Time series

Jialian Zhou

```
library(fpp2)
```

Exercise 4: Consider the daily IBM stock prices (data set ibmclose).

```
#  Daily IBM stock prices

ibmclose <- window(ibmclose)
head(ibmclose, 24)
```

```
## Time Series:
## Start = 1
## End = 24
## Frequency = 1
##   [1] 460 457 452 459 462 459 463 479 493 490 492 498 499 497 496 490 489 478 487
## [20] 491 487 482 479 478
```
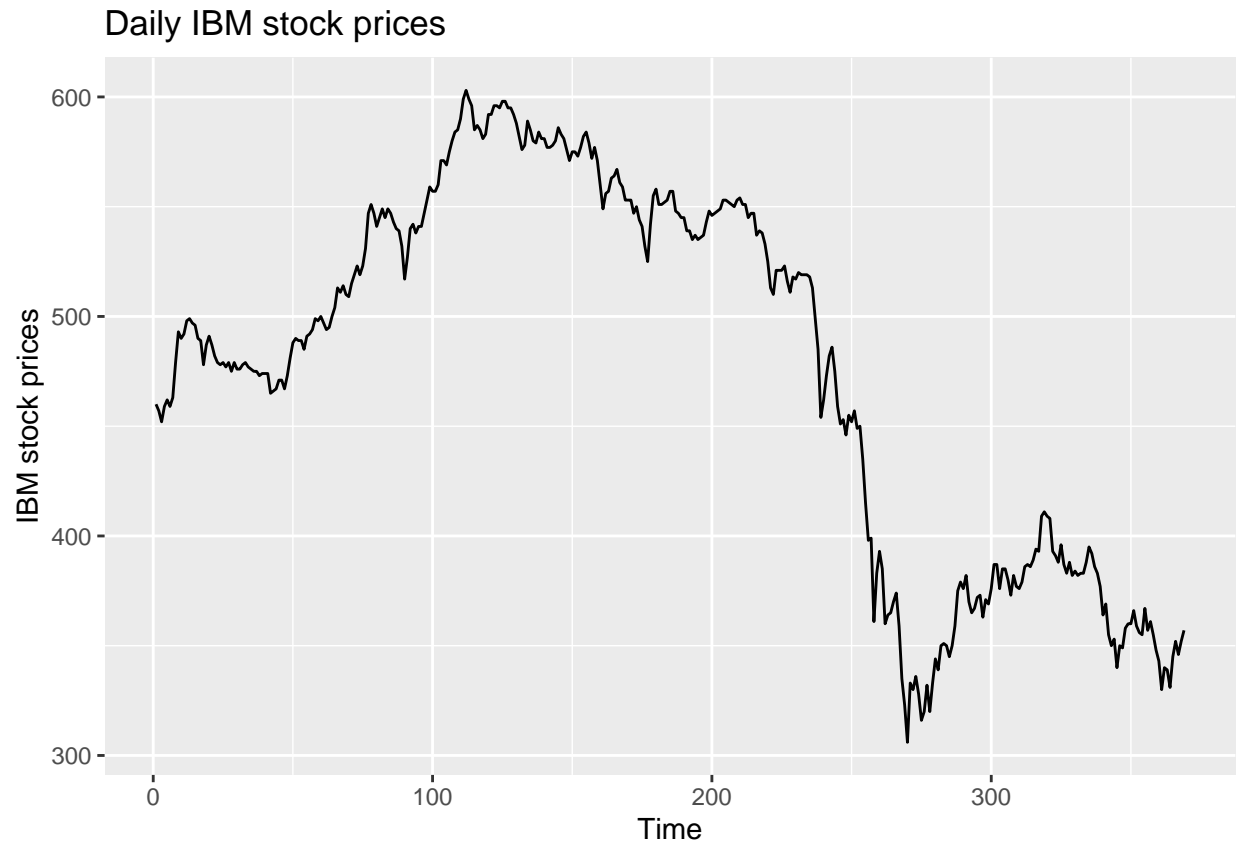
(a) Produce some plots of the data in order to become familiar with it.

