

# Assignment 2

*Shivakanth Thudi*

*11/16/2016*

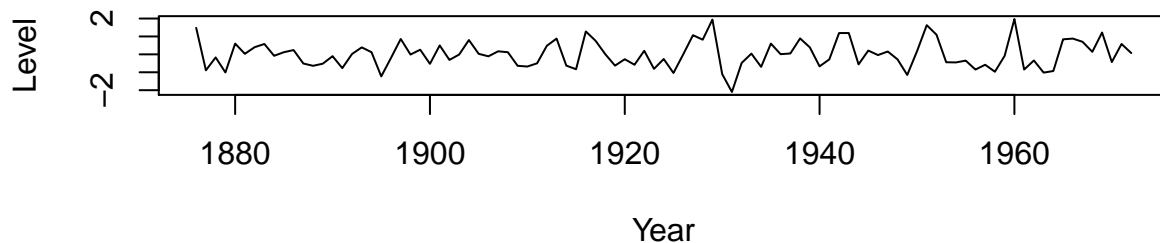
## Problem 1

### Part (a) - Picking $d$ as 1

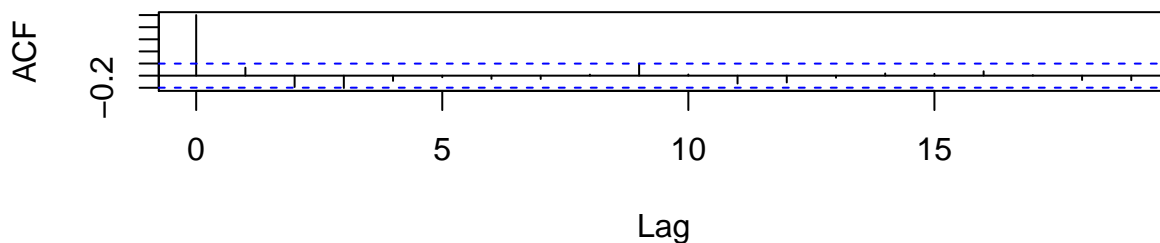
We look at the original LakeHuron data and see that it is not stationary. The ACF plot indicates slow decay (as opposed to rapid decay), a plot of the time series does not look flat, and an Augmented Dickey-Fuller Test indicates that there is not enough evidence to reject the null hypothesis that the times series is not stationary.

After differencing once, we find that the transformed time series is now stationary. In particular, the once differenced plot of the time series looks flat, the ACF plot indicates rapid decay, and the Augmented Dickey Fuller test implies that we can now reject the null hypothesis of non-stationarity, and instead conclude, that the transformed time series is now stationary. Hence we pick the **order of the ordinary differencing as  $d = 1$** .

### Differenced Annual Measurements



### Once Differenced (Ordinary) Measurements



Warning in `adf.test(dlake_huron)`: p-value smaller than printed p-value

Augmented Dickey-Fuller Test

```
data: dlake_huron
Dickey-Fuller = -5.4687, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

### Part(b) - Fitting an AR(1) model with MLE on the once-differenced time series

```
m1 <- arima(x = LakeHuron, order = c(1, 1, 0), method = "ML")  
# summary(m1)
```

Model	Error Variance, $\sigma^2$	Log likelihood	AIC
AR(1)	0.5452092	-108.2269995	220.453999

### Part(c) - Fitting an AR(2) model with MLE on the once-differenced time series

```
m2 <- arima(x = LakeHuron, order = c(2, 1, 0), method = "ML")  
# summary(m2)
```

Model	Error Variance, $\sigma^2$	Log likelihood	AIC
AR(2)	0.5188036	-105.8711095	217.742219

### Part (d) - Likelihood ratio test comparing the AR(1) and AR(2) models

We choose as our null model the simpler one, with fewer parameters - the AR(1) model. Our alternative model is the one with more parameters - the AR(2) model.

- **Null Hypothesis:** The null model AR(1) and the alternative model AR(2) fit equally well
- **Alternative Hypothesis:** The alternative model AR(2) fits better than the null model AR(1)

The test statistic is:

$$D = -2\log\left[\frac{L(\text{nullmodel})}{L(\text{altmodel})}\right] = -2[l(\text{nullmodel}) - l(\text{altmodel})] \quad (1)$$

Here, we have:

$$l(\text{nullmodel}) = -108.2269995 \quad (2)$$

$$l(\text{altmodel}) = -105.8711095 \quad (3)$$

So the statistic D is:

$$D = -2[l(\text{nullmodel}) - l(\text{alternativemodel})] = -2[(-108.2269995) - (-105.8711095)] = 4.71178 \quad (4)$$

and

$$pval = 1 - pchisq(D, 1) = 0.0299566 \quad (5)$$

The test statistic is 4.7118 and the corresponding p-value is 0.03. Hence, at the  $\alpha = 0.05$  level, we can **reject the null hypothesis that the null model and alternative models fit equally well**. We conclude that the alternative model, AR(2), fits the once-differenced time series better than the AR(1) model.

### Part (e) - Comparison of Models

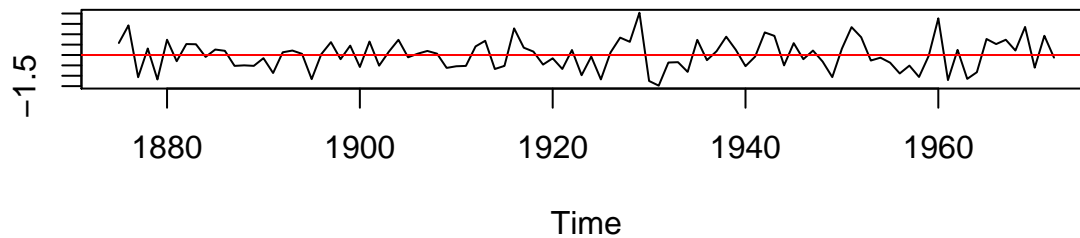
Model	Error Variance, $\sigma^2$	Log likelihood	AIC
AR(1)	0.5452092	-108.2269995	220.453999
AR(2)	0.5188036	-105.8711095	217.742219

We agree with the likelihood ratio test that the AR(2) model fits the once-differenced data better than the AR(1) model. The error variance,  $\sigma^2$ , is lower in the AR(2) model, the AIC goes down even after adding another parameter, and the log-likelihood goes up (which was a significant increase according to the likelihood ratio test). We deem that **the AR(2) model is the optimal choice here.**

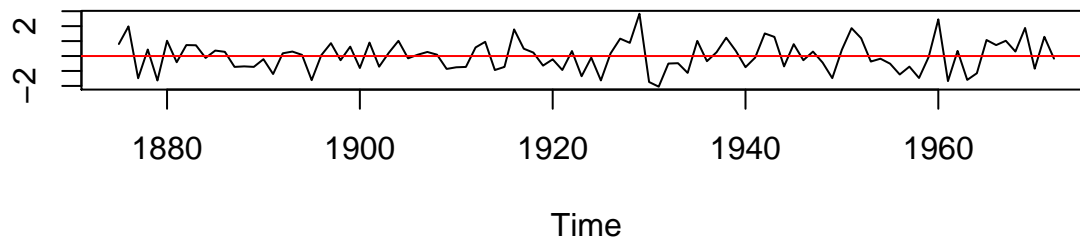
### Part(f) - Residual Diagnostics

#### Part 1 - Zero-Mean

#### Residuals vs t



#### Standardized Residuals vs t



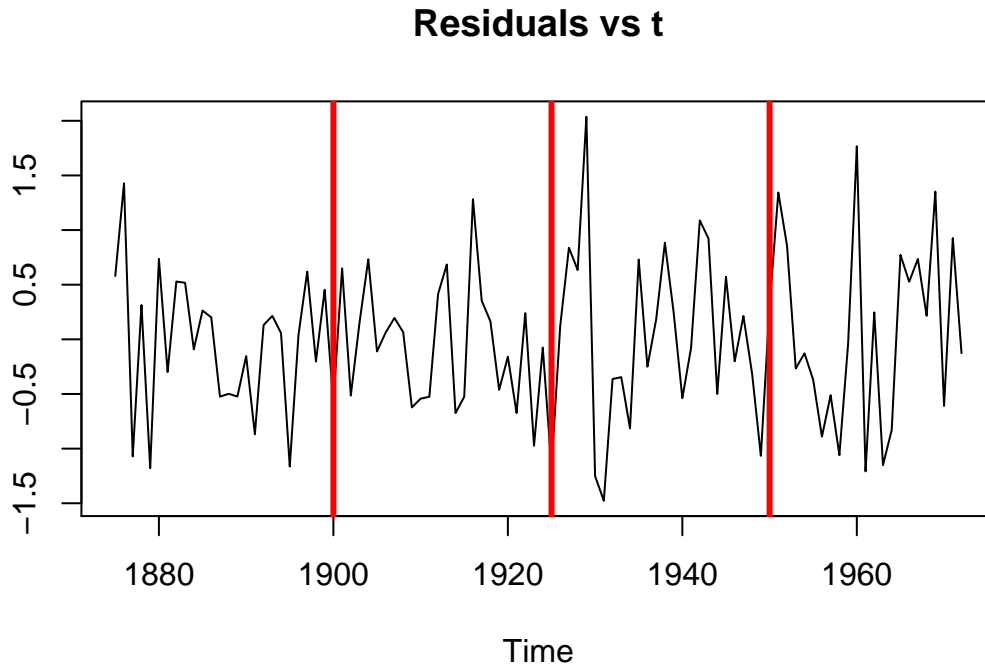
#### One Sample t-test

