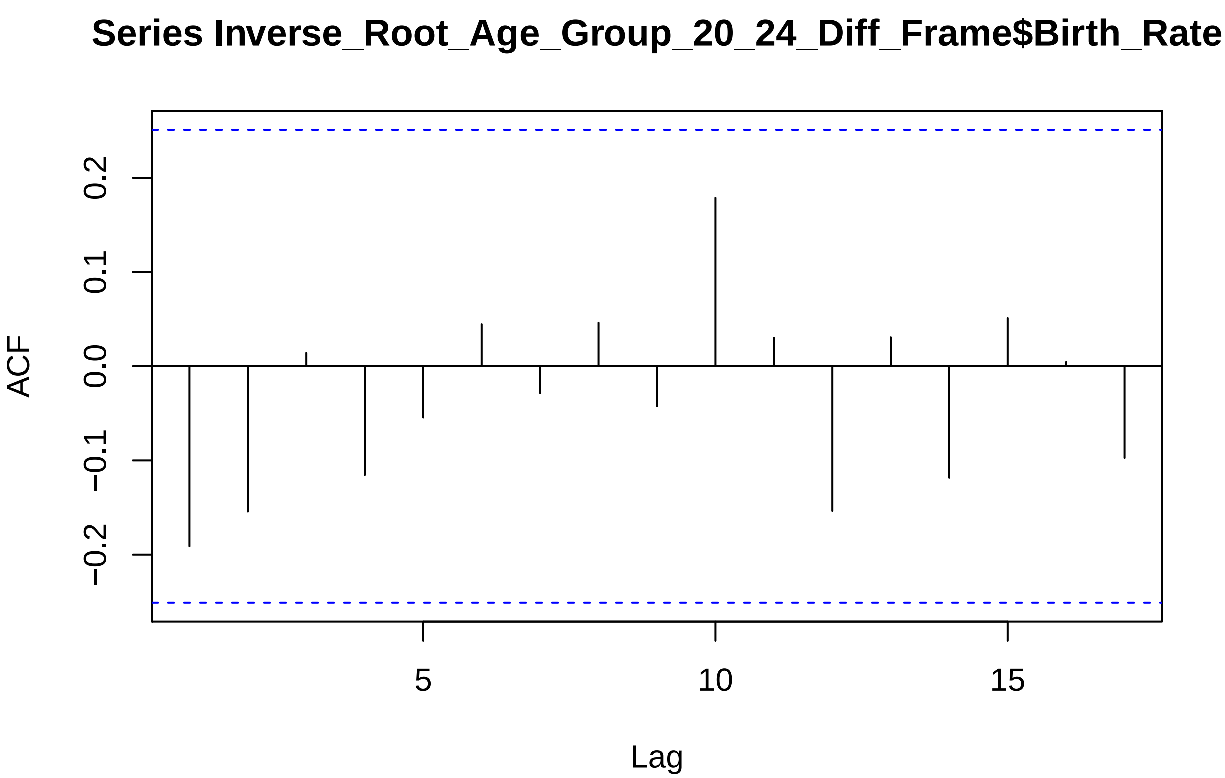
```
## 3 x o x o o o o o o o o  o  o  o
## 4 x o x o o o o o o o o  o  o  o
## 5 x x o o o o o o o o o  o  o  o
## 6 x x o o o o o o o o o  o  o  o
## 7 o x o o o o o o o o o  o  o  o
```

