MSMS - 106

Ananda Biswas

Practical 02

✎ Using the following bivariate data, obtain the following results.

X	Y	X	Y
12.4	11.2	17.3	15.1
14.3	12.5	18.4	16.1
14.5	12.7	19.2	16.8
14.9	13.1	17.4	15.2
16.1	14.1	17	14.9
16.9	14.8	17.9	15.6
16.5	14.4	18.8	16.4
15.4	13.4	20.3	17.7
22.4	19.6	19.5	17
19.4	16.9	19.7	17.2
15.5	14	21.2	18.6
16.7	14.6		

(a) Karl Pearson Correlation Coefficient,

(b) Spearman's Rank Correlation Coefficient,

(c) Regression line of Y on X,

(d) Regression line of X on Y,

(e) Scatterplot of X and Y with a regression line.

Loading the data and previous implementations

```
github_path <- 'https://raw.githubusercontent.com/sakunisgithub/data_sets/master/msc_semester_1/sonam_madam_practical_02_data.csv'
our_data <- read.csv(github_path)

source('https://raw.githubusercontent.com/sakunisgithub/R-programming/master/my_implementations.R')
```

⊕ *Karl Pearson's Correlation Coefficient*

Here we compute $r_{xy} = \dfrac{\text{cov}\ (x, y)}{\sqrt{\text{var}\ (x)} \cdot \sqrt{\text{var}\ (y)}}$, where $\text{cov}\ (x, y) = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$.

```
my_covariance_function <- function(x, y){

  x_bar <- my_mean_function(x)
  y_bar <- my_mean_function(y)

  temp <- 0
  for (i in 1:length(x)) {
    temp <- temp + (x[i] - x_bar) * (y[i] - y_bar)
  }
  return(temp/(length(x) - 1))
}
```

```
my_correlation_function <- function(x, y){

  var_x <- my_sample_central_moments_function(x, 2)
  var_y <- my_sample_central_moments_function(y, 2)
  cor_xy <- my_covariance_function(x, y) / sqrt(var_x * var_y)

  return(cor_xy)
}
```

```
my_correlation_function(our_data$X, our_data$Y)

## [1] 0.9985405

cor(our_data$X, our_data$Y, method = "pearson")

## [1] 0.9985405
```

⊘ Results matched !!

• Correlation coefficient is almost 1, implying a nearly perfect linear relationship between X and Y.

⊕ *Spearman's Rank Correlation Coefficient*

Here we compute $\rho = 1 - \dfrac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$ where d_i is the difference between the ranks of x_i and y_i.

```
my_rank_correlation_function <- function(x, y){

  x_sorted <- my_selection_sort(x)
  y_sorted <- my_selection_sort(y)

  rank_x <- rep(0, 23); rank_y <- rep(0, 23)

  for (i in 1:length(x_sorted)) {
    rank_x[i] <- my_mean_function(which(x_sorted == x[i]))
    rank_y[i] <- my_mean_function(which(y_sorted == y[i]))
  }

  sum_di_sq <- 0

  for (i in 1:length(x)) {
    sum_di_sq <- sum_di_sq + (rank_x[i] - rank_y[i])^2
  }

  rank_corr <- 1 - (6 * sum_di_sq) / (length(x) * (length(x)^2- 1))

  return(rank_corr)
}
```

```
my_rank_correlation_function(our_data$X, our_data$Y)

## [1] 1
```

```
cor(our_data$X, our_data$Y, method = "spearman")

## [1] 1
```

⊘ Results matched !!

- There is perfect agreement between X and Y.

Regression equation of Y on X

Let the regression equation of Y on X be $Y = aX + b$. Then $\hat{a} = r_{xy} \cdot \frac{s_y}{s_x}$ and $\hat{b} = \bar{y} - \hat{a}\bar{x}$.

```
x_bar <- my_mean_function(our_data$X)
y_bar <- my_mean_function(our_data$Y)

r_xy <- my_correlation_function(our_data$X, our_data$Y)
s_x <- sqrt(my_sample_central_moments_function(our_data$X, 2))
s_y <- sqrt(my_sample_central_moments_function(our_data$Y, 2))

a_yx <- r_xy * (s_y / s_x)
b_yx <- y_bar - a_yx * x_bar

b_yx; a_yx

## [1] 0.434585
## [1] 0.851144
```

```
fit_y_on_x <- summary(lm(Y ~ X, data = our_data))
fit_y_on_x$coefficients

##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 0.434585 0.17704865  2.454608 2.291188e-02
## X           0.851144 0.01004585 84.725922 4.148402e-28
```

✓ Results matched !!