```
data: e
t = 0.0040215, df = 97, p-value = 0.9968
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.1445959  0.1451831
sample estimates:
mean of x
0.0002935799
```

We see from the scatter plot that the error terms seem to be **symmetrically scattered around zero**. A **t-test confirms this** suspicion, returning a p-value of 0.996 which means that we have no evidence to reject the null hypothesis that the true mean is zero. The 95 percent confidence interval also contains zero for the mean.

## Part 2 - Homoscedasticity

```
# test for heteroscedasticity
par(mfrow = c(1, 1))
plot(e, main = "Residuals vs t", ylab = "")
abline(v = c(1900, 1925, 1950), lwd = 3, col = "red")
```



```
group <- c(rep(1, 25), rep(2, 25), rep(3, 25), rep(4, 23))
levene.test(e, group) #Levene
```

modified robust Brown-Forsythe Levene-type test based on the absolute deviations from the median

```
data: e
Test Statistic = 2.051, p-value = 0.112
```

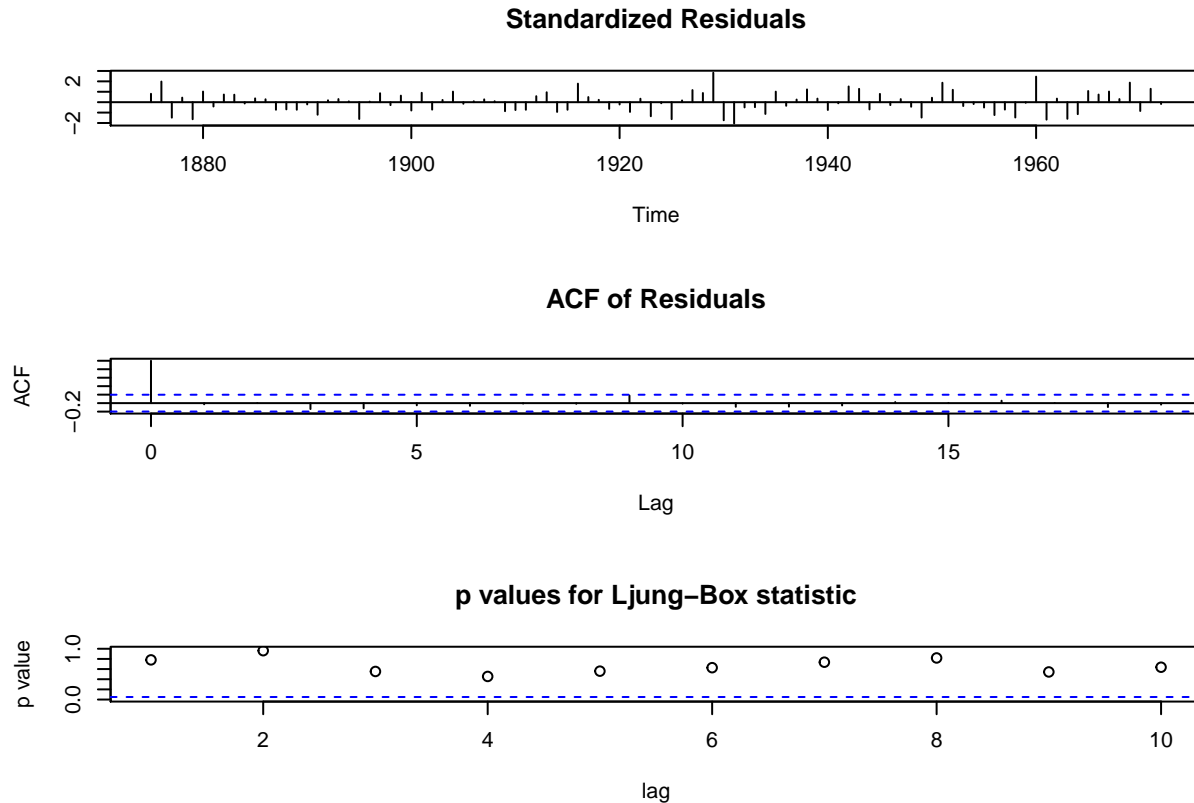
```
bartlett.test(e, group) #Bartlett
```

Bartlett test of homogeneity of variances