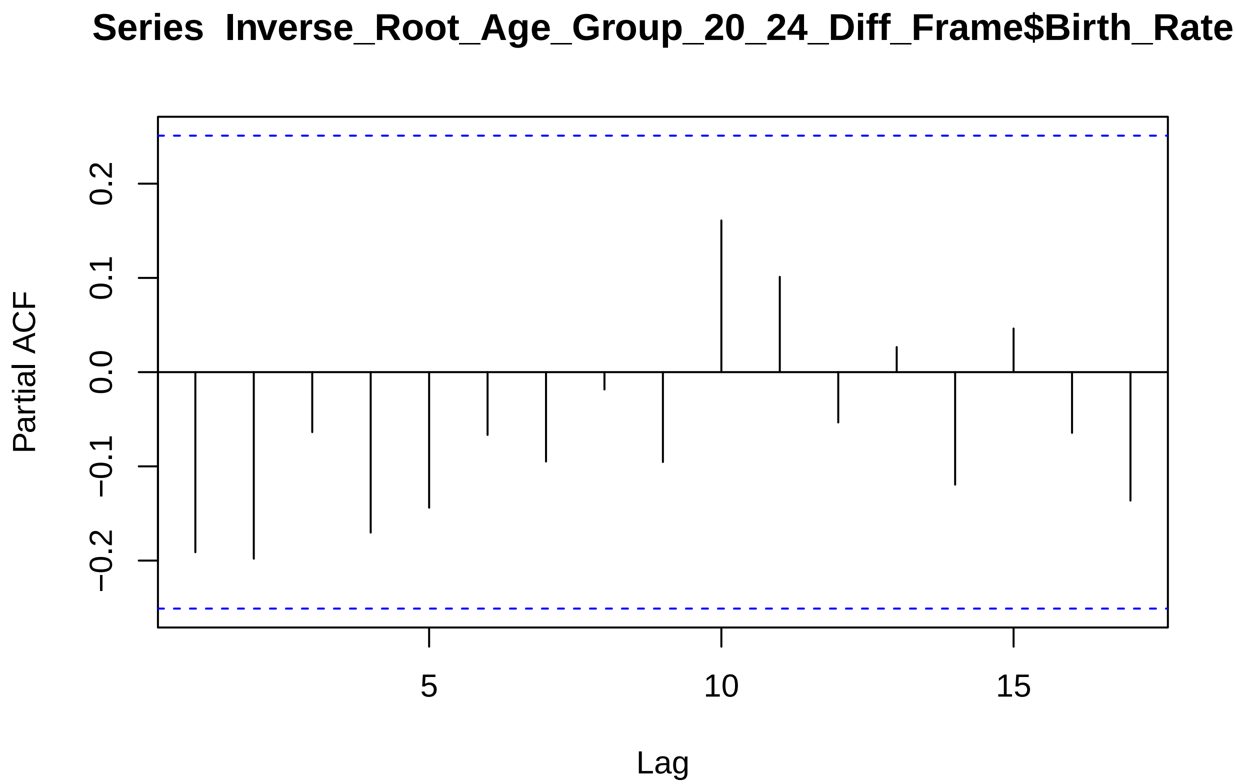
Inverse Root 2nd Difference:

```
## Warning in adf.test(Inverse_Root_Age_Group_20_24_Diff_Frame$Birth_Rate): p-
## value smaller than printed p-value
```

Table 24: Augmented Dickey-Fuller Test: `Inverse_Root_Age_Group_20_24_Diff_Frame$Birth_Rate`

Test statistic	Lag order	P value	Alternative hypothesis
-5.763	3	0.01 * *	stationary






```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 o o o o o o o o o o o  o  o  o
## 1 x o o o o o o o o o o  o  o  o
## 2 x x o o o o o o o o o  o  o  o
## 3 x x x o o o o o o o o  o  o  o
## 4 x o x o o o o o o o o  o  o  o
## 5 x x o o o o o o o o o  o  o  o
## 6 x x o o o o o o o o o  o  o  o
## 7 o x o o o o o o o o o  o  o  o
```