The data set contains the daily IBM stock's prices at closure time.

```
ibmclose %>%
autoplot() + xlab("Time")+ ylab("IBM stock prices")+ ggtitle("Daily IBM stock prices")
```
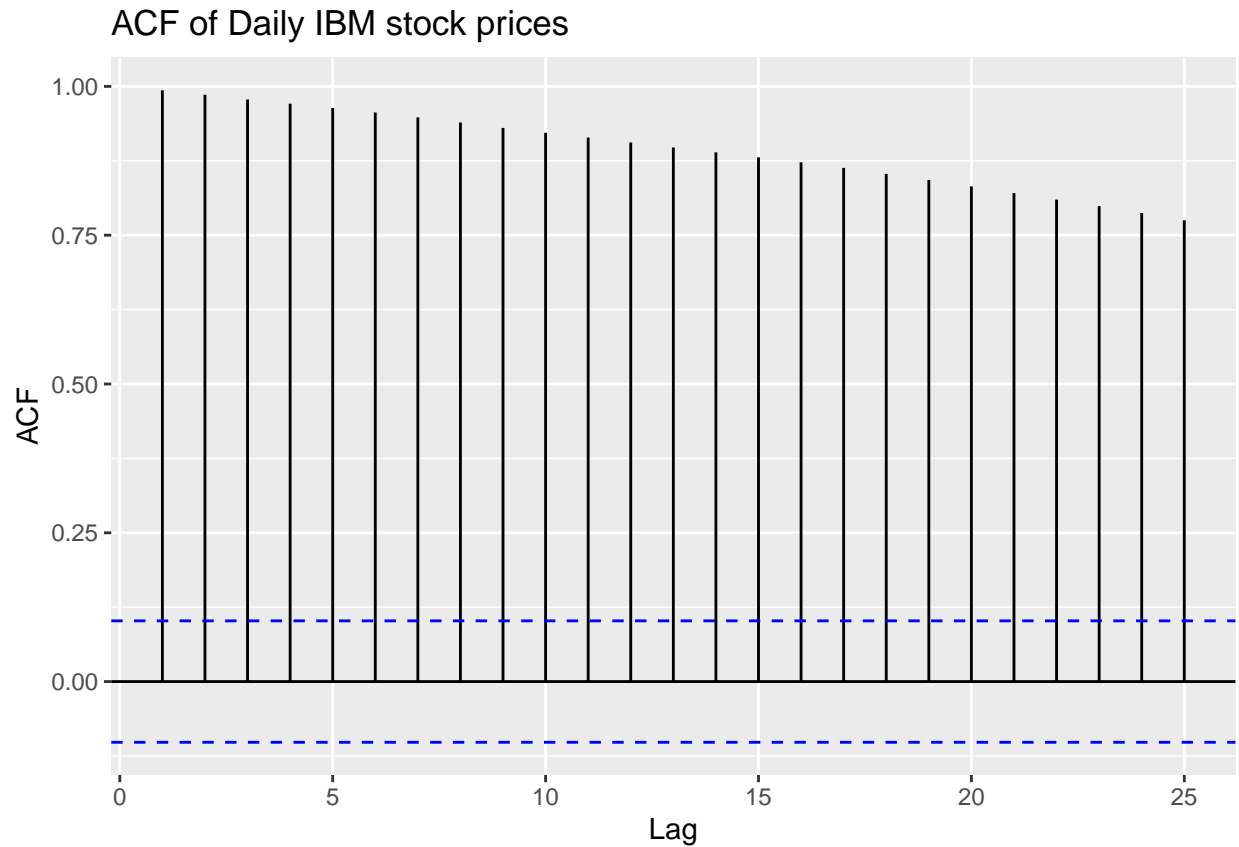
Daily IBM stock prices

The following conclusions can be drawn from this first plot:

- There is not a clear trend: the series has some ups and downs, which makes the series unpredictable.

- There is a big drop in the prices right after the 200 day mark.

- Not seasonal pattern can be detected as the frequency of the data is daily and only for a year.

- The mean method is not going to be the best option due to the heterogeneity and the naïve method will give the most conservative forecast.

The autocorrelation plot:

```
ggAcf(ibmclose) + ggtitle("ACF of Daily IBM stock prices")
```

ACF of Daily IBM stock prices

In the autocorrrelation plot we can conclude that the IBM stock prices are highly correlated with each other. That is to say, when the stock price rises, it tends to continue this way and when it falls, it keeps going downwards.

