Lab Experiment - 1

Prithviraj Guntha

16/12/2022

Question 1

```
#question 1
x <- c(0.01,0.48,0.71,0.95,1.19,0.01,0.48,1.44,0.71,1.96)
y <- c(127.6,124,110.8,103.9,101.5,130,122,92.3,113,83.7)
model <- lm(x~y)
sse <- sum((fitted(model) - x)^2)
sse
```

```
## [1] 0.1116899
```

```
ssr <- sum((fitted(model)-mean(x))^2)
ssr
```

```
## [1] 3.28695
```

```
sst <- ssr + sse
sst
```

```
## [1] 3.39864
```

```
mse <- mean((x - y)^2)
mse
```

```
## [1] 12352.75
```

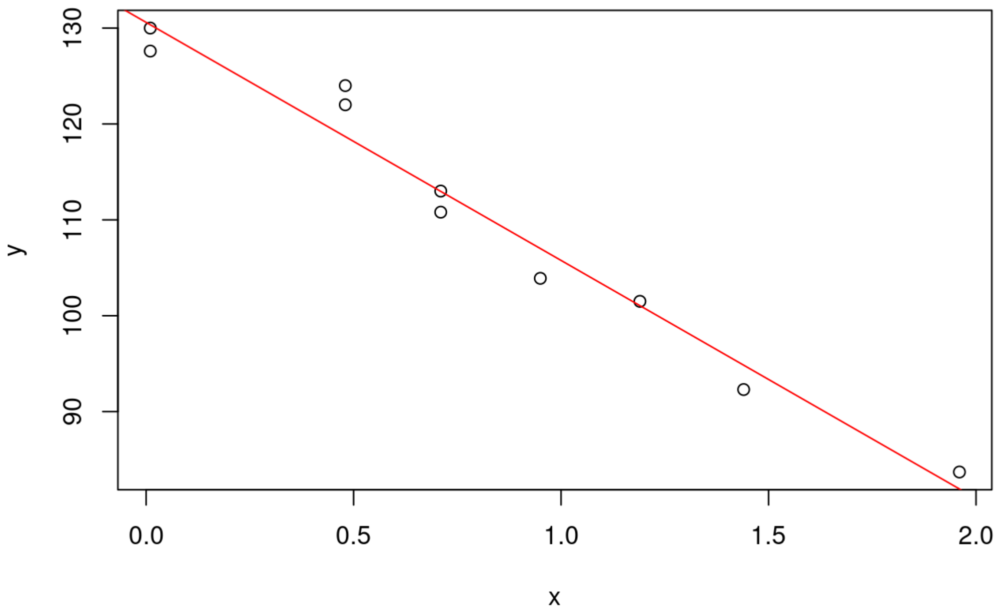
```
rmse <- sqrt(mse)
rmse
```

```
## [1] 111.1429
```

```
summary(model)
```

```
##
## Call:
## lm(formula = x ~ y)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13268 -0.08478 -0.02030  0.08806  0.19708
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.113276   0.283968   18.01 9.28e-08 ***
## y           -0.038955   0.002539  -15.34 3.23e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1182 on 8 degrees of freedom
## Multiple R-squared:  0.9671, Adjusted R-squared:  0.963
## F-statistic: 235.4 on 1 and 8 DF,  p-value: 3.232e-07
```

```
plot(x,y)
abline(lm(y~x),col='red')
```



inference

from the above model 96% variation which means that the points lie much closer to the regression line, this in turn means that there is a good relationship between the x and y value. the relationship between x and y is negative.

