Correlation and Simple Linear Regression

An Introduction

Subject matter

Correlation: Study of

1. existence
2. magnitude
3. direction

of the relation between two or more variables.

Regression:

4. their functional relationship

Methods of Determining Correlation

1. **Scatter plot**
   (The two variables are plotted on a graph paper. One is taken along the horizontal axis and the other along the vertical axis. The manner in which these points are scattered, suggest the degree and direction of correlation.)

2. **Karl Pearson's coefficient of correlation** (for quantitative variables)

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

3. **Spearman's rank correlation coefficient** (for qualitative variables)

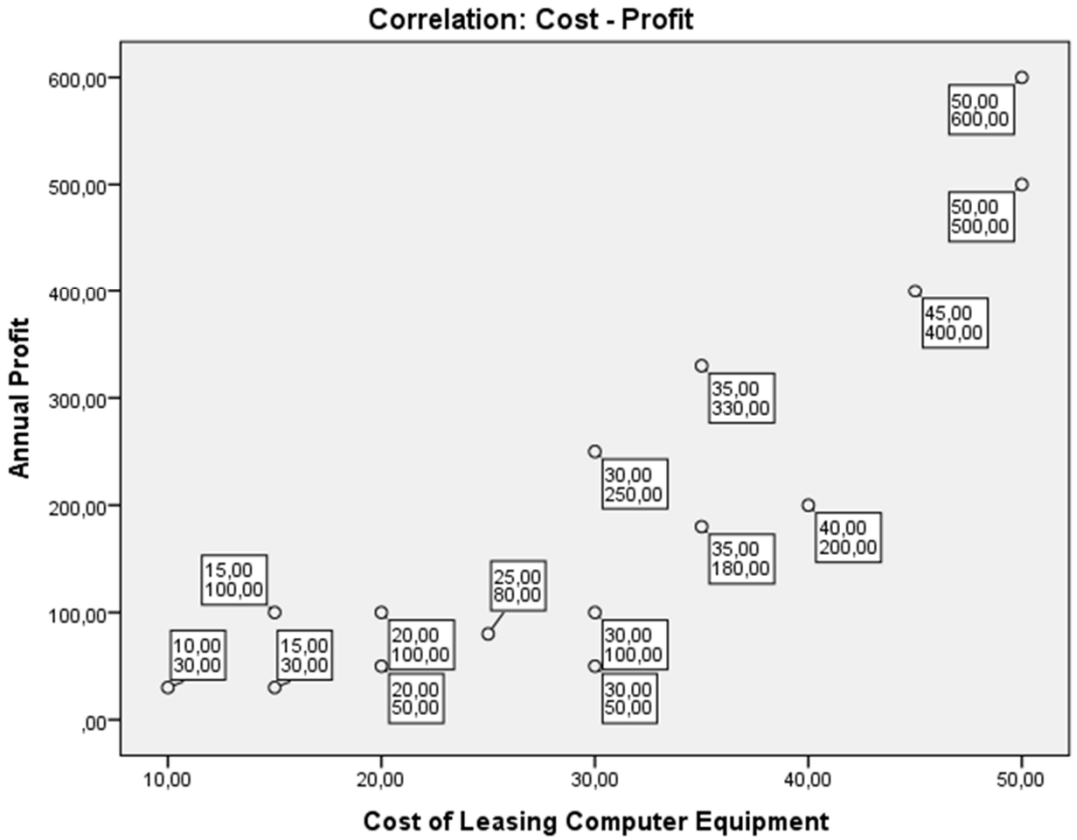
An Example

The following table shows the annual profit $y_i, i = 1,2,...,15$ [Mio €] and annual costs of leasing computer equipment $x_i, i = 1,2,...,15$ [1000 €] of 15 firms:

i	x_i	y_i
1	10	30
2	15	30
3	15	100
4	20	50
5	20	100
6	25	80
7	30	50
8	30	100
9	30	250
10	35	180
11	35	330
12	40	200
13	45	400
14	50	500
15	50	600

1. Plot the relation between the two factors as a scatter diagram.
2. Find and interpret the Pearson's coefficient of correlation.

1.
Scatter diagram



2. Coefficient of Correlation

Working Table

i	x_i	$(x_i - \bar{x})$	y_i	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
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

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

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

$\therefore$ There is a direct relatively strong relationship between the two factors.

Simple Linear Regression

Simple Linear Regression Model

A *simple linear regression model* is defined as follows:

$$y = \beta_0 + \beta_1 x + \varepsilon .$$

β_0 and β_1 are referred to as the *parameter* of the model, and ε is a random variable referred to as the *error* term.

