Regression

Subramani.M

7 May 2018

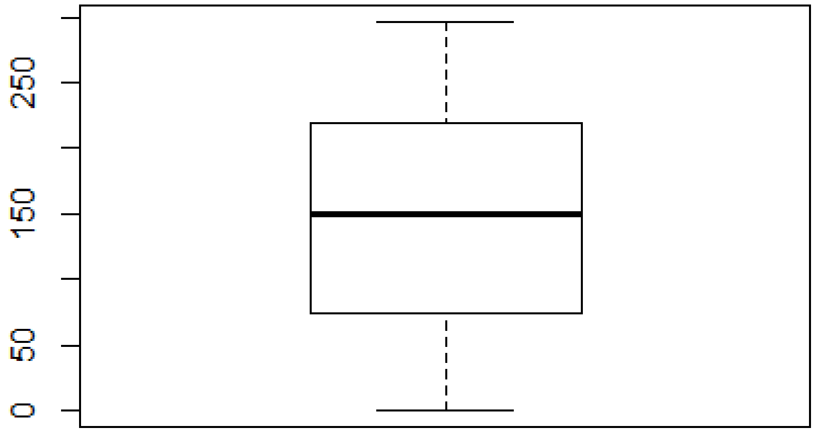
```
library(ISLR)

## Warning: package 'ISLR' was built under R version 3.4.4

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.3

setwd("C:/Users/Administrator/Desktop/Machine Learning/DATA SETS")
advertising <- read.csv("C:/Users/Administrator/Desktop/Machine Learning/DATA
SETS/Advertising.csv")
View(advertising)
dim(advertising)

## [1] 200   5
```

data Pre-processing

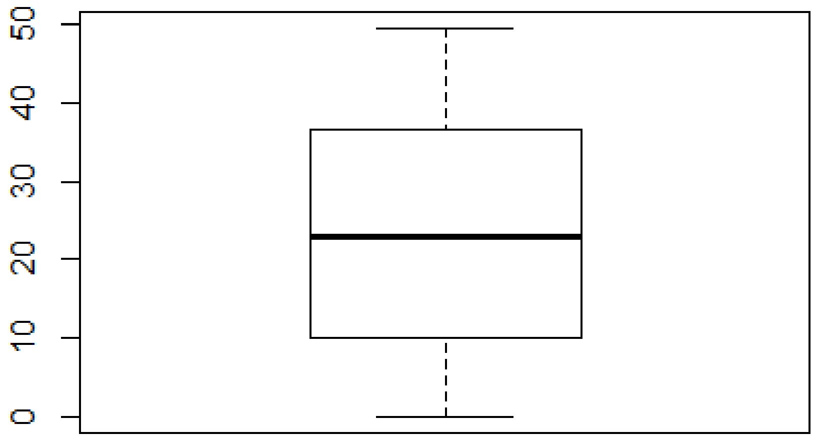
regression model

```
# select first 160 rows of data for training and 40 rows for testing
# for this you can select randomly by using sample function or select first
160 rows for training
# and 40 rows for testing

# Identify the count of missing values
colSums(is.na(advertising))

##         X        TV     radio newspaper     sales
##         0         0         0         0         0

colSums(is.na(airquality))

##   Ozone Solar.R    Wind    Temp   Month     Day
##      37       7       0       0       0       0

# percentage of missing values
colSums(is.na(advertising)) / nrow(advertising) * 100

##         X        TV     radio newspaper     sales
##         0         0         0         0         0
```

