Introduction to R Software

Introduction to Statistical Functions

:::

Correlation

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Descriptive statistics:

First hand tools which gives first hand information.

- **Central tendency of data**
- **Variation in data**
- **Structure and shape of data tendency**

- **Relationship study (correlation coefficient, rank correlation, corralation ratio, regression etc.)**

Bivariate Data

Quantitative measures provide quantitative measure of relationship.

Graphical plots provide first hand visual information about the nature and degree of relationship between two variables.

Relationship can be linear or nonlinear.

Bivariate Data

x, y: Two data vectors

Data $x = (x_1, x_2, \ldots, x_n)$ $y = (y_1, y_2, \ldots, y_n)$

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

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

Variance $\operatorname{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

var(x): Variance of x

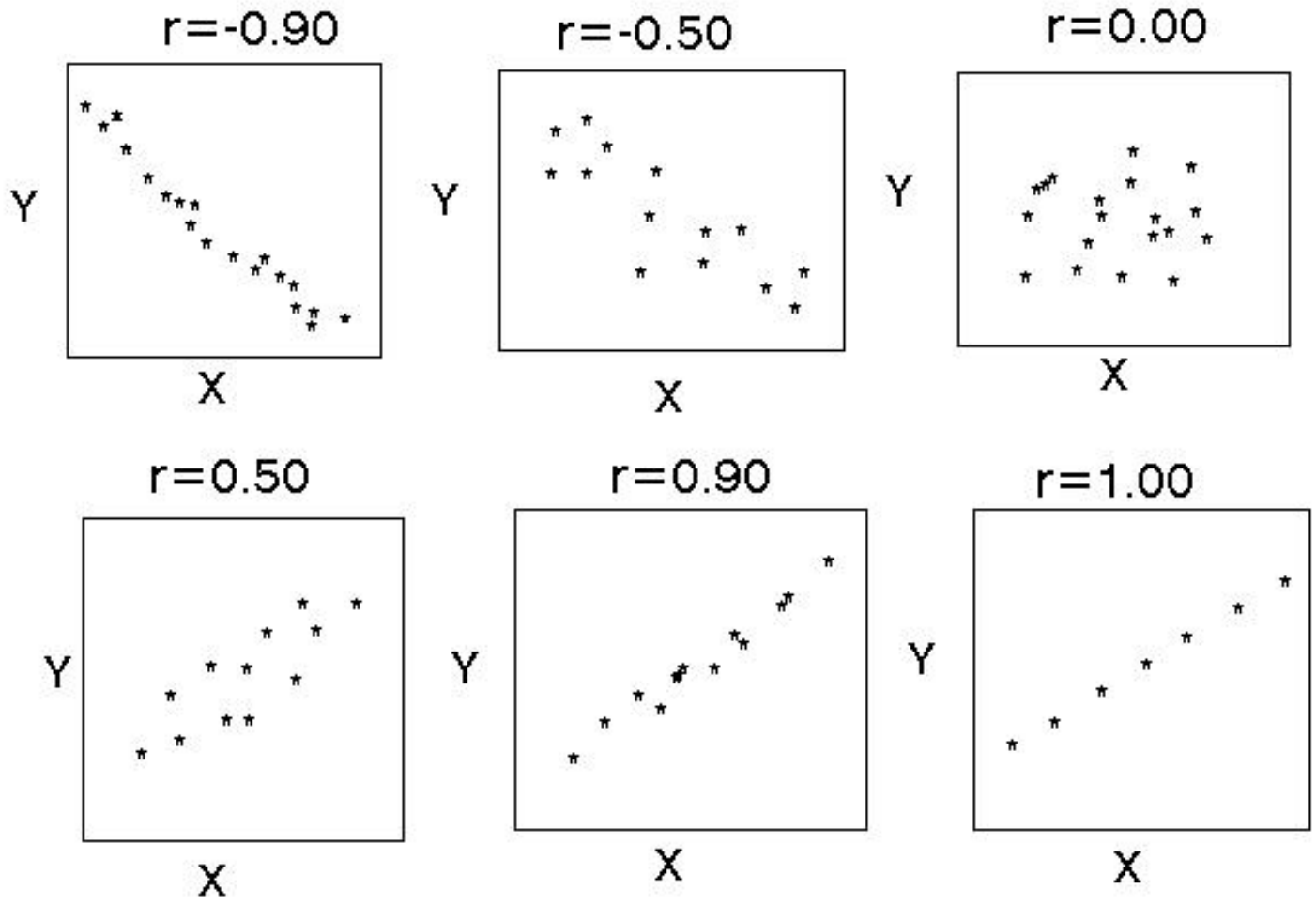
Correlation coefficient

Measures the degree of linear relationship between the two variables.

