# Week 03 Homework


### *Brian Ritz*


### *Monday, May 04, 2015*

The Excel data file is attached.

1. Use datasets from 1955 to 1968 to build an ARMA or ARIMA models for UN and GDP.
2. Justify why you chose (ARMA or ARIMA) one over the other. Note there will be 2 models, one for UN and another for GDP.
3. Use the chosen UN and GDP models to forecast the UN and the GDP for 1969.
4. Compare your forecasts with the actual values using error = actual - estimate and plot the errors.
5. Calculate the Sum of squared(error) for each UN and GDP models.
6. Regression - build regression models that use:

6a. UN as the independent variable and GDP as the dependent variable - use data from 1955 to 1968 to build the model. Forecast for 1969 and plot the errors and calculate the sum of squared(error) as previously. 6b. GDP as the independent variable and UN as the dependent variable - use data from 1955 to 1968 to build the model. Forecast for 1969 and plot the errors and calculate the sum of squared(error) as previously. 6c. Compare the 2 models - any reason to believe which should be the independent and the dependent variables.

```r
library(tseries)
```

```
## Warning: package 'tseries' was built under R version 3.1.3
```

```r
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 3.1.3
```

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

```
## Warning: package 'zoo' was built under R version 3.1.3
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
## Loading required package: timeDate
```

```
## Warning: package 'timeDate' was built under R version 3.1.3
```

```
## This is forecast 5.9
```

```r
library(ggplot2)
library(zoo)
library(xts)
```

```
## Warning: package 'xts' was built under R version 3.1.3
```

Import the data first

```
# import the data

raw.data <- read.csv("Unemployment_GDP_UK.csv", header=T)

unemployment <- ts(data=raw.data$UN, frequency=4, start=c(1955, 1))
gdp <- ts(data=raw.data$GDP, frequency=4, start=c(1955, 1))

training.gdp <- window(gdp, end=c(1968, 4))
training.unemployment <- window(unemployment, end=c(1968, 4))
```

### 1. Use datasets from 1955 to 1968 to build an ARMA or ARIMA models for UN and GDP.

UN Model

Test for stationarity – Augmented Dickey Fuller Test

```
adf.test(training.unemployment)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  training.unemployment
## Dickey-Fuller = -3.3336, Lag order = 3, p-value = 0.07538
## alternative hypothesis: stationary
```

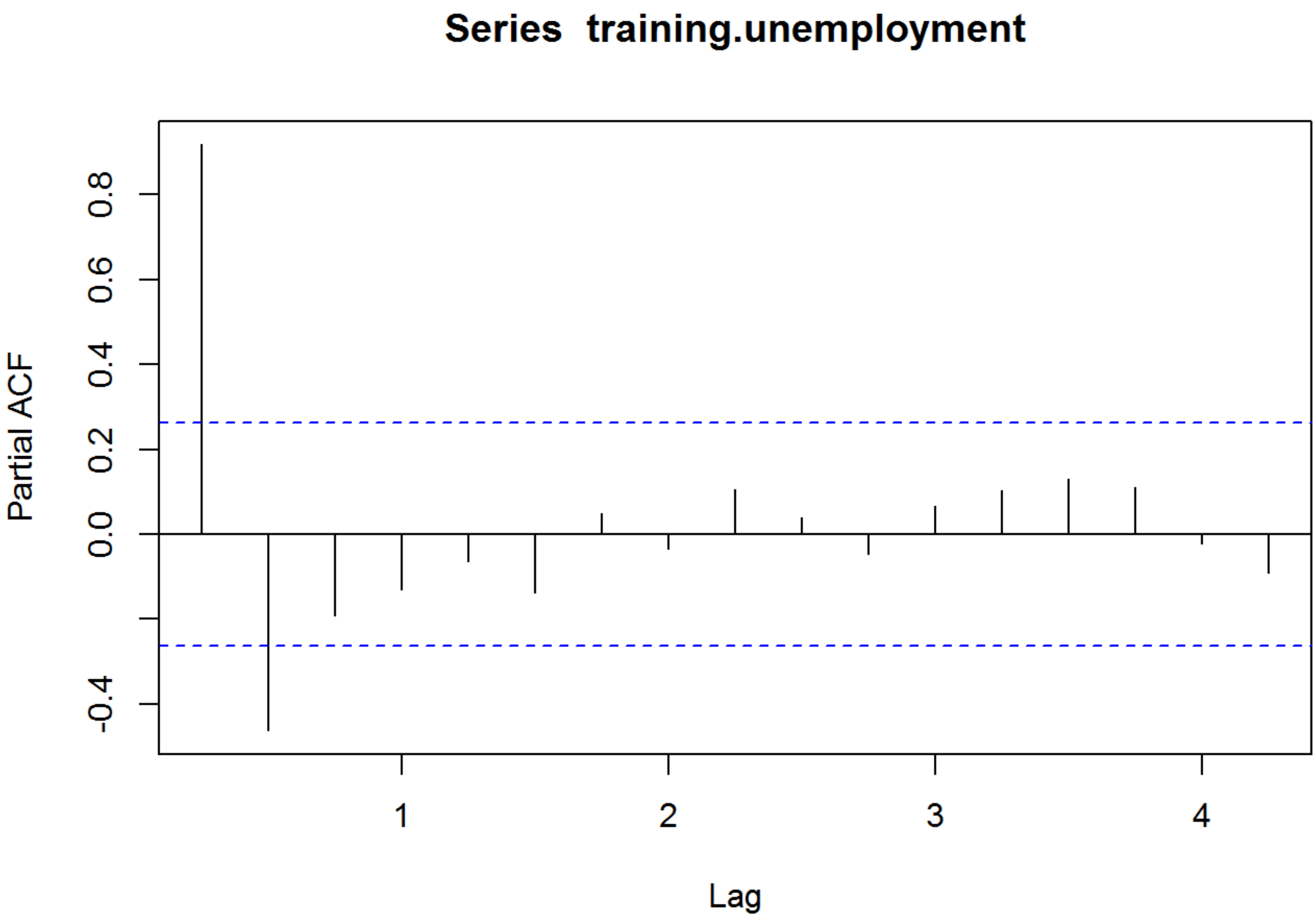
```
adf.test(training.gdp)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  training.gdp
## Dickey-Fuller = -2.9551, Lag order = 3, p-value = 0.1895
## alternative hypothesis: stationary
```

