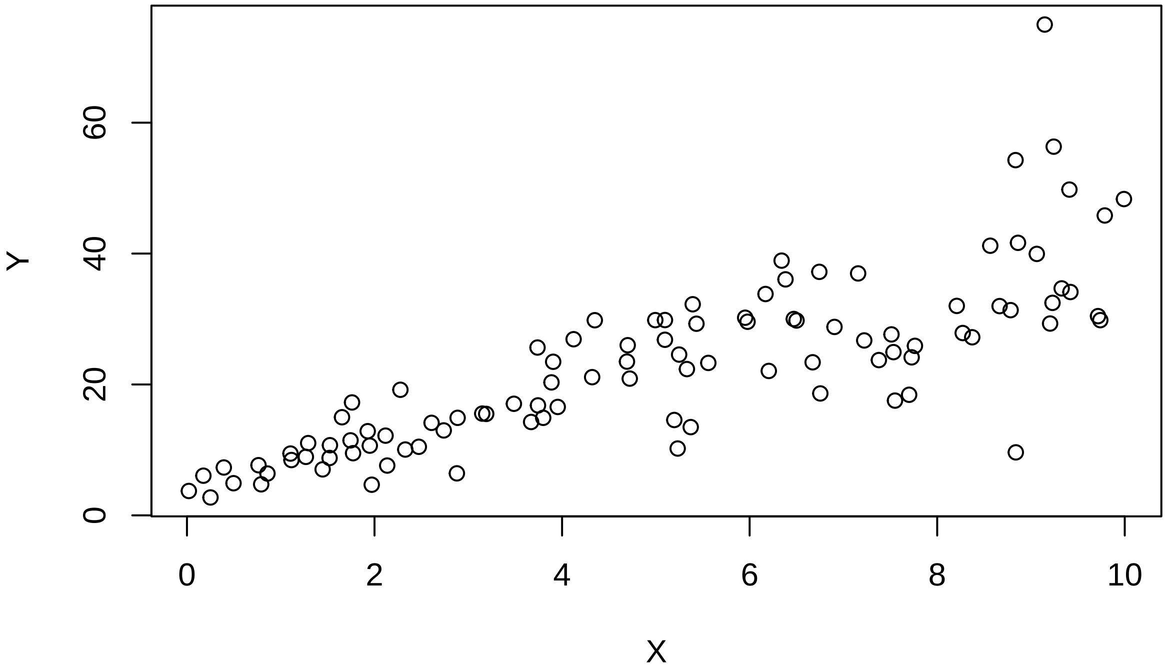
Regression assignment

Tongxiang Lu

2022-11-10

Question 1

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



Question 1a

Based on the plot, I think it fits a linear model because the graph shows positive relationship between X and Y, when Y increases, X increases.

```
Model =lm(Y~X)
Model
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)            X
##       4.465        3.611
```

```
summary(Model)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.755  -3.846  -0.387   4.318  37.503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4655     1.5537   2.874  0.00497 **
## X             3.6108     0.2666  13.542  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.756 on 98 degrees of freedom
## Multiple R-squared:  0.6517, Adjusted R-squared:  0.6482
## F-statistic: 183.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

Question 1b

The regression equation is Y=4.465 + 3.611*X. From summary, I get R2 0.65 which means the model explains 65% variability of Y variable, however, we still miss 35% variability, thus X is just a not-bad predictor of Y.

```
cor(X,Y)^2
```

```
## [1] 0.6517187
```

Question 1c

From summary, I get R2 65%, and I get correlation coefficient of X and Y is 65%, therefore, they are greatly related to each other.

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
lm(formula = mtcars$hp ~ mtcars$wt, data = mtcars)
```

```
## 
## Call:
## lm(formula = mtcars$hp ~ mtcars$wt, data = mtcars)
## 
## Coefficients:
## (Intercept)    mtcars$wt  
##      -1.821       46.160
```

```r
lm(formula = mtcars$hp ~ mtcars$mpg, data = mtcars)
```

```
## 
## Call:
## lm(formula = mtcars$hp ~ mtcars$mpg, data = mtcars)
## 
## Coefficients:
## (Intercept)   mtcars$mpg  
##      324.08        -8.83
```

