Chapter V

Correlation and Regression Analysis

R. 5. 1.
So far we have considered only univariate distributions. Many a time, however, we come across problems which involve two or more variables. This will be the subject matter of the current chapter.

D. 5. 1. (*Correlation*)
The *correlation* means the study of existence, magnitude and direction of the relation between two or more variables.

R. 5. 2. (*Types of Correlation*)
We distinguish

1. *Positive* and *negative* correlations
   If two variables change in the same direction, then this is called a *positive correlation*.
   (For example: price and supply)

   If two variables change in the opposite direction, then the correlation is called a *negative Correlation.*
   (For example: price and demand)

2. *Linear* and *non-linear* correlations.

R. 5. 3. (*Degree of Correlation*)

Degrees	Positive	Negative
Absence of correlation	0	0
Perfect correlation	+1	-1
High degree	$]0.75, \quad 1[$	$]-1, \quad -0.75[$
Moderate degree	$]0.25, \quad 0.75[$	$]-0.75, \quad -0.25[$
Low degree	$]0, \quad 0.25[$	$]-0.25, \quad 0[$

R. 5. 4. (*Methods of Determining Correlation*)
We shall consider the following most commonly used methods:

(1) Scatter Plot
(2) Karl Pearson's coefficient of correlation
(3) Spearman's rank correlation coefficient.

R. 5. 5. (***Scatter Plot or Dot Diagram***)
In this method the values of the two variables are plotted on a graph paper. One is taken along the horizontal ($x-$) axis and the other along the vertical ($y-$) axis. By plotting the data, we get points (dots) on the graph which are generally scattered and hence the name 'scatter plot'. The manner in which these points are scattered, suggest the degree and the direction of correlation.
Let the degree of correlation be denoted by r . Its direction is given by the signs positive and negative.

i) If all points lie on a rising straight line the correlation is perfectly positive and $r = +1$
ii) If all points lie on a falling straight line the correlation is perfectly negative and $r = -1$
iii) If the points lie in a narrow strip, rising upwards, the correlation is high degree of positive.
iv) If the points lie in a narrow strip, falling downwards, the correlation is high degree of negative.
v) If the points are spread widely over a broad strip, rising upwards, the correlation is low degree positive.
vi) If the points are spread widely over a broad strip, falling downwards, the correlation is low degree negative.
vii) If the points are spread (scattered) without any specific pattern, the correlation is absent, i.e. $r = 0$

Though this method is simple and gives a rough idea about the existence and the degree of correlation, it is not reliable. As it is not an exact mathematical method, it cannot measure the degree of correlation.

D. 5. 2. (***Karl Pearson's Coefficient of Correlation***)
Let X and Y be two variates.

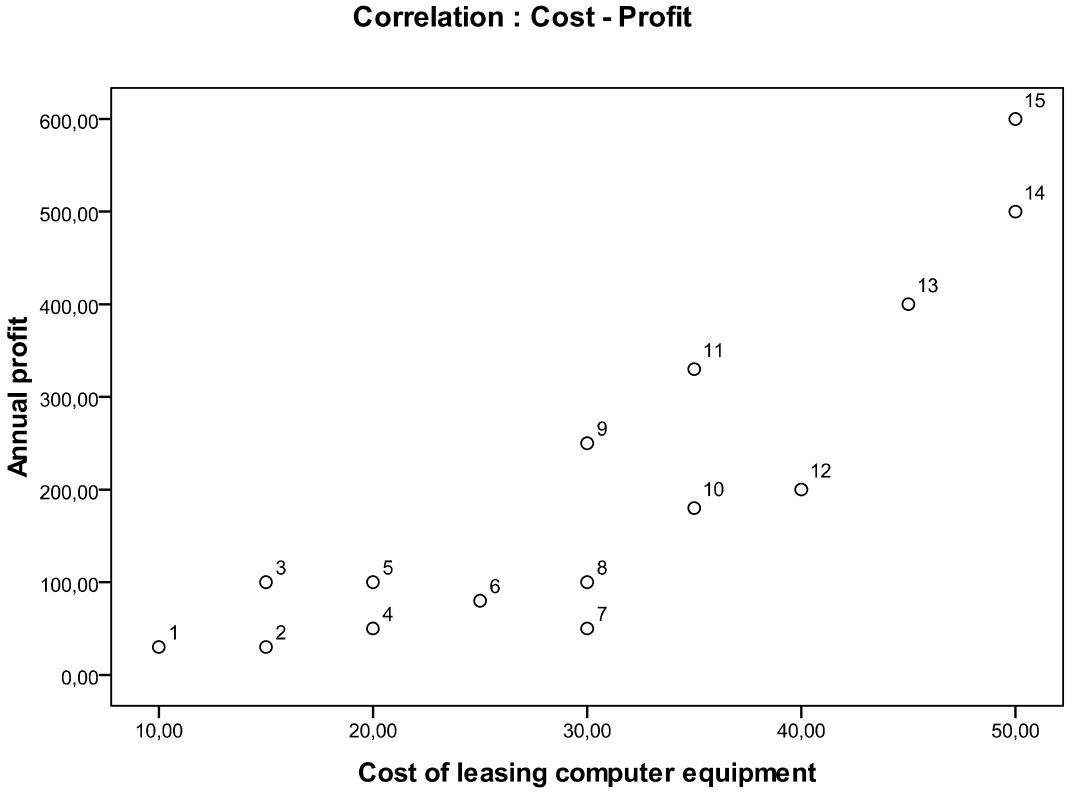
$$r := \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \cdot \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}.$$