(b) Split the data into a training set of 300 observations and a test set of 69 observations.

```r
# Split
train <- subset(ibmclose, end = 300)
test  <- subset(ibmclose, start = 301, end = length(ibmclose))

# Plot of the split
plot(ibmclose)
lines(train,col="red")
lines(test, col="blue")
```
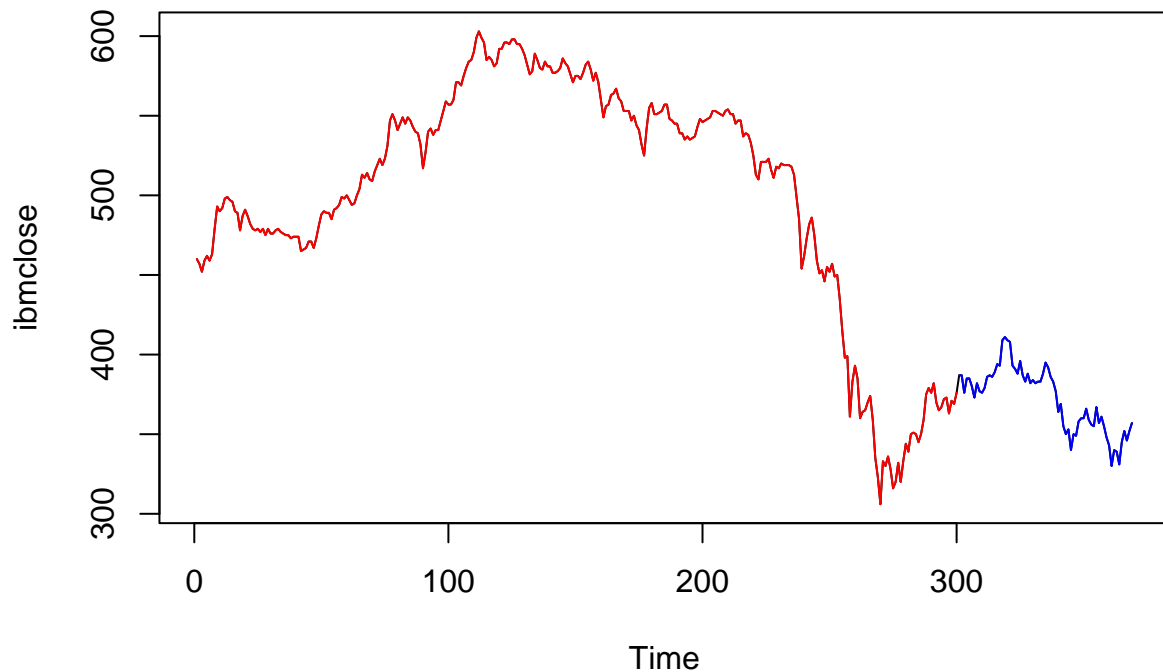
(c) Try various benchmark methods to forecast the training set and compare the results on the test set. Which method did best?

```r
ibmclosefit1 <- meanf(train,h=69)
ibmclosefit2 <- rwf(train,h=69)
ibmclosefit3 <- rwf(train, drift=TRUE,h=69)

autoplot(train) +
  autolayer(ibmclosefit1, series="Mean", PI=FALSE) +
  autolayer(ibmclosefit2, series="Naïve", PI=FALSE) +
  autolayer(ibmclosefit3, series="Drift", PI=FALSE) +
  xlab("Day") + ylab("Prices") +
  ggtitle("Forecasts for IBM stock prices") +
  guides(colour=guide_legend(title="Forecast"))
```