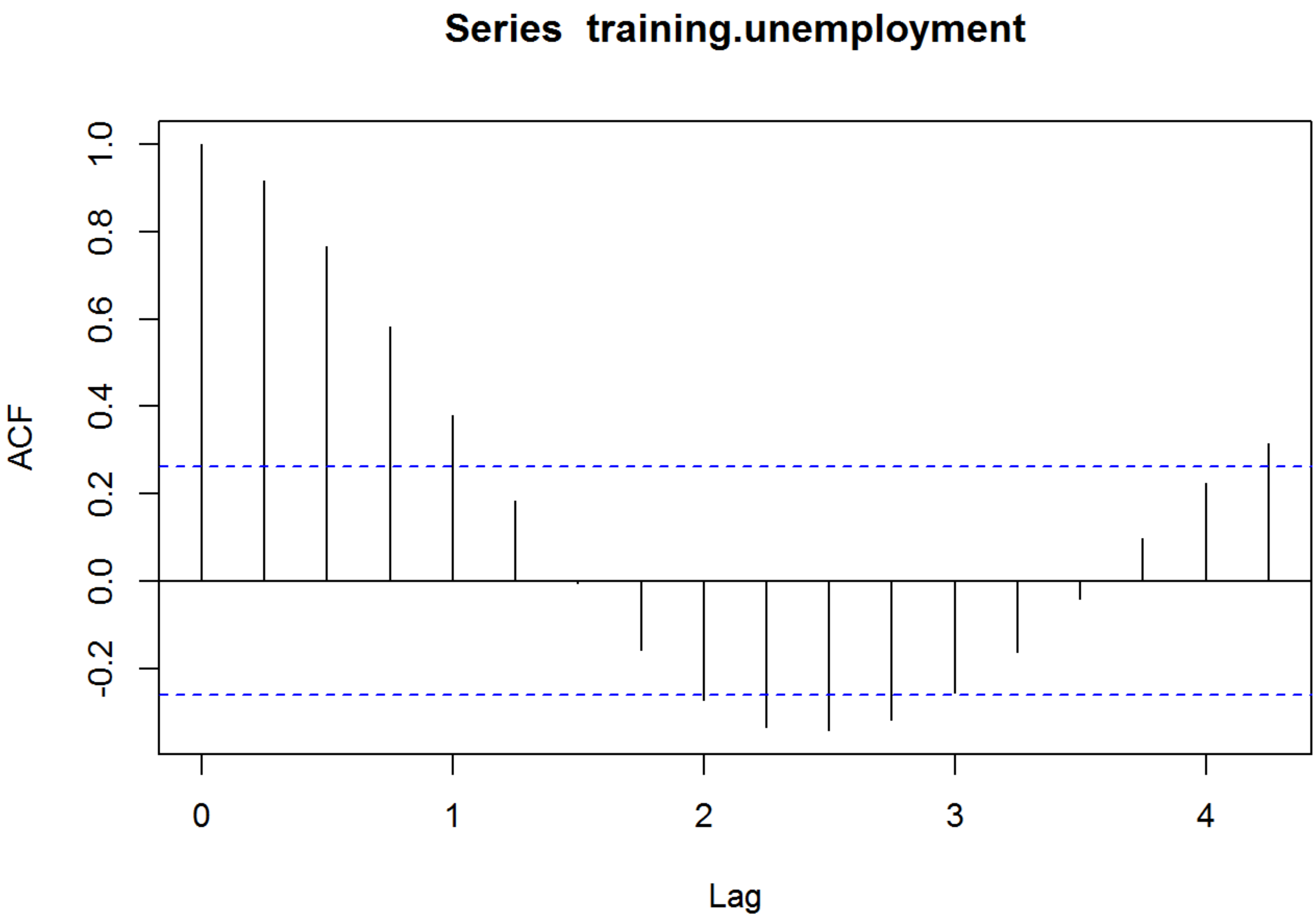
We see from the Augmented Dickey Fuller Test that we can reject the null hypothesis that the process is not stationary for the unemployment at a 10% confidence level, but we cannot reject the null hypothesis for the gdp. So, we cannot assume that gdp is stationary, so we will take the difference of the gdp and use that in our analysis. This is why we choose an arima model for GDP.

What order should the difference parameter for the unemployeemnt arima be?

```
# partial acf is for the ar parameter
# acf is for the moving average parameter
pacf(training.unemployment)
```




```
acf(training.unemployment)
```




The q parameter for the arma model for unemployment will be 4 and the p parameter will be 2.