Ex. 5. 1.
The following table shows the annual profit [Mio €] and annual costs of leasing computer equipment [1000 €] of 15 firms:

i	x_i	$x_i - \bar{x}$	y_i	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	10	-20	30	-170	3400	400	28900
2	15	-15	30	-170	2550	225	28900
3	15	-15	100	-100	1500	225	10000
4	20	-10	50	-150	1500	100	22500
5	20	-10	100	-100	1000	100	10000
6	25	-5	80	-120	600	25	14400
7	30	0	50	-150	0	0	22500
8	30	0	100	-100	0	0	10000
9	30	0	250	50	0	0	2500
10	35	5	180	-20	-100	25	400
11	35	5	330	130	650	25	16900
12	40	10	200	0	0	100	0
13	45	15	400	200	3000	225	40000
14	50	20	500	300	6000	400	90000
15	50	20	600	400	8000	400	160000
Sum	450	0	3000	0	28100	2250	457000

$$\bar{x} = \frac{450}{15} = 30, \qquad \bar{y} = \frac{3000}{15} = 200$$

$$r := \frac{28100}{\sqrt{2250 \cdot 457000}} \approx 0.88 .$$



D. 5. 3. (***Spearman's Rank Coefficient of Correlation***)

Let

$R_i, i = 1, 2, ..., n$: ranks of the characteristic X

$R_i^{'}, i = 1, 2, ..., n$: ranks of the characteristic Y,

$$\rho := 1 - \frac{6 \cdot \sum_{i=1}^{n} \left(R_i - R_i^{'} \right)^2}{(n-1) \cdot n \cdot (n+1)} :$$

Ex. 5. 2.

The following table shows the advertisement costs (Y) and the revenues (X) of a firm:

Advertisement Costs	Revenue
1.4	210
1.8	220
1.9	240
2.4	240
2.8	320
3.2	400
3.6	410
4.0	480

Find and interpret the Spearman's rank coefficient of correlation.

Solution:

Advertisement Costs	Revenue	R_i	$R_i^{'}$	$(R_i - R_i^{'})^2$
1.4	210	1	1.0	0.00
1.8	220	2	2.0	0.00
1.9	240	3	3.5	0.25
2.4	240	4	3.5	0.25
2.8	320	5	5.0	0.00
3.2	400	6	6.0	0.00
3.6	410	7	7.0	0.00
4.0	480	8	8.0	0.00
				0.50

$n = 8$, $\sum_{i=1}^{8} \left(R_i - R_i^{'} \right)^2 = 0.50$,

$$\rho := 1 - \frac{6 \cdot 0.50}{7.8.9 \cdot} \approx 0.994 .$$

There is therefore a strong degree of correlation between X and Y.

D. 5. 4. (***Regression Function***)
A *regression*[1] *function* describes the relationship between dependent variable Y and (at least one) *independent* or *explanatory* variable X :

$$y^* = f(x)$$

R. 5. 6. (***Least Square Method***)
The coefficients of a regression function can be estimated by using the *Least Square Method*:

$$S(...) = \sum_{i=1}^{n} \left(y_i - y^* \right)^2 \rightarrow Min!$$

R. 5. 7. (***Simple Linear Regression***)
The coefficients of a simple linear regression function

$$y^* = a_0 + a_1 x$$

can be obtained by solving the following system of linear equations:

$$\begin{aligned} n \cdot a_0 \quad &+ a_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \cdot \sum_{i=1}^{n} x_i &+ a_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i \end{aligned}$$

D. 5. 5. (***Coefficients of Correlation and Determination***)

1. The *coefficient of correlation of a linear regression function* is defined by

$$r := \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i \right) \cdot \left(n \cdot \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} y_i \right)}},$$

$$-1 \le r \le +1$$

2. r^2 $(0 \le r^2 \le 1)$ is called *coefficient of determination*.