Forecasts for IBM stock prices
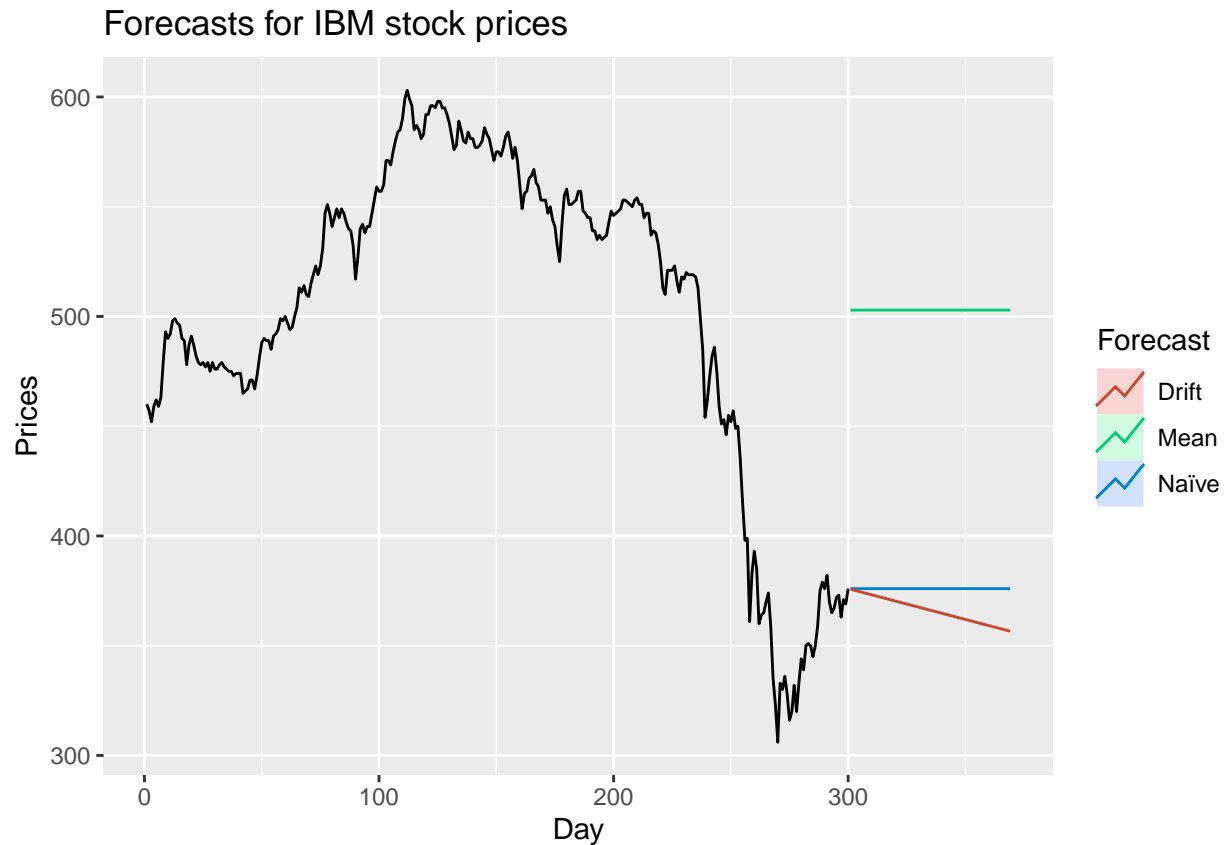
The mean method seems to perform the worst out of the three methods, which is to expected as the series is not stationary. And, as it has been stated before, we will be sticking to the naïve method as it gives the most reasonable forecast.

Now, we will be comparing the results with the real data.

```
autoplot(window(ibmclose)) +
  autolayer(ibmclosefit1, series="Mean", PI=FALSE) +
  autolayer(ibmclosefit2, series="Naïve", PI=FALSE) +
  autolayer(ibmclosefit3, series="Drift", PI=FALSE) +
  xlab("Day") + ylab("Prices") +
  ggtitle("Forecasts for IBM stock prices") +
  guides(colour=guide_legend(title="Forecast"))
```