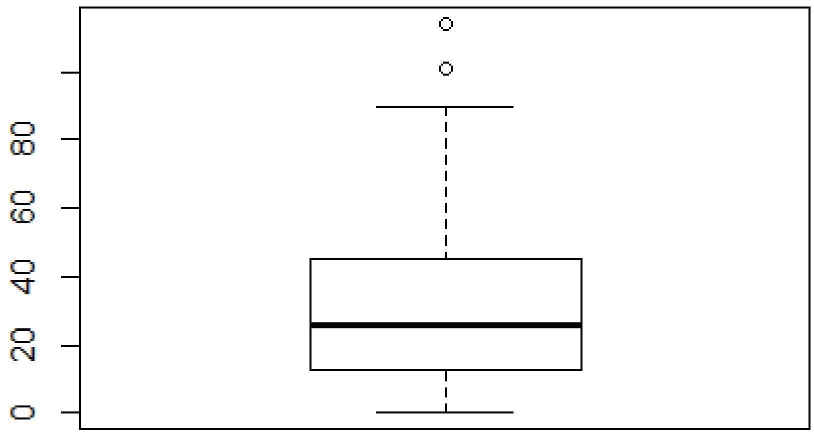
```r
# Detect outliers
boxplot(advertising$TV)
```



```r
boxplot(advertising$radio)
```

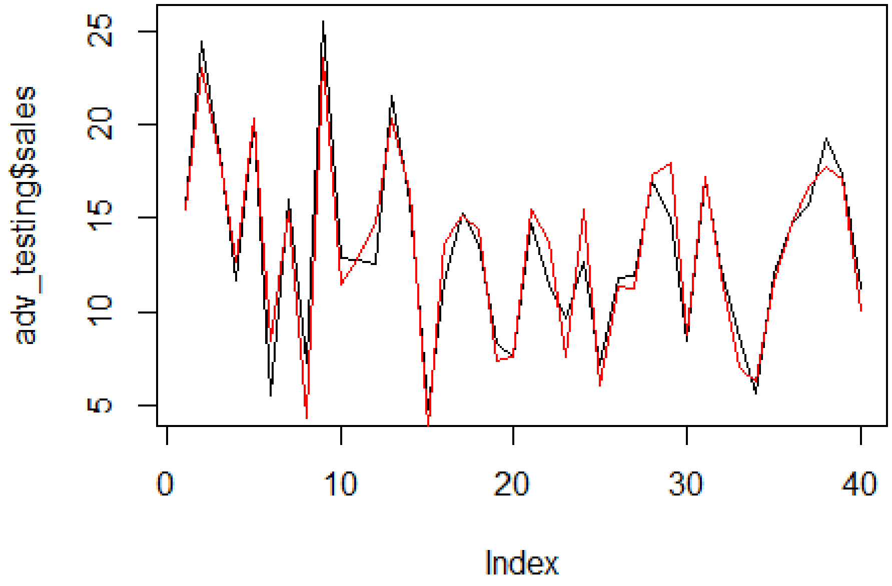


```
boxplot(advertising$newspaper)
```



```
# method 1
df_train <- advertising[1:162,]
df_testing <- advertising[163:200,]

# method 2 : randomly sampling with replacement
#example
sample(c(1,2,3,4,5),3)

## [1] 2 3 4

adv_train <- advertising[sample(seq(1,nrow(advertising)),160),]
adv_testing <- advertising[sample(seq(1,nrow(advertising)),40),]

names(adv_train)

## [1] "X"         "TV"        "radio"     "newspaper" "sales"

names(adv_testing)

## [1] "X"         "TV"        "radio"     "newspaper" "sales"

dim(adv_train)

## [1] 160   5

dim(adv_testing)

## [1] 40  5

# Feature selection ( Selection of input variables )
# use all imput variables i.e tv,radio and newspapers

# Fit a model
# it is multi linear regression model because there are more than 2 input
variables
advertising_model <- lm(sales ~ TV + radio + newspaper , data = adv_train)

# Predict sales for testing dataset
adv_testing$sales_predict <-  predict(advertising_model,
adv_testing[,c('TV','radio','newspaper')])

View(adv_testing)

# calculate error row-wise
adv_testing$error <- adv_testing$sales - adv_testing$sales_predict
View(adv_testing)

# to exclude negative values
adv_testing$sqr_error <- adv_testing$error ^ 2
View(adv_testing)

sum(adv_testing$sqr_error) # this must be less . Then it means the model is
perfect
```

```
## [1] 75.0022

# visually see the error
{plot(adv_testing$sales,type = 'l')
  lines(adv_testing$sales_predict, col = 'red')}
```


adv_testing$sales
Index