```
data: e and group
Bartlett's K-squared = 5.9332, df = 3, p-value = 0.1149
```

We see from the scatter plot of the residuals that the **residuals look homoscedastic**. We confirm this with the Levene test and the Bartlett test, which return **p-values of 0.113 and 0.1149 respectively**. We have no evidence to reject the null hypothesis, and thus we **conclude that the error terms are indeed homoscedastic**.

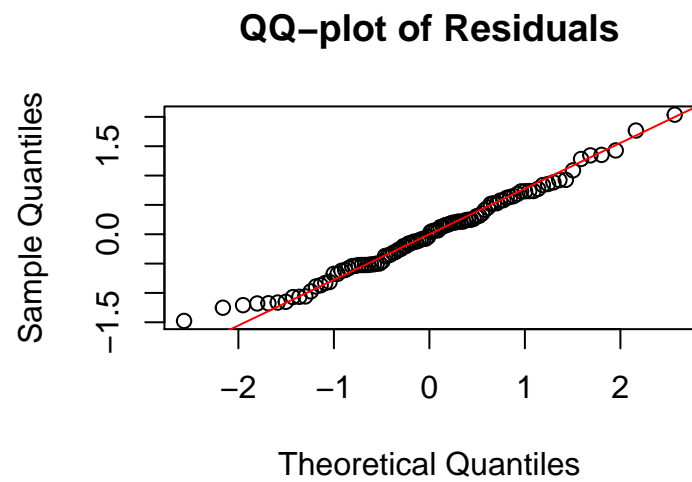
### Part 3 - Zero-Correlation



The ACF plot has **no significant spikes** so we can conclude that the error terms are **uncorrelated** or “**zero-correlated**”. The Ljung-Box test and plot confirms this statistic as the **p-values are all above the threshold**, so we can conclude that there is no evidence to reject the null hypothesis that the error terms are uncorrelated at any lag.

### Part 4 - Normality

The qq-plot suggests that the **error terms are normally distributed**, so the MLE assumption of normality holds. We **confirm this with the Shapiro-Wilk normality test** which gives a p-value of 0.4492, so we fail to reject the null hypothesis that the error terms are normally distributed.



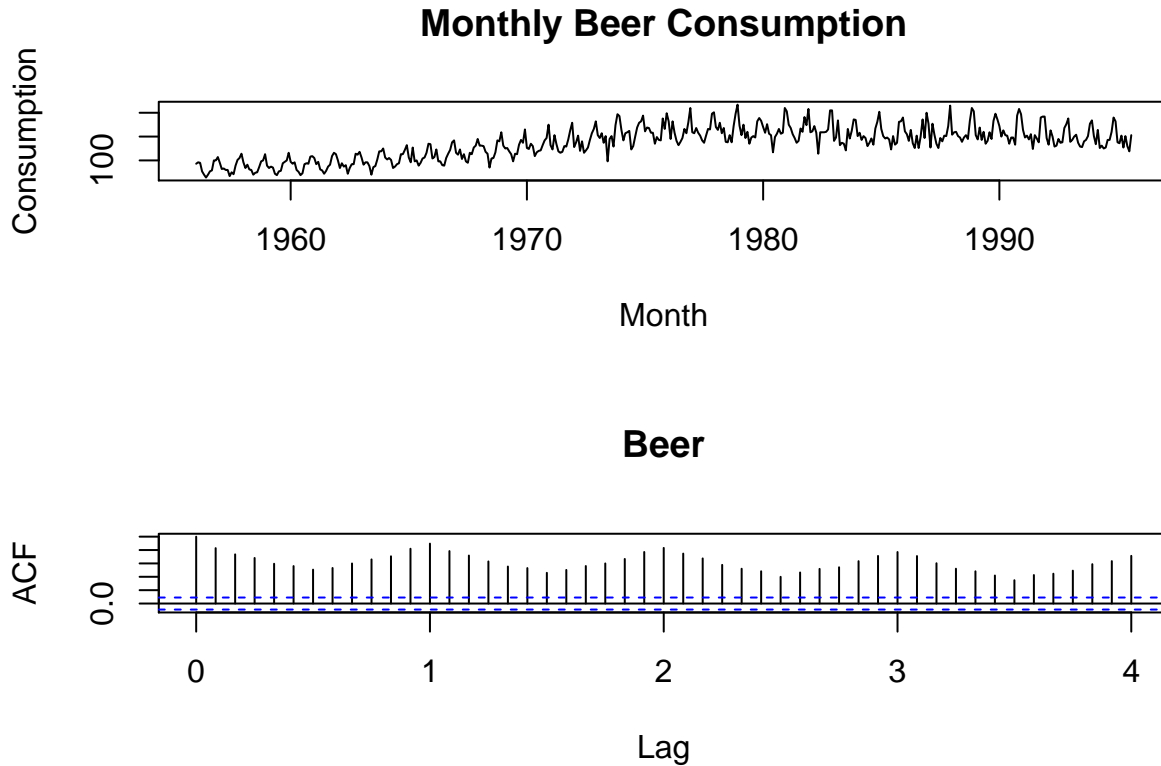
Shapiro-Wilk normality test

data: e

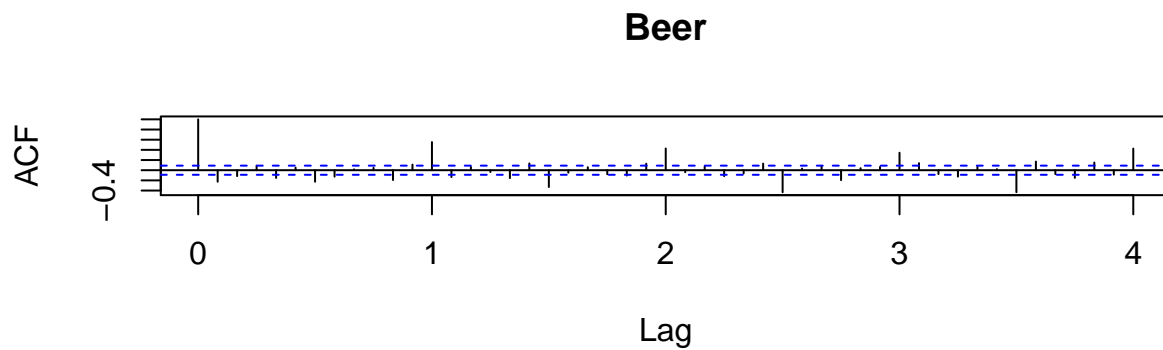
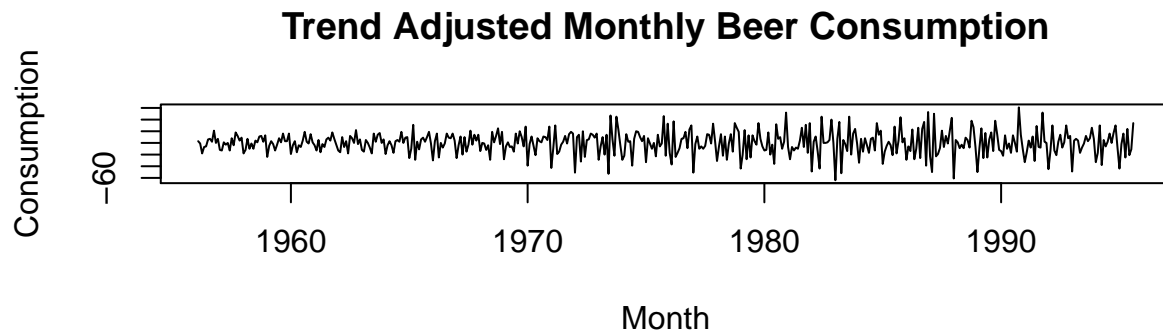