Forecasts for IBM stock prices
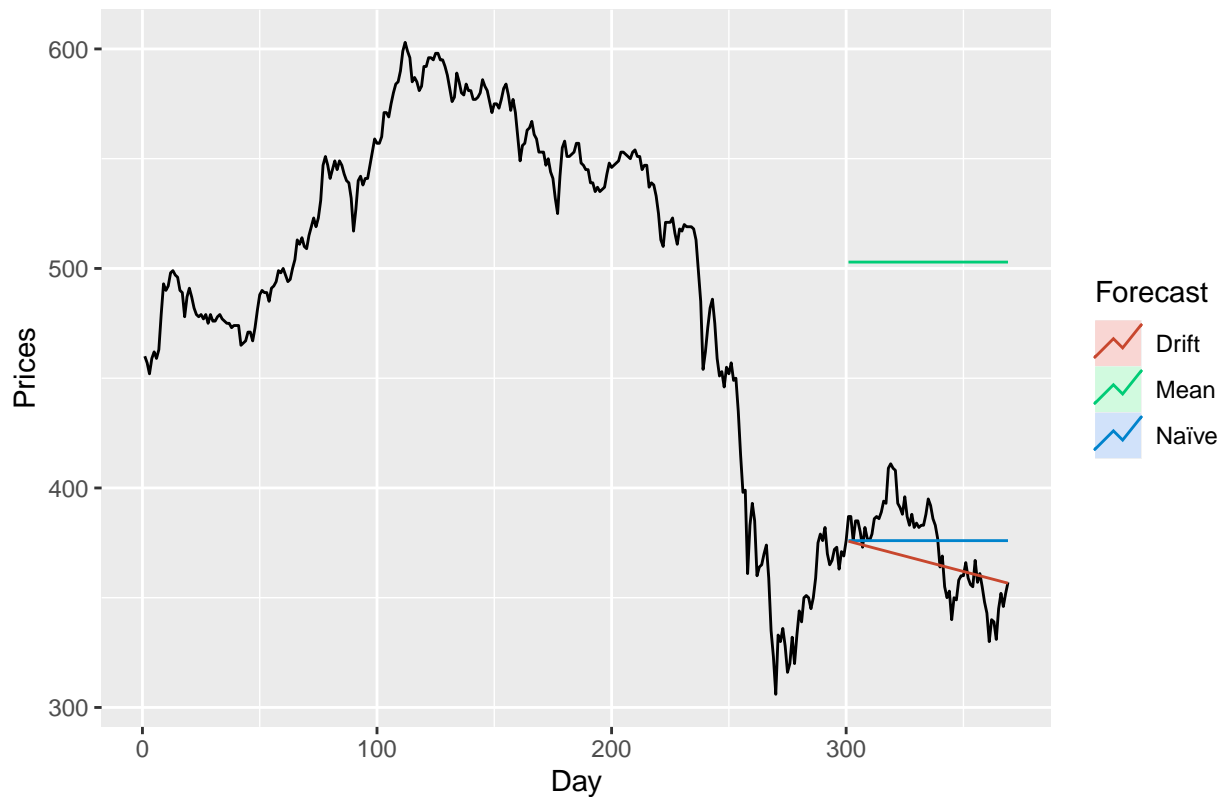
In order to evaluate the predictive performance, the **accuracy** been calculated based on the test set for each forecasting method.

```
accuracy(ibmclosefit1, test)
```

```
##                        ME      RMSE       MAE       MPE     MAPE     MASE
## Training set  1.660438e-14  73.61532  58.72231  -2.642058 13.03019 11.52098
## Test set     -1.306180e+02 132.12557 130.61797 -35.478819 35.47882 25.62649
##                   ACF1 Theil's U
## Training set 0.9895779        NA
## Test set     0.9314689  19.05515
```

```
accuracy(ibmclosefit2, test)
```

```
##                     ME      RMSE      MAE         MPE     MAPE     MASE
## Training set -0.2809365  7.302815  5.09699 -0.08262872 1.115844 1.000000
## Test set     -3.7246377 20.248099 17.02899 -1.29391743 4.668186 3.340989
##                   ACF1 Theil's U
## Training set 0.1351052        NA
## Test set     0.9314689  2.973486
```

```
accuracy(ibmclosefit3, test)
```

```
##                      ME      RMSE       MAE          MPE     MAPE     MASE
## Training set 2.870480e-14  7.297409  5.127996 -0.02530123 1.121650 1.006083
## Test set     6.108138e+00 17.066963 13.974747  1.41920066 3.707888 2.741765
##                   ACF1 Theil's U
## Training set 0.1351052        NA
## Test set     0.9045875  2.361092
```

The best results are obtained with the **drift method** for this test set, but the difference with the **naïve method** (the second best) is minimum. However, it cannot be assured that the data will follow the descent trend which is forecast by the drift method.