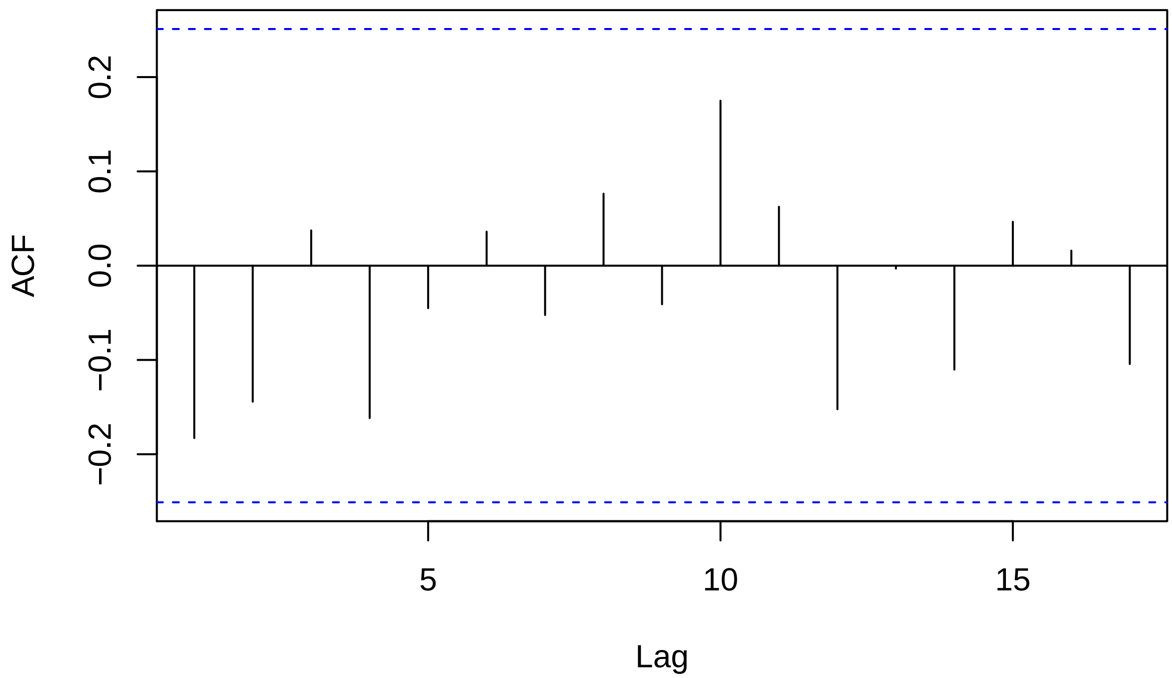
Reciprocal 2nd Difference:

```
## Warning in adf.test(Reciprocal_Age_Group_20_24_Diff_Frame$Birth_Rate): p-
## value smaller than printed p-value
```

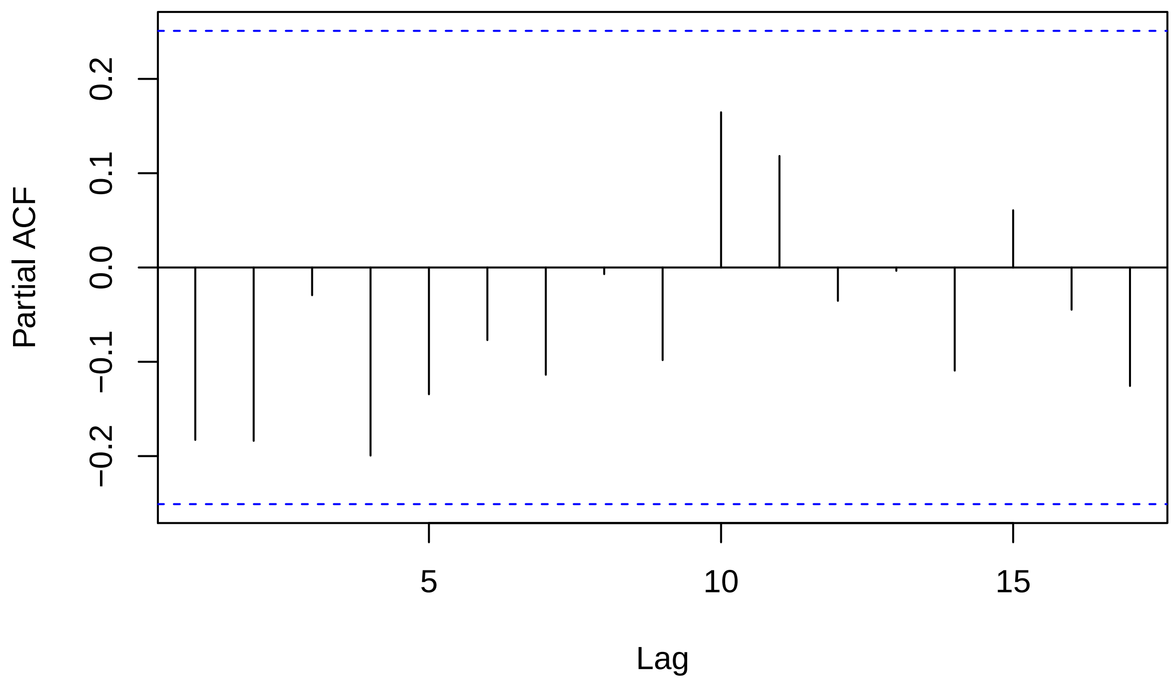
Table 25: Augmented Dickey-Fuller Test: `Reciprocal_Age_Group_20_24_Diff_Frame$Birth_Rate`