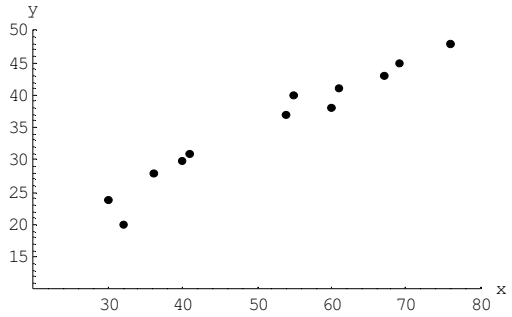
Ex. 5. 3.
In a study of how the productivity in 14 firms depends on the degree of automation , the following data have been made available :

[1] The term *regression* was first used by Sir Francis Galton (1822-1911), who studied the relationship between heights of children and heights of their parents.

Firm	Productivity	Degree of Automation (%)
1	20	32
2	24	30
3	28	36
4	30	40
5	31	41
6	33	47
7	34	56
8	37	54
9	38	60
10	40	55
11	41	61
12	43	67
13	45	69
14	48	76

1. Plot a scatter diagram.
2. Find an appropriate regression function.
3. Calculate the coefficients of correlation and determination and interpret them.
4. Predict the level of productivity for an automation degree of 80%.

Solution:
1.



2.

i	y_i	x_i	$x_i \cdot y_i$	x_i^2	y_i^2
1	20	32	640	1024	400
2	24	30	720	900	576
3	28	36	1008	1296	784
4	30	40	1200	1600	900
5	31	41	1271	1681	961
6	33	47	1551	2209	1089
7	34	56	1904	3136	1156
8	37	54	1998	2916	1369
9	38	60	2280	3600	1444
10	40	55	2200	3025	1600
11	41	61	2501	3721	1681
12	43	67	2881	4489	1849
13	45	69	3105	4761	2025
14	48	76	3648	5776	2304
total	492	724	26907	40134	18138

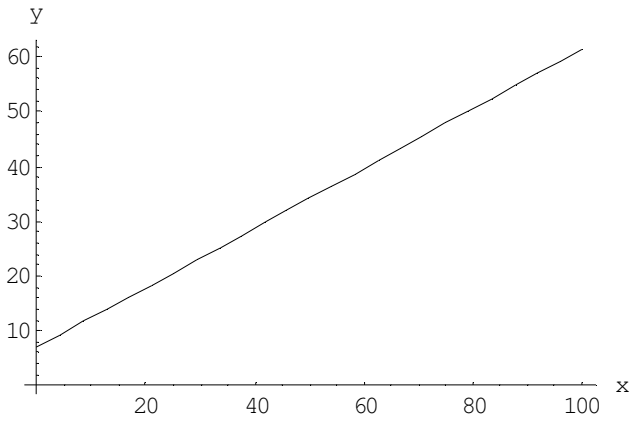
$$14a_0 + \quad 724a_1 = 492$$
$$724a_0 + 40134a_1 = 26907$$

The solution of the above systems yields:

$a_0 = 7.0356$, $a_1 = 0.5435$.

Therefore, we have the regression function:

$$y^* = 7.0356 + 0.5435x$$



3.

$$r = \frac{14 \cdot 26907 - 724 \cdot 492}{\sqrt{(14 \cdot 40134 - 724 \cdot 724) \cdot (14 \cdot 18134 - 492 \cdot 492)}} \approx 0.9687$$

$$r^2 = 0.9384$$

Because of $r > 0$ the productivity is directly dependent on the degree of automation. Changes in the productivity are up to 94% due to changes in the degree of automation.

3.

$$y^*(80) = 7.0356 + 0.5435 \cdot 80 = 51$$

D. 5. 6. (*Trend Function*)
A *trend function* is a special case of a regression function where the independent variable is the factor *time*.

