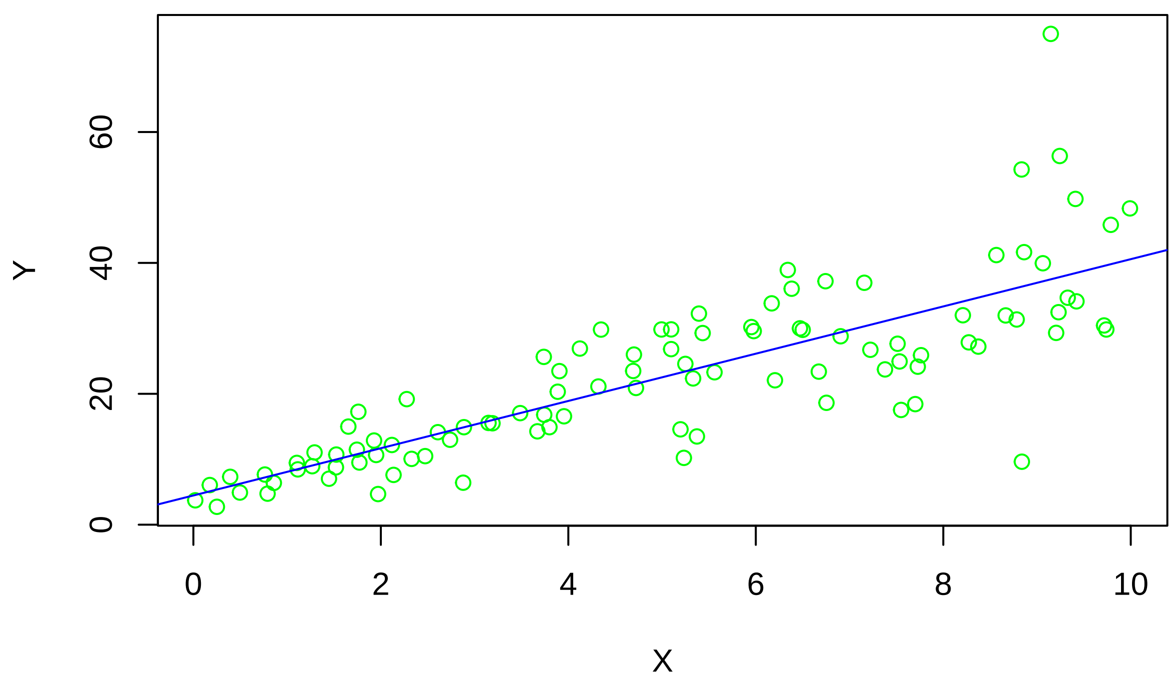
Business Analytics assignment 3

Khutso Ledwaba

2022-11-10

##Question 1

```
set.seed(2017)
X=runif(100)*10
Y=X*4+3.45
Y=rnorm(100)*0.29*Y+Y

#A)Plot Y against X. Include a screenshot of the plot in your submission. Using the File menu you can s
plot(Y~X,xlab='X',ylab='Y',col='green')
abline(lsfit(X, Y),col = "blue")
```




```
#B)Construct a simple linear model of Y based on X. Write the equation that explains Y based on X. What
```

```r
Y=4.4655+3.6108*X
# The accuracy is 65%
line_fit <- lm(Y ~ X)
summary(line_fit)
```

```
## Warning in summary.lm(line_fit): essentially perfect fit: summary may be
## unreliable
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -5.743e-15 -3.313e-15 -1.574e-15  8.700e-17  1.242e-13
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 4.466e+00  2.601e-15 1.717e+15   <2e-16 ***
## X           3.611e+00  4.463e-16 8.091e+15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.298e-14 on 98 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 6.546e+31 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
# C)How the Coefficient of Determination, R2, of the model above is related to the correlation coeffici
cor(X,Y)^2
```