(d) For the best method, compute the residuals and plot them. What do the plots tell you?

After the previous section, there are still be some doubts about which is the best method, the naïve or the drift. Therefore, we have computed the residuals for both methods.

```r
# Naïve
train_naive <- naive(train,h=69)
res_naive   <- residuals(train_naive)
summary(res_naive)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -38.0000  -4.0000   0.0000  -0.2809   4.0000  27.0000        1
```

```r
checkresiduals(train_naive)
```
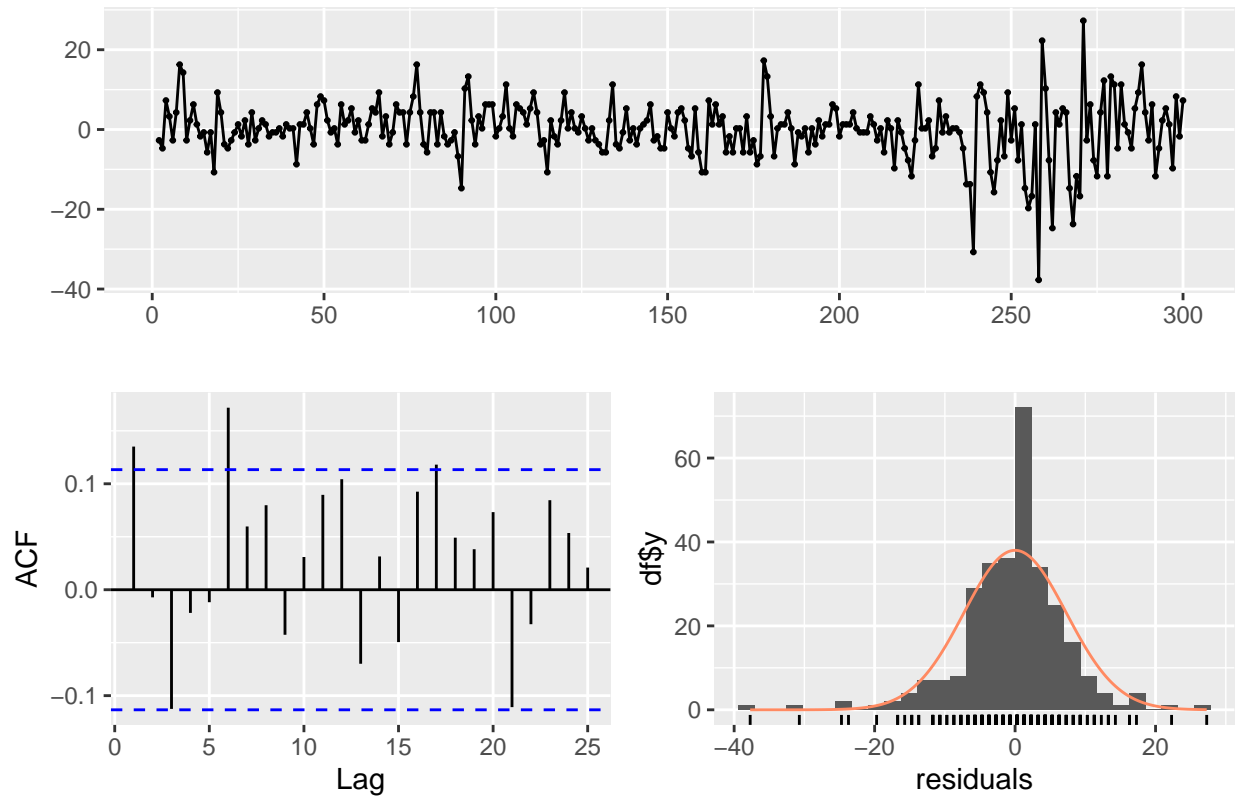
Residuals from Naive method
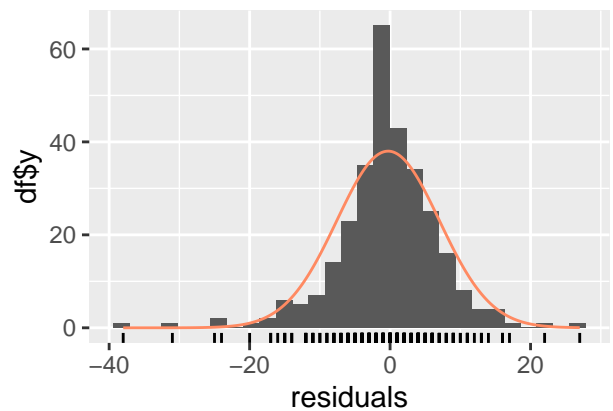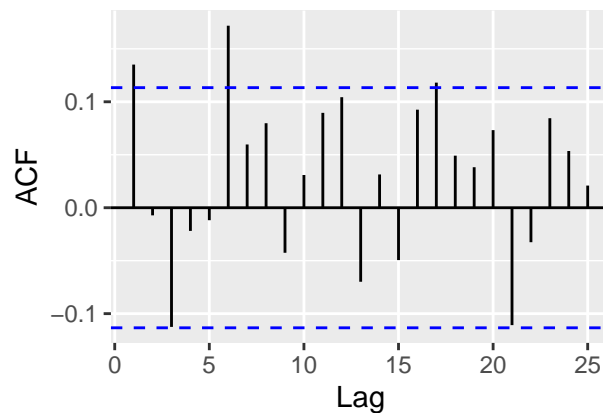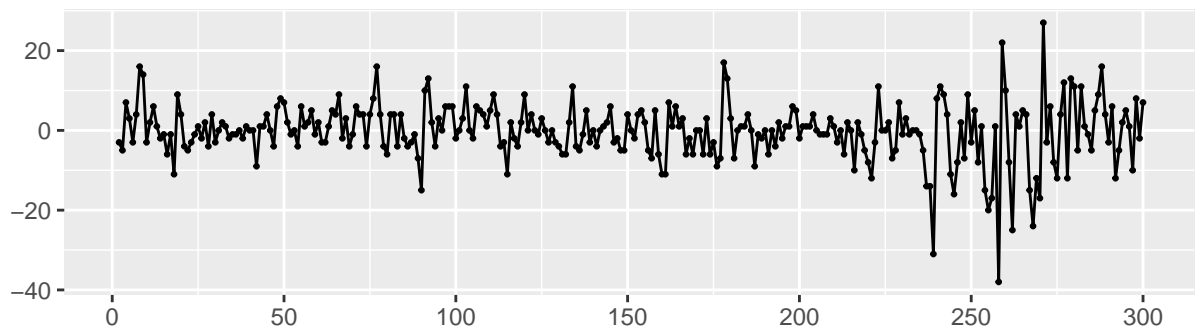

```
##
##  Ljung-Box test
##
## data:  Residuals from Naive method
## Q* = 22.555, df = 10, p-value = 0.01251
##
## Model df: 0.    Total lags used: 10
```

```r
# Drift method
train_drift <- rwf(train, drift=TRUE,h=69)
res_drift   <- residuals(train_drift)
summary(res_drift)
```

```
##     Min.  1st Qu.   Median    Mean  3rd Qu.     Max.    NA's
## -37.7191  -3.7191   0.2809  0.0000   4.2809  27.2809       1
```

```r
checkresiduals(train_drift)
```

Residuals from Random walk with drift

```
##
##  Ljung-Box test
##
## data:  Residuals from Random walk with drift
## Q* = 22.555, df = 9, p-value = 0.007278
##
## Model df: 1.    Total lags used: 10
```

Although both methods have some lags which exceed the 95% confidence interval, the residual plots do not show any kind of pattern. Therefore, they could behave as white noise.

However, the p-values obtained in the Ljung-Box test for both methods are smaller than 0.05, so the null hypothesis can be rejected: we cannot conclude that the data are not independently distributed.

$$H_0 = The \ data \ are \ independently \ distributed$$
$$H_1 = The \ data \ are \ not \ independently \ distributed$$

Finally, looking at the residuals and the normal distribution, we can see that the residuals obtained with the **naïve method** are better adjusted to the gaussian distribution. Hence, the **naïve method** was the correct one in this case.