STAT5703 HW2 Ex5

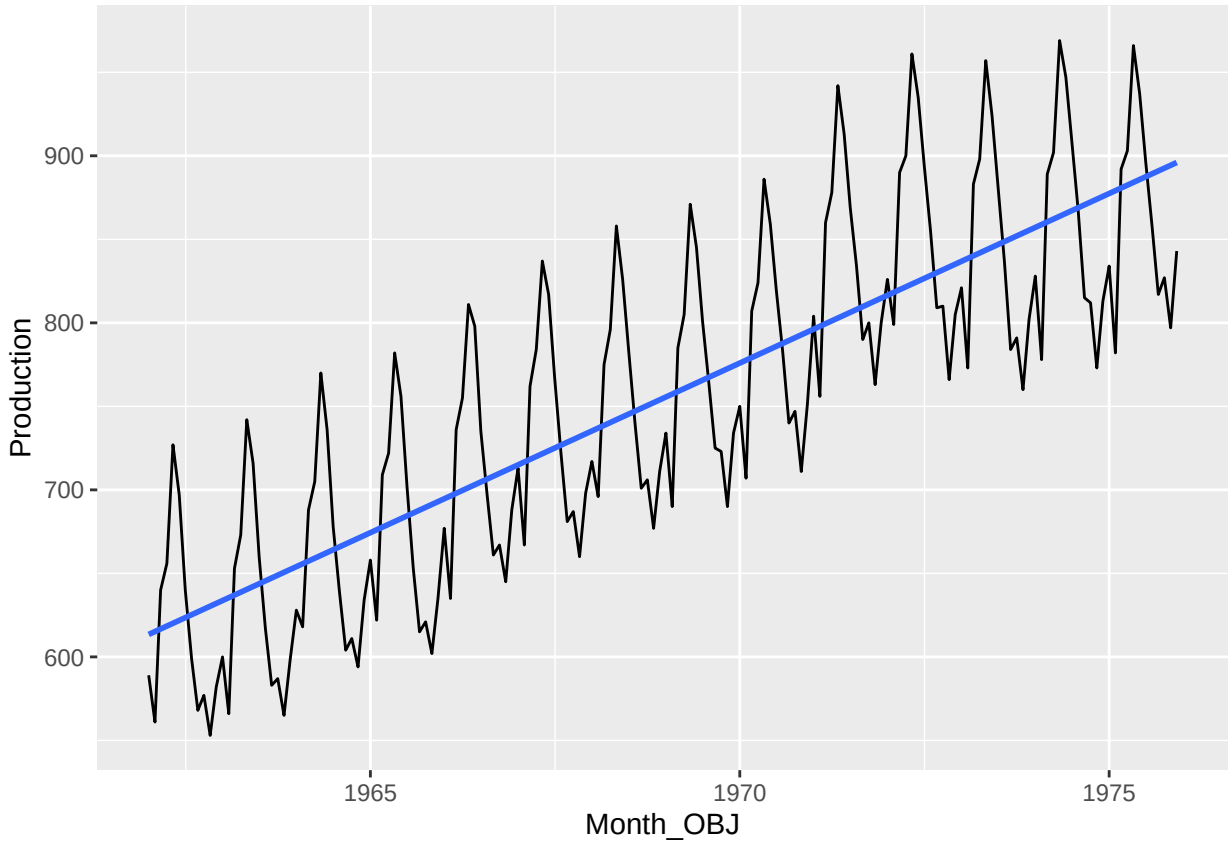
Chao Huang (ch3474), Wancheng Chen (wc2687), Chengchao Jin (cj2628)

Exercise 5

Q1

```
# manually export the data from xls file to a csv file
milk <- read.csv("./milk.csv", sep = "\t", header = F,
                 col.names = c("Month", "Production"),
                 as.is = F)
```