$$r_{xy} = \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}(x)\operatorname{var}(y)}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$-1 \le r_{xy} \le 1$$

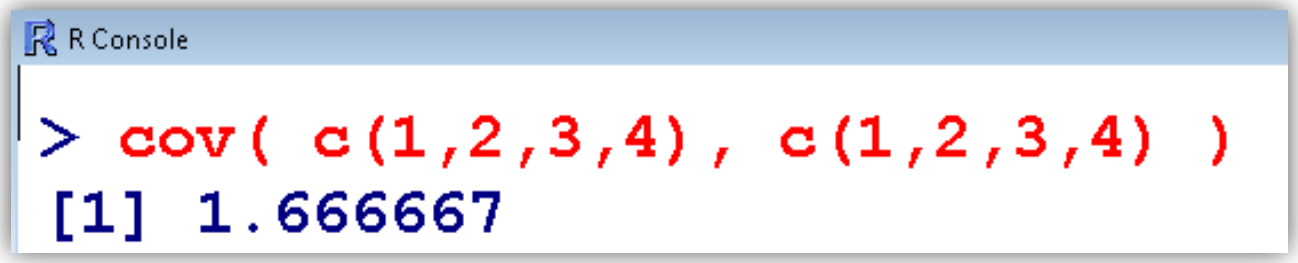
`cor(x,y)`: correlation between x and y


r=-0.90
Y
X
r=-0.50
Y
X
r=0.00
Y
X
r=0.50
Y
X
r=0.90
Y
X
r=1.00
Y
X

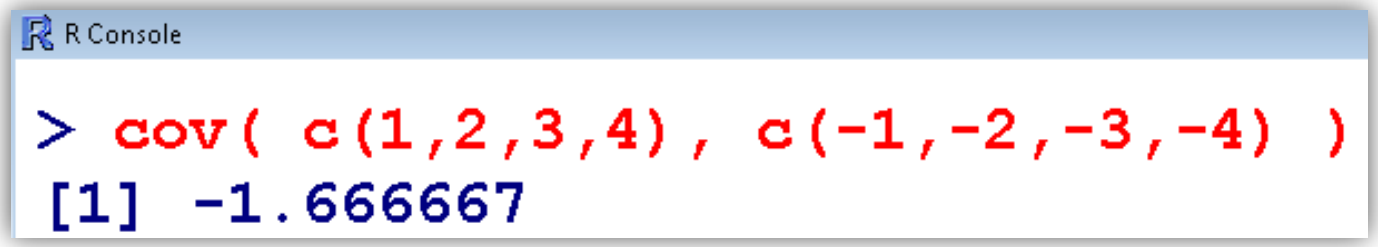
Example:

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
```

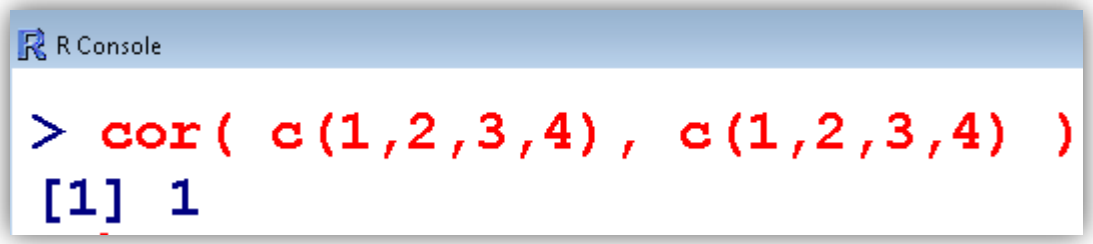


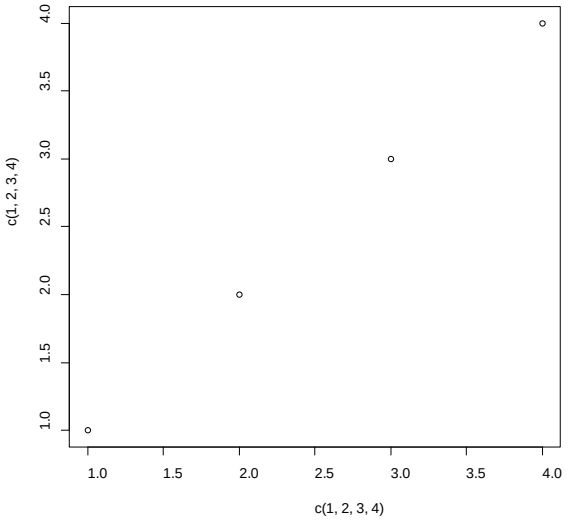
Example:

Correlation coefficient:

Exact positive linear dependence

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



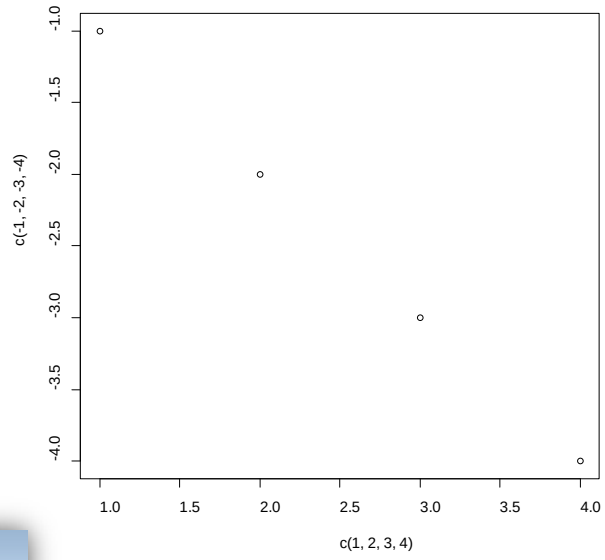


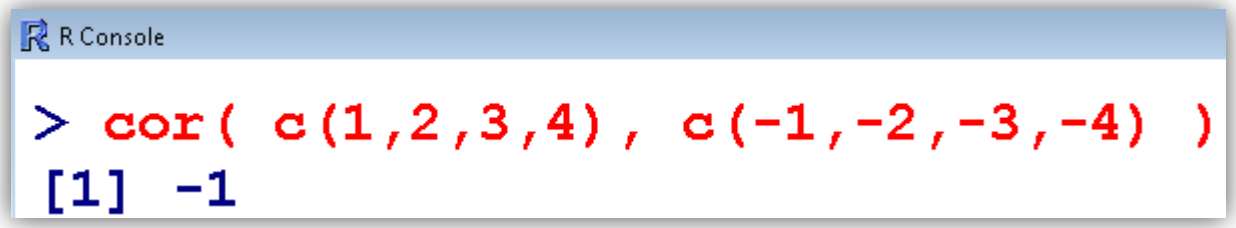
Example:

Correlation coefficient:

Exact negative linear dependence

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





Example:

Daily water demand in a city depends upon weather temperature.

We know from experience that water consumption increases as weather temperature increases.

Data on 27 days is collected as follows:

Daily water demand (in million litres)

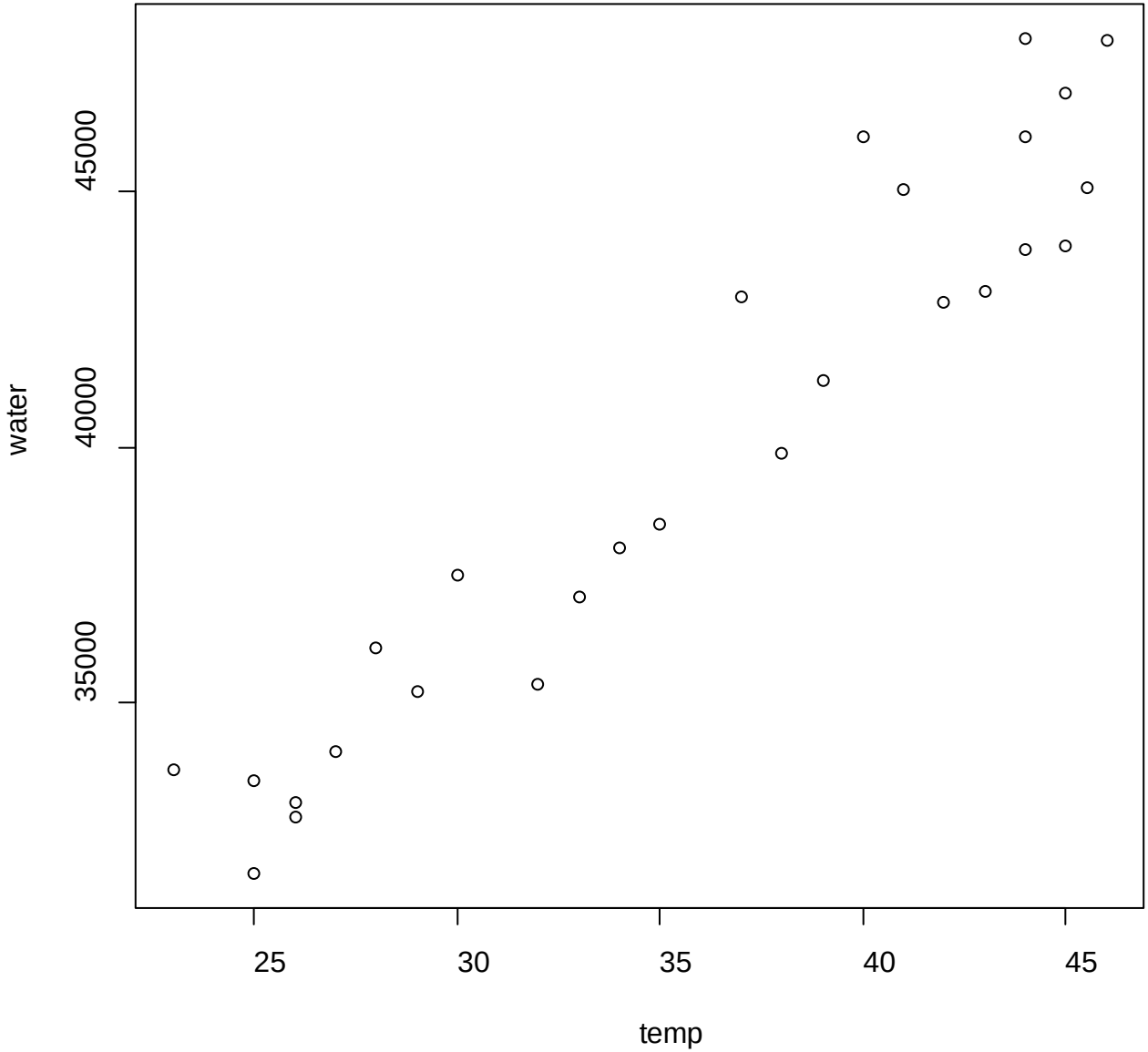
```
water <- c(33710,31666,33495,32758,34067,36069,
37497,33044,35216, 35383,37066,38037,38495,
39895,41311,42849,43038,43873,43923, 45078,
46935,47951,46085,48003,45050,42924,46061)
```

Temperature (in centigrade)