Find parameters for the GDP model:

```
# we will difference it until we can reject the null hypothesis that it is not stationary
```

```r
adf.test(diff(training.gdp))
```

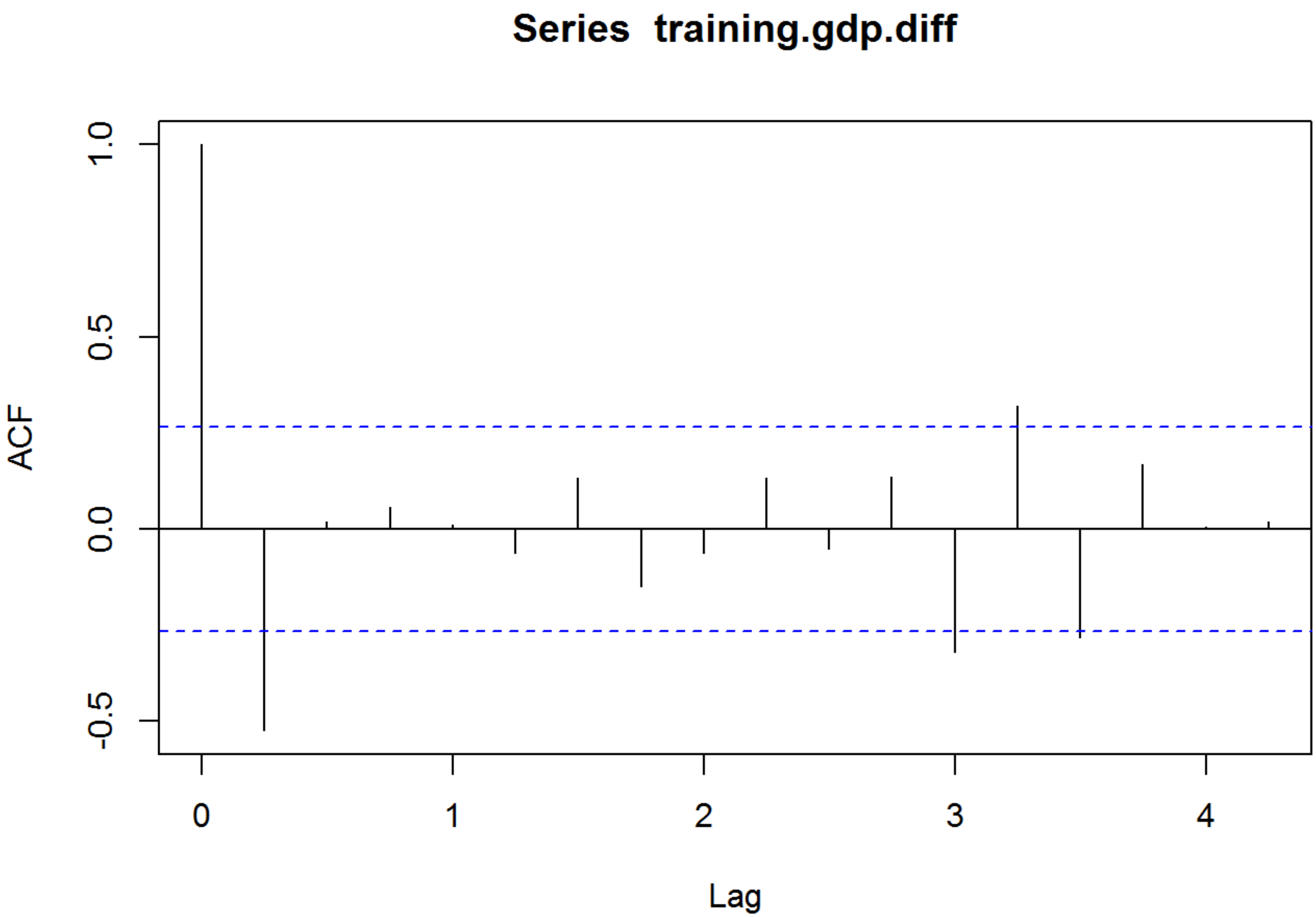
```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff(training.gdp)
## Dickey-Fuller = -2.8173, Lag order = 3, p-value = 0.2452
## alternative hypothesis: stationary
```

```r
adf.test(diff(training.gdp, differences=2))
```