Regressin equation of X on Y

Let the regression equation of X on Y be $X = aY + b$. Then $\hat{a} = r_{xy} \cdot \frac{s_x}{s_y}$ and $\hat{b} = \bar{x} - \hat{a}\bar{y}$.

```
a_xy <- r_xy * (s_x / s_y)
b_xy <- x_bar - a_xy * y_bar

b_xy; a_xy

## [1] -0.458156
## [1] 1.171462
```

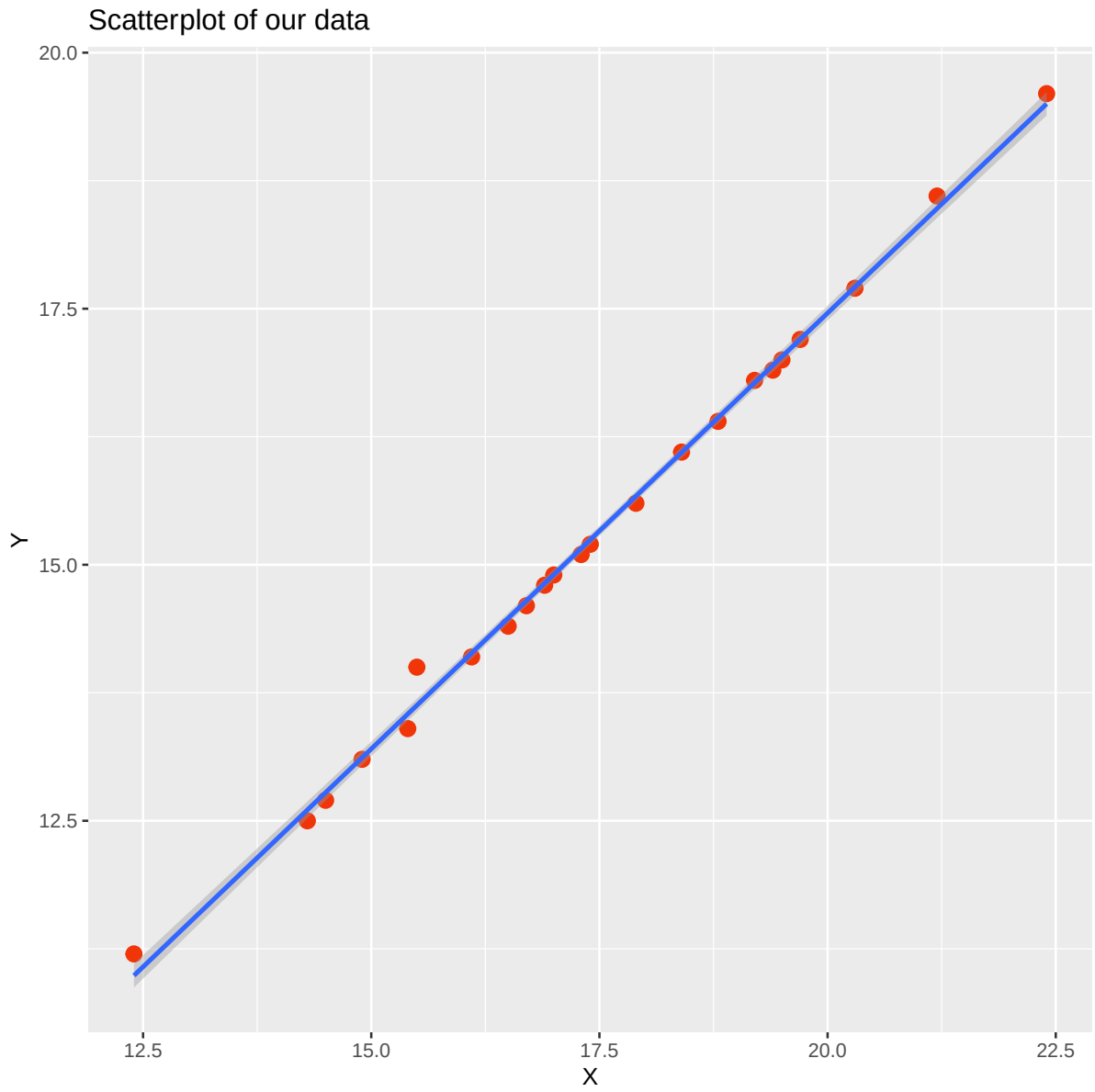
```
fit_x_on_y <- summary(lm(X ~ Y, data = our_data))
fit_x_on_y$coefficients

##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -0.458156 0.21336727 -2.147265 4.360262e-02
## Y            1.171462 0.01382649 84.725922 4.148402e-28
```

✓ Results matched !!

⊕ *Scatterplot of X and Y with a regression line*

```
library(tidyverse)
```

```
our_data %>%
  ggplot(aes(x = X, y = Y)) +
  geom_point(size = 3, col = "#f03608") +
  labs(title = "Scatterplot of our data") +
  geom_smooth(method = "lm", formula = y ~ x, level = 0.95)
```



- An almost perfect linear relation between X and Y is visible.