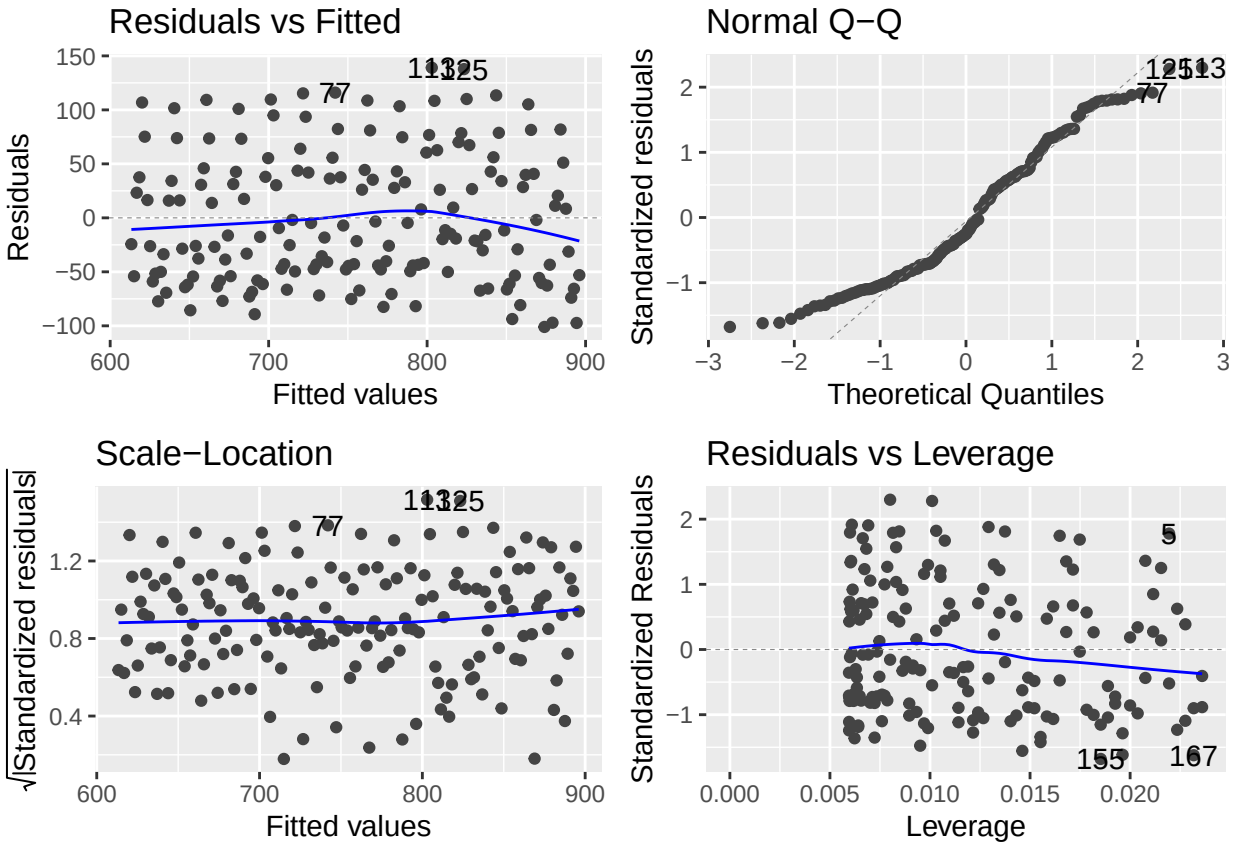
```
library(lubridate)
milk <- milk %>%
  mutate(Month_OBJ=ymd(Month, truncated = 1)) %>%
  mutate(Idx=row_number())

library(ggplot2)

milk %>% ggplot(aes(x=Month_OBJ, y=Production)) +
  geom_line() +
  geom_smooth(method = "lm", se = FALSE)
```



```r
# fit a linear model
lm_trend <- lm(Production ~ Idx, milk)
lm_trend
```

```
##
## Call:
## lm(formula = Production ~ Idx, data = milk)
##
## Coefficients:
## (Intercept)          Idx
##     611.682        1.693
```

```r
library(ggfortify)
autoplot(lm_trend)
```