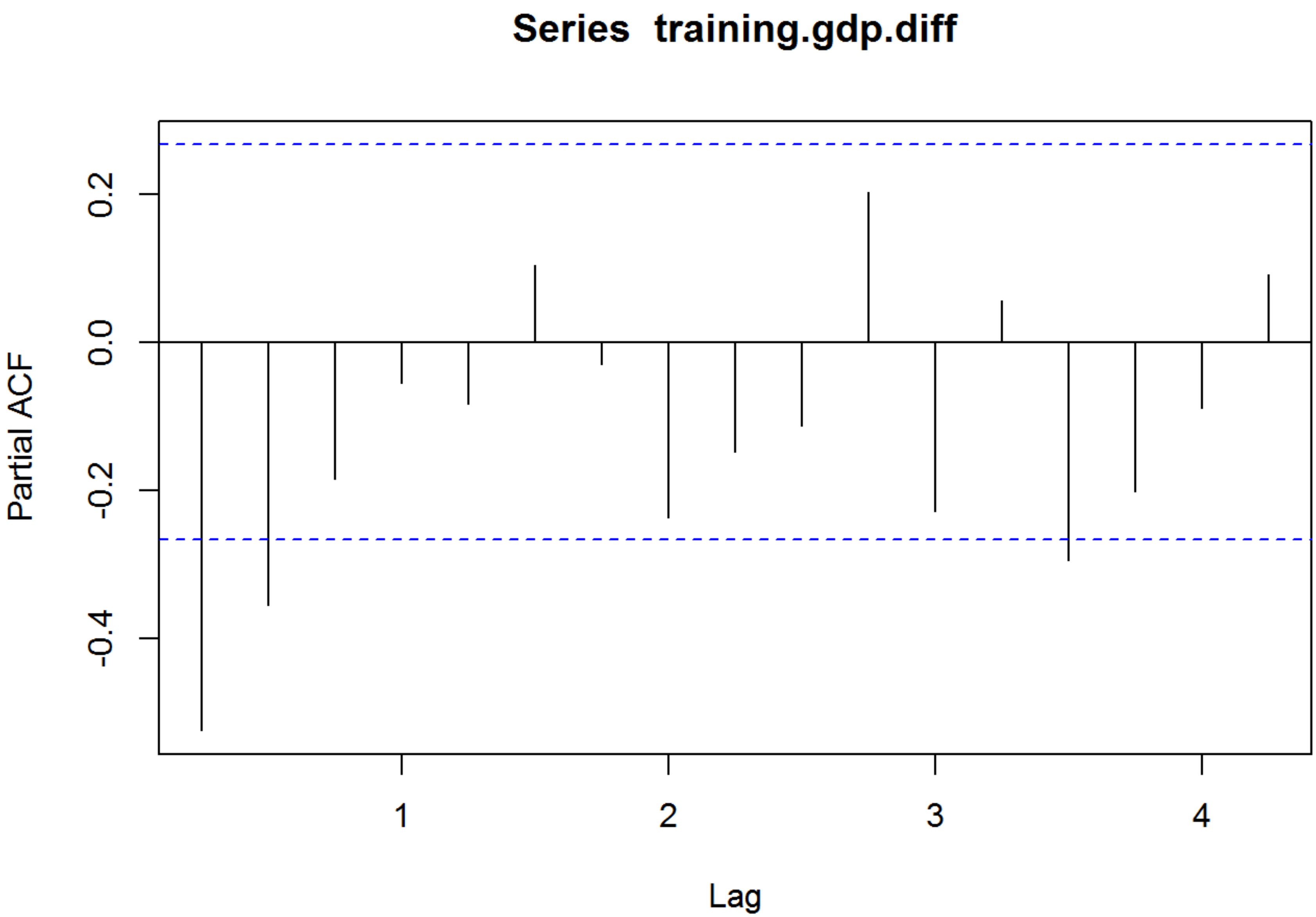
```
## Warning in adf.test(diff(training.gdp, differences = 2)): p-value smaller
## than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff(training.gdp, differences = 2)
## Dickey-Fuller = -4.9224, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

```r
# what is the order of the p parameter for gdp
training.gdp.diff <- diff(training.gdp, differences=2)
acf(training.gdp.diff)
```




```
pacf(training.gdp.diff)
```




A difference of order two rejects the hypothesis that the process is not stationary with a p-value of < .01. So, we will use the second order differencing when creating the arima model. Given the plots, the p(AR) parameter will be 1, and the q(MA) parameter will be 1.

We now build the ARMA and ARIMA models:

```
unemployment.model <- arima(training.unemployment, order=c(1,0,4))
gdp.model <- arima(training.gdp, order=c(1,2,1))
```

**2. Justify why you chose (ARMA or ARIMA) one over the other. Note there will be 2 models, one for UN and another for GDP.** I chose the ARMA model for the Unemployment data because we had indications that it was stationary with a p-value of 0.07 on the Augmented Dickey-Fuller test. The ARIMA model was chosen for the GDP model becuase we could not rule out that it was not stationary with a 10% confidence level. When I took the second order difference, we could rule out that it was stationary and therefore I conclude that an arima model with 2nd order difference term is appropriate.

**3. Use the chosen UN and GDP models to forecast the UN and the GDP for 1969.**

```
# forecast for GDP
gdp.preds.1969 <- predict(gdp.model, n.ahead = 4)

# forecast for Unemployment
unemployment.preds.1969 <- predict(unemployment.model, n.ahead=4)
print(gdp.preds.1969$pred)
```

```
##          Qtr1     Qtr2     Qtr3     Qtr4
## 1969 118.5719 119.3098 120.0458 120.7817
```

