Forecasting Retail and Wholesale Employment in Wisconsin (1961–1975)

Kerry Yu

Abstract

This project analyzes and forecasts monthly employment levels in the wholesale and retail trade sectors in Wisconsin between 1961 and 1975.. Using the Box-Jenkins methodology, we apply time series transformation, differencing, autocorrelation diagnostics, and model selection based on AICc. After validating residuals through diagnostic tests, we use the chosen model to forecast values and interpret patterns of labor market participation in wholesale and retail. This projects aim to provide understanding labor market forecasting and economic planning.

Introduction

The dataset, compiled by R.B. Miller, on monthly employment data for the wholesale and retail sectors in Wisconsin from 1961 to 1975. The main goal of this project is to model this time series in order to forecast future employment levels and identify any underlying patterns, including trends and seasonality. We apply transformation to stabilize variance, differencing for stationarity, and ACF/PACF diagnostics to identify suitable ARIMA or SARIMA models. After estimating and validating candidate models, we evaluate the model's forecasting performance and assess its adequacy through residual analysis.

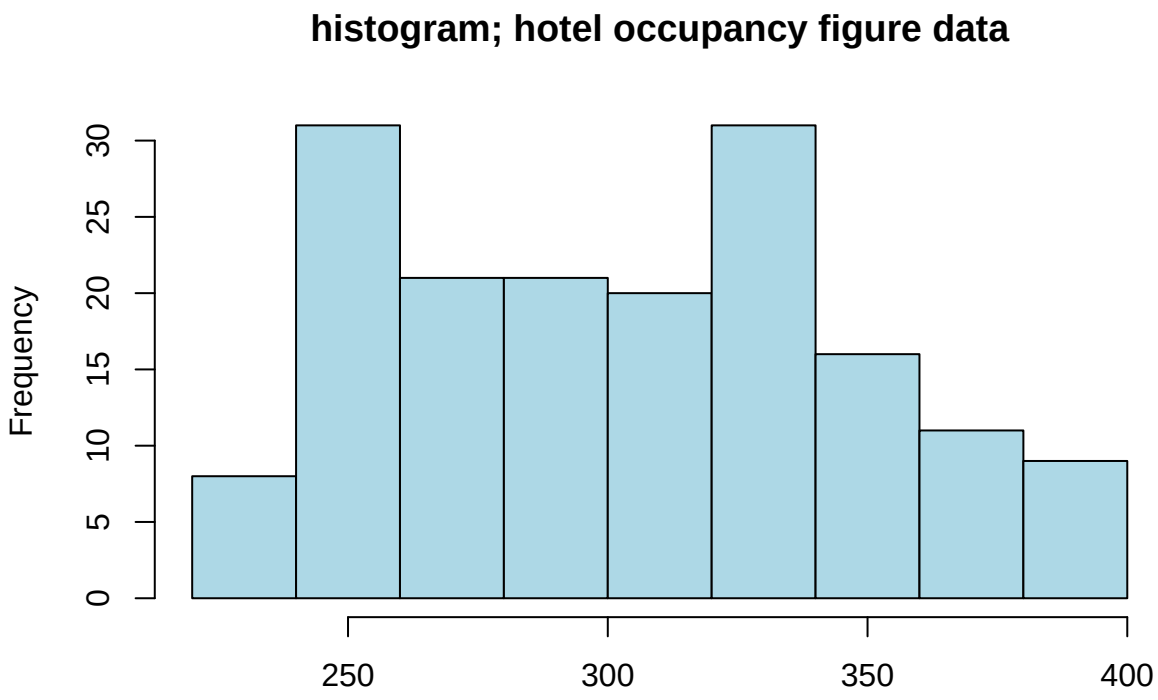
The dataset was retrieved from the Time Series Data Library via tsdl R package and is appropriately cited.

Since there is no new data available, we will split our data into training and test datasets.

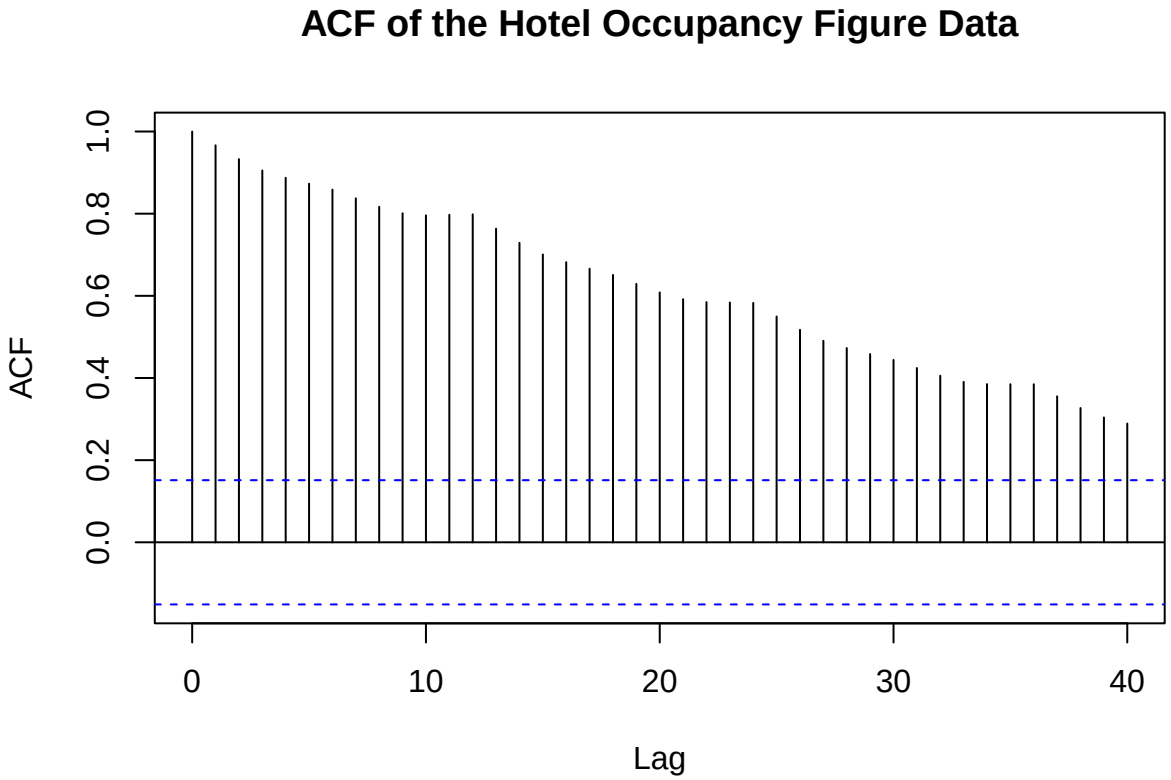
```
library(ggplot2)
library(ggfortify)
library(forecast)
library(MASS)
library(tsdl)

#456 565 545 544
meta_tsdl$description[[544]]
```

```
## [1] "Monthly empolyees wholes./retail Wisconsin -61--75 R.B.Miller"
```

```
employee.ts <- tsdl[[544]]


# splitting into training and testing sets
et <- employee.ts[c(1:(14*12))]
e.test <- employee.ts[c(((14*12)+1):(15*12))]

# histogram
hist(et, col="light blue", xlab="", main="histogram; hotel occupancy figure data")
```



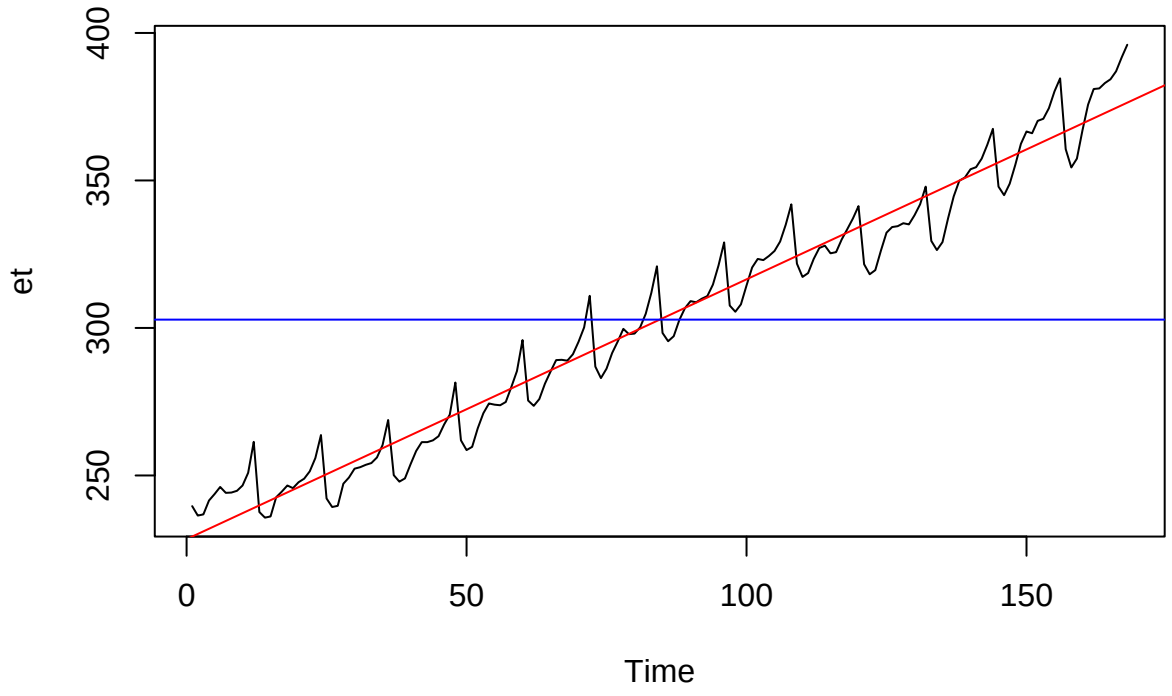

```
# acf
acf(et, lag.max=40, main="ACF of the Hotel Occupancy Figure Data")
```



Analysis and Transformations

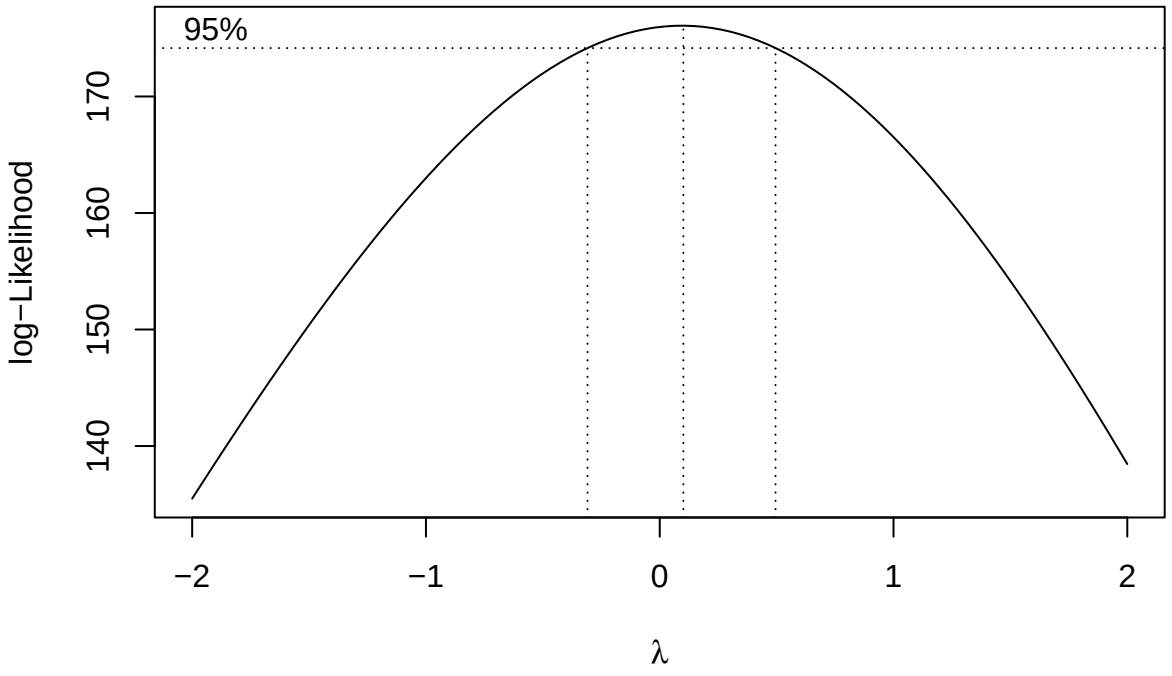
Now we'll analyze the time series by plotting it

```
plot.ts(et)
fit <- lm(et ~ as.numeric(1:length(et))); abline(fit, col="red")
abline(h=mean(et), col="blue")
```



