Descriptive Statistics With R Software

Association of Variables

::

Correlation Coefficient using R Software

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

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

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

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

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

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

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

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

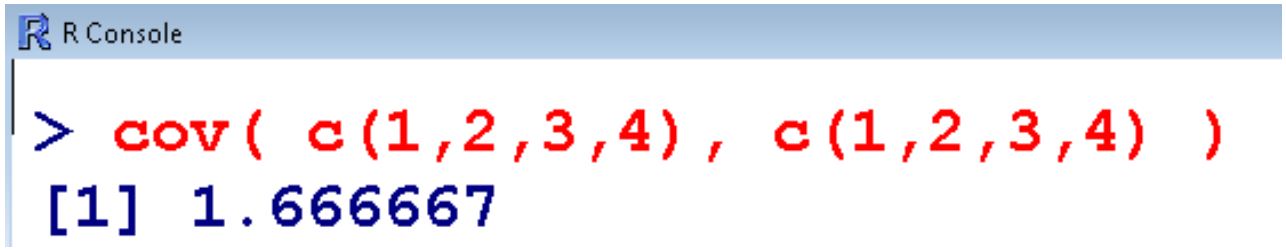
`use` : an optional character string giving a method for computing covariances in the presence of missing values. This must be (an abbreviation of) one of the strings `"everything"`, `"all.obs"`, `"complete.obs"`, `"na.or.complete"`, or `"pairwise.complete.obs"`.