```
print(unemployment.preds.1969$pred)
```

```
##          Qtr1     Qtr2     Qtr3     Qtr4
## 1969 511.3963 488.6112 466.8458 453.4144
```

### 4. Compare your forecasts with the actual values using error = actual - estimate and plot the errors.

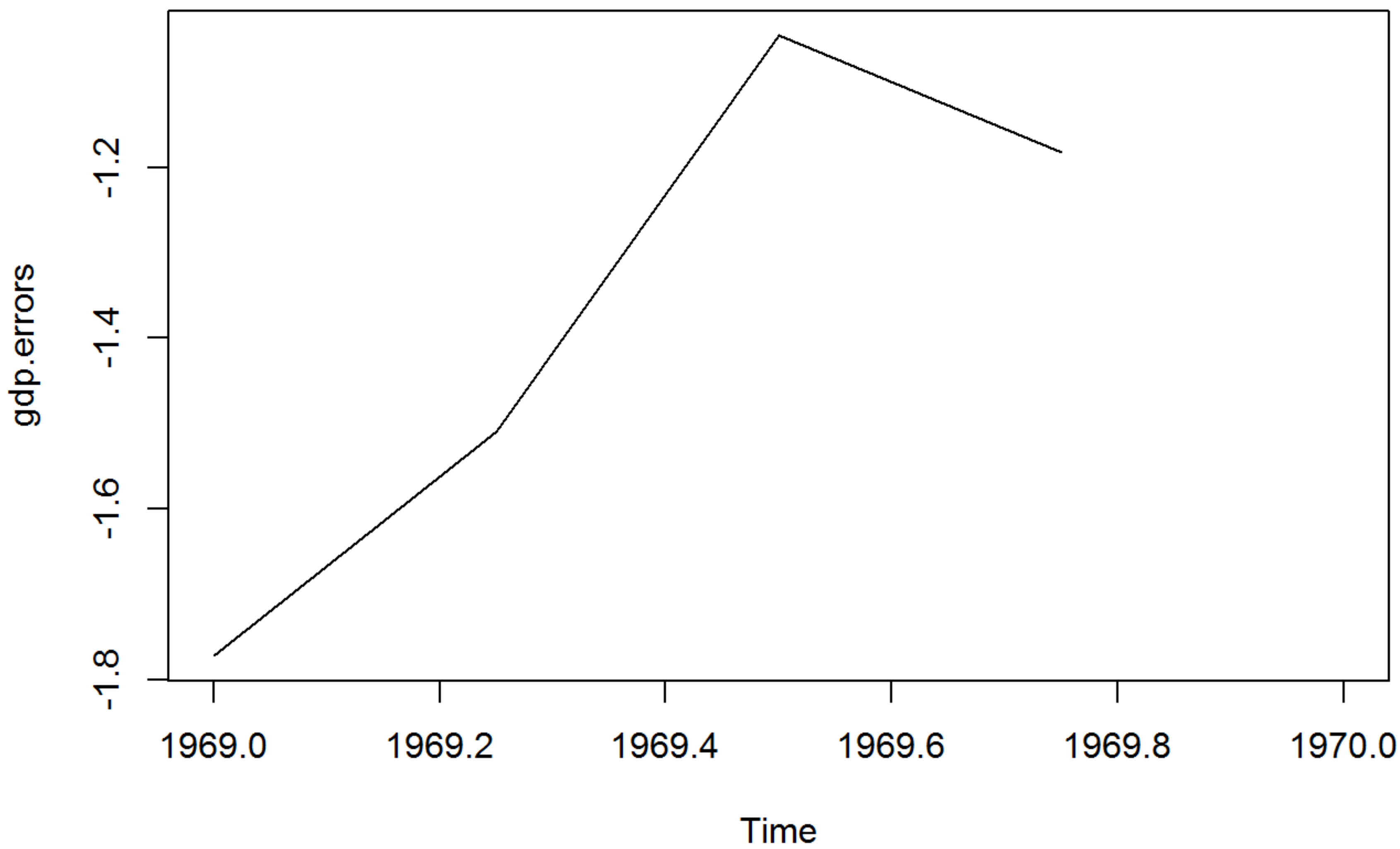
```
# comparison of actuals

(gdp.errors <- window(gdp, start=c(1969,1), end=c(1969,4)) - gdp.preds.1969$pred)
```

```
##            Qtr1      Qtr2      Qtr3      Qtr4
## 1969 -1.771947 -1.509802 -1.045807 -1.181712
```

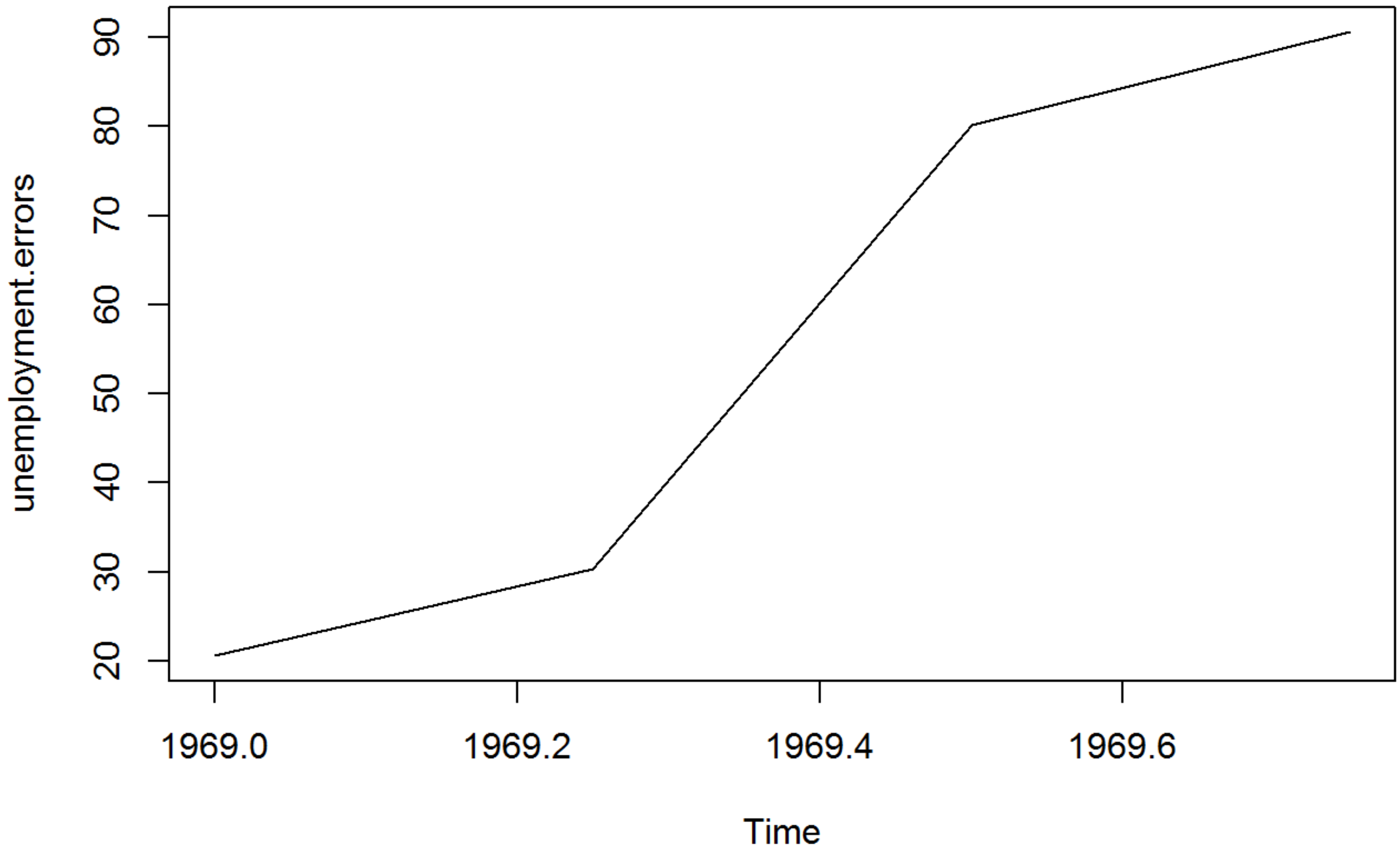
```
(unemployment.errors <- window(unemployment, start=c(1969,1), end=c(1969,4)) - unemployment.preds.1969$pred
)
```

