Exploratory Statistical Data Analysis With R Software (ESDAR)

Swayam Prabha

Lecture 36

Correlation Coefficient using R and Rank Correlation Coefficient

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Covariance

***X*, *Y* : Two variables**

***n* pairs of observations are available as (x_1,y_1), (x_2,y_2),..., (x_n,y_n)**

The covariance between the variables *x* and *y* is defined as

$$\text{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Similar definition is available for grouped data in frequency table.

Covariance

R command:

`x,y` : Two data vectors

`cov(x,y)`: covariance between x and y.

Command `cov(x,y)` calculates the covariance with divisor (*n* – 1)

$$\operatorname{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Coefficient of Correlation

Also called as Karl Pearson Coefficient of Correlation

$$r \equiv r(x,y) = \frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x).\operatorname{var}(y)}}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\ \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{n}x_i y_i - n\,\bar{x}\,\bar{y}}{\sqrt{\left(\sum_{i=1}^{n}x_i{}^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n}y_i{}^2 - n\bar{y}^2\right)}}$$

Coefficient of Correlation

R Command

`cor(x,y)` **computes the correlation between x and y**

```
cor(x, y, use = "everything", method =
c("pearson", "kendall", "spearman"))
```

`x` **: a numeric vector, matrix or data frame.**

`y` **: a numeric vector, matrix or data frame with compatible dimensions to x.**

Coefficient of Correlation

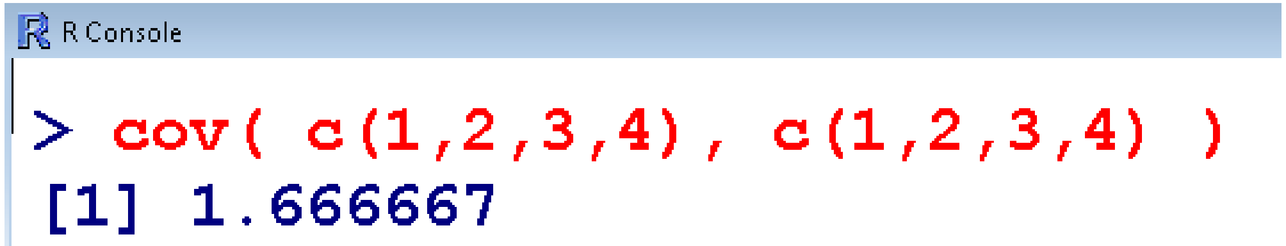
`use` : an optional character string giving a method for computing covariances in the presence of missing values. This must be (an abbreviation of) one of the strings `"everything"`, `"all.obs"`, `"complete.obs"`, `"na.or.complete"`, or `"pairwise.complete.obs"`.

`method` : a character string indicating which correlation coefficient (or covariance) is to be computed. One of `"pearson"` (default), `"kendall"`, or `"spearman"` can be abbreviated.

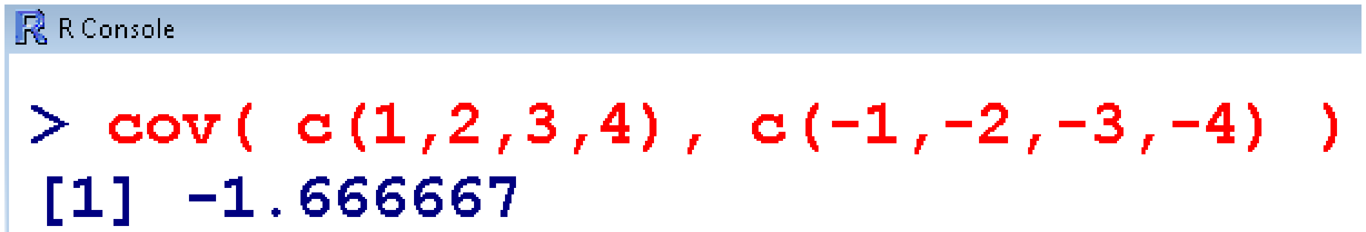
Example

Covariance

```
> cov( c(1,2,3,4), c(1,2,3,4) )
[1]  1.666667
```




```
> cov( c(1,2,3,4), c(-1,-2,-3,-4) )
 [1] -1.666667
```