Test statistic	Lag order	P value	Alternative hypothesis
-6.053	3	0.01 * *	stationary

Series Reciprocal_Age_Group_20_24_Diff_Frame$Birth_Rate



Series Reciprocal_Age_Group_20_24_Diff_Frame$Birth_Rate




```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 o o o o o o o o o o o  o  o  o
## 1 x o o o o o o o o o o  o  o  o
## 2 o x o o o o o o o o o  o  o  o
```

```
## 3 o x x o o o o o o o o o  o  o  o
## 4 x o x o o o o o o o o o  o  o  o
## 5 x o o o o o o o o o o o  o  o  o
## 6 x x x o o o o o o o o o  o  o  o
## 7 o x o o o o o o o o o o  o  o  o
```

Lets just quickly try to fit one model to the untransformed data to see what happens.

ARIMA(0 , 2 , 1)

Call: arima(x = Reved_Age_Group_20_24$reved_Birth_Rates, order = c(0, 2, 1), method = "ML")

Table 26: Coefficients

	ma1
	-0.3689
s.e.	0.1449

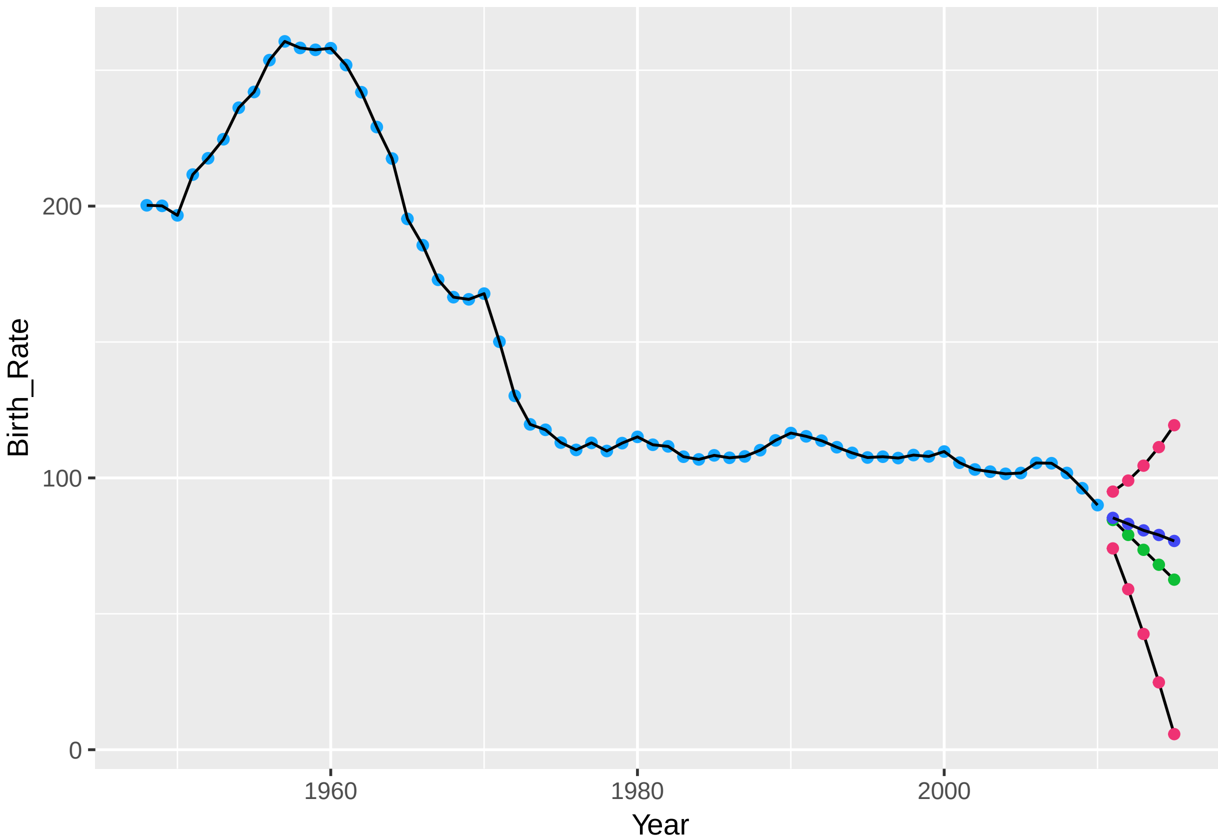
sigma^2 estimated as 28.43: log likelihood = -188.72, aic = 379.45