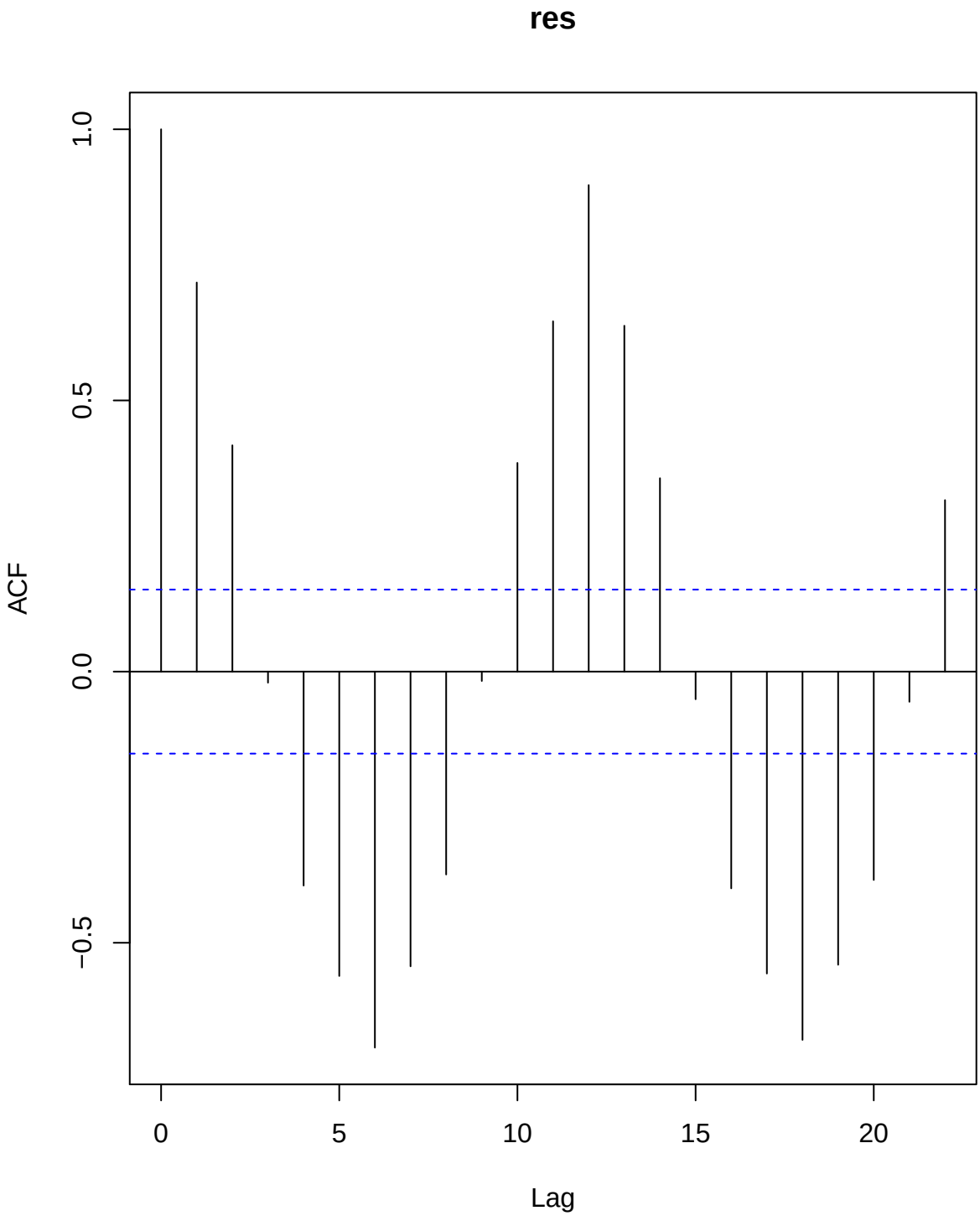
Here the regression model is fitted using the index of each month instead of the timestamp. It can be interpreted as the production of first month is around 611 pounds per cow, and later on , for each month, there will be an increase of 1.693 pound of the production per cow, which indicates a trend component.

The residual plots show that the residuals seem to scatter randomly with a mean value of 0 and a similar variance with different fitted value. So it's possible to view the residuals as a stationary time series.

Q2

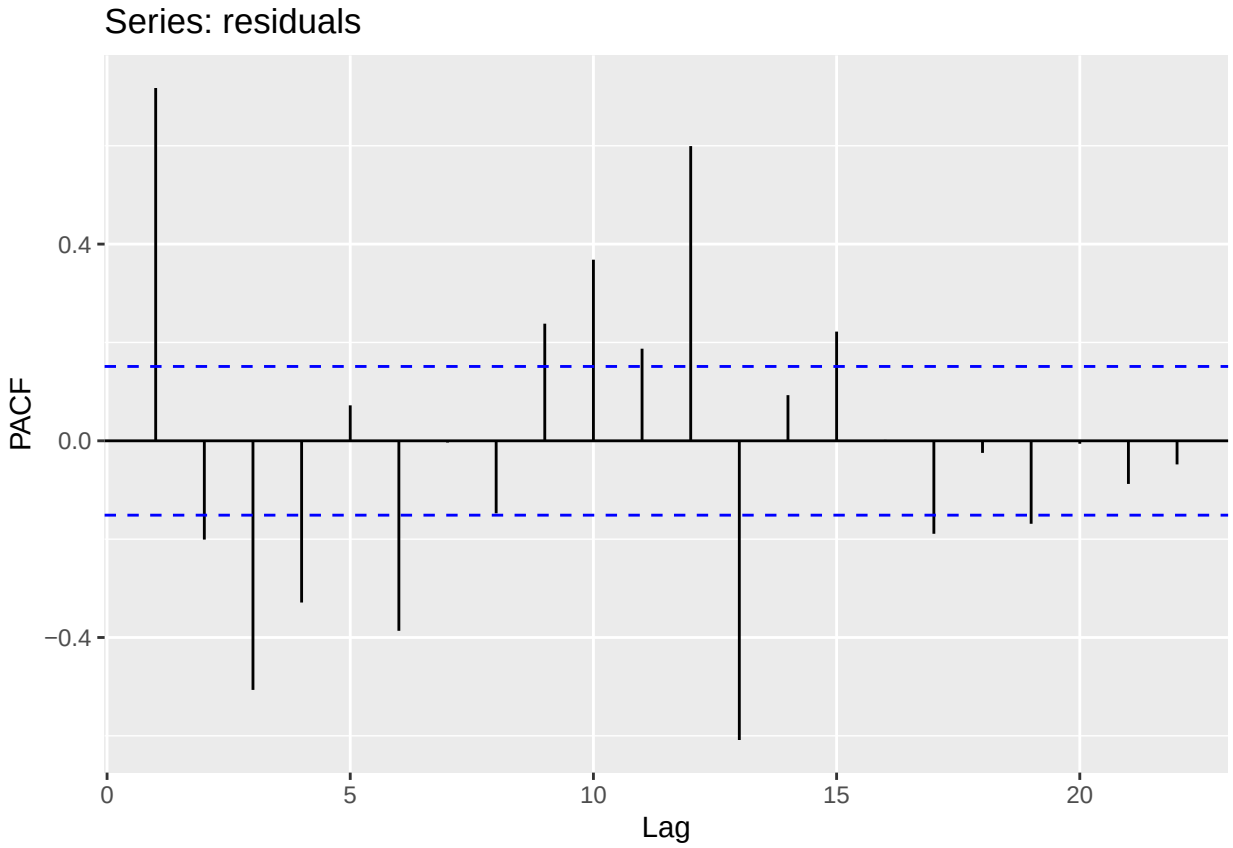
```r
residuals <- milk %>%
  mutate(res=lm_trend$residuals) %>%
  select(res) %>%
```

```
  as.ts()

acf(residuals)
```