W = 0.98695, p-value = 0.4492

## Problem 2

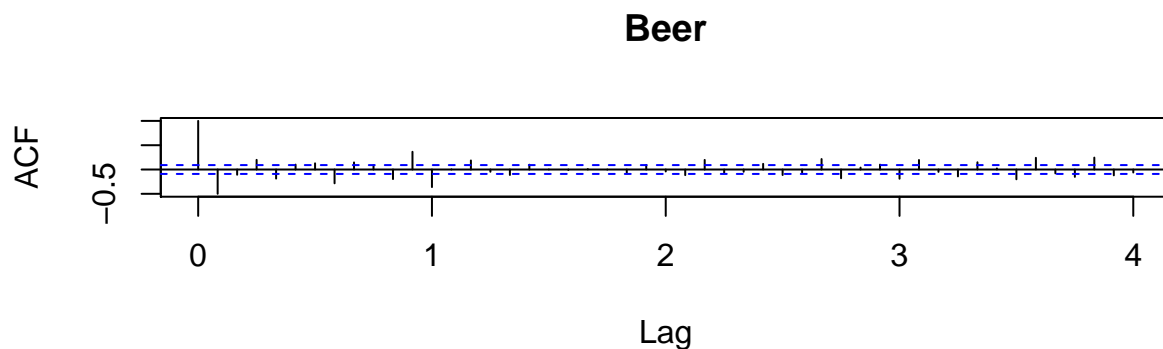
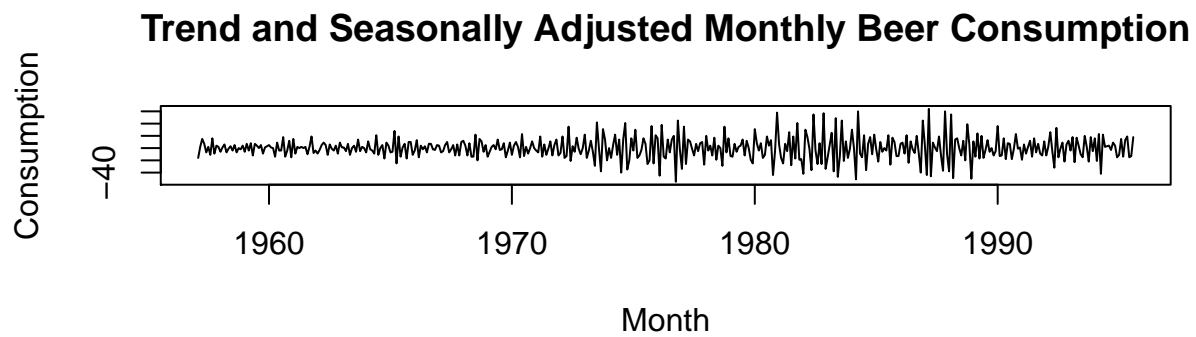
Choosing  $d$  and  $D$



We see that the time series is not stationary and we will need to do some differencing. The ACF plot shows slow decay, and a seasonal trend as well. While the Augmented Dickey-Fuller test implied that the original time series was stationary, we chose not to believe this result as the plots indicated that there was non-stationarity (The KPSS test indicated that we needed to difference once, however). We difference once (ordinary) and look at the transformed time-series and ACF plot afterwards:



We observe that after ordinary differencing once, the observed time series looks flat. The ACF plot also depicts rapid decay after we difference once, as opposed to slow decay. The ACF plot also indicates that there is seasonality present, due to the recurring nature of the peaks when the lag is 12. So we do seasonal differencing as well.

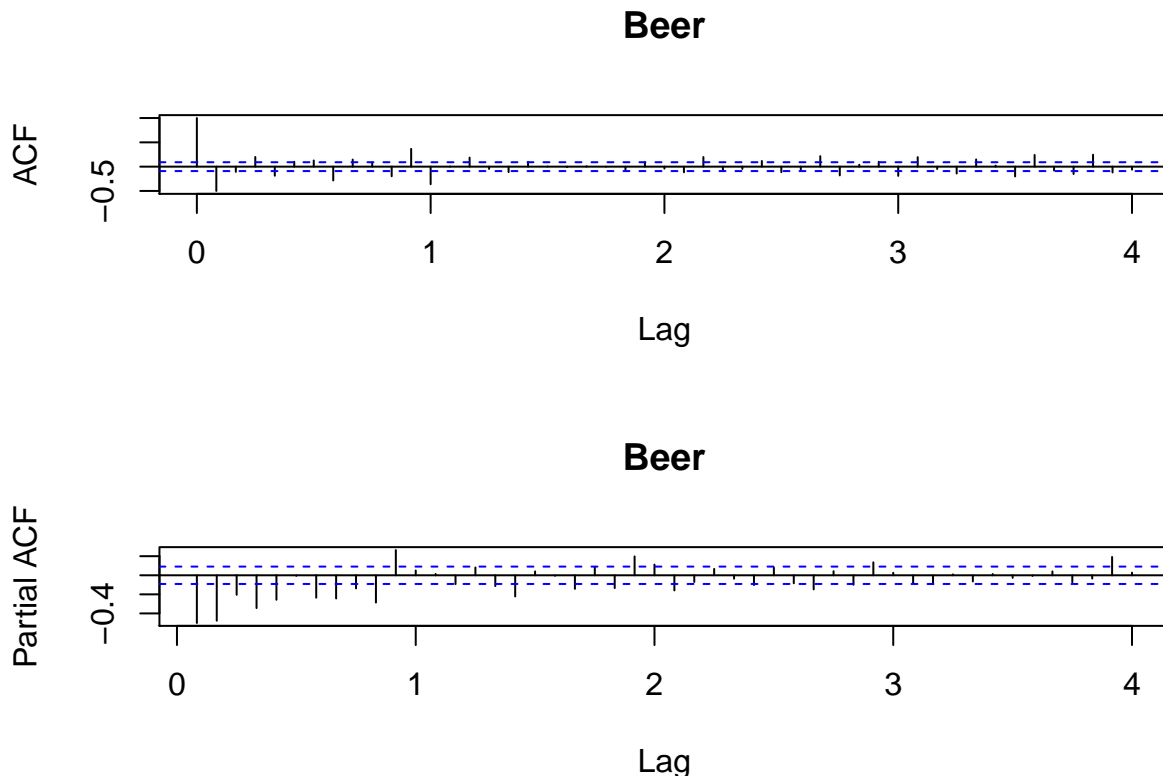




After seasonally differencing once, the ACF plot does not indicate that any more seasonal differencing is required. So we choose the order of the ordinary differencing,  $d$  as 1 and the order of the seasonal differencing,  $D$  as 1 too.

### Choosing $p, q, P,$ and $Q$