Question 2a

Based on the above data, I think James is right for his correlation coefficient (hp and weight)is 46.160, which is much bigger than Chris's correlation coefficient(hp and mlp) which is -8.83. Thus, weight is more related to HP than mlp.

```r
Model=lm(mtcars$hp ~ mtcars$cyl + mtcars$mpg, data = mtcars)
Model
```

```
## 
## Call:
## lm(formula = mtcars$hp ~ mtcars$cyl + mtcars$mpg, data = mtcars)
## 
## Coefficients:
## (Intercept)   mtcars$cyl   mtcars$mpg  
##      54.067       23.979       -2.775
```

```r
HP=54.067+ 23.979*4 - 2.775*22
HP
```

```
## [1] 88.933
```

Question 2b

The estimated Horse Power of a car with 4 calendar and mpg of 22 is 88.933.

Question 3

```
library(mlbench)
```

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

```
data(BostonHousing)
Model=lm(BostonHousing$medv~BostonHousing$crim + (BostonHousing$zn > 25) +
         BostonHousing$ptratio+BostonHousing$chas)
summary(Model)
```

```
##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$crim + (BostonHousing$zn >
##     25) + BostonHousing$ptratio + BostonHousing$chas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.806  -4.657  -1.081   2.691  32.560
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                52.57192    3.17329  16.567  < 2e-16 ***
## BostonHousing$crim         -0.26724    0.04053  -6.593  1.1e-10 ***
## BostonHousing$zn > 25TRUE   3.11260    0.98593   3.157  0.00169 **
## BostonHousing$ptratio      -1.61708    0.16933  -9.550  < 2e-16 ***
## BostonHousing$chas1         4.29214    1.32198   3.247  0.00125 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.466 on 501 degrees of freedom
## Multiple R-squared:  0.3462, Adjusted R-squared:  0.341
## F-statistic: 66.32 on 4 and 501 DF,  p-value: < 2.2e-16
```

Question 3a

From summary, I get R2 34%, while we miss 66% data, my conclusion is that this is not an accurate model.

Set Model1 as houses bounds the Chas River, Model 2 as houses bounds no river.

```
Model1=lm(BostonHousing$medv~BostonHousing$crim + (BostonHousing$zn > 25) +
         BostonHousing$ptratio+BostonHousing$chas)
Model1
```

```
##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$crim + (BostonHousing$zn >
##     25) + BostonHousing$ptratio + BostonHousing$chas)
##
## Coefficients:
##              (Intercept)         BostonHousing$crim
##                  52.5719                    -0.2672
## BostonHousing$zn > 25TRUE      BostonHousing$ptratio
##                   3.1126                    -1.6171
##       BostonHousing$chas1
##                   4.2921
```

```
Model2=lm(BostonHousing$medv~BostonHousing$crim + (BostonHousing$zn > 25) +
         BostonHousing$ptratio)
Model2
```

```
##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$crim + (BostonHousing$zn >
##     25) + BostonHousing$ptratio)
##
## Coefficients:
##              (Intercept)         BostonHousing$crim
##                   54.215                     -0.271
## BostonHousing$zn > 25TRUE      BostonHousing$ptratio
##                    2.870                     -1.687
```

Question 3b

From the data shown, the house bounds the river is more expensive than the house bounds no river, by US$4292.1 per square.

Question 3c

I get smallest p-value < 2.2e-16 with BostonHousing$ptratio variable, that is, pupil-teacher ratio by town has a biggest influence to Boston housring compared

with the other three variables such as crime(per capita crime rate by town), chas(Charles River) and zn(proportion of residential land zoned for lots over 25,000 sq.ft).

Question 3d

```
anova(Model)
```

```
## Analysis of Variance Table
##
## Response: BostonHousing$medv
##                          Df  Sum Sq Mean Sq F value    Pr(>F)
## BostonHousing$crim      1  6440.8  6440.8 115.537 < 2.2e-16 ***
## BostonHousing$zn > 25   1  2131.8  2131.8  38.242 1.298e-09 ***
## BostonHousing$ptratio   1  5627.1  5627.1 100.941 < 2.2e-16 ***
## BostonHousing$chas      1   587.6   587.6  10.541  0.001245 **
## Residuals             501 27928.9    55.7
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

From anova analysis, we learn the sum sq of crim, zn, ptratio are all big

values whereas the chas is small value, in other words, crim has the first important influence to Bostonhousing price, ptratio the second, zn the third, whereas chas has the least influence to Bostonhousing price.