There is a clear upward trend; the variance appears to be nonconstant; clear seasonality. Because the variance is nonconstant, apply a transformation (e.g. Box-Cox, log).

```
# box-cox transformation
bcTransform <- boxcox(et ~ as.numeric(1:length(et)))
```

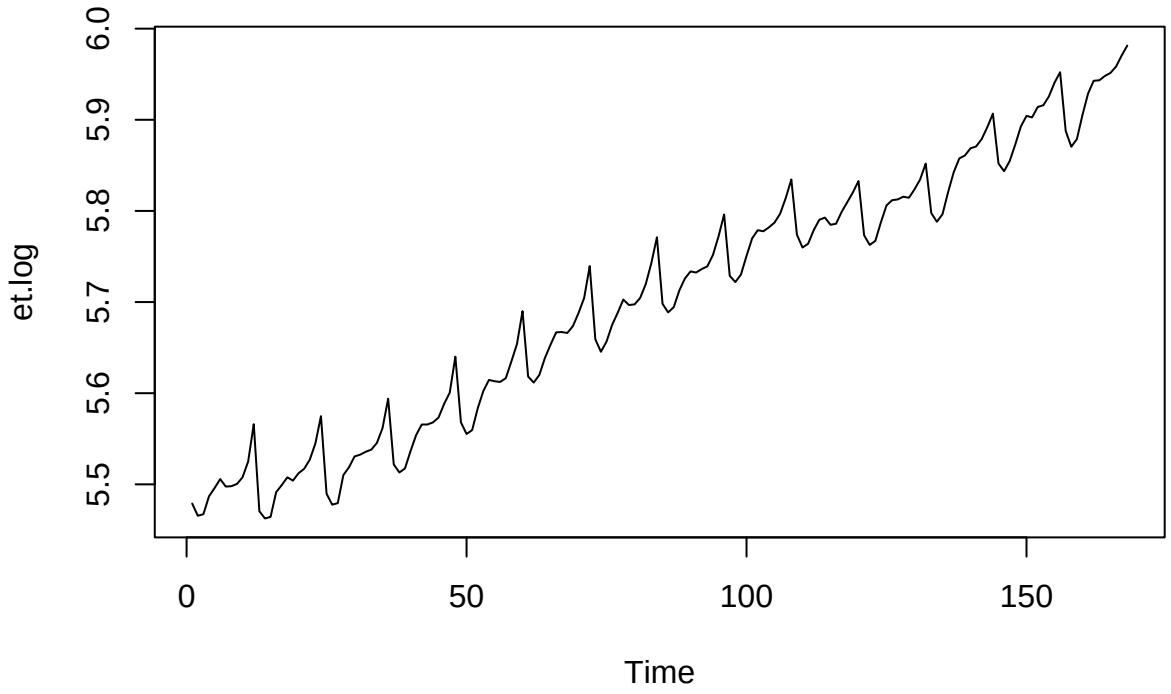


```
bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
```

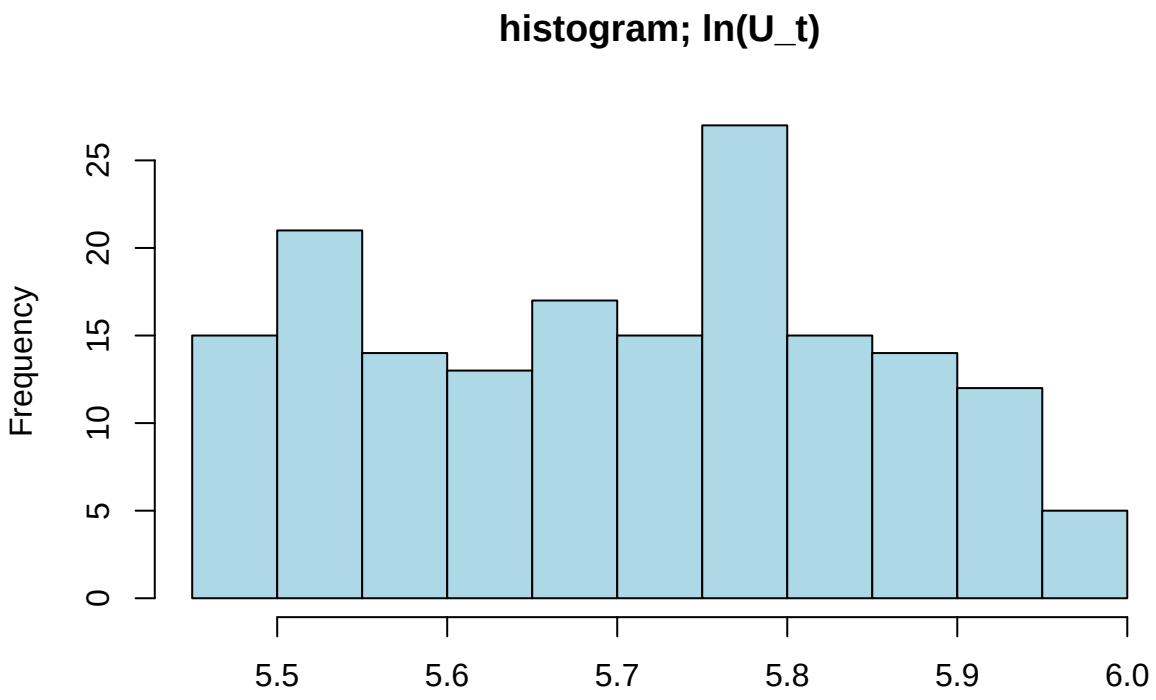
```
## [1] 0.1010101
```

In the above Box-Cox plot, the confidence interval for λ includes 0. Thus, a log transformation is more appropriate than a Box-Cox transformation.

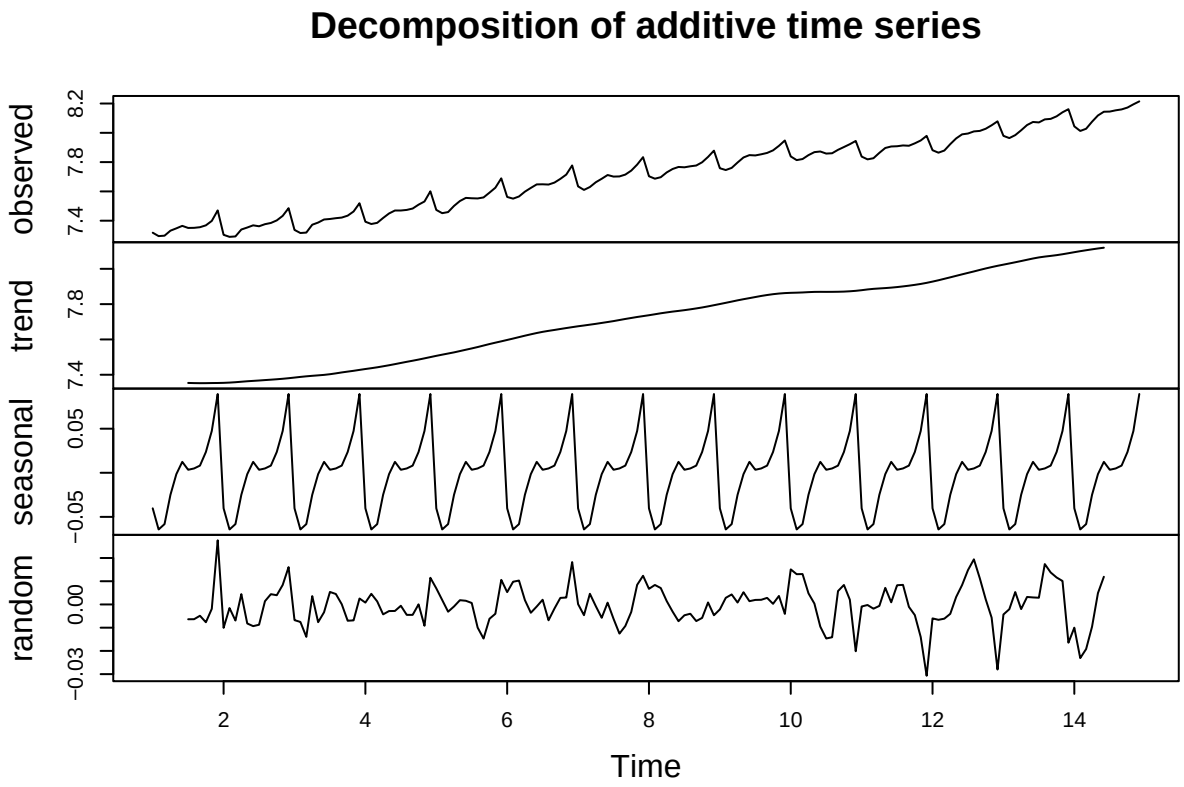
```
# perform transformation
lambda=bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
et.bc = (1/lambda)*(et^lambda-1)
et.log = log(et)

# plot transformed data
plot.ts(et.log)
```




```
# histogram of transformed data
hist(et.log, col="light blue", xlab="", main="histogram; ln(U_t)")
```




```
y <- ts(as.ts(et.bc), frequency = 12)
decomp <- decompose(y)
plot(decomp)
```

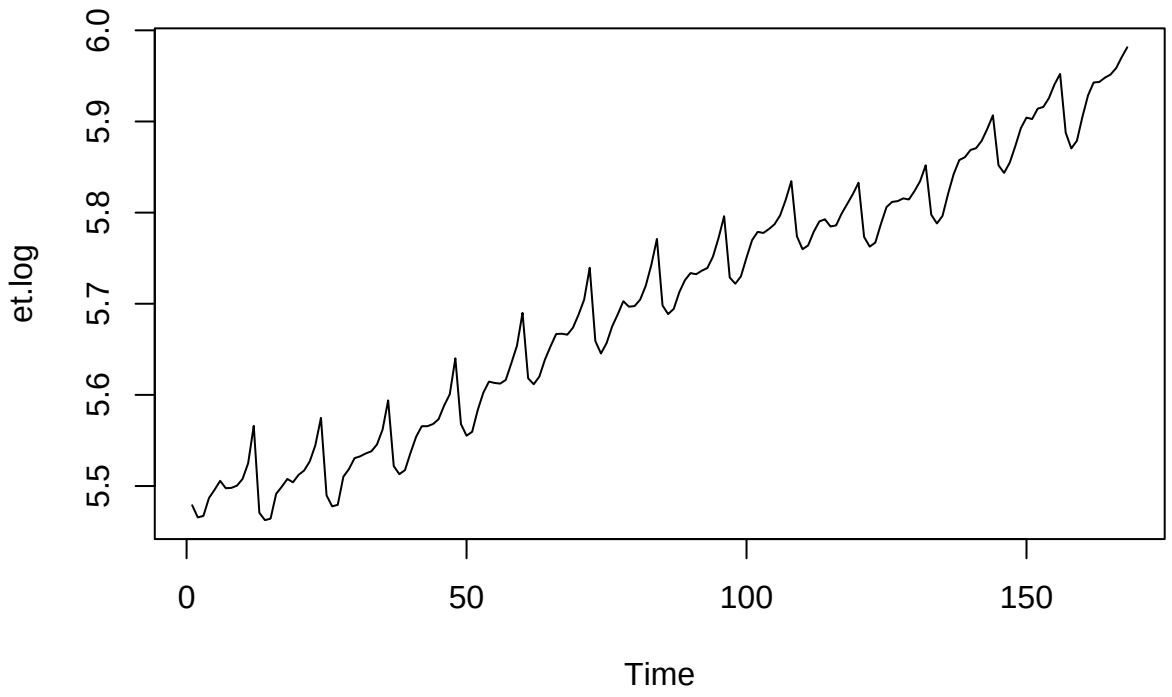


The decomposition of $\mathrm{bc}(U_t)$ shows seasonality and a slight linear trend; thus, we will continue with differencing.