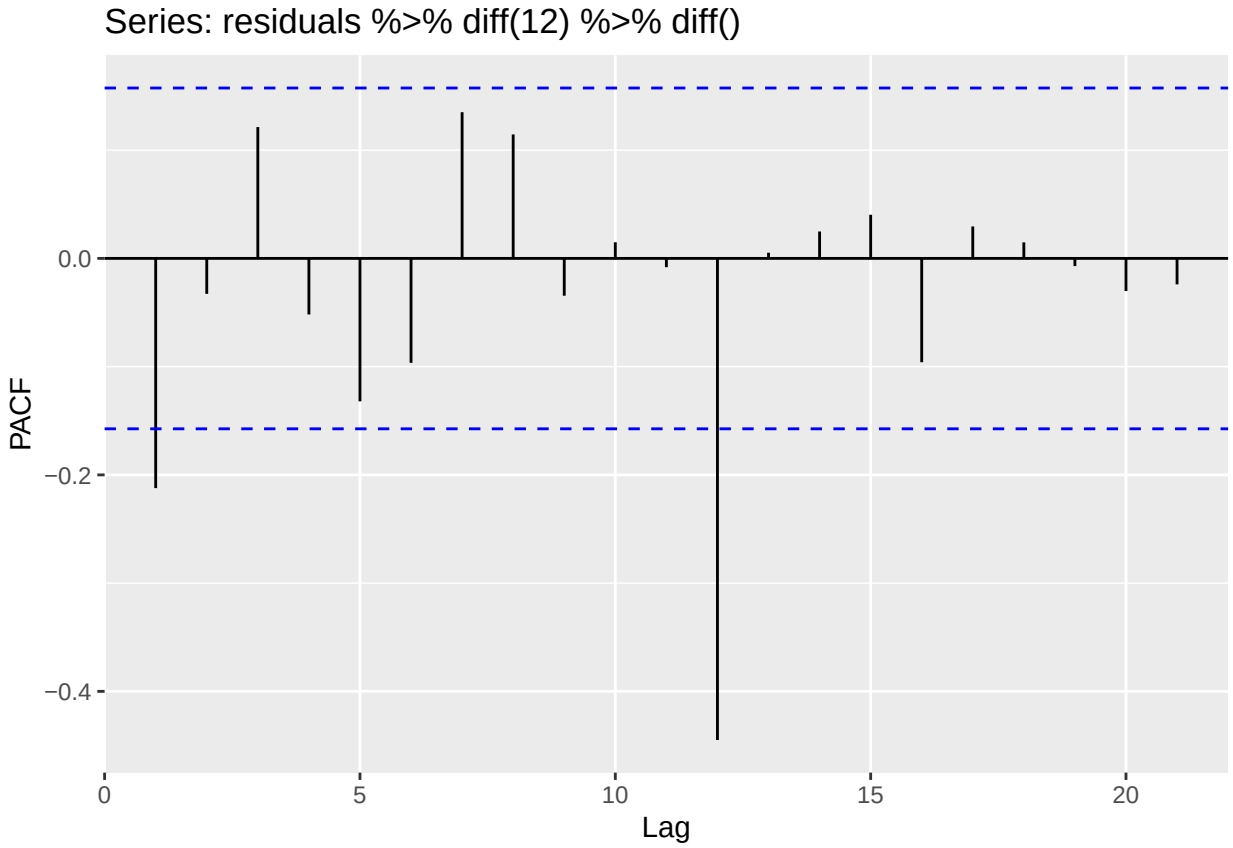
From the ACF plot, we can see that there is a strong seasonal component with the period of 12 months in this dataset.

```
forecast::ggPacf(residuals)
```