Example

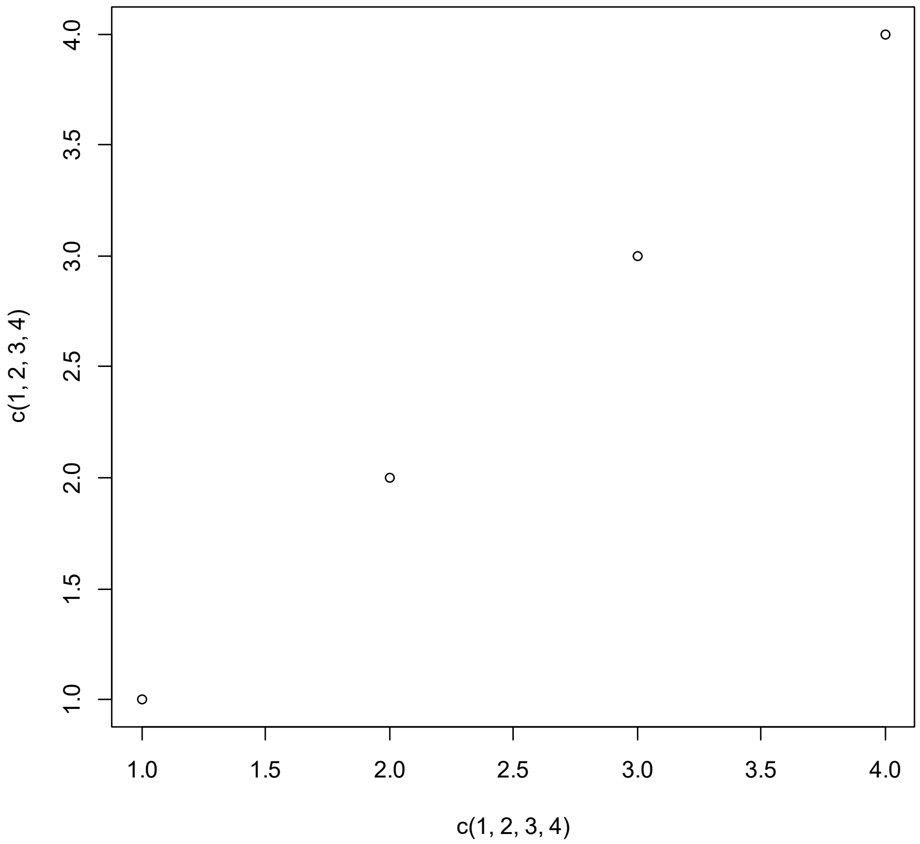
Correlation coefficient

Exact positive linear dependence

```
> cor( c(1,2,3,4), c(1,2,3,4) )
[1] 1
```

R Console

```
> cor( c(1,2,3,4), c(1,2,3,4) )
[1] 1
```

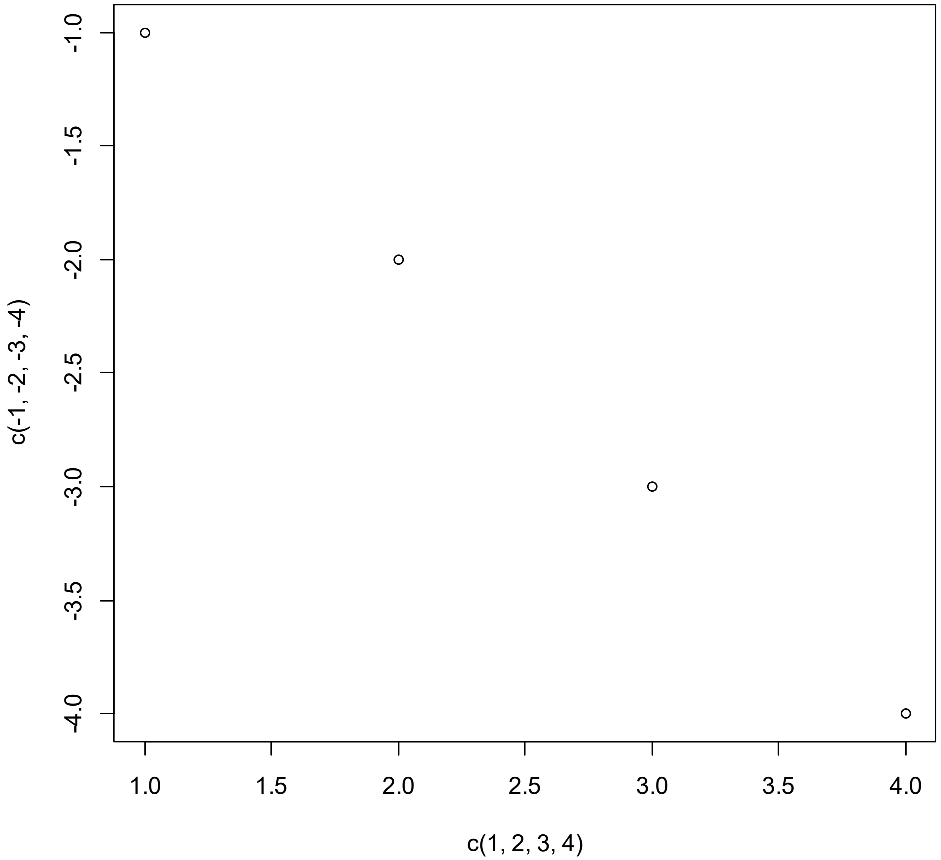


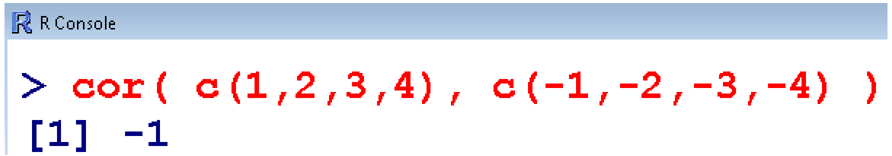
Example

Correlation coefficient

Exact negative linear dependence

```
> cor( c(1,2,3,4), c(-1,-2,-3,-4) )
[1] -1
```





Coefficient of Correlation

Example

Data on marks obtained by 20 students out of 500 marks and the number of hours they studied per week are recorded as follows:

We know from experience that marks obtained by students increase as the number of hours increase.

Marks	337	316	327	340	374	330	352	353	370	380