D. 5. 7 (*Coefficient of Determination*)
The *coefficient of determination* is defined as:

$$r^2 := \frac{SSR}{SST}$$

where:

$SSR = \sum_{i=1}^{n} (y_i^* - \bar{y})^2$: sum of square due to regression

$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$: total sum of squares

R. 5. 8.
The formula given for the coefficient of correlation given in D. 5. 5. is a special case of the formula in D. 5. 7. for *linear regression*.

Ex. 5. 4.
The following table shows the consumption per head of butter (in kg) in a certain country in the years 1998-2004:

Year	1998	1999	2000	2001	2002	2003	2004
Consumption per head	9.2	9.5	9.7	9.6	9.9	10.5	10.8

1. Fit a linear trend by regression analysis.
2. Predict the consumption per head of butter in the year 2005.

Solution
1.

year	x_i	y_i	x_i^2	$x_i \cdot y_i$
1998	-3	9.2	9	-27.6
1999	-2	9.5	4	-19.0
2000	-1	9.7	1	-9.7
2001	0	9.6	0	0.0
2002	1	9.9	1	9.9
2003	2	10.5	4	21.0
2004	3	10.8	9	32.4
	0	69.2	28	7

$$a_0 = \frac{69.2}{7} = 9.886\,, \qquad a_1 = \frac{7}{28} = 0.25\,.$$

We, therefore, have the trend function:

$$y^* = 0.885 + 0.25x\,.$$

2.

$$y^*(4) = 0.885 + 0.25 \cdot 4 = 10.9$$

Ex. 5. 4. (*Linearisation of a Non-Linear Function*)
The following table shows the number of cattle in a certain country in the years 1992-2004:

Year	1992	1994	1995	1998	2000	2004
No. of cattle ((10^5)	11.0	13.3	14.6	19.6	23.7	33.1

1. Fit a trend by the regression function

$$y^* = a_0 \cdot a_1^{\,x}$$

.
2. Predict the number of cattle in the year 2005.

3. Calculate and interpret the coefficient of determination for the regression.

Solution:

$$y^* = a_0 \cdot a_1^{\,x}$$
$$\lg y^* = \lg a_0 + \lg a_1^{\,x}$$
$$\lg y^* = \lg a_0 + x \cdot \lg a_1\,.$$

Let

$$Y^* := \lg y^*, \qquad A_0 = \lg a_0\,, \qquad A_1 = \lg a_1\,.$$

Then we have:

$$Y^* = A_0 + A_1 \cdot x$$

$$n \cdot A_0 + A_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i$$

$$A_0 \cdot \sum_{i=1}^{n} x_i + A_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot Y_i$$

year	x_i	y_i	$\lg y_i$	${x_i}^2$	$x_i \cdot \lg y_i$
1992	1	11.0	1.04139	1	1.04139
1994	3	13.3	1.12385	9	3.37155
1995	4	14.6	1.16435	16	4.65740
1998	7	19.6	1.29226	49	9.04582
2000	9	23.7	1.37475	81	12.37275
2004	13	33.1	1.51983	169	19.75779
	37	115.3	7.51643	325	50.24670

$$6A_0 + 37A_1 = 7.52$$
$$37A_0 + 325A_1 = 50.25$$

The solution of the above systems yields:

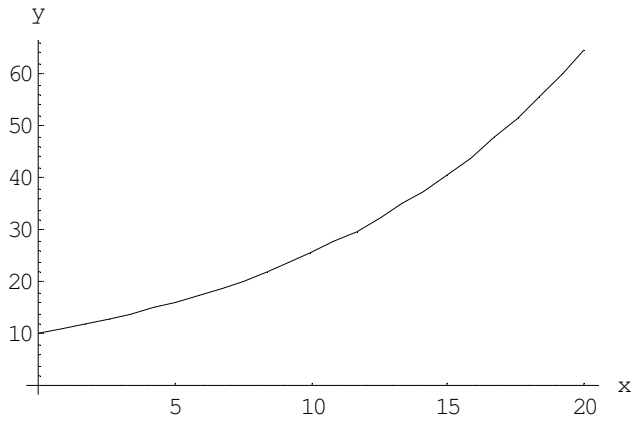
$$A_0 = 1.006454392, \qquad A_1 = 0.040034423,$$

i. e.

$$a_0 = 10.14972772, \qquad a_1 = 1.096565109$$

Therefore, we have the regression function:

$$y^* = 10.15 \cdot 1.097^x$$



2.

$$y^*(14) = 10.15 \cdot 1.097^{14} = 37.10 \ (\cdot 10^5)$$

(Last revised: 07.11.10)