Eliminate the seasonal effect and stationarize it using a first order difference,

```
forecast::ggPacf(residuals %>% diff(12) %>% diff())
```



From PACF model after eliminating the seasonal effect, we can see that there seem to be a AR(1) or AR(2) model since PACF cuts off quickly after the first two bars (the second bar is almost as large as the first one). Since there is a spike at lag=12 after differencing, it also indicates a seasonal AR(1) component.

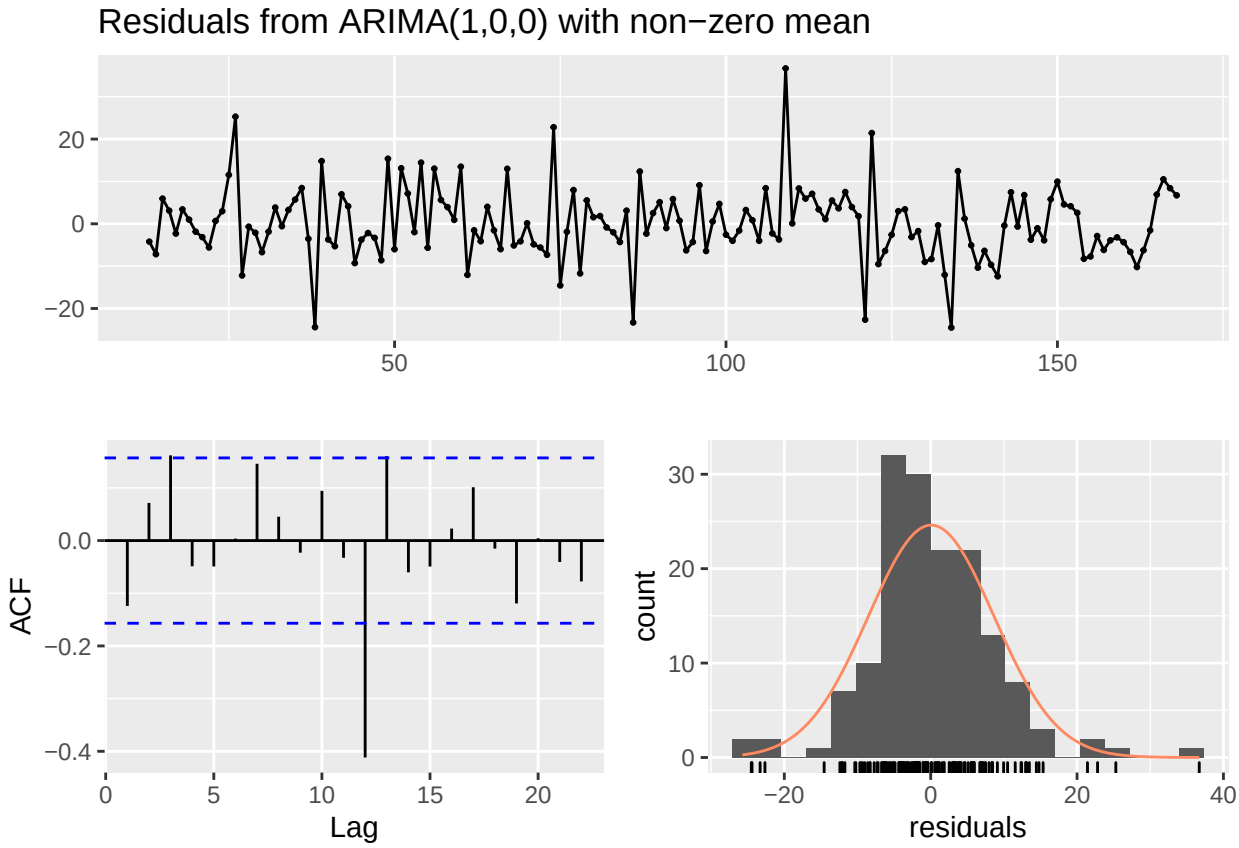
Q3

Before fitting a AR model, we first eliminate the seasonal effect by using diff(12).

```
library(forecast)
fitAR1 <- Arima(residuals %>% diff(12), order=c(1,0,0))
fitAR1
```

```
## Series: residuals %>% diff(12)
## ARIMA(1,0,0) with non-zero mean
##
## Coefficients:
##          ar1     mean
##       0.8543  -1.2121
## s.e.  0.0404   4.5427
##
## sigma^2 estimated as 74.42:  log likelihood=-557.17
## AIC=1120.33   AICc=1120.49   BIC=1129.48
```

```
checkresiduals(fitAR1)
```