```
# This seems fine now. Since we seasonally differenced, we
# are fitting a SARIMA model and need to choose  $p, q, P, Q$ .
# Let's look at the ACF/PACF plots for this
par(mfrow = c(2, 1))
acf(dbeer.12, lag.max = 48, main = "Beer")
pacf(dbeer.12, lag.max = 48, main = "Beer")
```



We observe that there are 5 significant spikes in the PACF. Considering seasonal lags separated by 12 (at 1s, 2s, 3s, etc. with  $s=12$ ), we see 2 spikes. So  $p$  might be less than/equal to 5 and  $P$  less than/equal to 2. In the ACF, we see 11 significant spikes (can only take  $q \leq s-1$  i.e.  $q \leq 11$ ), and with seasonal lags separated by 12 we see 1 significant spike. So  $q$  might be less than 12, and  $Q$  might be less than 1.

```
#  $p \leq 5, q \leq 11, P \leq 2, Q \leq 1$ 

m1 <- arima(beer, order = c(1, 1, 0), seasonal = list(order = c(1,
  1, 0), period = 12), method = "ML")
m2 <- arima(beer, order = c(1, 1, 5), seasonal = list(order = c(1,
  1, 0), period = 12), method = "ML")
m3 <- arima(beer, order = c(1, 1, 11), seasonal = list(order = c(1,
  1, 0), period = 12), method = "ML")
m4 <- arima(beer, order = c(1, 1, 11), seasonal = list(order = c(2,
  1, 1), period = 12), method = "ML")
m5 <- arima(beer, order = c(5, 1, 0), seasonal = list(order = c(1,
  1, 0), period = 12), method = "ML")
```

```

sigma2 <- c(m1$sigma2, m2$sigma2, m3$sigma2, m4$sigma2, m5$sigma2)
loglik <- c(m1$loglik, m2$loglik, m3$loglik, m4$loglik, m5$loglik)
AIC <- c(m1$aic, m2$aic, m3$aic, m4$aic, m5$aic)
d <- data.frame(pq = c("(1,1,0), (1,1,0) [12]", "(1,1,5), (1,1,0) [12]",
  "(1,1,11), (1,1,0) [12]", "(1,1,11), (2,1,1) [12]", "(5,1,0), (1,1,0) [12]"),
  sigma2, loglik, AIC)