Differencing

Original plot:

```
plot.ts(et.log)
```




```
var(et.log) # original variance
```

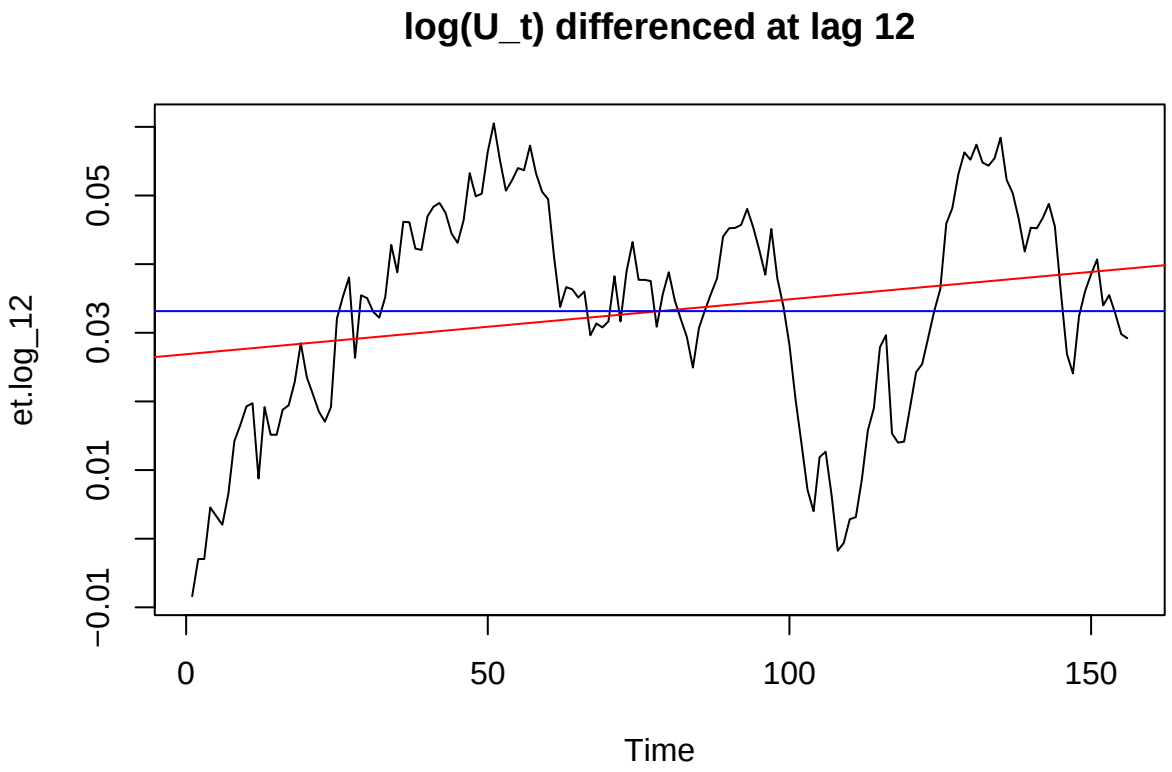
```
## [1] 0.0209324
```

The transformed data is highly nonstationary. The first step toward stationarity is to address seasonality by differencing once at lag 12.

```
et.log_12 <- diff(et.log, lag=12)
var(et.log_12)
```

```
## [1] 0.000245481
```

```
plot.ts(et.log_12, main="log(U_t) differenced at lag 12")
fit <- lm(et.log_12 ~ as.numeric(1:length(et.log_12))); abline(fit, col="red")
#mean(et.log_12)
abline(h=mean(et.log_12), col="blue")
```

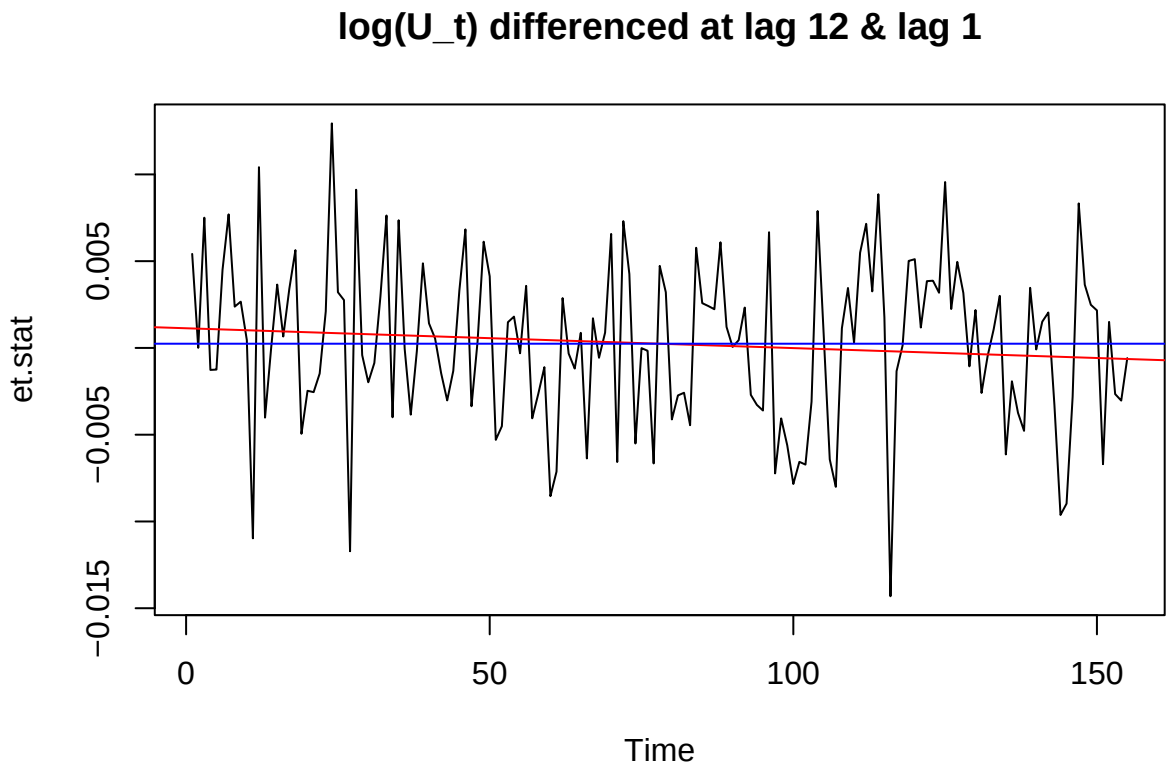


The variance decreased significantly after seasonally differencing once at lag 12, indicating the difference was a good decision. A slight positive trend is still present, so we move on to differencing once at lag 1, twice if needed.

```
et.stat <- diff(et.log_12, lag=1)
var(et.stat)
```

```
## [1] 2.281867e-05
```

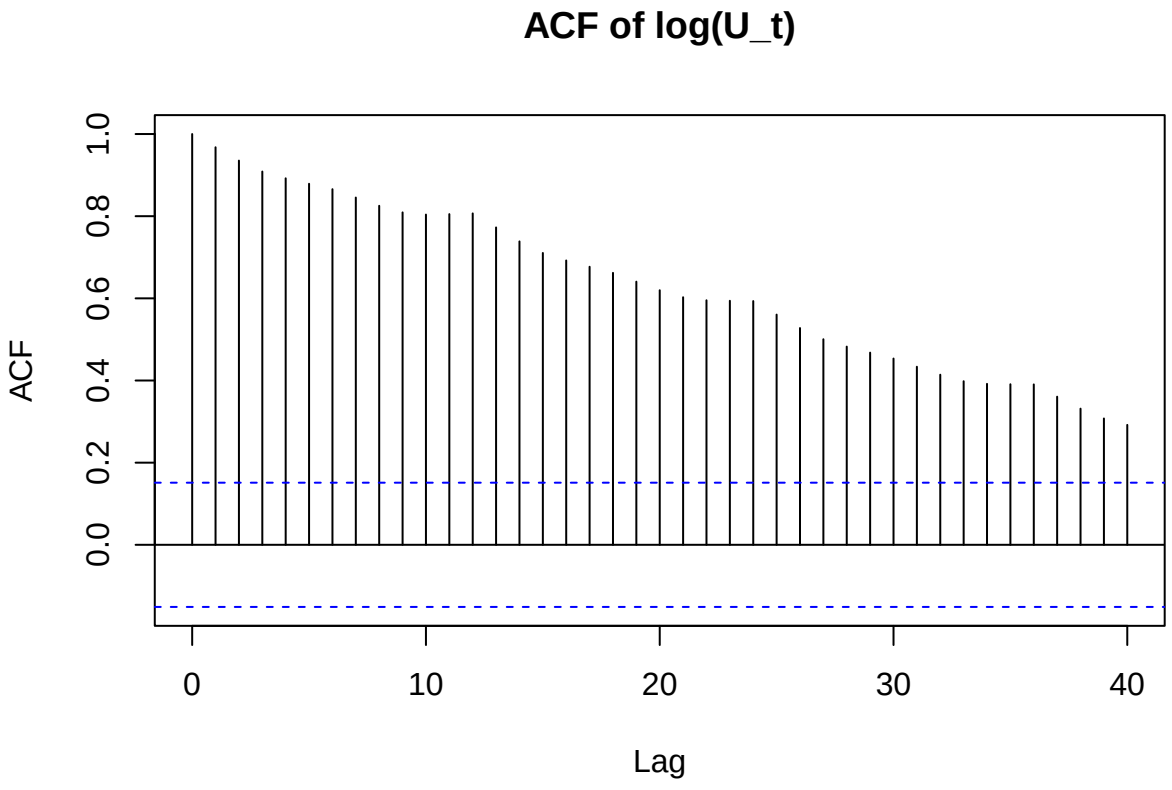
```
plot.ts(et.stat, main="log(U_t) differenced at lag 12 & lag 1")
fit <- lm(et.stat ~ as.numeric(1:length(et.stat))); abline(fit, col="red")
#mean(et.stat)
abline(h=mean(et.stat), col="blue")
```



After differencing once at lag 12 and once at lag 1, the process looks much more stationary. The variance has also gone down.