```
temp <- c(23,25,25,26,27,28,30,26,29,32,33,34,
35,38,39,42,43,44, 45,45.5,45,46,44,44,41,37,40)
```

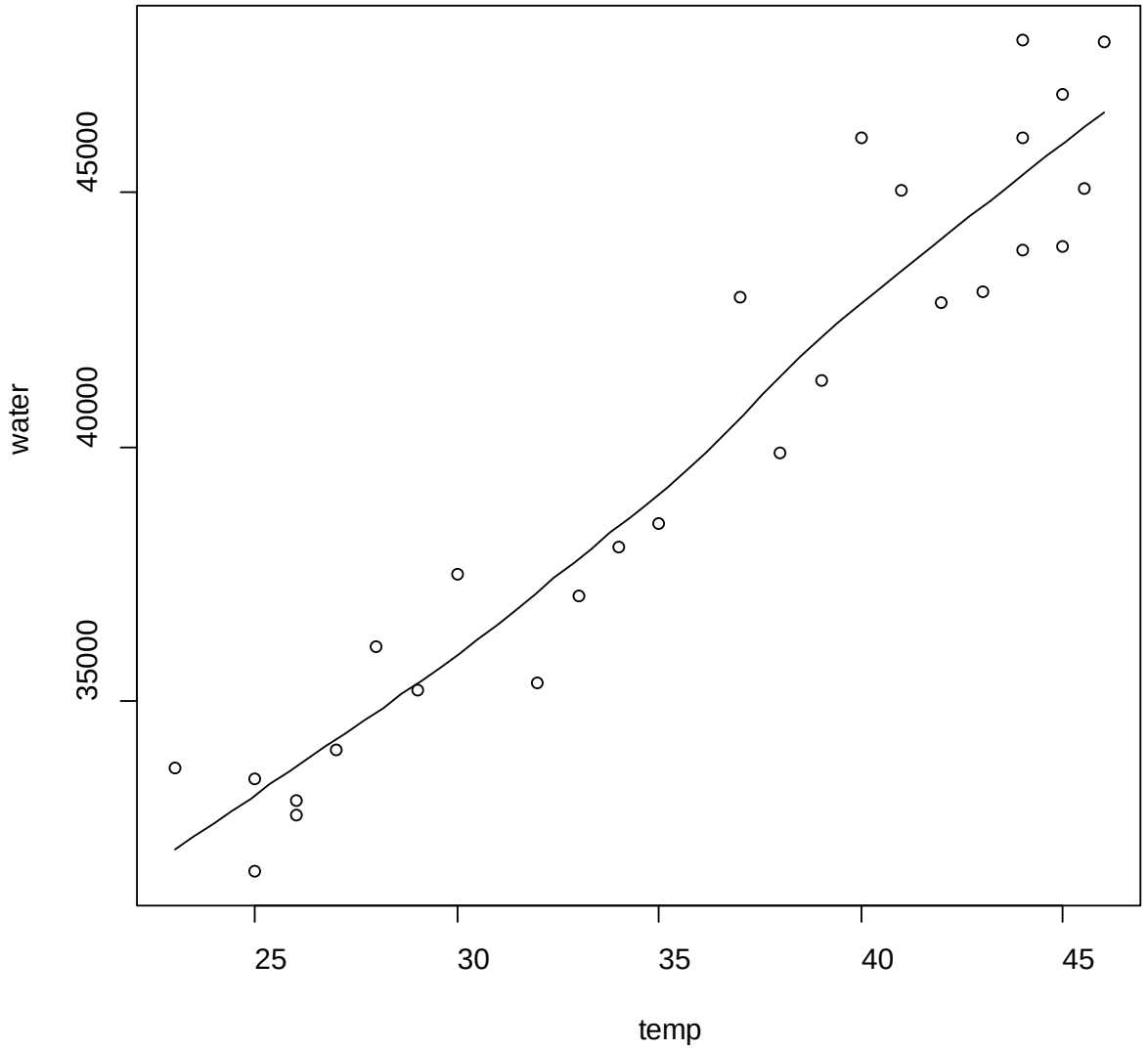
Example:

```
> plot(temp, water)
```



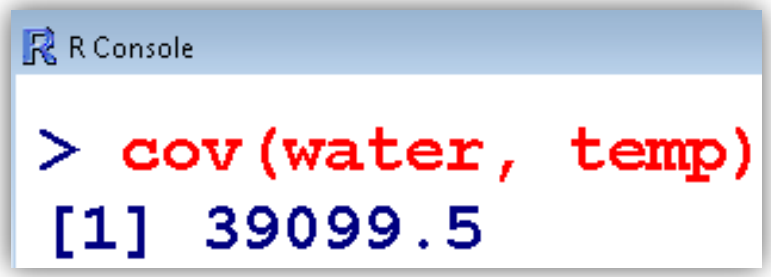
Example:

```
> scatter.smooth(temp,water)
```



Data on Daily water demand

```
> cov(water, temp)
[1] 39099.5
```




```
> cor(water, temp)
[1] 0.9567599
```