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

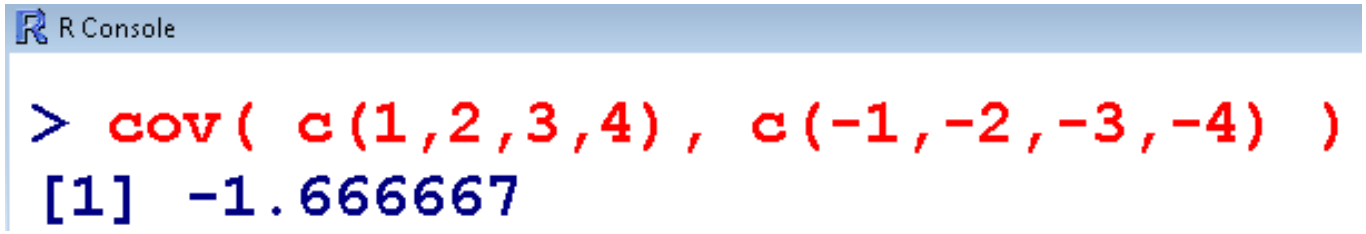
Example

Covarianve

```
> 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
```

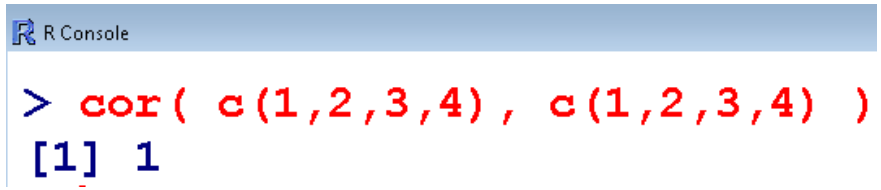


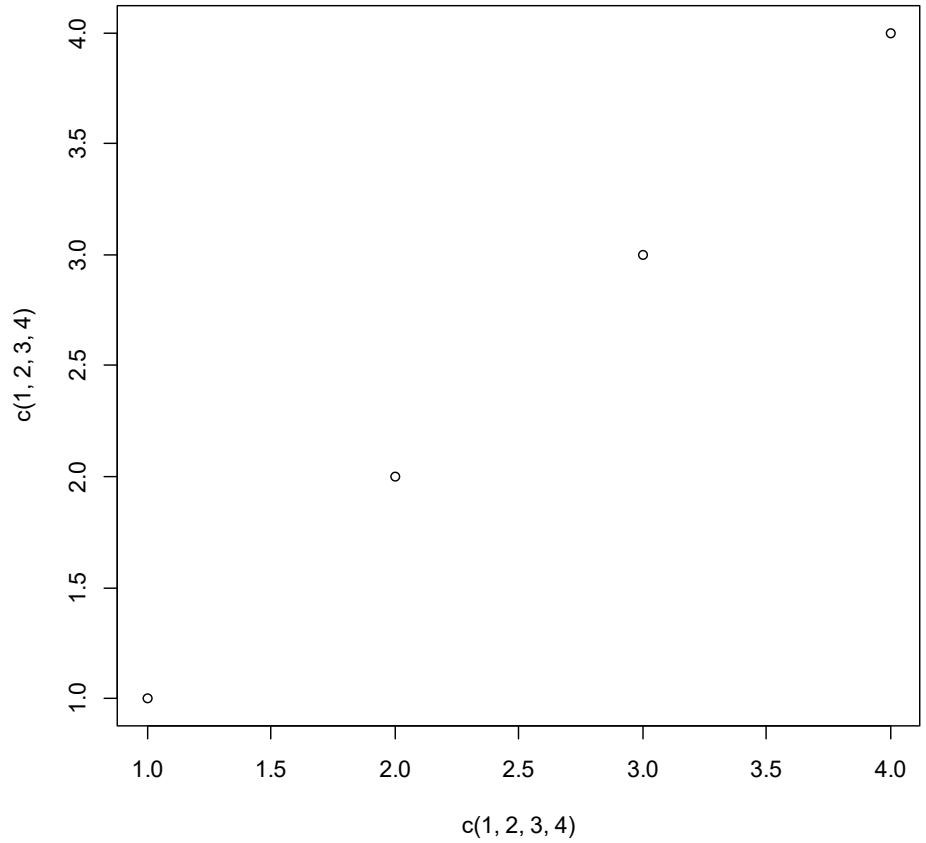
Example

Correlation coefficient

Exact positive linear dependence

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