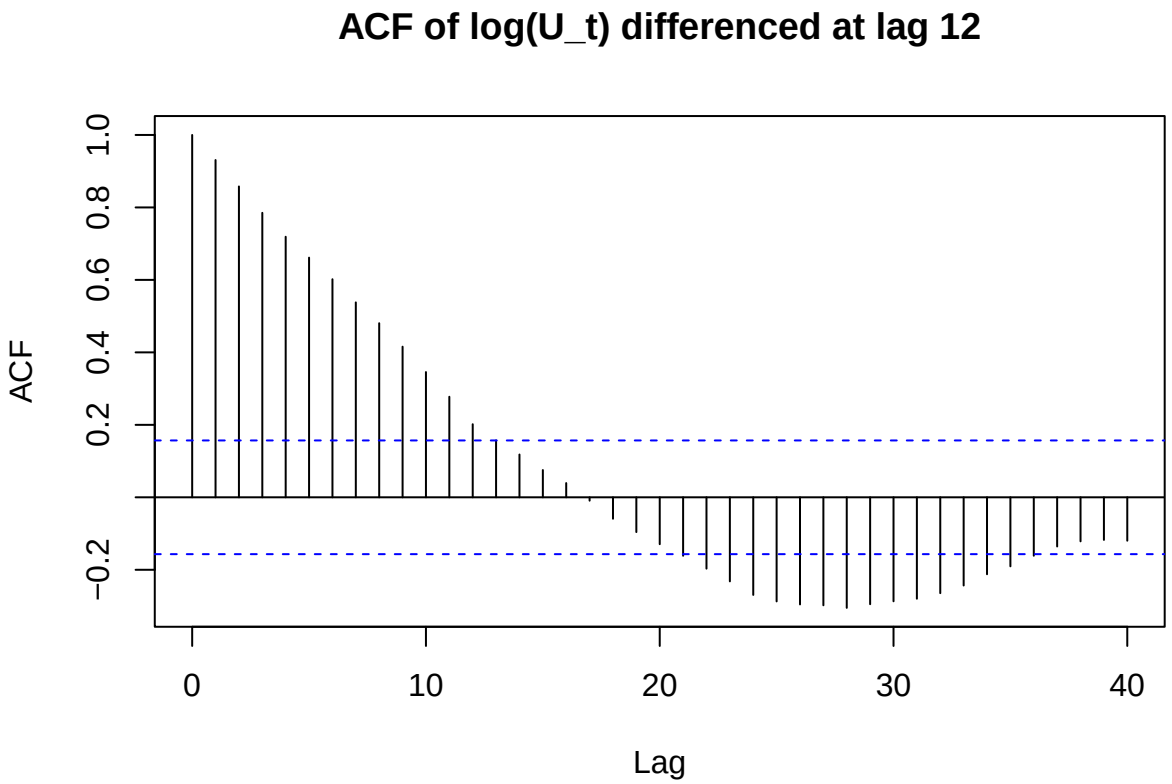
Let's confirm stationarity by looking at the ACF plots.

```
acf(et.log, lag.max=40, main="ACF of log(U_t)")
```



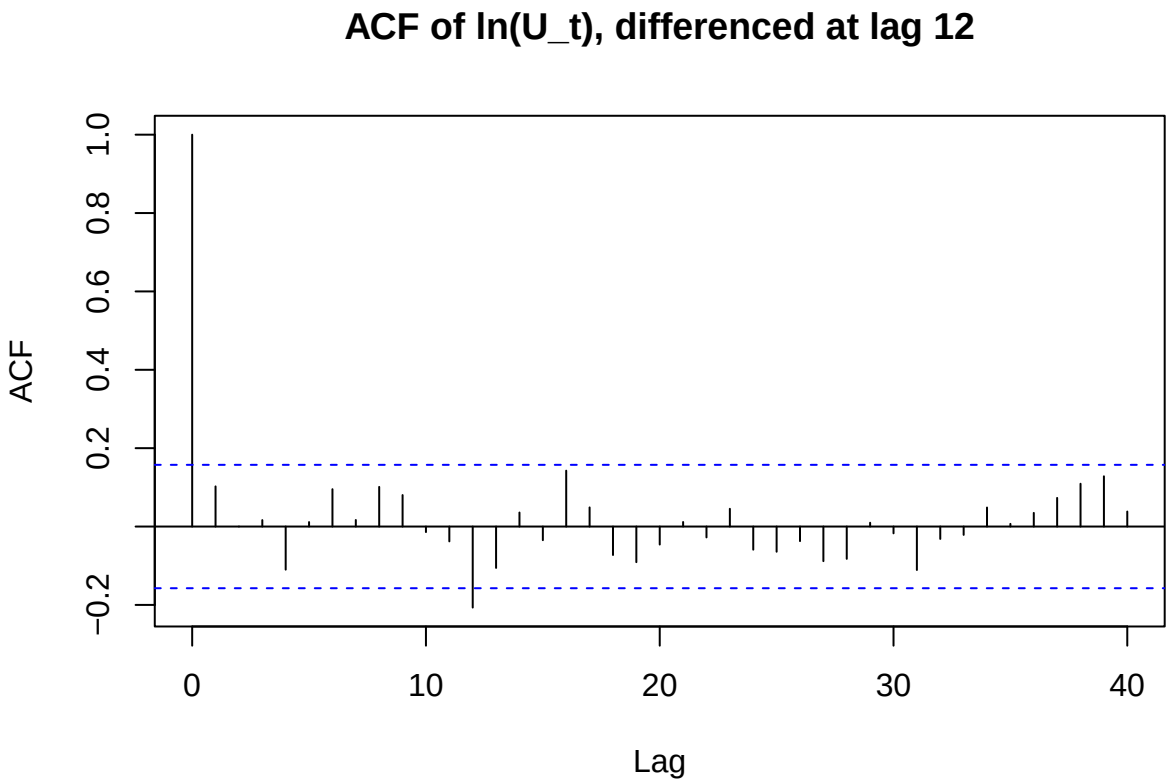
A slow decay with seasonal spikes in the ACF plot of $\log(U_t)$ indicates nonstationarity.

```
acf(et.log_12, lag.max=40, main="ACF of log(U_t) differenced at lag 12")
```



After differencing at lag 12, seasonality is no longer seen in the above plot; however, the slow decay is still present, indicating nonstationarity.

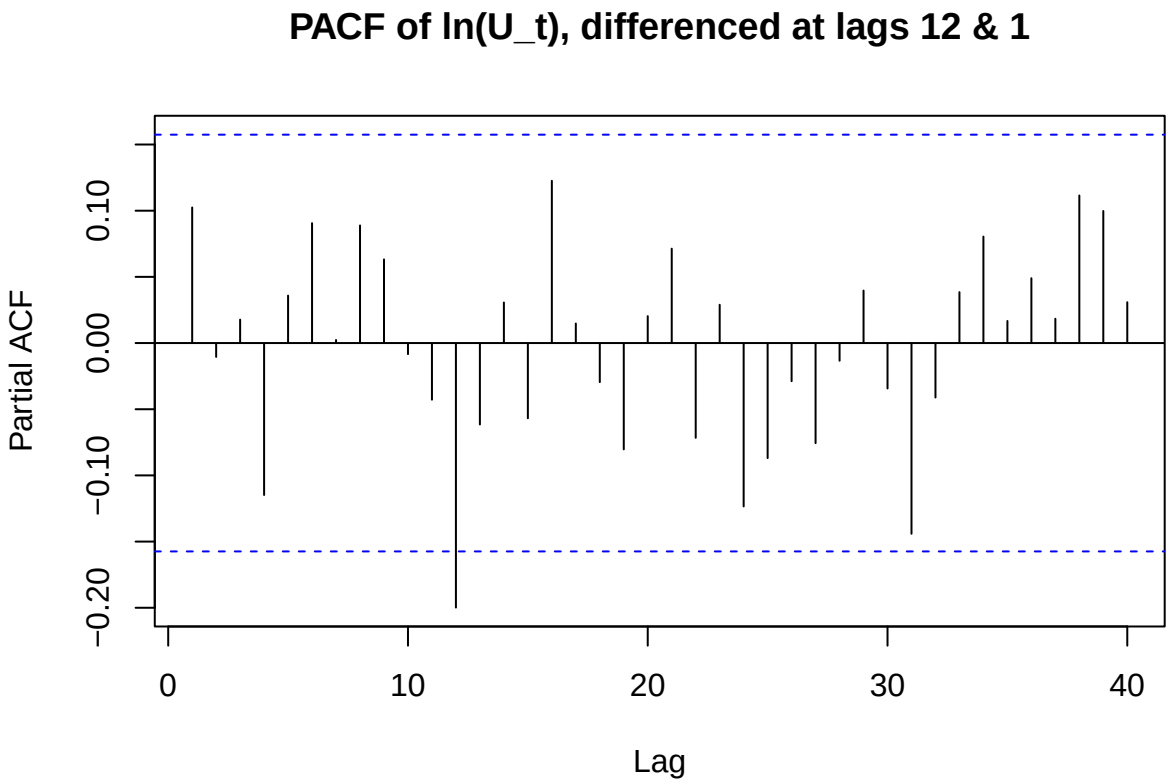
```
acf(et.stat, lag.max=40, main="ACF of ln(U_t), differenced at lag 12")
```



The ACF pattern above after differencing again at lag 1 shows a stationary process. Thus, we will work with the data: log transformed and differenced once at lags 12 & 1, $\nabla_1 \nabla_{12} \log(U_t)$.

The ACF at lag 12 appears to be statistically significant (nonzero). This indicates a SAR(1) process $(q = 0, Q = 1)$.

```
pacf(et.stat, lag.max=40, main="PACF of ln(U_t), differenced at lags 12 & 1")
```



The PACF at only lag 12. With both ACF and PACF plots analyzed, we can say possible parameters for U_t are:

$$\text{SARIMA for bc}(U_t):$$
$$s = 12, D = 1, d = 1,$$
$$Q = 0 \text{ or } 1, q = 0, P = 0 \text{ or } 1, p = 0$$

Fitting Models

```
# p=0, q=0, P=0, Q=1
Arima(et.log, order = c(0,1,0), seasonal = list(order = c(0,1,1), period = 12), method = "ML")
```

```
## Series: et.log
## ARIMA(0,1,0)(0,1,1)[12]
##
## Coefficients:
##          sma1
##       -0.2943
## s.e.   0.0898
##
## sigma^2 = 2.392e-05:  log likelihood = 613.48
## AIC=-1222.97   AICc=-1222.89   BIC=-1216.88
```

```
# p=0, q=0, P=1, Q=0
Arima(et.log, order = c(0,1,0), seasonal = list(order = c(1,1,0), period = 12), method = "ML")
```

```
## Series: et.log
## ARIMA(0,1,0)(1,1,0)[12]
##
## Coefficients:
```

```
##          sar1
##       -0.2364
## s.e.   0.0837
##
## sigma^2 = 2.424e-05:  log likelihood = 612.53
## AIC=-1221.06   AICc=-1220.98   BIC=-1214.98
```

The AICc decreased from -1222.89 ($Q = 1, P = 0$) to -1220.98 ($Q = 0, P = 1$). Therefore, the SMA1 component is more significant.

```
# p=0, q=0, P=1, Q=1
Arima(et.log, order = c(0,1,0), seasonal = list(order = c(1,1,1),  period = 12),  method = "ML")
```

```
## Series: et.log
## ARIMA(0,1,0)(1,1,1)[12]
##
## Coefficients:
##         sar1     sma1
##       0.1814  -0.4597
## s.e.  0.2458   0.2206
##
## sigma^2 = 2.4e-05:  log likelihood = 613.72
## AIC=-1221.45   AICc=-1221.29   BIC=-1212.32
```

The 95% confidence interval for the SAR1 coefficient contains 0; thus, the coefficient is unreliable. Revert back to $P = 0$

Introducing Non-seasonal Components

```
# p = 0, q = 1, P = 0, Q = 1
Arima(et.bc, order = c(0,1,1), seasonal = list(order = c(0,1,1), period = 12), method = "ML")
```

```
## Series: et.bc
## ARIMA(0,1,1)(0,1,1)[12]
##
## Coefficients:
##          ma1     sma1
##       0.0765  -0.2963
## s.e.  0.0808   0.0927
##
## sigma^2 = 7.168e-05:  log likelihood = 525.29
## AIC=-1044.58   AICc=-1044.42   BIC=-1035.45
```

```
# p = 1, q = 0, P = 0, Q = 1
Arima(et.bc, order = c(0,1,1), seasonal = list(order = c(0,1,1), period = 12), method = "ML")
```

```
## Series: et.bc
## ARIMA(0,1,1)(0,1,1)[12]
##
## Coefficients:
##          ma1     sma1
##       0.0765  -0.2963
## s.e.  0.0808   0.0927
##
## sigma^2 = 7.168e-05:  log likelihood = 525.29
## AIC=-1044.58   AICc=-1044.42   BIC=-1035.45
```

Introducing any non-seasonal MA or AR components increases the AICc.