Question 2

```
data <- data.frame(x1 = c(95.2,85.1,80.6,70.5,60.2,70.2,75.1),
                   x2 = c(88.1,76.5,79.2,85.4,90.2,74.3,67.7),
                   y = c(85.9,85.2,70.3,65.4,70.4,66,71.1))
head(data)
```

```
##     x1   x2    y
## 1 95.2 88.1 85.9
## 2 85.1 76.5 85.2
## 3 80.6 79.2 70.3
## 4 70.5 85.4 65.4
## 5 60.2 90.2 70.4
## 6 70.2 74.3 66.0
```

```
model1 <- lm(data$x1~data$y)
sse <- sum((fitted(model1) - data$x1)^2)
sse
```

```
## [1] 281.1143
```

```
ssr <- sum((fitted(model1)-mean(data$x1))^2)
ssr
```

```
## [1] 502.4057
```

```
sst <- ssr + sse
sst
```

```
## [1] 783.52
```

```
mse <- mean((data$x1 - data$y)^2)
mse
```

```
## [1] 50.89714
```

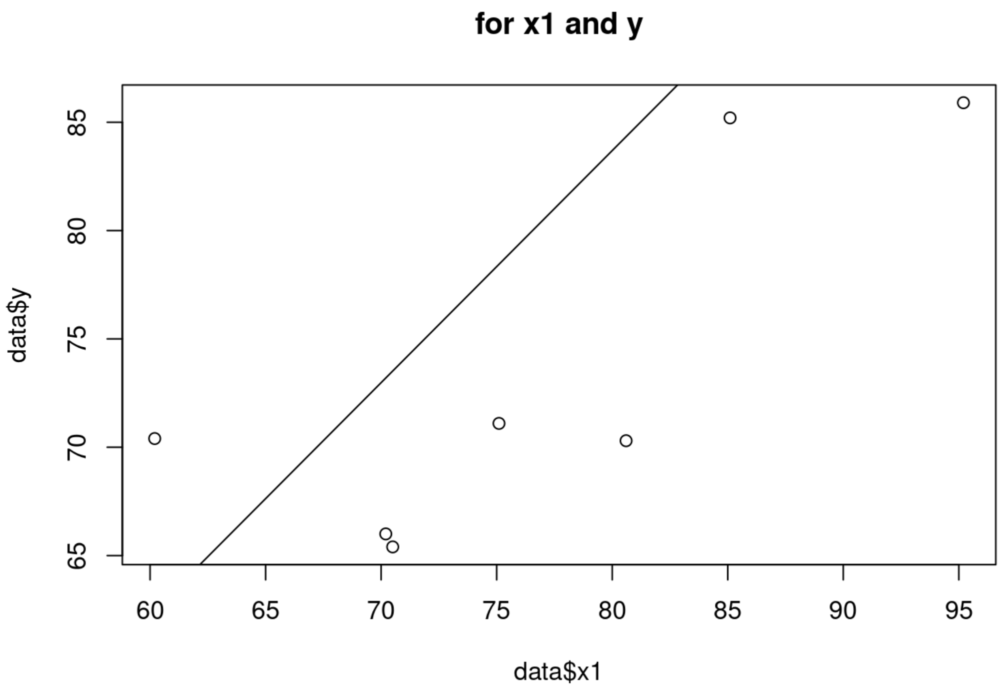
```
rmse <- sqrt(mse)
rmse
```

```
## [1] 7.134223
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = data$x1 ~ data$y)
##
## Residuals:
##        1        2        3        4        5        6        7
##   5.1907  -4.1597   7.2962   2.4434 -13.2109   1.5009   0.9395
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.9778    26.4719  -0.075   0.9433
## data$y        1.0709     0.3582   2.989   0.0305 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.498 on 5 degrees of freedom
## Multiple R-squared:  0.6412, Adjusted R-squared:  0.5695
## F-statistic: 8.936 on 1 and 5 DF,  p-value: 0.03047
```

```
plot(data$x1,data$y,main = "for x1 and y")
abline(lm(data$x1~data$y))
```




```
model2 <- lm(data$x2~data$y)
sse <- sum((fitted(model2) - data$x2)^2)
sse
```

```
## [1] 387.2546
```

```
ssr <- sum((fitted(model2)-mean(data$x2))^2)
ssr
```

```
## [1] 7.945415
```

```
sst <- ssr + sse
sst
```

```
## [1] 395.2
```

```
mse <- mean((data$x2 - data$y)^2)
mse
```

```
## [1] 147.4614
```