# Order this by loglik
d[order(-d$loglik), ]

```

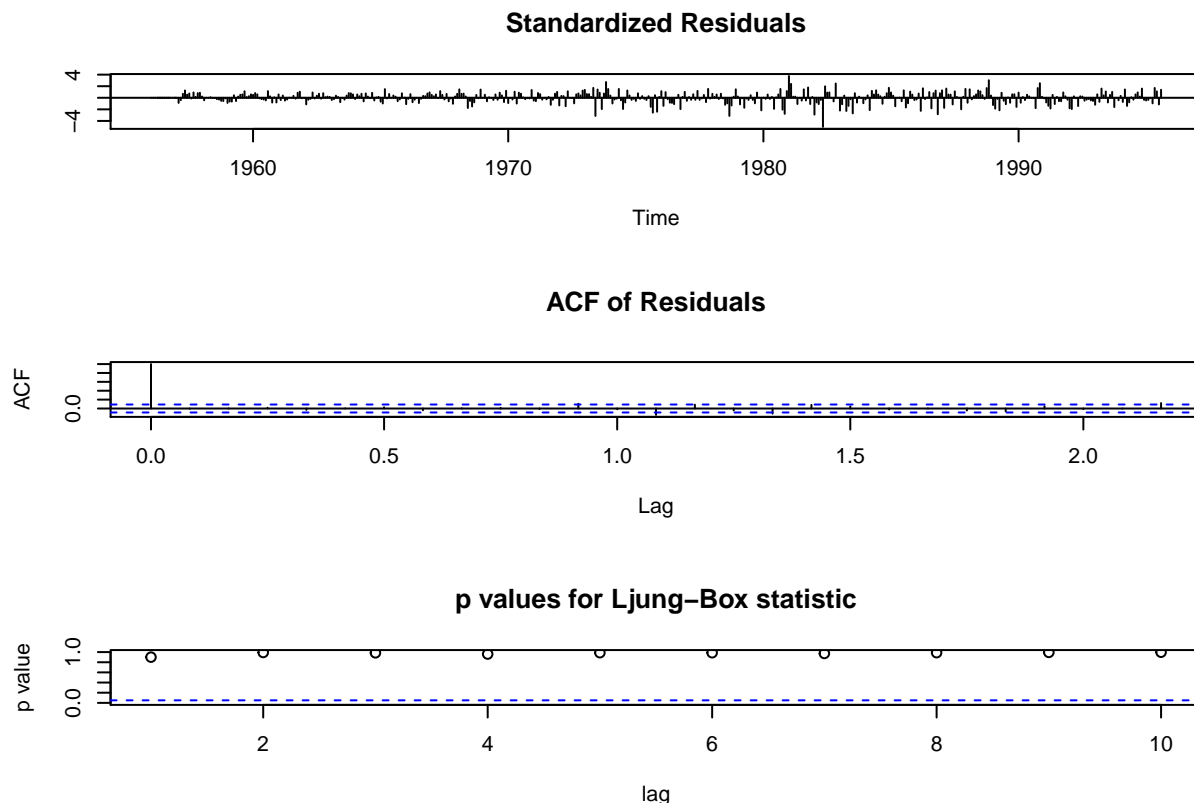
	pq	sigma2	loglik	AIC
4	(1,1,11), (2,1,1) [12]	85.81056	-1695.919	3423.837
3	(1,1,11), (1,1,0) [12]	105.95172	-1741.168	3510.335
2	(1,1,5), (1,1,0) [12]	121.18106	-1769.748	3555.495
5	(5,1,0), (1,1,0) [12]	124.83744	-1776.616	3567.233
1	(1,1,0), (1,1,0) [12]	219.56947	-1906.108	3818.216

While there may be a huge range of models considering the values of  $p, q, P$ , and  $Q$  we picked, we only look at 5 candidate models.

We observe that model 4, with  $(p, d, q), (P, D, Q) [s] = "(1,1,11), (2,1,1) [12]"$ , out of the models we considered, seems to be the best in terms of all the metrics - it has the lowest error variance and AIC, and the highest log-likelihood among the candidate models we considered.

## Residual Diagnostics for Model 4 - The (1,1,11), (2,1,1) [12] model

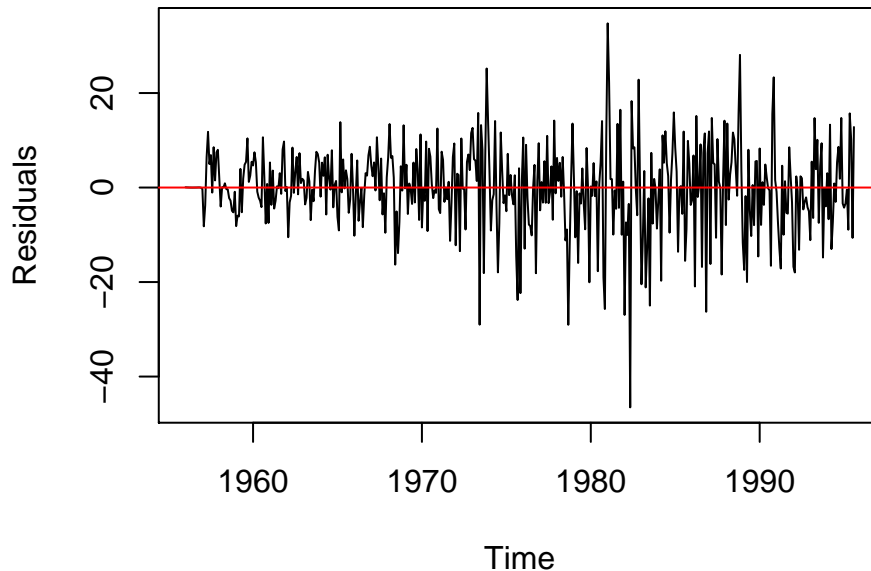
### 1) Zero Mean and Zero-Correlation



The plot of the residuals indicates that the error terms have a zero mean. The ACF plot does not have any significant spikes and the Ljung-Box plot confirms that the error residuals are zero-correlated/uncorrelated since the p-values are all above the threshold.

## 2) Homoscedasticity

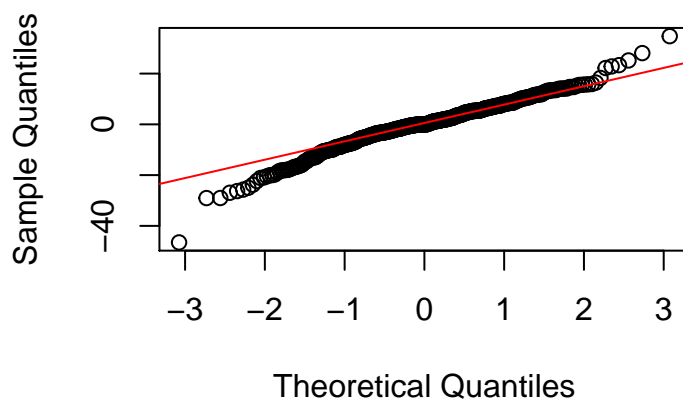
### Residuals vs. Time



We see signs of heteroscedasticity in the plot, and this prompts us to consider log-transforming the original time-series data.

## 3) Normality

### Normal Q-Q Plot

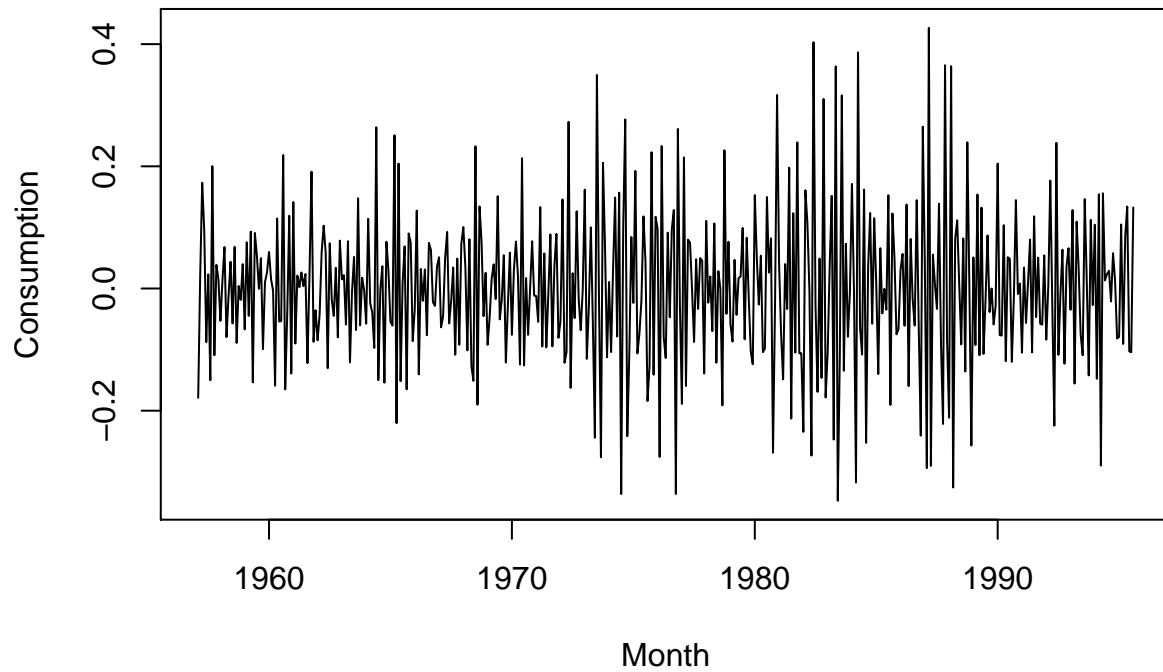


The qqplot indicates non-normality.

The heteroscedasticity and non-normality issues above prompt us to consider log-transforming the time-series and also fitting it with Least Squares.

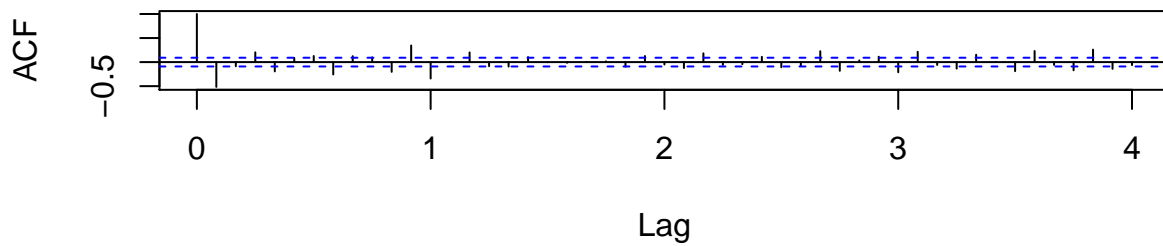
## Transformed Log-Model

### Trend and Seasonally Adjusted Monthly Beer (Log-Transformed) Consumption

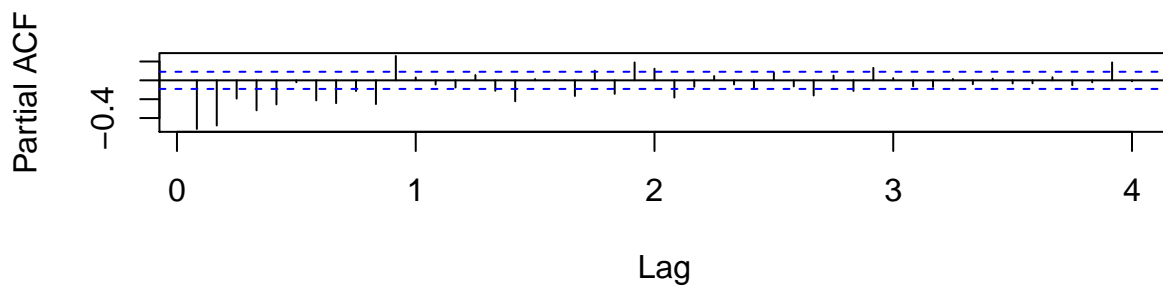


We go through the same steps as before and determine the order of the ordinary differencing,  $d$  as 1 and the order of the seasonal differencing,  $D$  as 1.

### Beer



### Beer



We observe a similar number of spikes as we did previously in both the ACF and PACF plots, and make the same choices for  $p, q, P$  and  $Q$  as before.