Models

Our best models are:

```
# p=0, q=0, P=0, Q=1
Arima(et.log, order = c(0,1,0), seasonal = list(order = c(0,1,1), period = 12), method = "ML")
```

```
## Series: et.log
## ARIMA(0,1,0)(0,1,1)[12]
##
## Coefficients:
##          sma1
##       -0.2943
## s.e.   0.0898
##
## sigma^2 = 2.392e-05:  log likelihood = 613.48
## AIC=-1222.97   AICc=-1222.89   BIC=-1216.88
```

```
# p=0, q=0, P=1, Q=0
Arima(et.log, order = c(0,1,0), seasonal = list(order = c(1,1,0), period = 12), method = "ML")
```

```
## Series: et.log
## ARIMA(0,1,0)(1,1,0)[12]
##
## Coefficients:
##          sar1
##       -0.2364
## s.e.   0.0837
##
## sigma^2 = 2.424e-05:  log likelihood = 612.53
## AIC=-1221.06   AICc=-1220.98   BIC=-1214.98
```

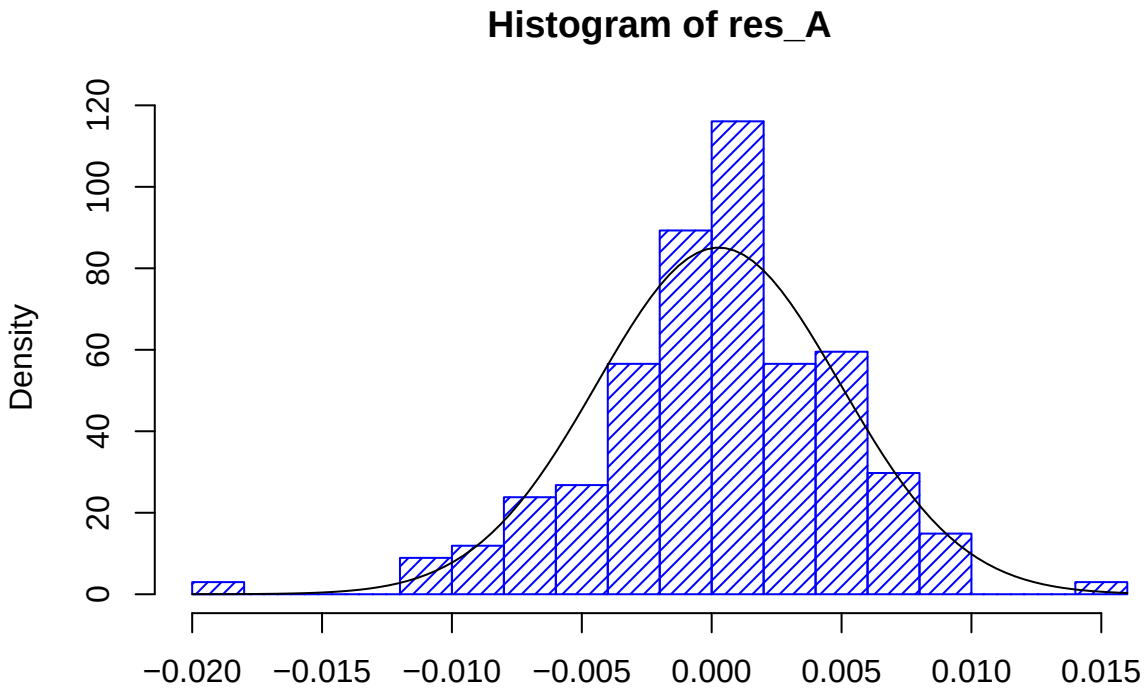
$$\text{(A) } \nabla_1\nabla_{12}\ln(U_t) = (1 - 0.2943_{(0.0898)}B^{12})Z_t$$
$$\hat{\sigma}_Z^2 = 2.392\text{e-}05$$
$$\text{(B) } \nabla_1\nabla_{12}\ln(U_t)(1 - 0.2364_{(0.0837)}B^{12}) = Z_t$$
$$\hat{\sigma}_Z^2 = 2.424\text{e-}05$$

Model (A) is stationary because it is a pure MA model. Model (A) is invertible because $|\Theta_1| = 0.2943 < 1$. Model (B) is invertible because it is a pure AR model. Model (B) is stationary because $|\Phi_1| = 0.2364 < 1$.

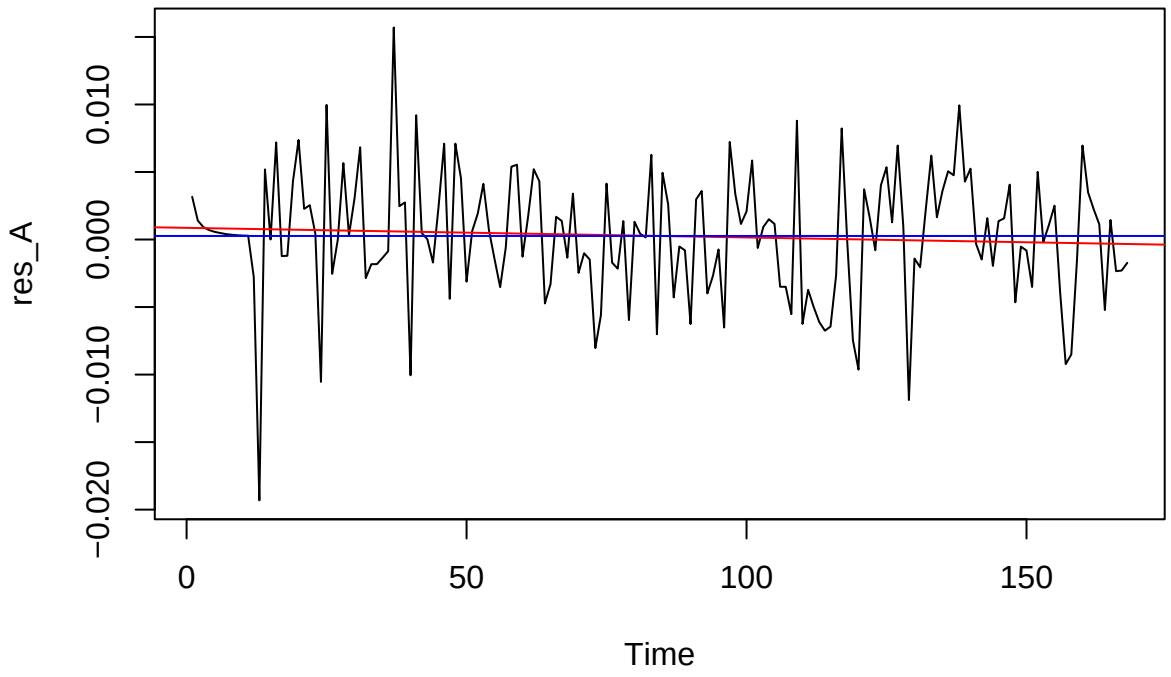
Diagnostic Checking

Model (A)

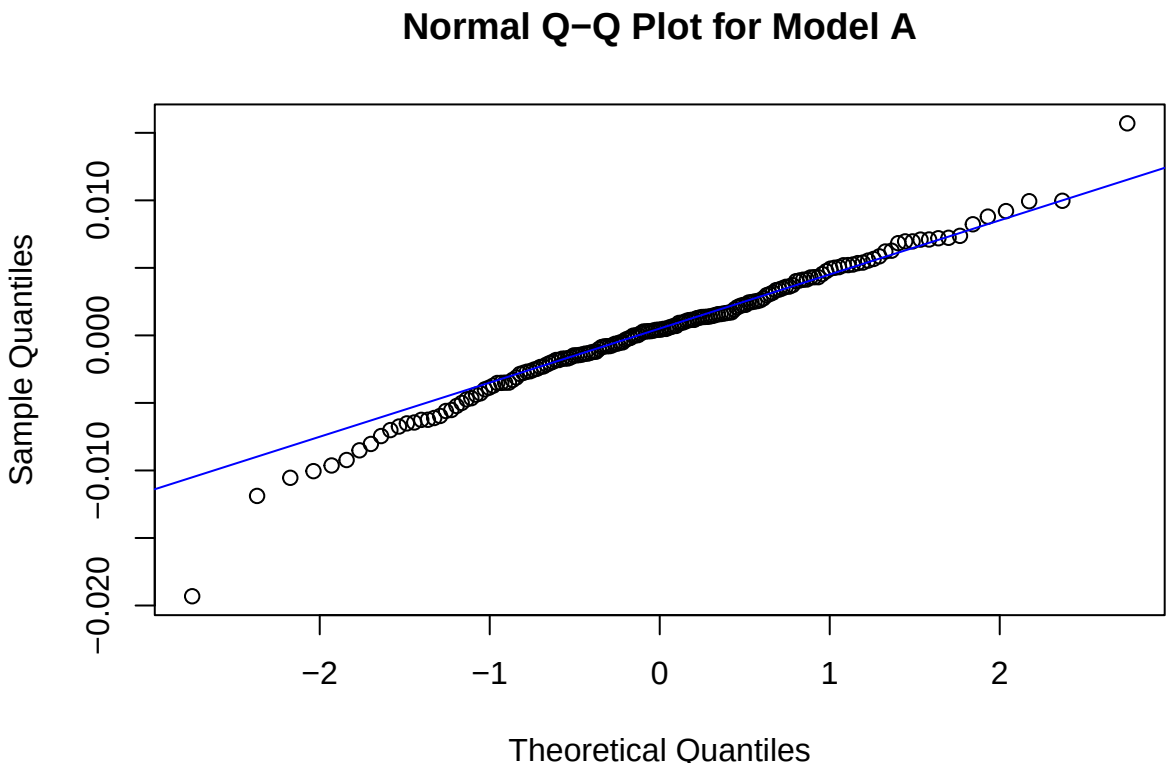
```
fit_A <- arima(et.log, order = c(0,1,0), seasonal = list(order = c(0,1,1), period = 12),
               method = "ML")
res_A <- residuals(fit_A)
hist(res_A,density=20,breaks=20, col="blue", xlab="", prob=TRUE) # histogram
m <- mean(res_A)
std <- sqrt(var(res_A))
curve(dnorm(x,m,std), add=TRUE ) # density curve
```




```
plot.ts(res_A) # plotting residuals
fitt_A <- lm(res_A ~ as.numeric(1:length(res_A))); abline(fitt_A, col="red") # trend line
abline(h=mean(res_A), col="blue") # mean line
```