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

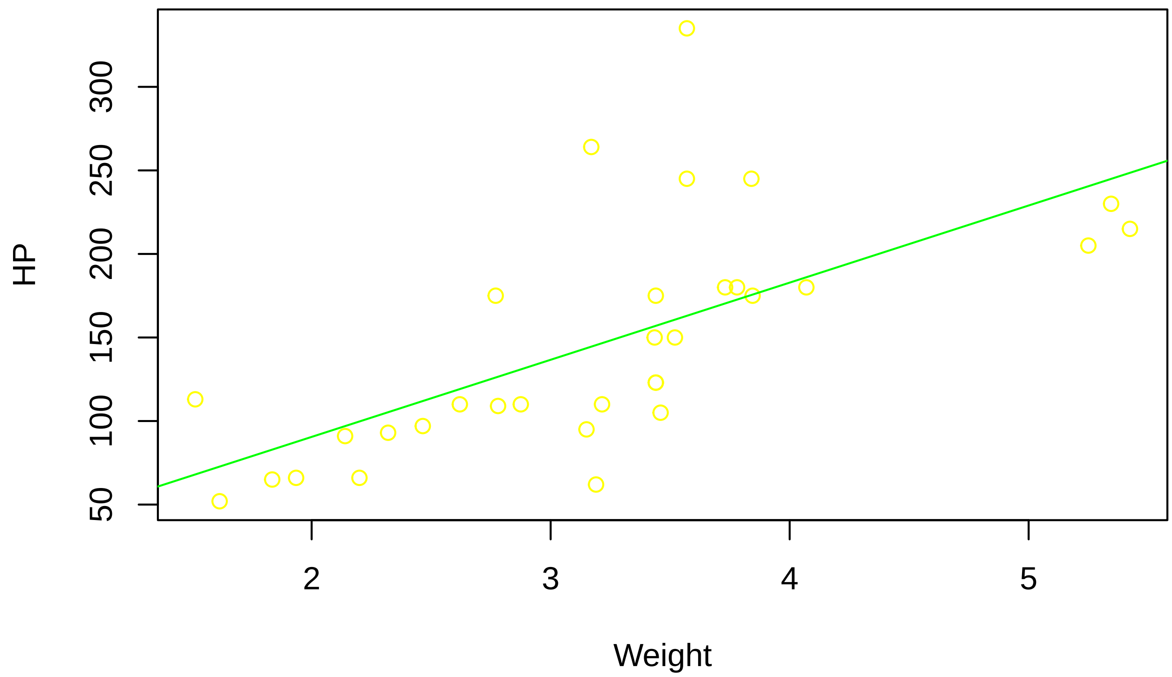
```r
#The square of correlation coefficient is same as coefficient of determination which is 65.17%
```

#Question 2

```r
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

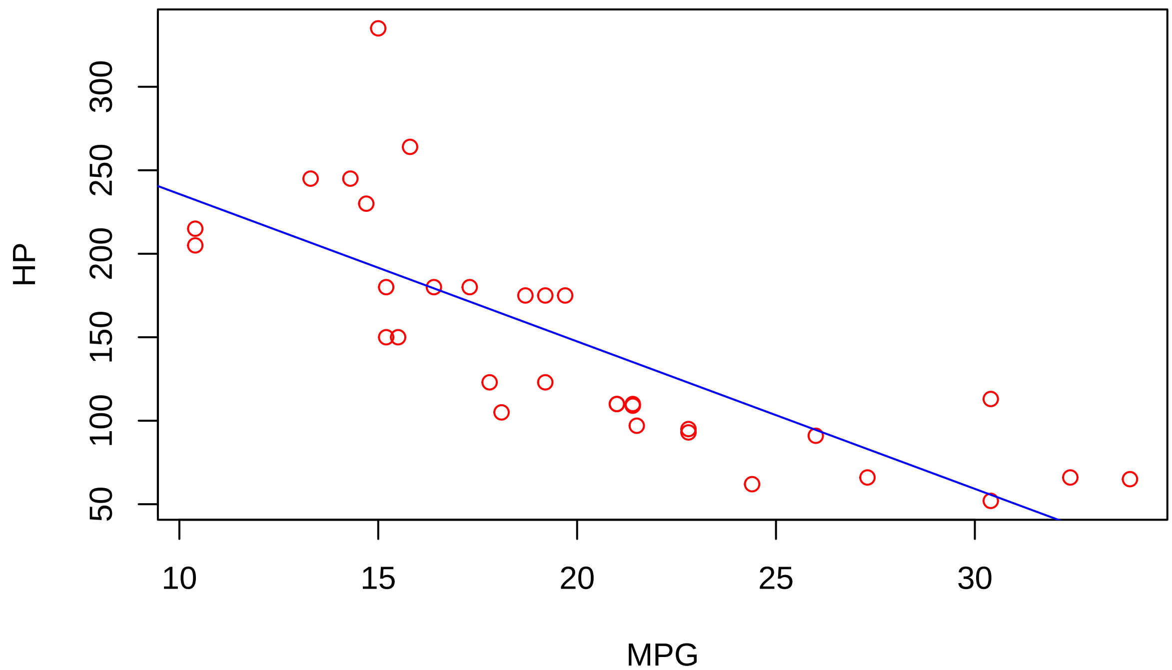
```r
#A)James wants to buy a car. He and his friend, Chris, have different opinions about the Horse Power (h