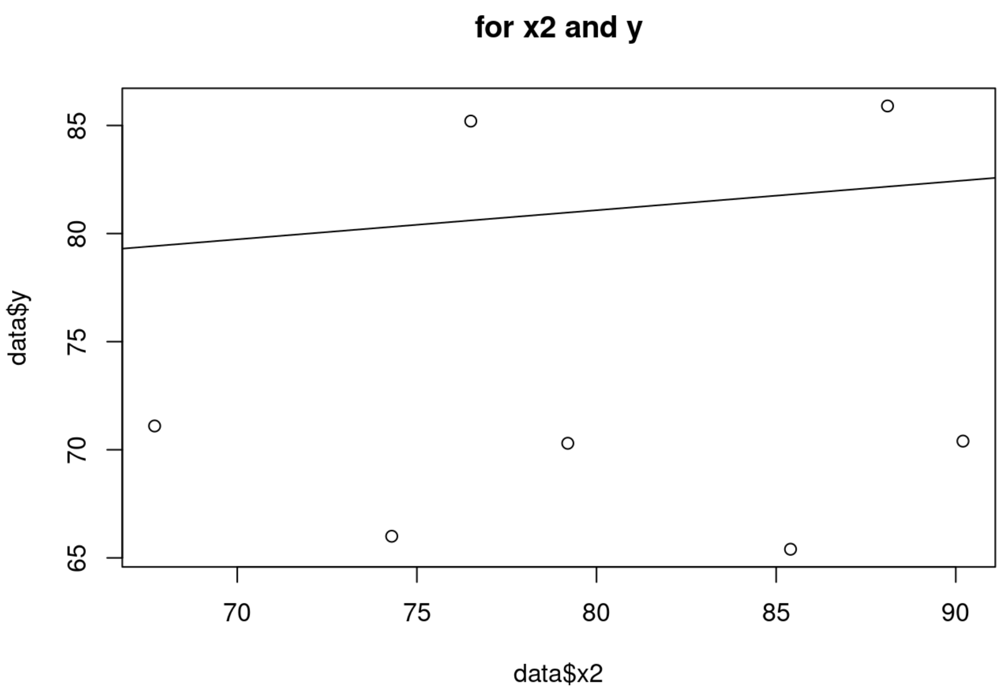
```
rmse <- sqrt(mse)
rmse
```

```
## [1] 12.14337
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = data$x2 ~ data$y)
##
## Residuals:
##       1        2        3        4        5        6        7
##   6.2263  -5.2795  -0.5729   6.2870  10.4136  -4.8938 -12.1806
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.3057    31.0700   2.263   0.0731 .
## data$y        0.1347     0.4205   0.320   0.7617
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.801 on 5 degrees of freedom
## Multiple R-squared:  0.0201, Adjusted R-squared:  -0.1759
## F-statistic: 0.1026 on 1 and 5 DF,  p-value: 0.7617
```

```
plot(data$x2,data$y,main = "for x2 and y")
abline(lm(data$x2~data$y))
```




```
anova(model1,model2)
```

```
## Warning in anova.lmlist(object, ...): models with response '"data$x2"' removed
## because response differs from model 1
```

```
## Analysis of Variance Table
##
## Response: data$x1
##           Df Sum Sq Mean Sq F value  Pr(>F)
## data$y     1 502.41  502.41   8.936 0.03047 *
## Residuals  5 281.11   56.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

inference

in model 1(x1 and y), there is 56% variation which means that the points lie moderately to the regression line, this in turn means that there isn't a good relationship between x1 and y. the relationship between x1 and y is positive.

in model 2(x2 and y), there is very low variation which means that the points lie very far from the regression line, this in turn means that there isn't a relationship between x2 and y.

hence, model1(x1 and y) shows a better relationship than model2(x2 and y).

Question 3

we have obtained real time data of number of hours student studied and the marks they got