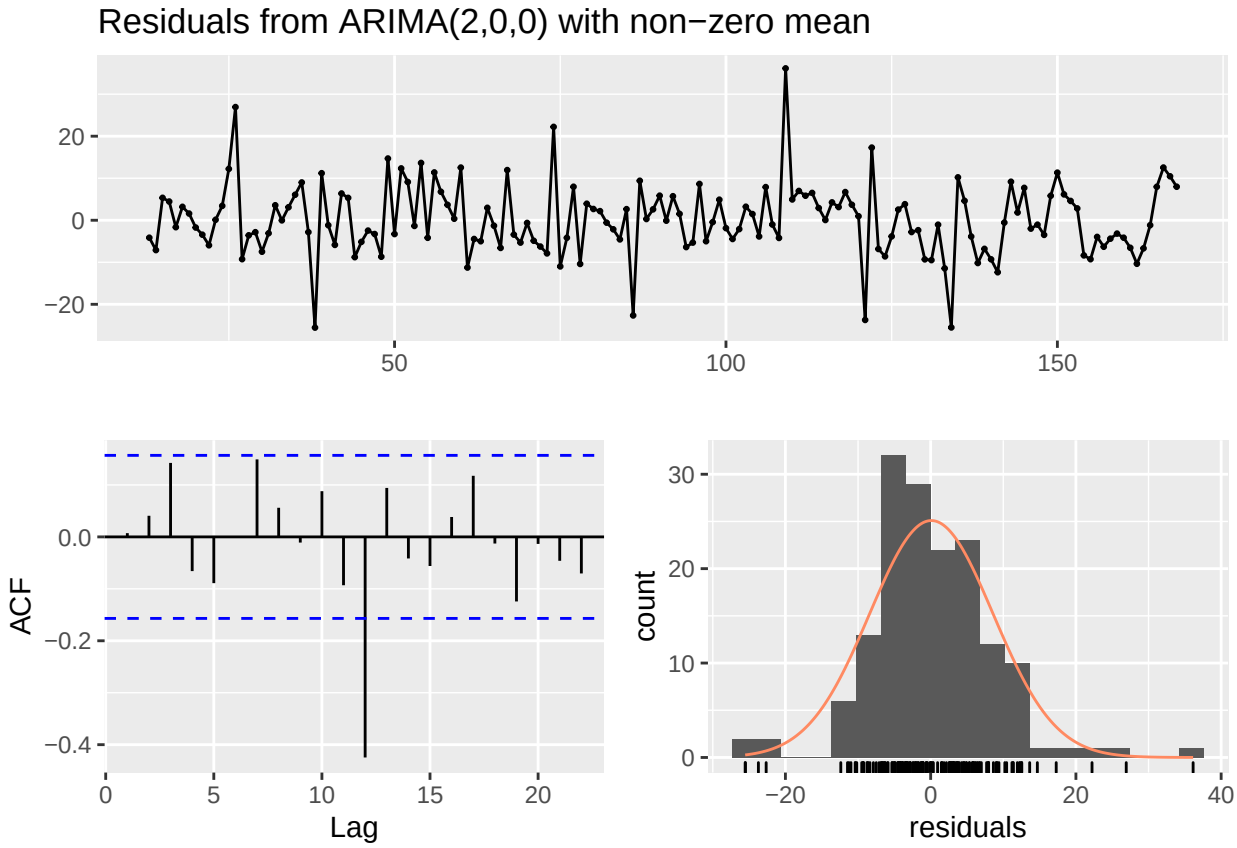



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,0) with non-zero mean
## Q* = 13.707, df = 8, p-value = 0.08972
##
## Model df: 2.   Total lags used: 10
```

```
fitAR2 <- Arima(residuals %>% diff(12), order=c(2,0,0))

fitAR2
```

```
## Series: residuals %>% diff(12)
## ARIMA(2,0,0) with non-zero mean
##
## Coefficients:
##          ar1     ar2     mean
##       0.7242  0.1512  -1.2623
## s.e.  0.0789  0.0790   5.1753
##
## sigma^2 estimated as 73.18:  log likelihood=-555.36
## AIC=1118.72   AICc=1118.98   BIC=1130.92
```