95 Percent CI for Parameters:

- -0.3689 $\pm$ 0.2840 $\longrightarrow$ ✓

```
## Time Series:
## Start = 64
## End = 68
## Frequency = 1
## [1] 84.513 79.026 73.539 68.052 62.565
```

```
## Time Series:
## Start = 64
## End = 68
## Frequency = 1
## [1]  5.332 10.201 15.797 22.080 28.994
```



Lets try Overfitting Now!:

Call: arima(x = Reved_Age_Group_20_24$reved_Birth_Rates, order = c(0, 2, 2), method = "ML")

Table 27: Coefficients

	ma1	ma2
	-0.3329	-0.1721
s.e.	0.1268	0.141

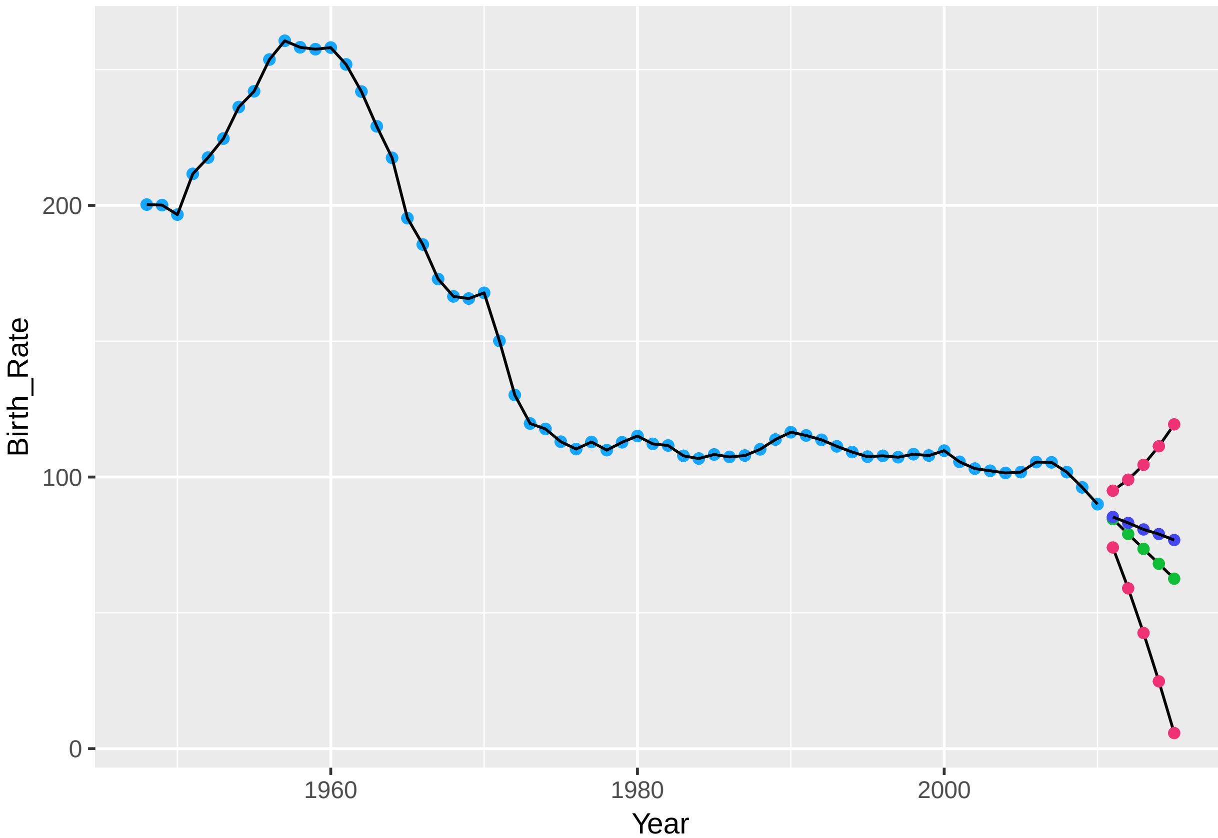
sigma^2 estimated as 27.74: log likelihood = -188.02, aic = 380.04