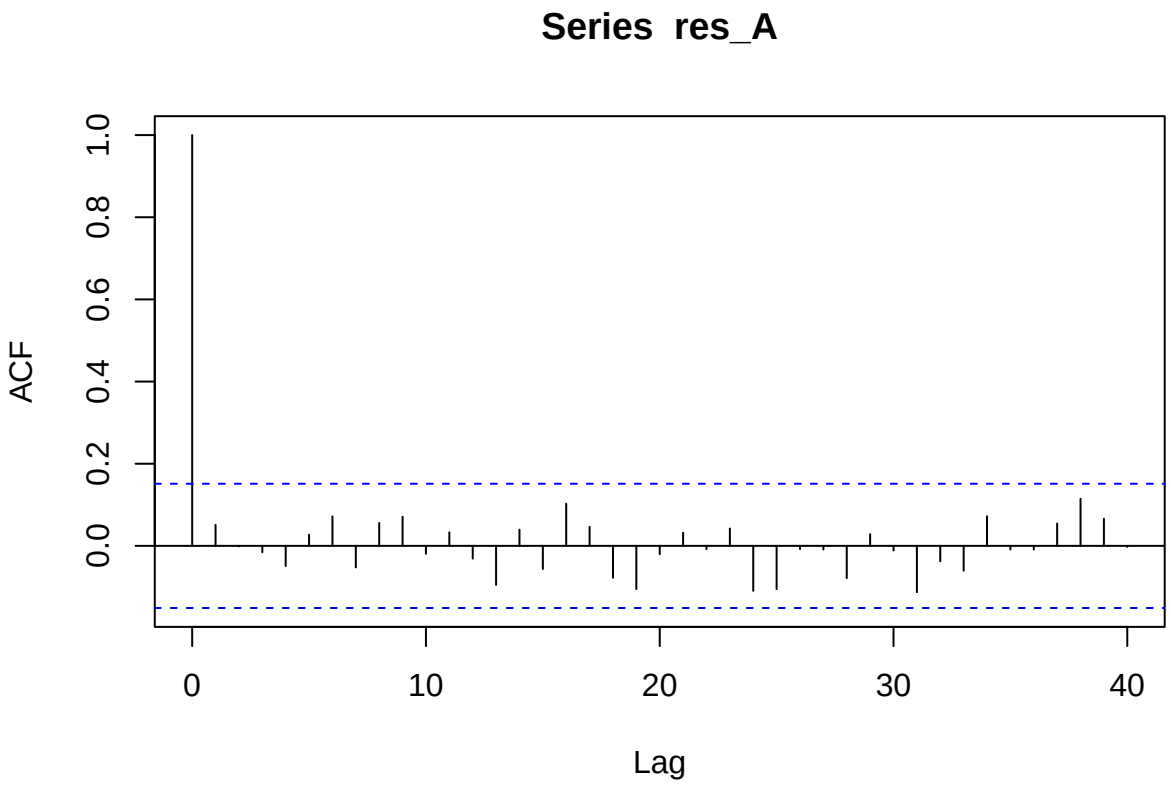



```
qqnorm(res_A,main= "Normal Q-Q Plot for Model A") #qq plot
qqline(res_A,col="blue") #qq line
```

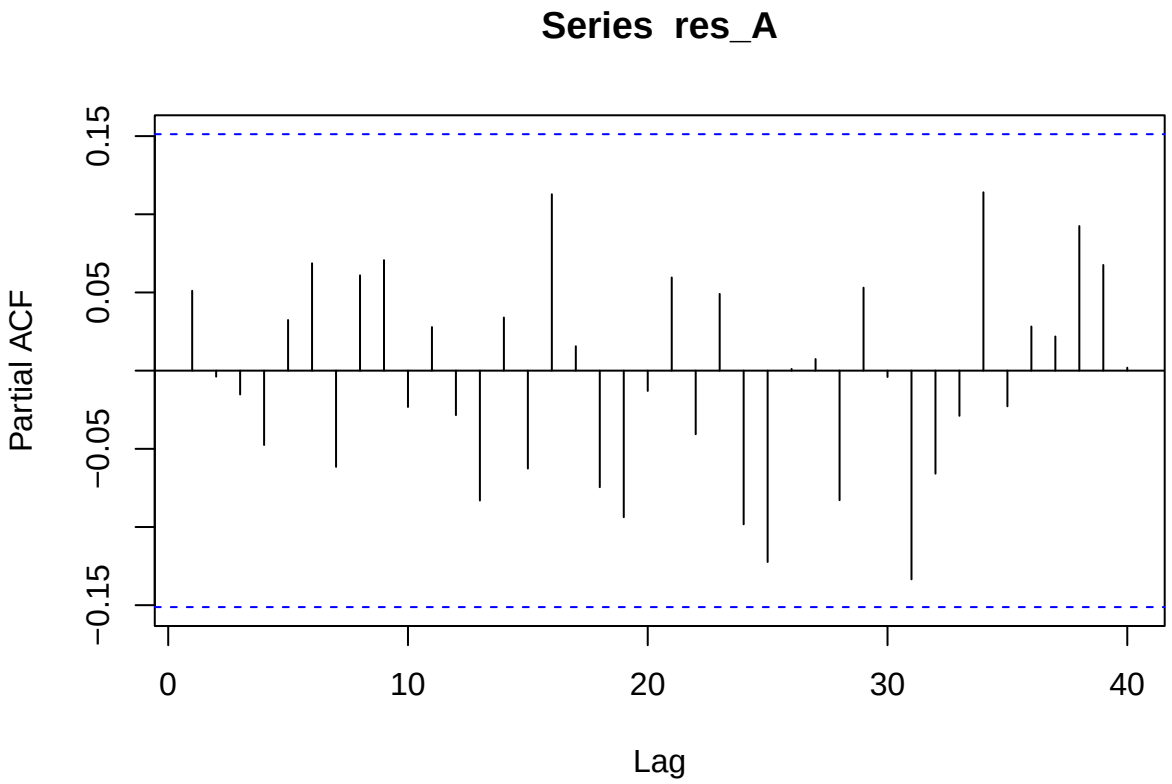


The histogram of the residuals from model A is slightly skewed. Nothing looks out of the ordinary in residual plot and Q-Q plot.

```
acf(res_A, lag.max=40)
```



```r
pacf(res_A, lag.max=40)
```



All residual ACFs and PACFs are contained within the 95% confidence interval.

```r
shapiro.test(res_A)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_A
## W = 0.98082, p-value = 0.02006
```

```r
Box.test(res_A, lag = 12, type = c("Box-Pierce"), fitdf = 1)
```

```
##
##  Box-Pierce test
##
## data:  res_A
## X-squared = 4.1078, df = 11, p-value = 0.9667
```

```r
Box.test(res_A, lag = 12, type = c("Ljung-Box"), fitdf = 1)
```

```
##
##  Box-Ljung test
##
## data:  res_A
## X-squared = 4.3311, df = 11, p-value = 0.9592
```

```r
Box.test(res_A^2, lag = 12, type = c("Ljung-Box"), fitdf = 0)
```

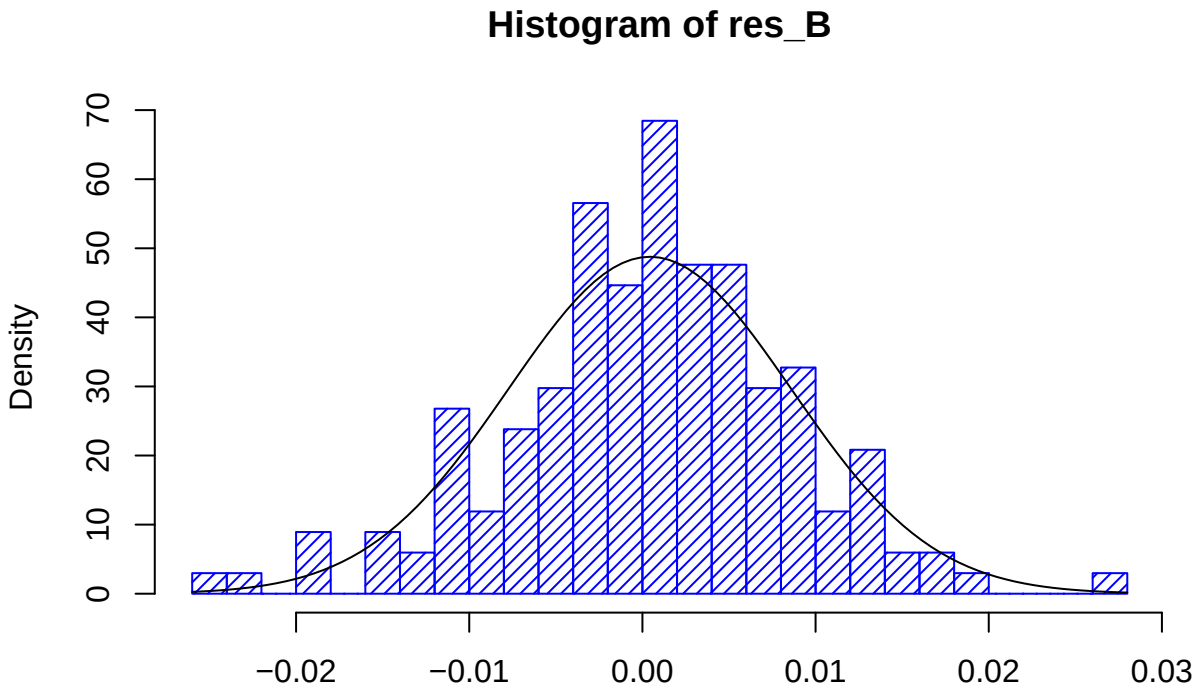
```
##
```

```
##  Box-Ljung test
##
## data:  res_A^2
## X-squared = 11.077, df = 12, p-value = 0.5224
```

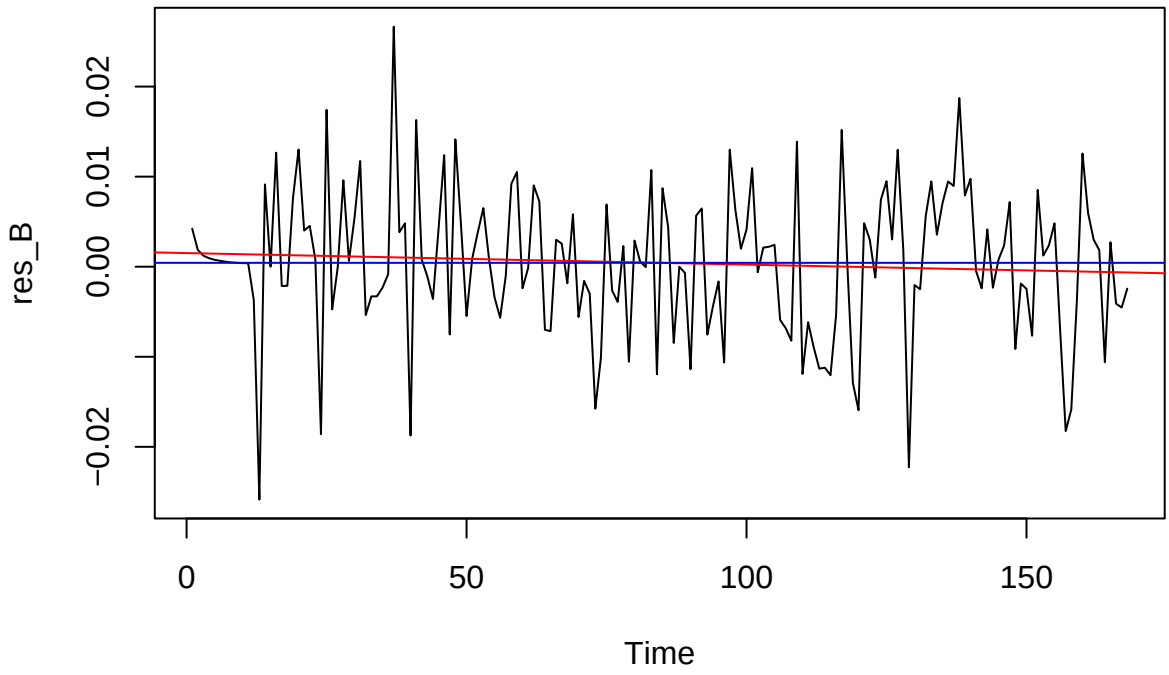
Model (A) rejects the null hypothesis of normality since $p = 0.02006 < 0.05$. Whereas every other p-value is greater than 0.05, indicating no autocorrelation and no heteroskedacity.