```
stud <- read.csv("/home/student1/Downloads/score.csv")
head(stud)
```

```
##   Hours Scores
## 1   2.5     21
## 2   5.1     47
## 3   3.2     27
## 4   8.5     75
## 5   3.5     30
## 6   1.5     20
```

```
mod <- lm(stud$Hours~stud$Scores)
sse <- sum((fitted(mod) - stud$Hours)^2)
sse
```

```
## [1] 7.200168
```

```
ssr <- sum((fitted(mod)-mean(stud$Hours))^2)
ssr
```

```
## [1] 145.8262
```

```
sst <- ssr + sse
sst
```

```
## [1] 153.0264
```

```
mse <- mean((stud$Hours - stud$Scores)^2)
mse
```

```
## [1] 2659.569
```

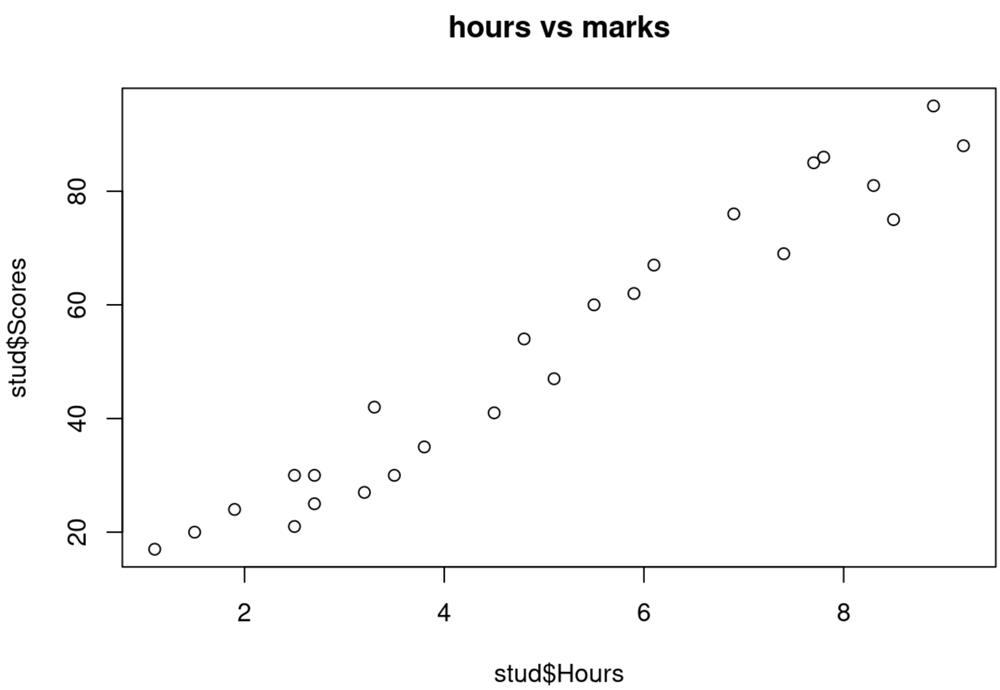
```
rmse <- sqrt(mse)
rmse
```

```
## [1] 51.57101
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = stud$Hours ~ stud$Scores)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7879 -0.4433 -0.2181  0.5096  1.1953
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.006286   0.258038  -0.024    0.981
## stud$Scores  0.097480   0.004517  21.583   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5595 on 23 degrees of freedom
## Multiple R-squared:  0.9529, Adjusted R-squared:  0.9509
## F-statistic: 465.8 on 1 and 23 DF,  p-value: < 2.2e-16
```

```
plot(stud$Hours,stud$Scores,main = "hours vs marks")
abline(lm(stud$Hours~stud$Scores))
```



inference

from the above model 95% variation which means that the points lie much closer to the regression line, this in turn means that there is a good relationship between the x and y value. the relationship between x and y is positive. if a student studies for longer hours, he/she can score more marks.