DA5030.A2.Parpattedar

Shruti Parpattedar

January 27, 2019

Question 1

Determining outliers in terms of Murders. Outliers are those lying beyond 1.5 times standard deviation from mean.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------ ti
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data <- USArrests
sd <- sd(data$Murder)
#data[data$Murder>1.5*sd+mean(data$Murder) | data$Murder<mean(data$Murder)-1.5*sd,]
paste("Outliers are - ")
```

```
## [1] "Outliers are - "
```

```
matrix(row.names(data[data$Murder > 1.5*sd+mean(data$Murder) |
                        data$Murder < mean(data$Murder)-1.5*sd, ]))
```

```
##      [,1]
## [1,] "Florida"
## [2,] "Georgia"
## [3,] "Louisiana"
## [4,] "Mississippi"
## [5,] "North Dakota"
## [6,] "South Carolina"
```

Question 2

Finding the correlation between urban population and murder. The Pearson coefficient of correlation is ~0.069 which is a very weak correlation. So, we can conclude that Urban population and murder are weakly correlated i.e. if one goes up, it is not necessary that the other will go up as well.

```
cor(data$UrbanPop, data$Murder)
```

```
## [1] 0.06957262
```

Question 3

Calculating forecasts using weighted moving average (using the most recent two years), exponential smoothing (alpha = 0.4) and linear regression trendline models.

```r
q3data <- read.csv("PhoneGrowth.csv", header = TRUE)
names(q3data) <- c("Year", "Subscribers")
n <- nrow(q3data)

# Weighted Moving Averages
sample <- q3data[n:(n-1), 2]
w <- c(5, 2)
prod <- w * sample
F.wma <- sum(prod) / sum(w)

# Exponential Smoothing
q3data$F.es <- 0
q3data$E.es <- 0
q3data$Sq.E.es <- 0
a <- 0.4
q3data$F.es[1] <- q3data$Subscribers[1]
for(i in 2:n)
{
  q3data$F.es[i] <- q3data$F.es[i-1] + a * q3data$E.es[i-1]
  q3data$E.es[i] <- q3data$Subscribers[i] - q3data$F.es[i]
}
F.es <- q3data$F.es[n] + a*q3data$E.es[n]

# Linear Regression
model <- lm(q3data$Subscribers ~ q3data$Year)
#Using the exact coefficients gives us the exact number including any decimals.
F.lr <- model$coefficients[1] + model$coefficients[2]*12

forecasts <- data.frame(F.wma, F.es, F.lr)
row.names(forecasts) <- "Year 12 Subscribers Forecast"
forecasts
```

```
##                                 F.wma      F.es      F.lr
## Year 12 Subscribers Forecast 194662700 165168214 203610220
```

Question 4

Calculating the Mean Squared Error (MSE) for the three models used above. On comparing the three MSEs we observe that the linear regression trendline model has the lowest MSE.

```r
# For the Weighted Moving Average model
q3data$F.wma <- 0
q3data$Sq.E.wma <- 0
q3data$F.wma[1:2] <- NA
q3data$Sq.E.wma[1:2] <- NA
w <- c(5, 2)
for(i in 3:nrow(q3data))
{
  q3data$F.wma[i] <- sum(w*q3data$Subscribers[(i-1):(i-2)])/sum(w)
  q3data$Sq.E.wma[i] <- (q3data$Subscribers[i] - q3data$F.wma[i])^2
}
MSEforWMA <- mean(q3data$Sq.E.wma[3:11])

# For the Exponential Smoothing model
```

```
q3data$Sq.E.es <- q3data$E.es^2
MSEforES <- mean(q3data$Sq.E.es)

# For the Linear Regression Trendline model
q3data$F.lr <- 0
q3data$Sq.E.lr <- 0
for(i in 1:nrow(q3data))
{
  q3data$F.lr[i] <- model$coefficients[1] + model$coefficients[2]*i
  q3data$Sq.E.lr[i] <- (q3data$Subscribers[i] - q3data$F.lr[i])^2
}
MSEforLR <- mean(q3data$Sq.E.lr)

models <- c("Weighted Moving Average", "Exponential Smoothing", "Linear Regression")
MSEs <- c(MSEforWMA, MSEforES, MSEforLR)
result <- data.frame(Models = models, MSE = as.numeric(MSEs))
result
```

```
##                     Models          MSE
## 1 Weighted Moving Average 6.650647e+14
## 2   Exponential Smoothing 1.473838e+15
## 3       Linear Regression 1.265347e+14
```

```
result %>%
  filter(MSE == min(MSE))
```

```
##              Models          MSE
## 1 Linear Regression 1.265347e+14
```

Question 5

Calculating a weighted average forecast by averaging out the three forecasts calculated in question 3 using the weights 4 for trendline, 2 for exponential smoothing and 1 for weighted moving average.

```
forecasts <- c(F.lr, F.es, F.wma)
w <- c(4, 2, 1)
forecasts.wma <- sprintf("%.4f",sum(w * forecasts)/sum(w))
forecasts.wma
```

```
## [1] "191348572.7334"
```