```
##          Qtr1     Qtr2     Qtr3     Qtr4
## 1969 20.60371 30.38881 80.15416 90.58556
```

```
# plot the errors
par(mfrow=c(1,1))
plot(gdp.errors, xlim=c(1969, 1970))
```

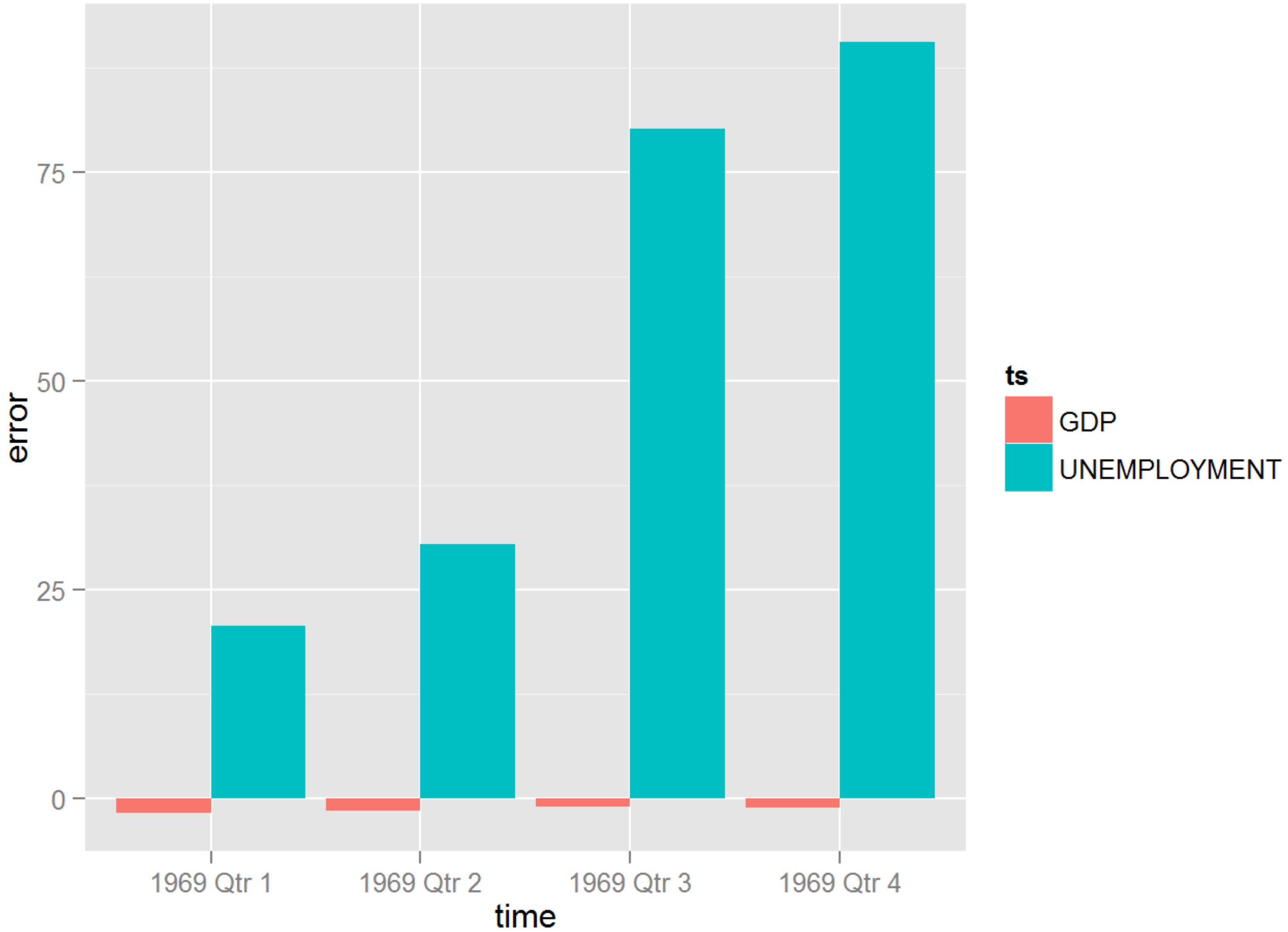



```
plot(unemployment.errors)
```




```
data.for.plot <- data.frame(ts=c(rep("GDP", times=4), rep("UNEMPLOYMENT", times=4)), time=rep(c("1969 Qtr 1
", "1969 Qtr 2", "1969 Qtr 3", "1969 Qtr 4"), times=2), error = c(gdp.errors, unemployment.errors))
```

```
ggplot(data=data.for.plot, aes(x=time, y=error, fill=ts) ) + geom_bar(stat="identity", position="dodge")
```




#### 5. Calculate the Sum of squared(error) for each UN and GDP models.