```
checkresiduals(fitAR2)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,0,0) with non-zero mean
## Q* = 11.072, df = 7, p-value = 0.1355
##
## Model df: 3.   Total lags used: 10
```

AR(1) and AR(2) doesn't seem to have a huge difference, the ACF plot seems to have a large spike at lag=12, so it may indicate another seasonal MA(1) component there. AR(2) has a slightly higher AICc compared with AR(1).

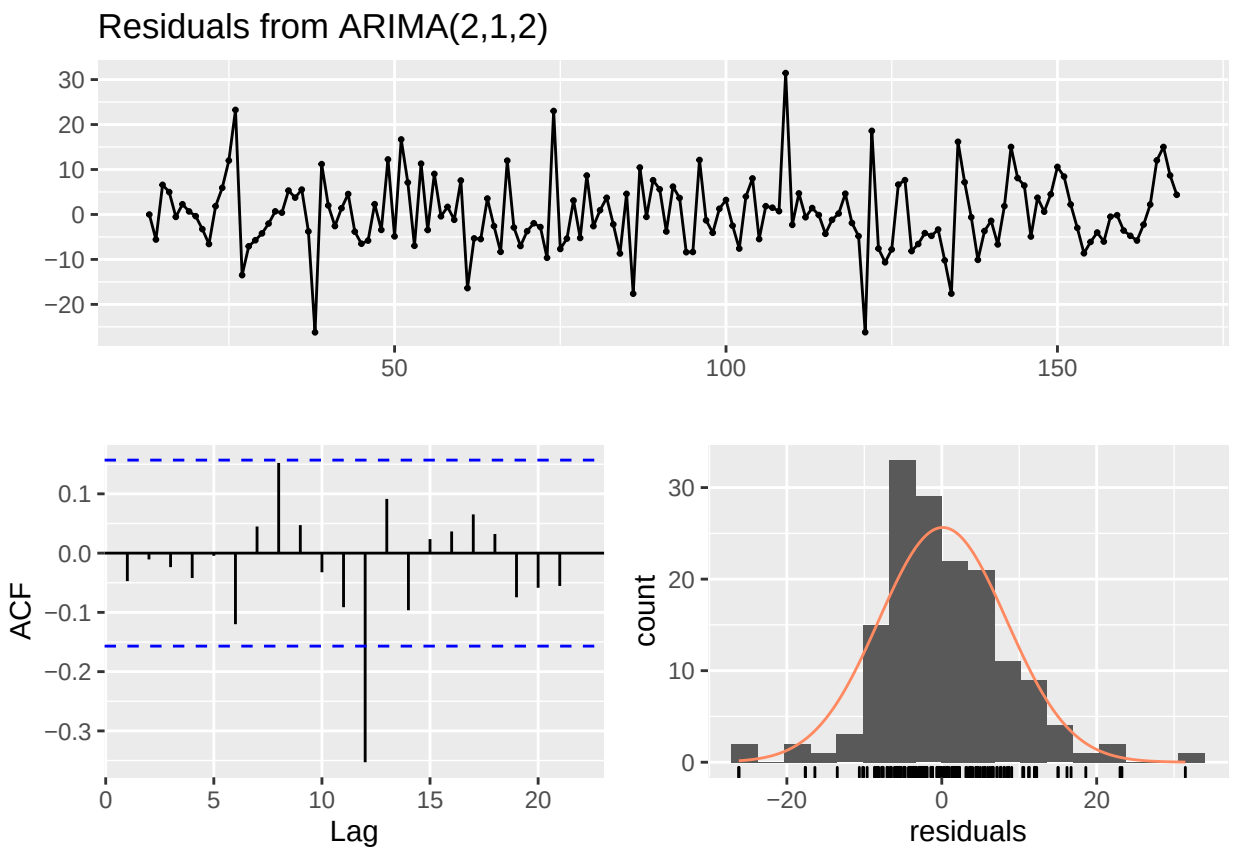
Q4

```
auto <- auto.arima(residuals %>% diff(12), seasonal = T,
                   approximation = F, stepwise = F)
auto
```

```
## Series: residuals %>% diff(12)
## ARIMA(2,1,2)
##
## Coefficients:
##           ar1      ar2     ma1     ma2
##       -0.3418  -0.8660  0.1841  0.9564
## s.e.   0.0538   0.0542  0.0334  0.0536
##
## sigma^2 estimated as 69.59:  log likelihood=-547.69
```