95 Percent CI for Parameters:

- -0.3329 $\pm$ 0.2485 $\longrightarrow$ ✓
- -0.1721 $\pm$ 0.2763 $\longrightarrow$ $\otimes$

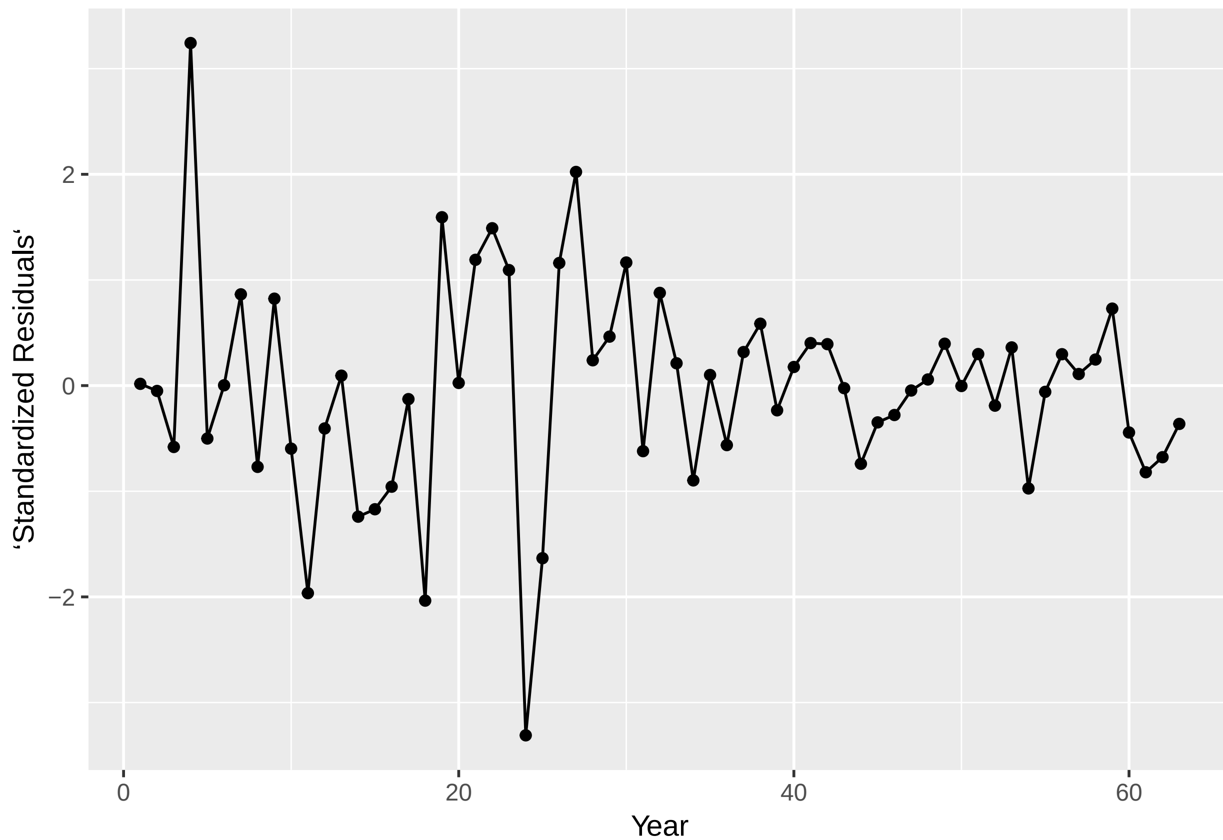
```
## Time Series:
## Start = 64
## End = 68
## Frequency = 1
## [1] 84.513 79.026 73.539 68.052 62.565
```

```
## Time Series:
## Start = 64
## End = 68
## Frequency = 1
## [1]  5.332 10.201 15.797 22.080 28.994
```



After overfitting the model the second MA term is not significant and think this is why the two graphs are very similar to each other.