```
(sse.unemployment<- sum(sapply(unemployment.errors, function(x) x**2)))
```

```
## [1] 15978.43
```

```
(sse.gdp <- sum(sapply(gdp.errors, function(x) x**2)))
```

```
## [1] 7.909455
```

### 6. Regression - build regression models that use:

### 6a. UN as the independent variable and GDP as the dependent variable - use data from 1955 to 1968 to build the model. Forecast for 1969 and plot the errors and calculate the sum of squared(error) as previously.

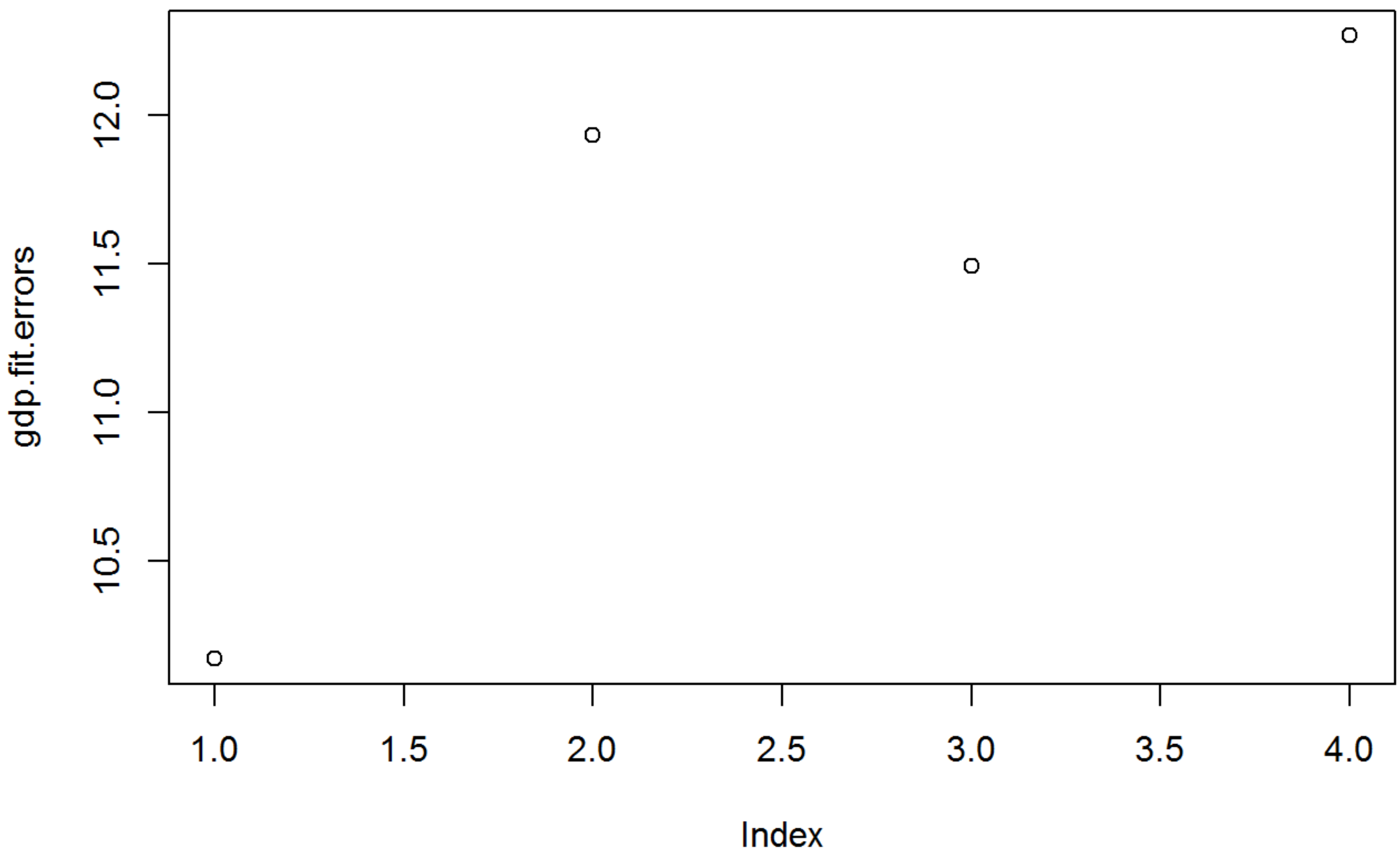
```
gdp.on.unemployment.fit <- lm(data = raw.data[raw.data$Year != 1969,], GDP ~ UN )
gdp.validation <-raw.data[raw.data$Year == 1969,]
lm.gdp.predictions <- predict(gdp.on.unemployment.fit, newdata=gdp.validation, type = "response")
```

Error Plot:

```
(gdp.fit.errors <- gdp.validation[, "GDP"] - lm.gdp.predictions)
```

```
##       57       58       59       60
## 10.16947 11.93207 11.48954 12.26553
```

```
plot(gdp.fit.errors)
```




SSE:

```
(gdp.fit.sse <- sum(sapply(gdp.fit.errors, function(x) x**2)))
```

```
## [1] 528.2449
```

**6b. GDP as the independent variable and UN as the dependent variable - use data from 1955 to 1968 to build the model. Forecast for 1969 and plot the errors and calculate the sum of squared(error) as previously.**

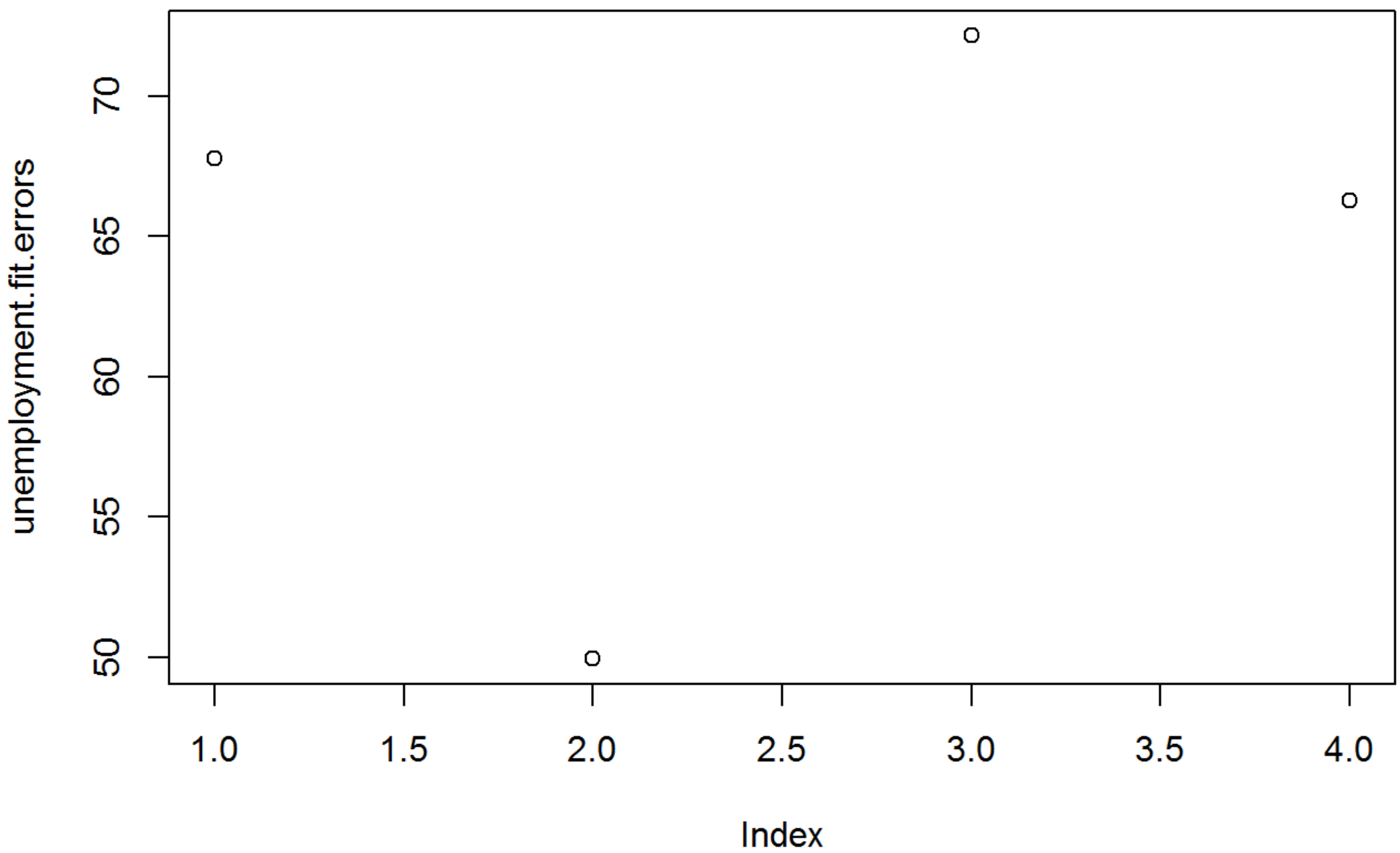
```
unemployment.on.gdp.fit <- lm(data = raw.data[raw.data$Year != 1969,], UN ~ GDP )
unemployment.validation <- raw.data[raw.data$Year == 1969,]
lm.unemployment.predictions <- predict(unemployment.on.gdp.fit, newdata=unemployment.validation, type = "re
sponse")
```

Error Plot:

```
(unemployment.fit.errors <- unemployment.validation[, "UN"] - lm.unemployment.predictions)
```

```
##       57       58       59       60
## 67.76199 49.93849 72.15029 66.25619
```

```
plot(unemployment.fit.errors)
```




SSE:

```
(unemployment.fit.sse <- sum(sapply(unemployment.fit.errors, function(x) x**2)))
```

```
## [1] 16681.09
```

**6c. Compare the 2 models - any reason to believe which should be the independent and the dependent variables.** To compare the two models, we will calculate the out-of-sample MAPE and the out-of-sample R-squared for both models.

```
(mape.gdp <- mean(abs(gdp.fit.errors / gdp.validation$GDP)))
```

```
## [1] 0.09686589
```

```
sst.gdp <- sum(sapply(gdp.validation$GDP - mean(gdp.validation$GDP), function(x) x**2))
(r.squared.gdp.on.unemployment <- 1 - (gdp.fit.sse / sst.gdp))
```

```
## [1] -111.8728
```

```
(mape.unemployment <- mean(abs(unemployment.fit.errors / unemployment.validation$UN)))
```

```
## [1] 0.1193223
```

```
sst.unemployment <- sum(sapply(unemployment.validation$UN - mean(unemployment.validation$UN), function(x) x
**2))
(r.squared.unemployment.on.gdp <- 1 - (unemployment.fit.sse / sst.unemployment))
```

```
## [1] -33.11265
```

Because the GDP on unemployment linear model has a slightly better fit out of sample (as shown by the mape), I conclude that GDP is a function of of unemployment.