Model (B)

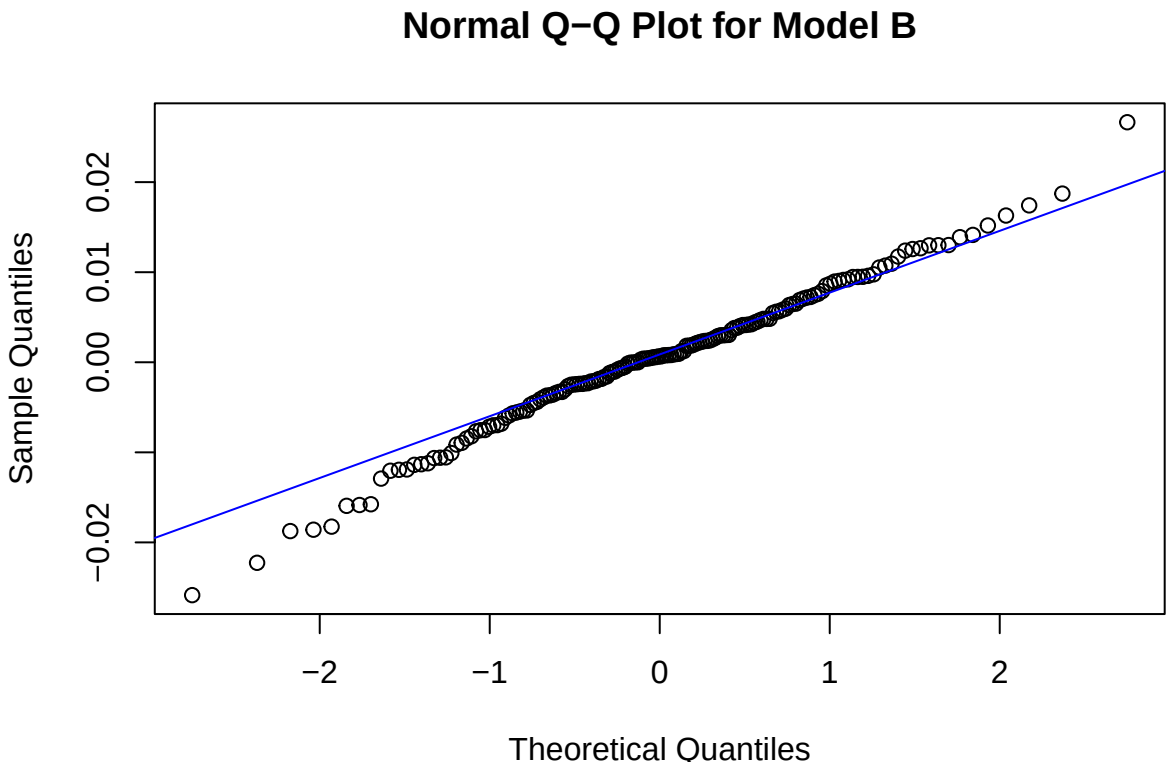
```
fit_B <- arima(et.bc, order = c(0,1,0), seasonal = list(order = c(1,1,0), period = 12), method = "ML")
res_B <- residuals(fit_B)
hist(res_B,density=20,breaks=20, col="blue", xlab="", prob=TRUE)
m <- mean(res_B)
std <- sqrt(var(res_B))
curve(dnorm(x,m,std), add=TRUE )
```




```
plot.ts(res_B)
fitt_B <- lm(res_B ~ as.numeric(1:length(res_B))); abline(fitt_B, col="red")
abline(h=mean(res_B), col="blue")
```

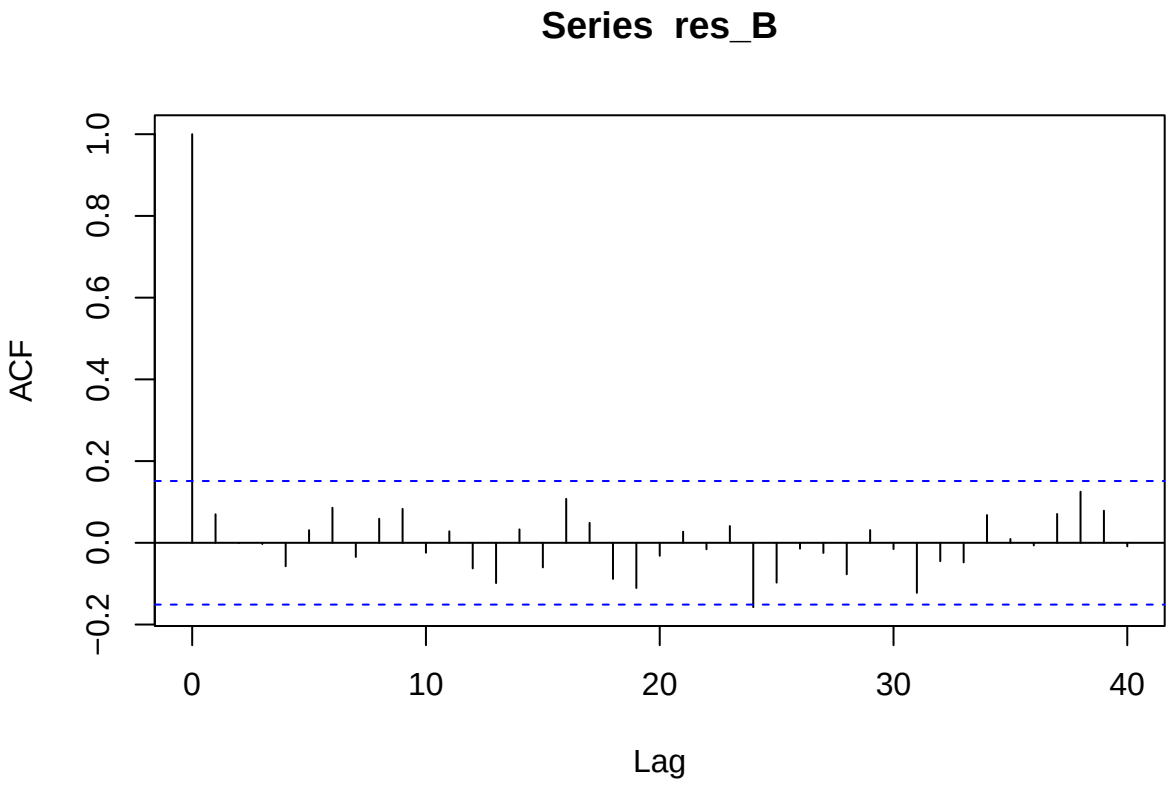



```
qqnorm(res_B,main= "Normal Q-Q Plot for Model B")
qqline(res_B,col="blue")
```

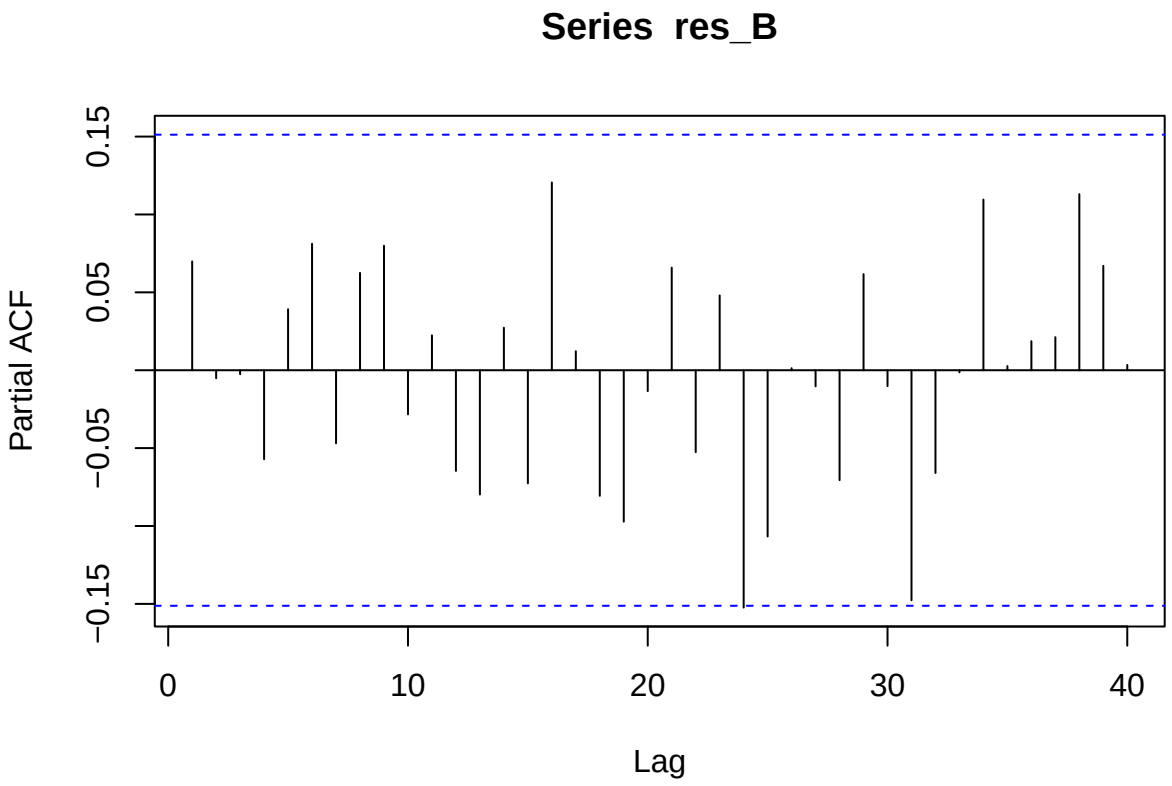


The distribution of the residuals of model B look normal (histogram). Nothing looks out of the ordinary in residual plot and Q-Q plot.

```
acf(res_B, lag.max=40)
```




```
pacf(res_B, lag.max=40)
```