plot(mtcars$hp~mtcars$wt,xlab='Weight',ylab='HP',col='yellow')
abline(lsfit(mtcars$wt,mtcars$hp),col = "green")
```




```
JamesModel<-lm(formula =hp~wt, data = mtcars )
summary(JamesModel)
```

```
##
## Call:
## lm(formula = hp ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.430 -33.596 -13.587   7.913 172.030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.821     32.325  -0.056    0.955
## wt            46.160      9.625   4.796 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.44 on 30 degrees of freedom
## Multiple R-squared:  0.4339, Adjusted R-squared:  0.4151
## F-statistic:    23 on 1 and 30 DF,  p-value: 4.146e-05
```

```
plot(mtcars$hp~mtcars$mpg,xlab='MPG',ylab='HP',col='red')
abline(lsfit(mtcars$mpg, mtcars$hp),col = "blue")
```




```
ChrisModel<-lm(formula =hp~mpg, data = mtcars )
summary(ChrisModel)
```

```
##
## Call:
## lm(formula = hp ~ mpg, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.26 -28.93 -13.45  25.65 143.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   324.08      27.43  11.813 8.25e-13 ***
## mpg            -8.83       1.31  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.95 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

```
#James' estimation is 43%, while Chris estimation is 60%. Therefore Chris is correct.

#B)Build a model that uses the number of cylinders (cyl) and the mile per gallon (mpg) values of a car
```

```r
HpModel<-lm(hp~cyl+mpg,data = mtcars)
summary(HpModel)
```

```
##
## Call:
## lm(formula = hp ~ cyl + mpg, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.72 -22.18 -10.13  14.47 130.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.067     86.093   0.628  0.53492
## cyl           23.979      7.346   3.264  0.00281 **
## mpg           -2.775      2.177  -1.275  0.21253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.22 on 29 degrees of freedom
## Multiple R-squared:  0.7093, Adjusted R-squared:  0.6892
## F-statistic: 35.37 on 2 and 29 DF,  p-value: 1.663e-08
```

```r
estimated_HP<-predict(HpModel,data.frame(cyl=4,mpg=22))
estimated_HP
```

```
##        1
## 88.93618
```

```r
predict(HpModel,data.frame(cyl=4,mpg=22),interval = "prediction",level = 0.85)
```

```
##        fit      lwr      upr
## 1 88.93618 28.53849 149.3339
```