Number of hours per week	23	25	26	27	30	26	29	32	33	34

Marks	384	398	413	428	430	438	439	479	460	450
Number of hours per week	35	38	39	42	43	44	45	46	44	41

Coefficient of Correlation

Example

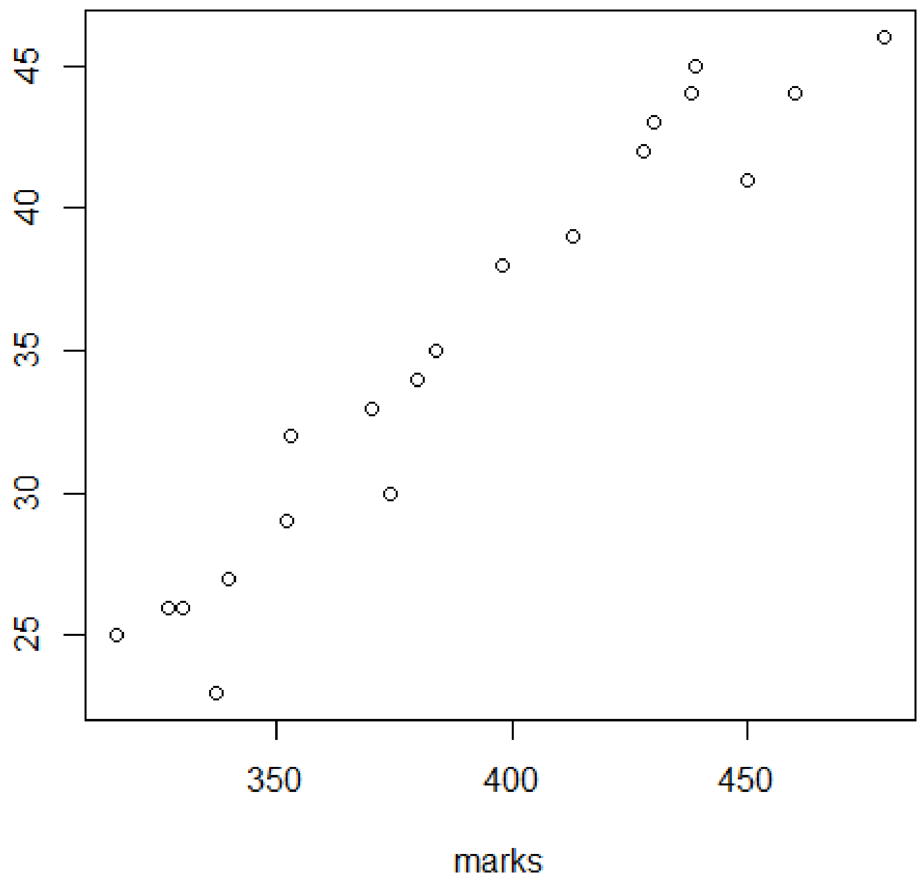
```
marks =
c(337,316,327,340,374,330,352,353,370,380,384,39
8,413,428,430,438,439,479,460,450)

hours =
c(23,25,26,27,30,26,29,32,33,34,35,38,39,42,43,4
4,45,46,44,41)
```

Coefficient of Correlation

Example

```
> plot(marks, hours)
```



Coefficient of Correlation

Example

```
> cor(marks, hours)
[1] 0.9679961

> cor(hours,marks)
[1] 0.9679961
```

Sign of correlation coefficient is positive.

As number of hours of study per week are increasing, marks obtained are also increasing.