All residual ACFs and PACFs are contained within the 95% confidence interval.

```
shapiro.test(res_B)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_B
## W = 0.98943, p-value = 0.2436
```

```
Box.test(res_B, lag = 12, type = c("Box-Pierce"), fitdf = 1)
```

```
##
##  Box-Pierce test
##
## data:  res_B
## X-squared = 5.5964, df = 11, p-value = 0.8989
```

```
Box.test(res_B, lag = 12, type = c("Ljung-Box"), fitdf = 1)
```

```
##
##  Box-Ljung test
##
## data:  res_B
## X-squared = 5.9044, df = 11, p-value = 0.8797
```

```
Box.test(res_B^2, lag = 12, type = c("Ljung-Box"), fitdf = 0)
```

```
##
##  Box-Ljung test
##
## data:  res_B^2
## X-squared = 10.724, df = 12, p-value = 0.5527
```

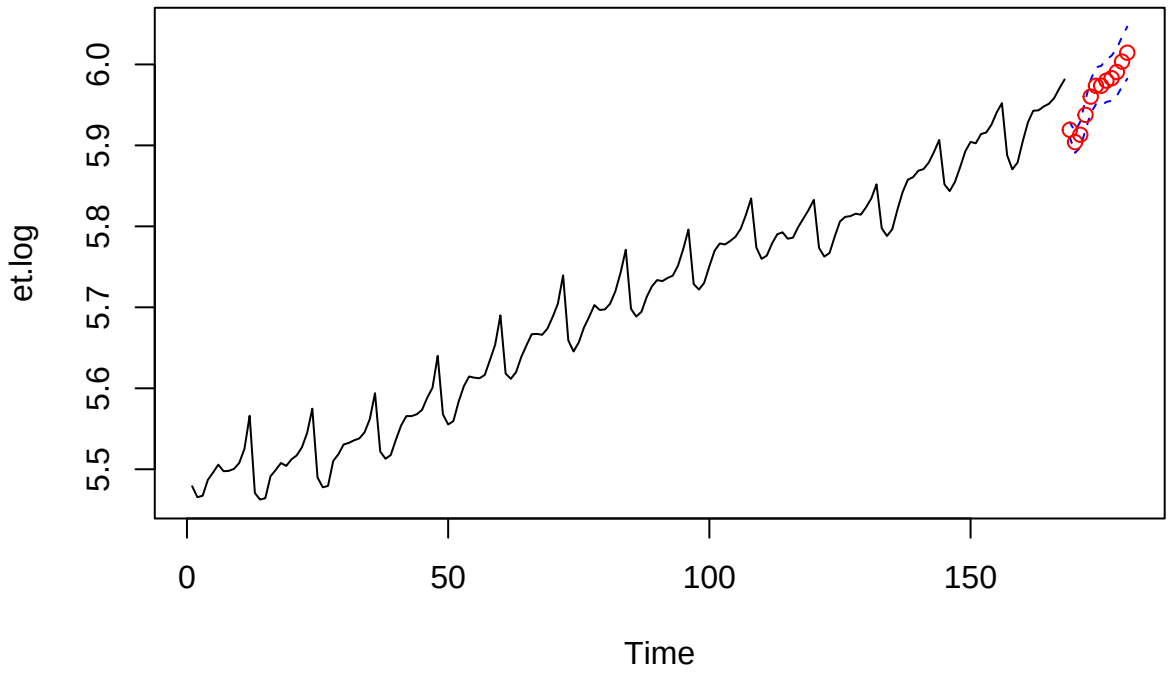
Model (B) fails to reject every null hypothesis tested, indicating normality, no autocorrelation, and no heteroskedacity.

Since model (A) failed the Shapiro Normality test and model (B) passed every diagnostic check, we choose model (B) as the final model.

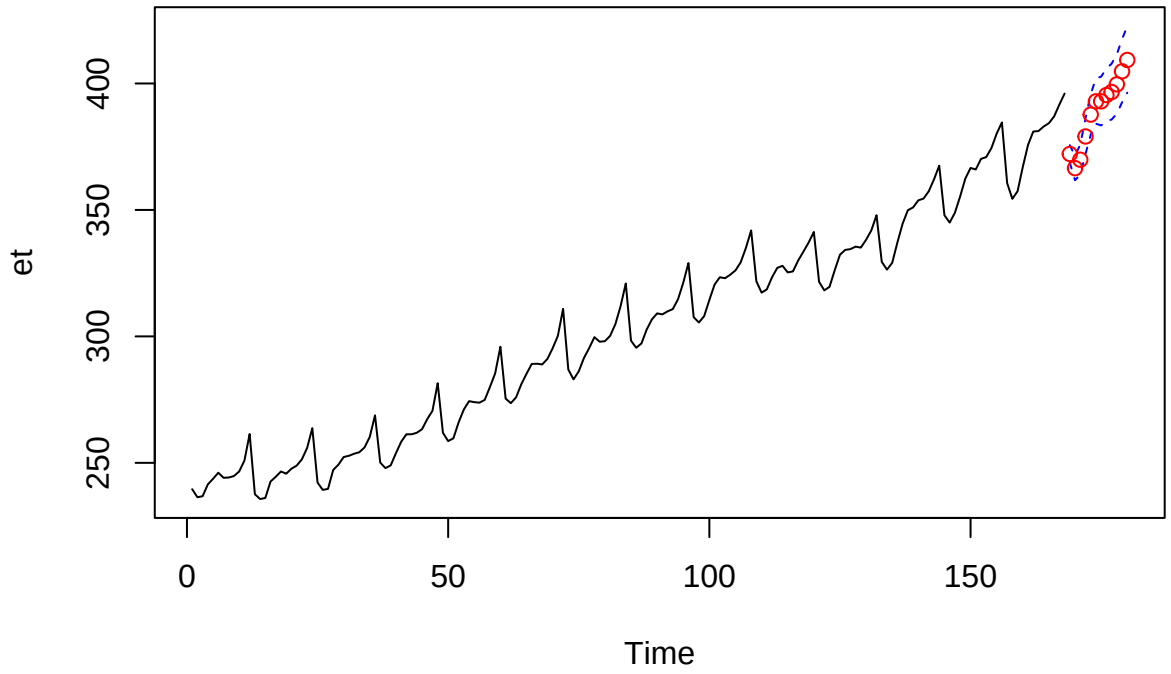
Forecasting

```
fit.B <- arima(et.log, order=c(0,1,0), seasonal = list(order = c(1,1,0), period = 12), method="ML")
```

```
pred.tr <- predict(fit.B, n.ahead = 12)
U.tr= pred.tr$pred + 2*pred.tr$se # upper bound of prediction interval
L.tr= pred.tr$pred - 2*pred.tr$se # lower bound
ts.plot(et.log, xlim=c(1,length(et.log)+12), ylim = c(min(et.log),max(U.tr)))
lines(U.tr, col="blue", lty="dashed")
lines(L.tr, col="blue", lty="dashed")
points((length(et.log)+1):(length(et.log)+12), pred.tr$pred, col="red")
```

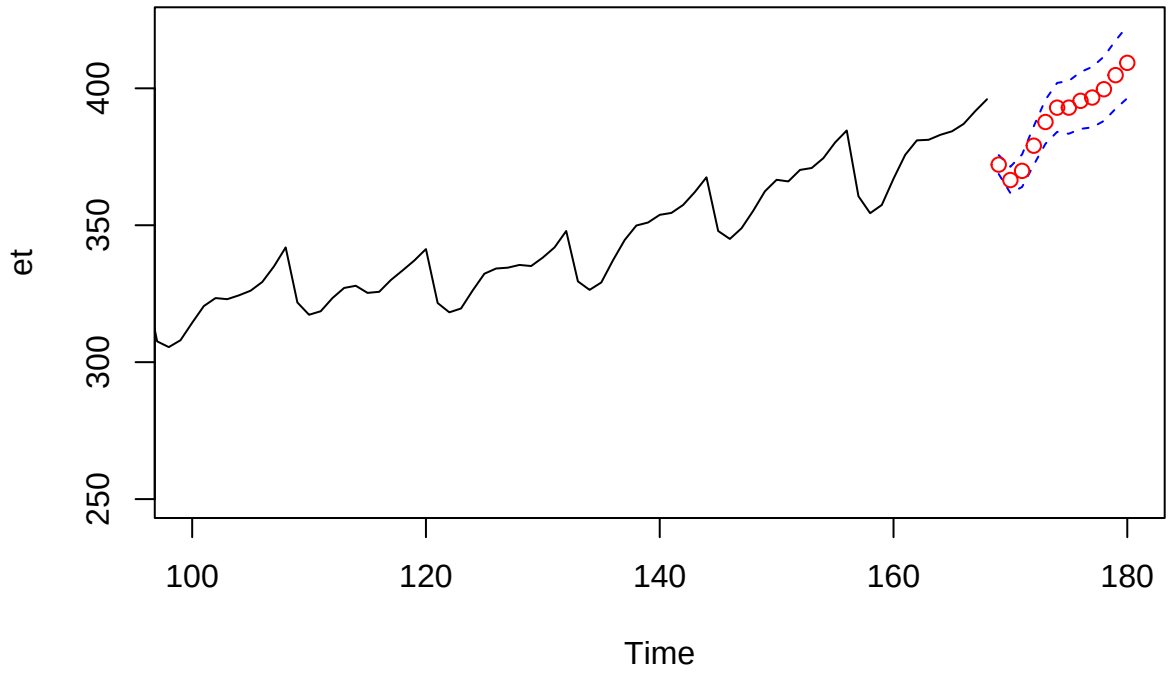



```
pred.orig <- exp(pred.tr$pred)
U= exp(U.tr)
L= exp(L.tr)
ts.plot(et, xlim=c(1,length(et)+12), ylim = c(min(et),max(U)))
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
points((length(et)+1):(length(et)+12), pred.orig, col="red")
```




```
ts.plot(et, xlim = c(100,length(et)+12), ylim = c(250,max(U)))
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
```

```
points((length(et)+1):(length(et)+12), pred.orig, col="red")
```



Conclusion

The goal of this project was to model and forecast monthly employment in the wholesale and retail sectors in Wisconsin from 1961 to 1975. We applied a log transformation to stabilize the variance and seasonal and regular differencing to achieve stationarity. Visual diagnostics and autocorrelation plots suggested a SARIMA structure with one seasonal difference and one regular difference.

Two candidate models were fitted, SARIMA(0,1,0)(0,1,1)[12] (Model A) and SARIMA(0,1,0)(1,1,0)[12] (Model B).

Model B

$$\nabla_1 \nabla_{12} \ln(U_t)(1 - 0.2364_{(0.0837)} B^{12}) = Z_t$$

was selected as the final model due to better residual diagnostics. While both models were invertible and stationary, Model A failed the Shapiro-Wilk normality test, whereas Model B satisfied all diagnostic checks, including tests for autocorrelation, heteroskedasticity, and normality.

Using Model B, we produced 12-month forecasts with 95% prediction intervals. The forecasts were back-transformed to the original scale and plotted alongside the original series. The model effectively captures the seasonal fluctuations and long-term upward trend in employment, making it a reliable tool for short-term forecasting in this field.

Thank you to Professor Feldman for making concepts in this class clear and for providing a very useful example as a reference!

References

Time Series Data Library (TSDL). (n.d.). Dataset #544: “Monthly employees wholesale/retail Wisconsin –61–75 R.B. Miller”. Retrieved via tsdl R package.

Professor Raisa Feldman. Lectures 1-15.