Checking Assumptions of ARIMA(0,2,1)

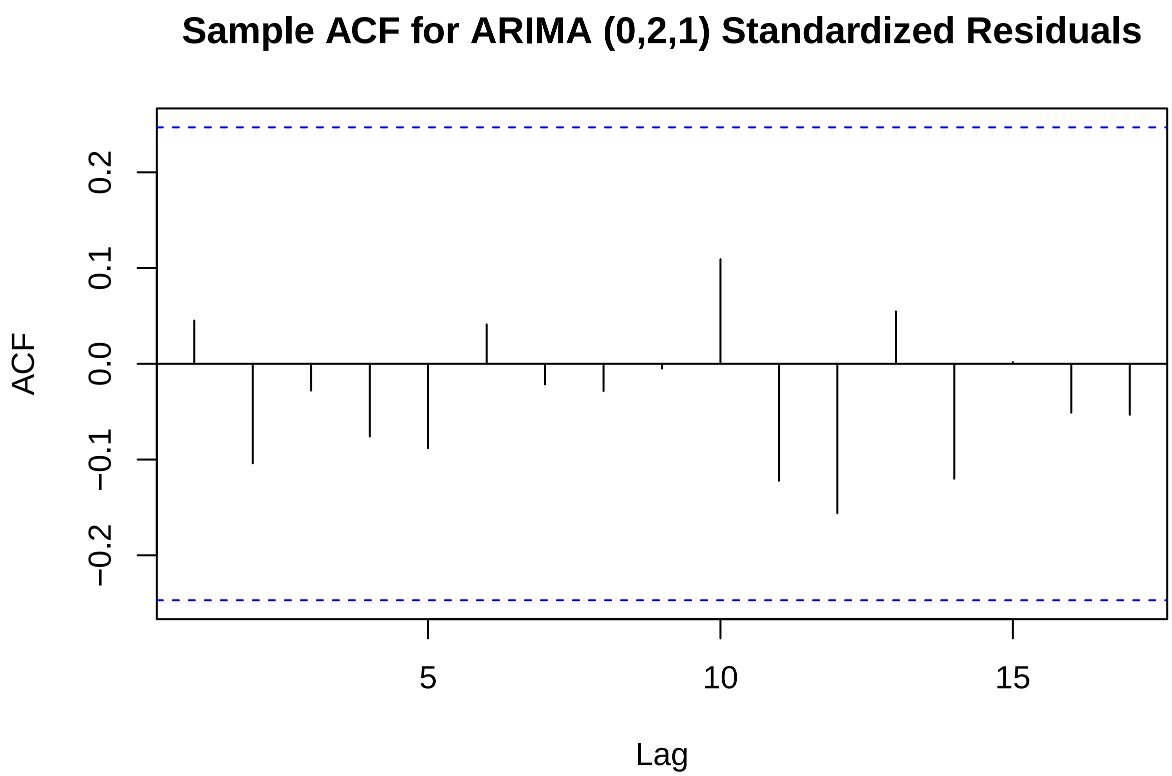


- **pvalue**: *0.614*
- **observed.runs**: *30*
- **expected.runs**: *32.49*
- **n1**: *31*
- **n2**: *32*
- **k**: *0*