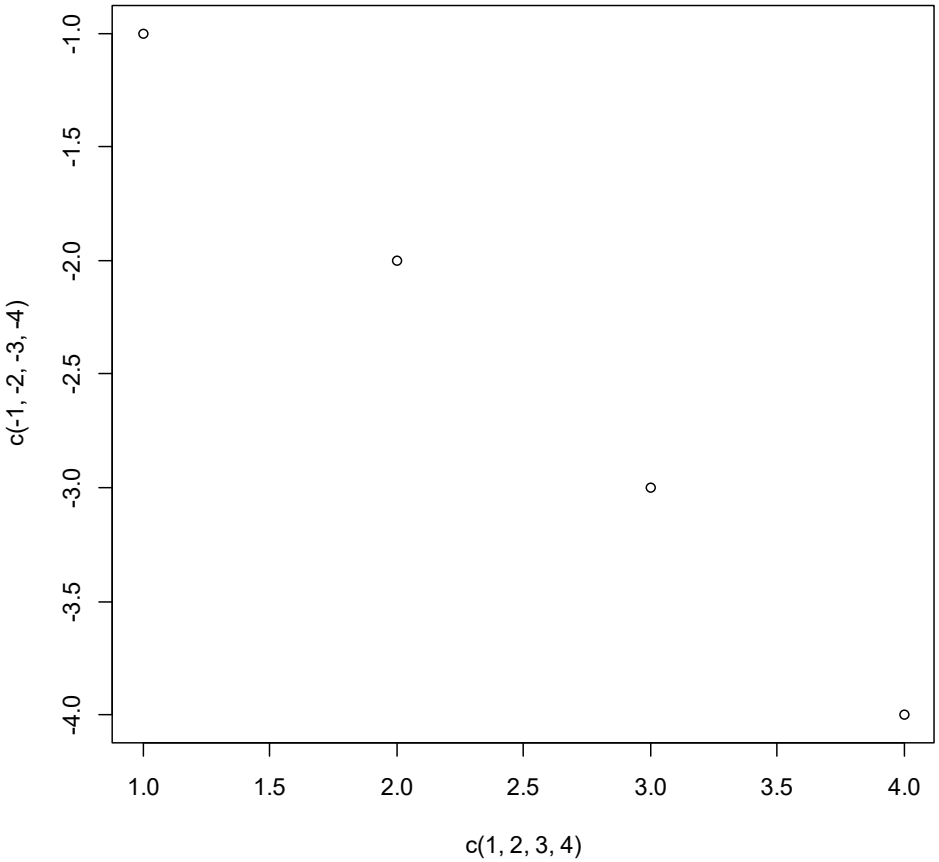


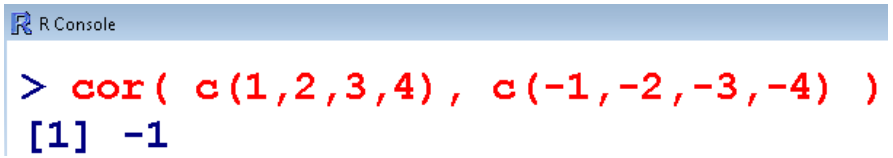
Example

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

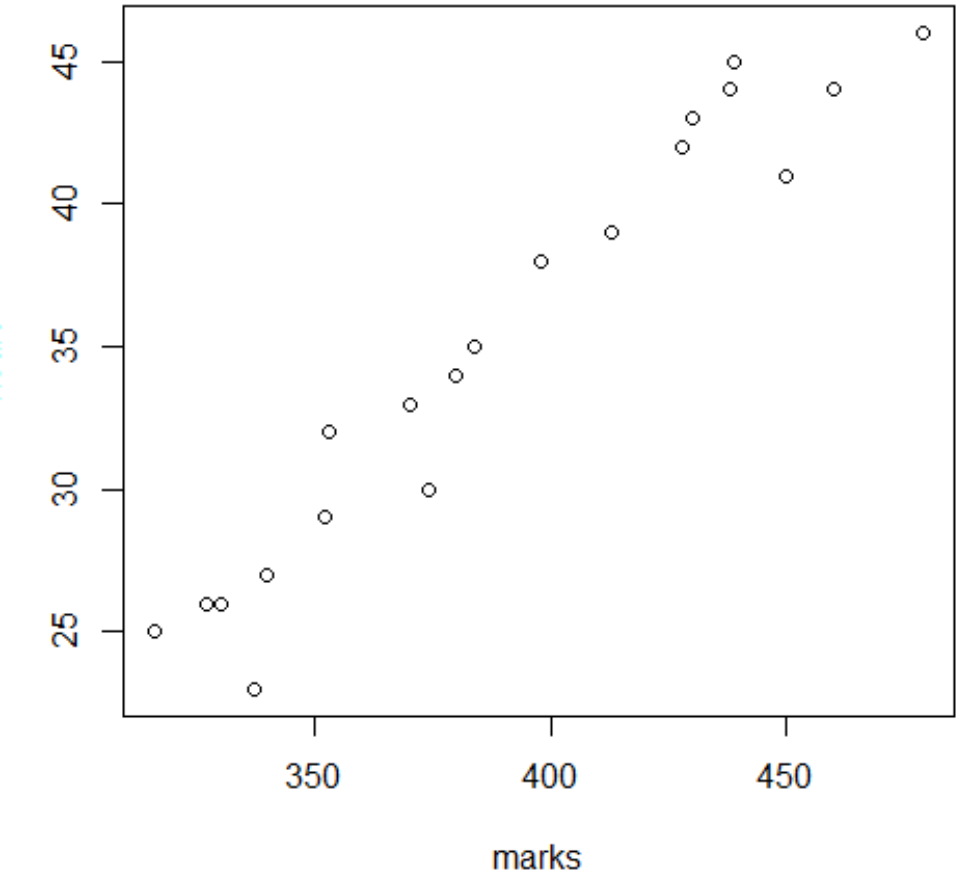
Example

```
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

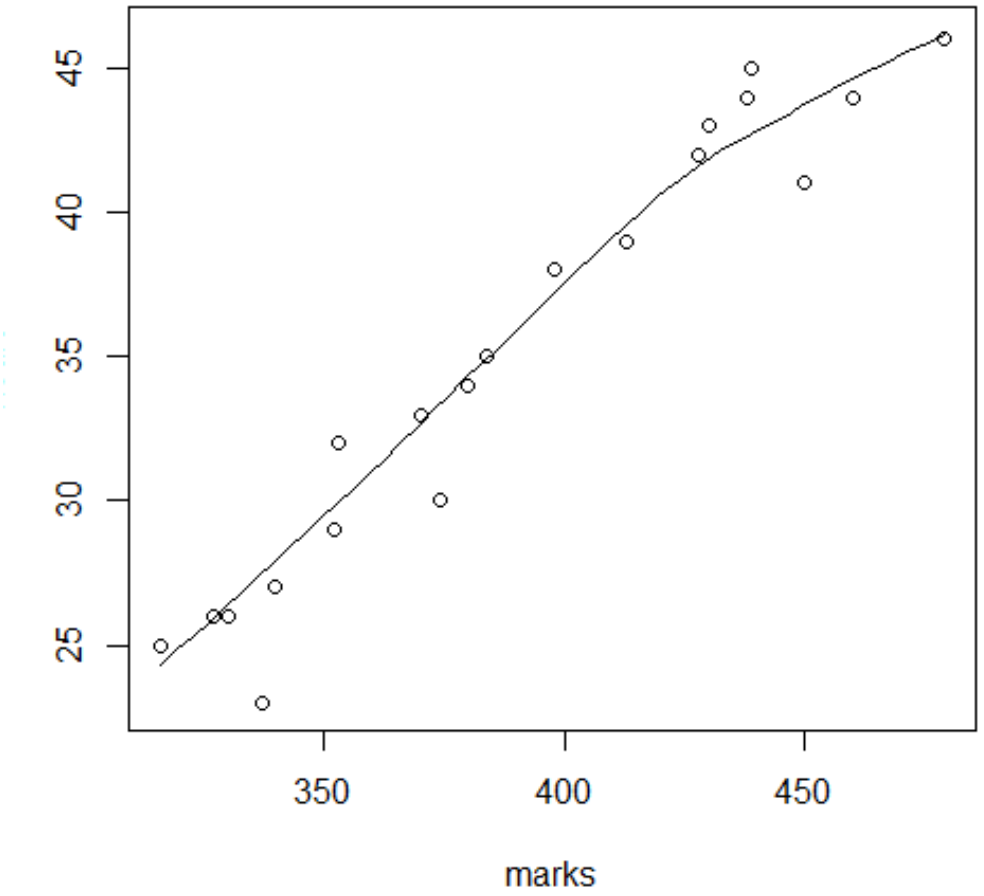
```
> plot(marks, hours)
```



Coefficient of Correlation

Example