Coefficient of Correlation
Example

```
> marks
 [1] 337 316 327 340 374 330 352 353 370 380 384 398 413 428 430
[16] 438 439 479 460 450
> hours
 [1] 23 25 26 27 30 26 29 32 33 34 35 38 39 42 43 44 45 46 44 41
> cor(marks, hours)
[1] 0.9679961
> cor(hours,marks)
[1] 0.9679961
```

Association between Ranks
Spearman‘s Rank Correlation Coefficient

Example: **Ranked observations**

- **Two judges give ranks to a fashion model.**
- **Two persons give ranks to food prepared or their scores are ranked.**

These observations are the ranks of two variables (two judges).

Association between Ranks
Spearman‘s Rank Correlation Coefficient

Two variables : *X, Y*

***n* observations on *X* and *Y* are available.**

n* observations are ranked with respect to *X* and *Y.

Ranks of the *n* observations are recorded.

Association between Ranks
Spearman's Rank Correlation Coefficient

Judge *X* gives ranks to *n* candidates as

- **Rank 1 to worst candidate with lowest score x_i**
- **Rank 2 to candidate with second lowest score x_i**
- ***....***
- **Rank *n* to the best candidate with highest score x_i.**

Similarly, judge *Y* give ranks to *n* candidates and gives ranks 1,2,...,*n* based on scores y_1, y_2,..., y_n. In the same way as judge *X* gave.

Association between Ranks
Spearman‘s Rank Correlation Coefficient

Every participant has two ranks given by two different judges.

We expect that both the judges give

- **higher ranks to good candidates and**
- **lower ranks to bad candidates.**

We want to measure the degree of association between the two different judgements, i.e., the two different set of ranks.

Association between Ranks
Spearman‘s Rank Correlation Coefficient

Measure the degree of agreement between the ranks of two judges.

Use Spearman‘s rank correlation coefficient.

Uses ranks of the values and not the values themselves.

Association between Ranks
Spearman‘s Rank Correlation Coefficient

$Rank(x_i)$: Rank of i^{th} observation on X.

: Rank of x_i among ordered values $x_1, x_2, \ldots, x_n$ of X.

$Rank(y_i)$: Rank of i^{th} observation on Y.

: Rank of y_i among ordered values $y_1, y_2, \ldots, y_n$ of Y.

$d_i = Rank(x_i) - Rank(y_i)$

Spearman‘s rank correlation coefficient (R) is defined as

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad ; \quad -1 \le R \le 1$$

Association between Ranks
Spearman‘s Rank Correlation Coefficient

It does not matter whether the ascending or descending order of ranks is used.

When both the judges assign exactly the

- **same ranks to all the candidates then $R = +1$**

- **opposite ranks to all the candidates then $R = -1$**

Association between Ranks
Spearman's Rank Correlation Coefficient

Example: Scores given by two judges to 5 candidates are as follows:

Candidates	Judge1		Judge2		$d_i = Rank(x_i) - Rank(y_i)$
	Scores (x_i)	Rank(x_i)	Scores(y_i)	Rank(y_i)	
1	75	4	70	4	0
2	25	1	80	5	-4
3	35	2	60	3	-1
4	95	5	30	1	4
5	50	3	40	2	1

Association between Ranks
Spearman's Rank Correlation Coefficient

$$n = 5$$

$$R = \frac{1 - 6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2 - 1)} = \frac{1 - 6\sum_{i=1}^{5} d_i^{\,2}}{5(5^2 - 1)} = -0.7$$

Association between Ranks
Spearman‘s Rank Correlation Coefficient

R Command

`cor(x,y)` computes the correlation between `x` and `y`

```
cor(x, y, use = "everything", method =
c("spearman"))
```

`x` : a numeric vector, matrix or data frame.

`y` : a numeric vector, matrix or data frame with compatible dimensions to `x`.

Association between Ranks
Spearman‘s Rank Correlation Coefficient

`use` : an optional character string giving a method for computing covariances in the presence of missing values. This must be (an abbreviation of) one of the strings `"everything"`, `"all.obs"`, `"complete.obs"`, `"na.or.complete"`, or `"pairwise.complete.obs"`.

Association between Ranks
Spearman‘s Rank Correlation Coefficient

Example: Scores given by two judges to 5 candidates are as follows:

Candidates	Judge1	Judge2
	Scores (x_i)	Scores(y_i)
1	75	70
2	25	80
3	35	60
4	95	30
5	50	40

```
> x = c(75, 25, 35, 95, 50)
> y = c(70, 80, 60, 30, 40)
```

Association between Ranks
Spearman‘s Rank Correlation Coefficient

```
> judge1 = c(75, 25, 35, 95, 50)
> judge2 = c(70, 80, 60, 30, 40)


> cor(judge1, judge2, use = "everything",
method = c("spearman"))
[1] -0.7
```

R Console

```
> judge1 = c(75, 25, 35, 95, 50)
> judge2 = c(70, 80, 60, 30, 40)
> cor(judge1, judge2, use = "everything", method = c("spearman"))
[1] -0.7
```