#Question 3

```r
#call package mlbench
library(mlbench)
```

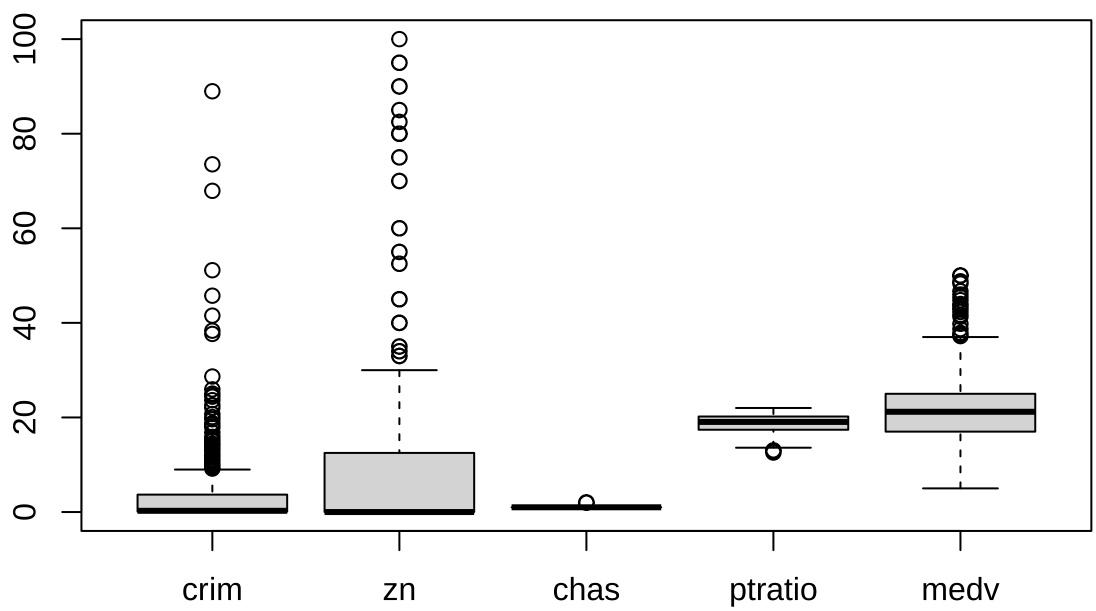
```
## Warning: package 'mlbench' was built under R version 4.2.2
```

```r
data(BostonHousing)
str(BostonHousing)
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : num  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ b      : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

```r
#boxplot of values
boxplot(BostonHousing[,c(1,2,4,11,14)])
```




```r
#A) Build a model to estimate the median value of owner-occupied homes (medv)based on the following var

set.seed(123)
MModel<-lm(medv~crim+zn+ptratio+chas,data = BostonHousing)
summary(MModel)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + ptratio + chas, data = BostonHousing)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.91868    3.23497  15.431  < 2e-16 ***
## crim        -0.26018    0.04015  -6.480 2.20e-10 ***
## zn           0.07073    0.01548   4.570 6.14e-06 ***
## ptratio     -1.49367    0.17144  -8.712  < 2e-16 ***
## chas1        4.58393    1.31108   3.496 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.388 on 501 degrees of freedom
## Multiple R-squared:  0.3599, Adjusted R-squared:  0.3547
## F-statistic: 70.41 on 4 and 501 DF,  p-value: < 2.2e-16
```

```
#No the model has an accuracy of 36% and is therefore not accurate enough.

#B) Use the estimated coefficient to answer these questions?
#I). Imagine two houses that are identical in all aspects but one bounds the Chas River and the other d

# The Chas1 bound river has a coefficient of 4.583 and the median value of the homes are in the 1000s.


#C) Which of the variables are statistically important (i.e. related to the house price)? Hint: use the

#Yes, the p-values of all the variables are not equal to zero that means that we can very comfortably r

#Statistically, all the variables are important because none of the p-values equal to zero which proves

# D) Use the anova analysis and determine the order of importance of these four variables. (18% of tota
anova(MModel)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## crim        1  6440.8  6440.8 118.007 < 2.2e-16 ***
## zn          1  3554.3  3554.3  65.122 5.253e-15 ***
## ptratio     1  4709.5  4709.5  86.287 < 2.2e-16 ***
## chas        1   667.2   667.2  12.224 0.0005137 ***
## Residuals 501 27344.5    54.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#The order of importance is crim, ptratio,zn, chas. Because the sum squared value of the crim is way hi
```