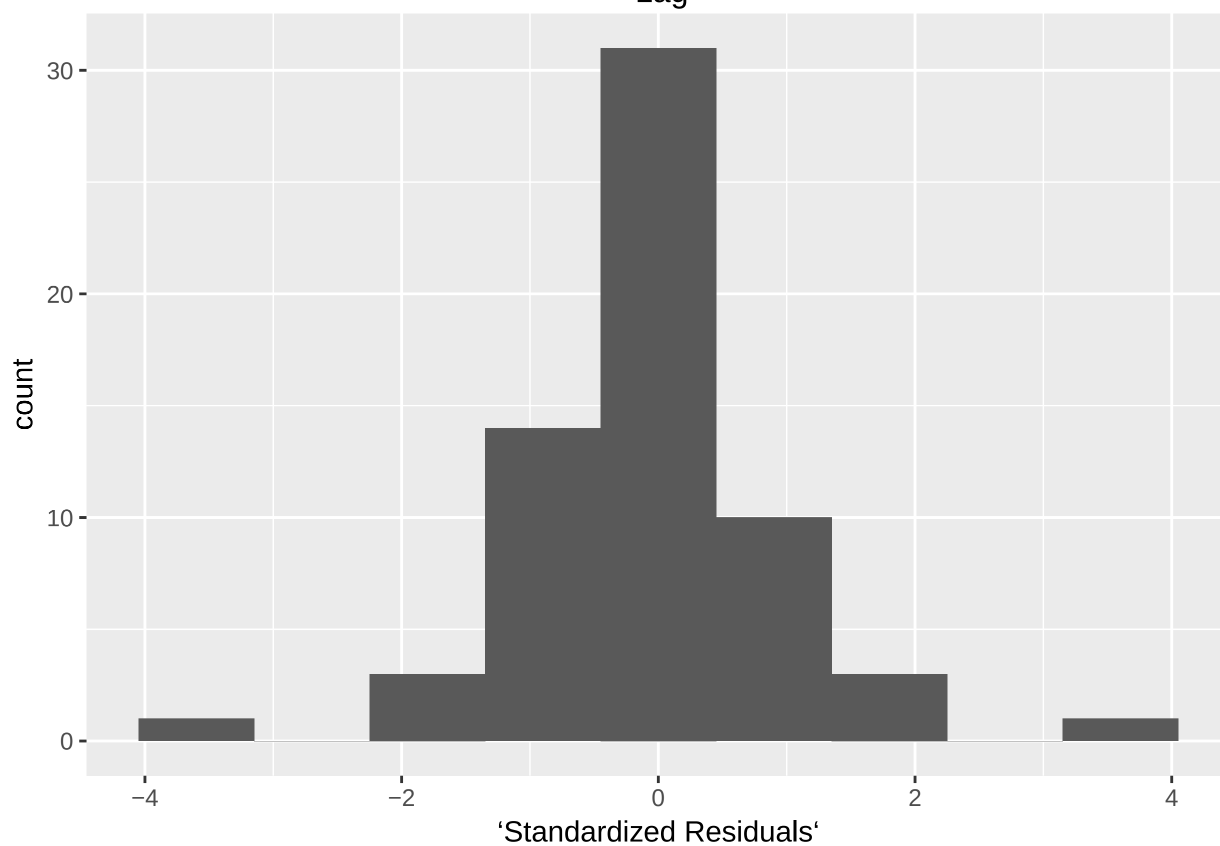
```
## Warning in adf.test(ARIMA_0_2_1$residuals): p-value smaller than printed p-
## value
```

Table 28: Augmented Dickey-Fuller Test: `ARIMA_0_2_1$residuals`

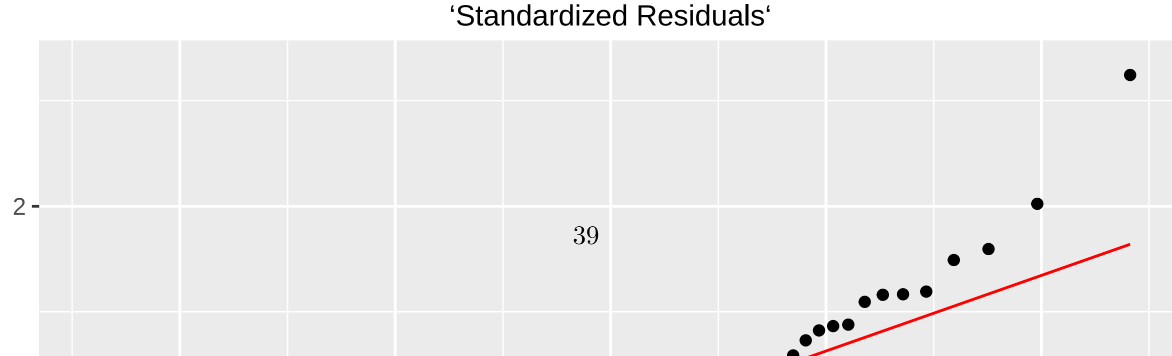
Test statistic	Lag order	P value	Alternative hypothesis
-4.466	3	0.01 * *	stationary



Sample ACF for ARIMA (0,2,1) Standardized Residuals
ACF
0.2
0.1
0.0
−0.1
−0.2
5
10
15
Lag



count
30
20
10
0
−4
−2
0
2
4
‘Standardized Residuals‘



2

Table 29: Shapiro-Wilk normality test: `ARIMA_0_2_1_Frame$`Standardized Residuals“

Test statistic	P value
0.9588	0.03394 *

```r
MMSE_Forecast_Error = Actual_Values$Birth_Rate - Predicted_Values$Birth_Rate

pander(mean(MMSE_Forecast_Error))
```

7.441