```
# p <= 5, q <= 11, P <= 2, Q <= 1

m1 <- arima(beer, order = c(1, 1, 0), seasonal = list(order = c(1,
  1, 0), period = 12), method = "CSS")
m2 <- arima(beer, order = c(1, 1, 5), seasonal = list(order = c(1,
  1, 0), period = 12), method = "CSS")
m3 <- arima(beer, order = c(1, 1, 11), seasonal = list(order = c(1,
  1, 0), period = 12), method = "CSS")
m4 <- arima(beer, order = c(1, 1, 11), seasonal = list(order = c(2,
  1, 1), period = 12), method = "CSS")
m5 <- arima(beer, order = c(5, 1, 11), seasonal = list(order = c(1,
  1, 0), period = 12), method = "CSS")

sigma2 <- c(m1$sigma2, m2$sigma2, m3$sigma2, m4$sigma2, m5$sigma2)
loglik <- c(m1$loglik, m2$loglik, m3$loglik, m4$loglik, m5$loglik)
AIC <- c(m1$aic, m2$aic, m3$aic, m4$aic, m5$aic)
d <- data.frame(pq = c("(1,1,0), (1,1,0) [12]", "(1,1,5), (1,1,0) [12]",
  "(1,1,11), (1,1,0) [12]", "(1,1,11), (2,1,1) [12]", "(5,1,11), (1,1,0) [12]"),
  sigma2, loglik, AIC)