```
> scatter.smooth(marks, hours)
```



Coefficient of Correlation

Example

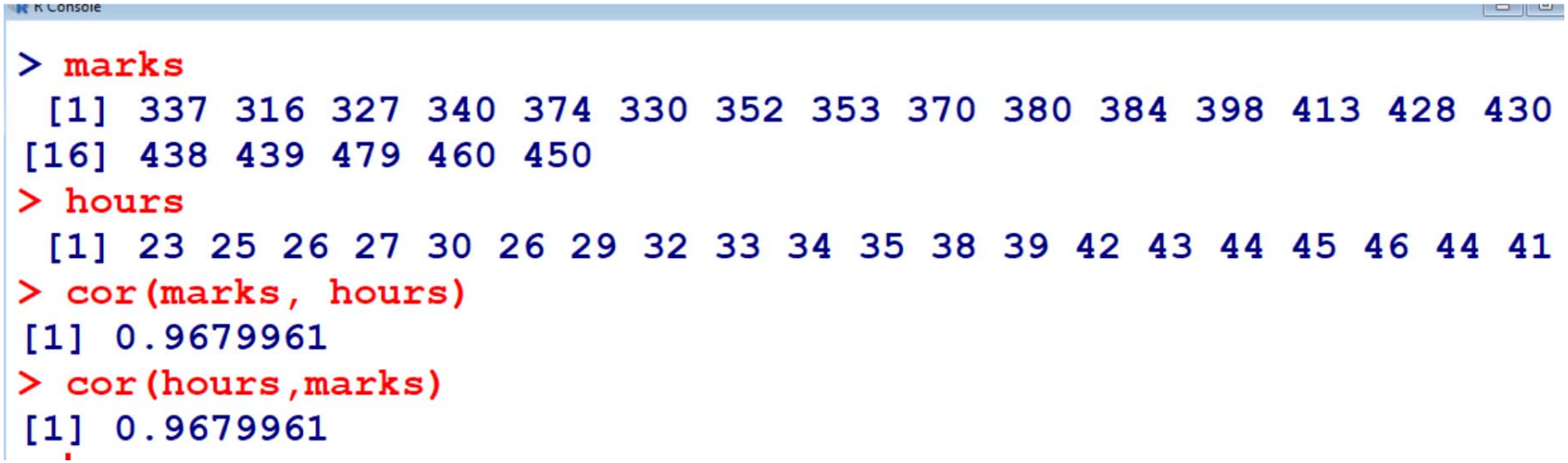
```
> cor(marks, hours)
[1] 0.9679961

> 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



Coefficient of Correlation

Example

A medicine was given to 10 patients. Its quantity (in mg.) and the time (in hours) taken in showing the affect was recorded as follows:

We want to know the effect of medicine on time taken to affect.