```
## AIC=1105.37   AICc=1105.78   BIC=1120.59
```

```
checkresiduals(auto)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,2)
## Q* = 7.8532, df = 6, p-value = 0.2491
##
## Model df: 4.   Total lags used: 10
```

The auto chosen model has a much lower AICc of 1105, better than AR(1) and AR(2) model. The model includes both the MA(2) and AR(2) component with a first order difference, which is similar to what we have observed in the previous question.

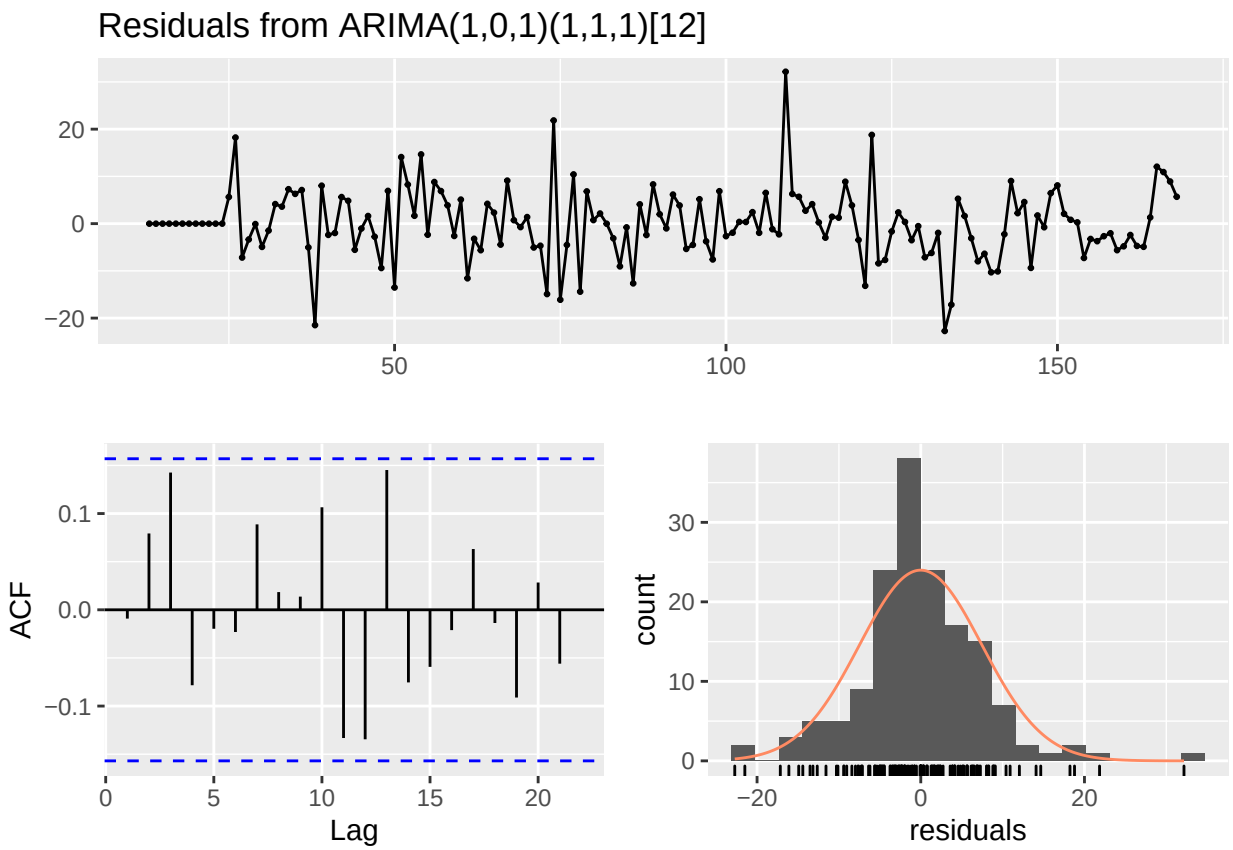
By including a seasonal ARIMA component and manually searching the orders and selecting them using AICc score, we have a better model ARIMA(2,0,1)(1,1,1)[12] with a lower AICc score of 1050.98. And now the large spike at lag=12 in ACF plot has been removed.

```
arima <- Arima(residuals %>% diff(12),order=c(1,0,1),
               seasonal = list(order = c(1, 1, 1), period = 12))
arima
```

```
## Series: residuals %>% diff(12)
## ARIMA(1,0,1)(1,1,1)[12]
##
## Coefficients:
##          ar1      ma1     sar1     sma1
##       0.9400  -0.1949  -0.4202  -1.0000
```

```
## s.e.  0.0337    0.0800    0.0761    0.0746
##
## sigma^2 estimated as 62.18:  log likelihood=-520.27
## AIC=1050.54   AICc=1050.98   BIC=1065.39
```

```
checkresiduals(arima)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,1)(1,1,1)[12]
## Q* = 8.7474, df = 6, p-value = 0.1883
##
## Model df: 4.   Total lags used: 10
```