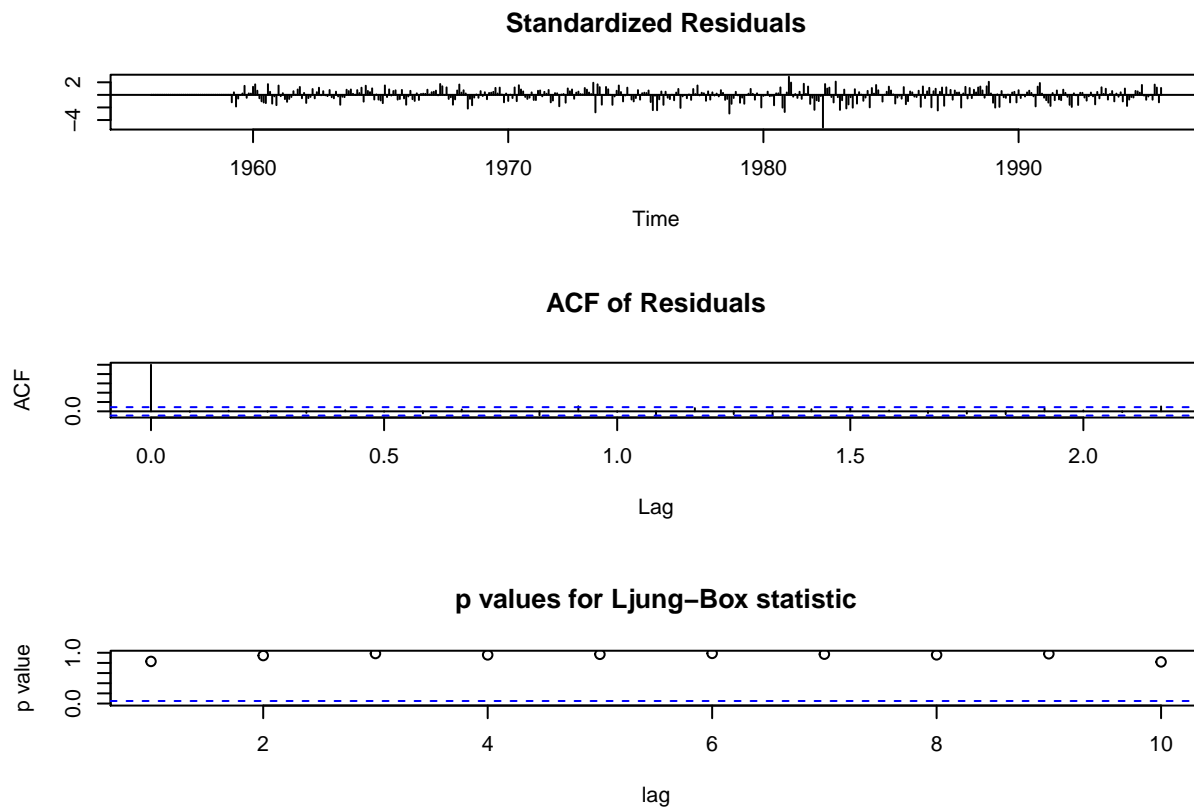
# Order this by sigma2
d[order(d$sigma2), ]
```

	pq	sigma2	loglik	AIC
4	(1,1,11), (2,1,1) [12]	0.004191433	610.4274	NA
5	(5,1,11), (1,1,0) [12]	0.004310933	603.9196	NA
3	(1,1,11), (1,1,0) [12]	0.005249204	558.3321	NA
2	(1,1,5), (1,1,0) [12]	0.005827106	534.1533	NA
1	(1,1,0), (1,1,0) [12]	0.010424201	399.5107	NA

We observe that the same model, model 4, with  $(p,d,q), (P,D,Q)$  [s] = “(1,1,11), (2,1,1) [12]”, was the best out of all candidate models. Since we fit using Least Squares, we use only the metric of error variance here.

## Residual Diagnostics for Model 4 - The (1,1,12), (2,1,1) [12] model (Log-Transformed)

### 1) Zero Mean and Zero-Correlation

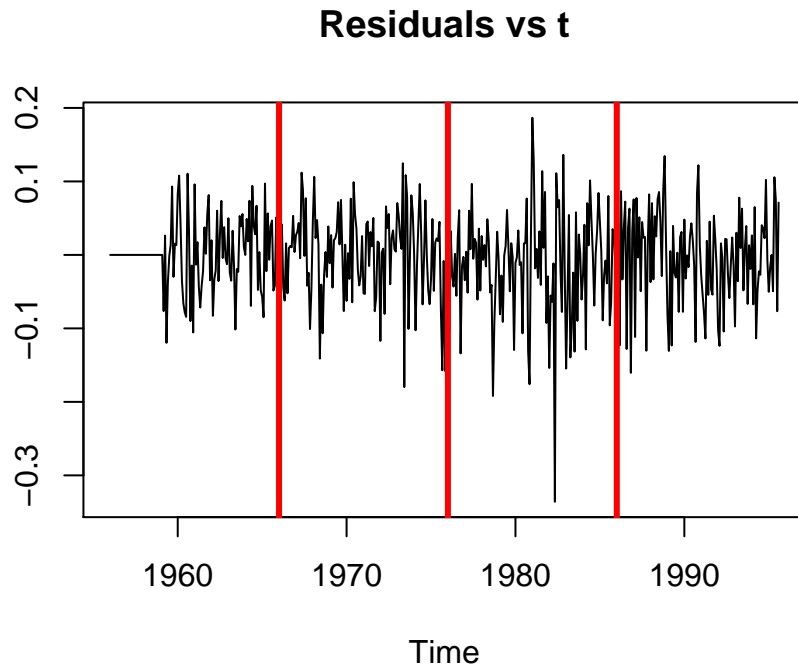


As before, the plot of the residuals indicates that the error terms have a zero mean. The ACF plot does not have any significant spikes and the Ljung-Box plot confirms that the error residuals are zero-correlated/uncorrelated since the p-values are all above the threshold.

### 2) Homoscedasticity

```
e <- m4$residuals # residuals
r <- e/sqrt(m4$sigma2) # standardized residuals

par(mfrow = c(1, 1))
plot(e, main = "Residuals vs t", ylab = "")
abline(v = c(1966, 1976, 1986), lwd = 3, col = "red")
```



```
group <- c(rep(1, 120), rep(2, 120), rep(3, 120), rep(4, 116))
levene.test(e, group) #Levene
```

modified robust Brown-Forsythe Levene-type test based on the absolute deviations from the median

```
data: e
Test Statistic = 7.6473, p-value = 5.32e-05
```

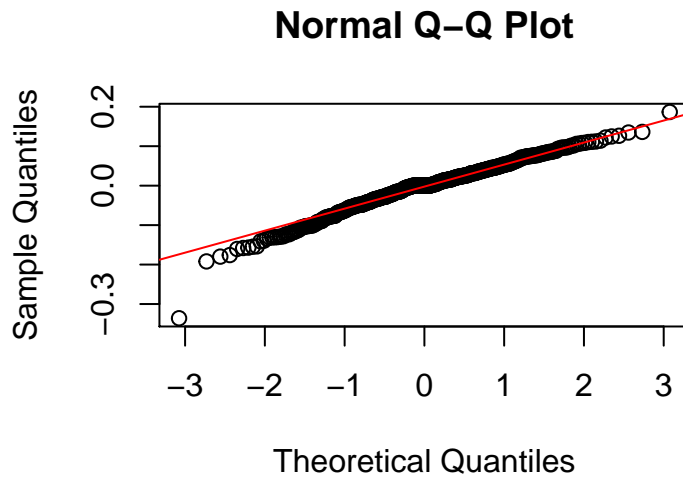
```
bartlett.test(e, group) #Bartlett
```

Bartlett test of homogeneity of variances

```
data: e and group
Bartlett's K-squared = 27.241, df = 3, p-value = 5.24e-06
```

The error terms **look homoscedastic now, after the log-transformation**. However, when we perform a Levene test and Bartlett test, we end up rejecting the hypothesis that the error terms are homoscedastic. Considering the way we grouped the observations, we observe that the third group has a lot of variation at one observed time point, but that roughly the observations seem to display constant variance. So while there is a discrepancy between the informal diagnostics, and the formal diagnostics (tests), one could still argue that the assumption of homoscedasticity was met.

## Normality



## Shapiro-Wilk normality test

```
data: e
W = 0.97734, p-value = 9.433e-07
```

The qqplot indicates non-normality of the error terms. The Shapiro Test also indicates that the error terms are not normally distributed. However, since we fit the model using Least Squares this time, we do not need to satisfy the assumption of normality.

## Model Validity

While the log-transformed model satisfies the assumptions of zero-mean and zero-correlation, it fails to satisfy the assumptions of homoscedasticity and normality of the error terms. The normality of the error terms should not be an issue since we fit the log-transformed model using the least squares approach in the second iteration. As for the heteroscedasticity issues, I would argue that while the Levene Test and Bartlett Test indicate non-constant variance, the plot of the residuals indicates that the error terms have mostly constant variance. So while it doesn't "rigidly" meet all assumptions, it could still be used to make good predictions and prediction intervals.