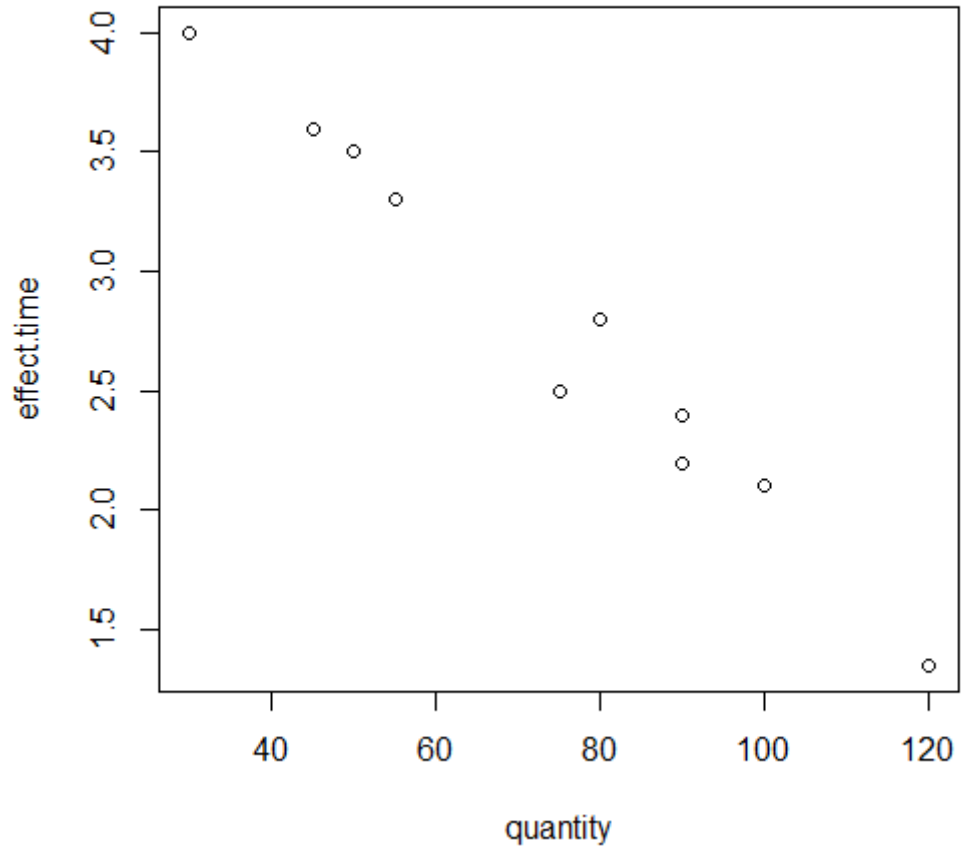
Quantity (in mg.)	30	45	80	120	90	75	55	90	50	100
Time (in hours)	4	3.6	2.8	1.35	2.4	2.5	3.3	2.2	3.5	2.1

```
> quantity = c(30, 45, 80, 120, 90, 75, 55, 90, 50, 100)
> effect.time = c(4, 3.6, 2.8, 1.35, 2.4, 2.5, 3.3, 2.2, 3.5, 2.1)
```

Coefficient of Correlation

Example

```
> plot(quantity, effect.time)
```



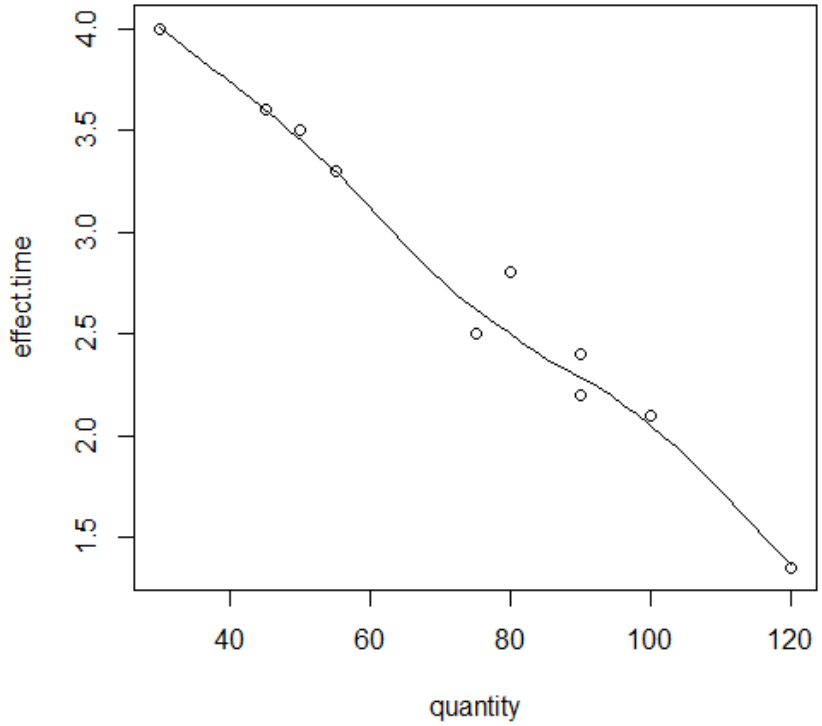
Coefficient of Correlation

Example

```
> scatter.smooth(quantity, effect.time)

> cor(quantity, effect.time)
[1] -0.9885454
```

Sign of correlation coefficient is negative.

As quantity of medicine is increasing, the number of hours to affect are decreasing.



Coefficient of Correlation

Example