Simple Linear Regression Equation

The equation that describes how the expected value of y, denoted by $E(y)$, is related to x is called the *regression equation*:

$$E(y) = \beta_0 + \beta_1 x.$$

Estimated Linear Regression Equation

Substituting the values of the sample statistics b_0 and b_1 for β_0 and β_1, we obtain the *estimated linear regression equation*:

$$y^* = b_0 + b_1 x .$$

Least Square Method

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

For the **simple linear regression:** $y^*=b_0+b_1x$

$$S(b_0,b_1)=\sum_{i=1}^{n}\left(y_i-b_0-b_1x\right)^2\rightarrow Min!$$

Normal Equations:

$$\begin{cases} n\cdot b_0 + b_1\cdot\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i \\ b_0\cdot\sum_{i=1}^{n}x_i+b_1\cdot\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_i\cdot y_i \end{cases}$$

Coefficient of determination:

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

$$SST = SSR + SSE$$

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

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

$SSE = \sum_{i=1}^{n} (y_i - y_i^*)^2$: sum of squares due to error

r^2 $(0 \le r^2 \le 1)$ can be interpreted as the percentage of the total sum of squares that can be explained by using the estimated regression equation.

Correlation Coefficient:

$$r := (\text{sign of } b_1)\sqrt{r^2}, \quad -1 \le r \le 1.$$

The coefficient of determination for a simple linear regression function can alternatively be calculated according to the following formula:

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

Example:

Suppose the following data were collected from a sample of 10 pizzerias of a certain firm:

Restaurant	Student Population (1000s)	Quarterly Sales(1000s)
i	x_i	y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

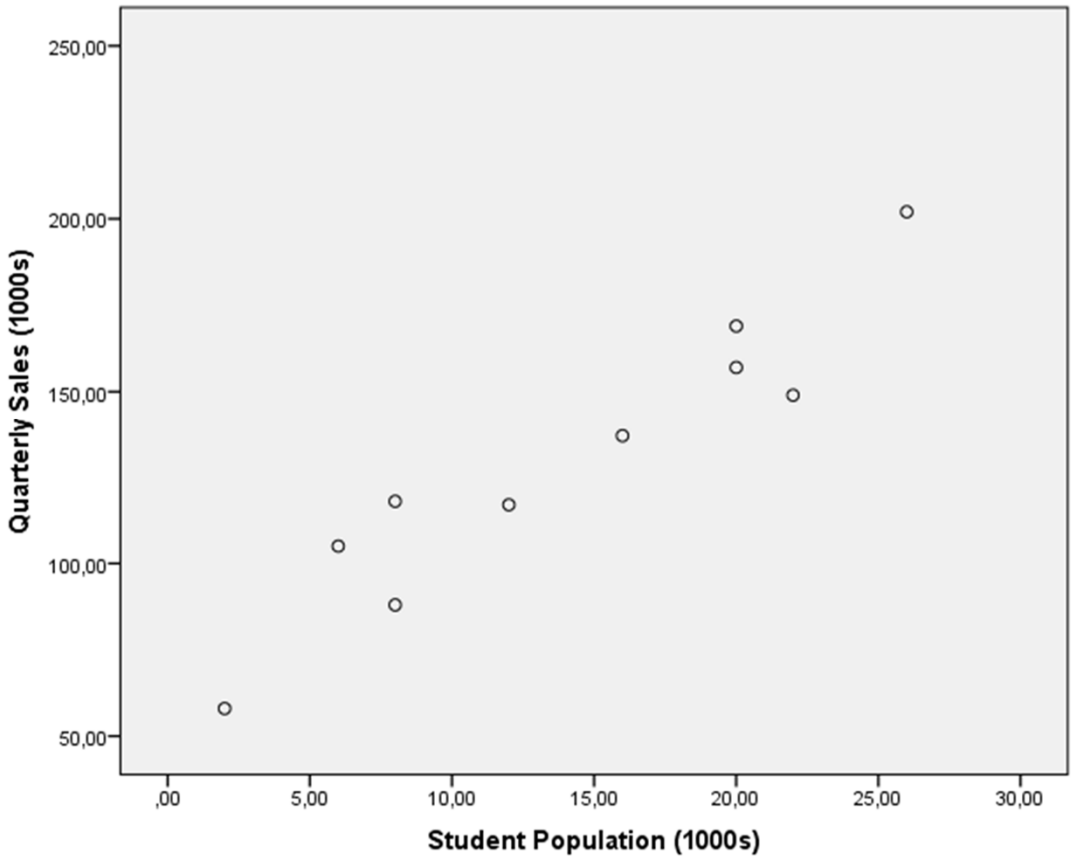
1. Develop a scatter diagram.

2. Find a regression function describing the quarterly sales as a linear function of the student population.

3. Calculate and interpret the coefficients of correlation and determination.

4. Predict the quarterly sales of the pizzerias if the student population rises to 27000.

1.


250,00
200,00
150,00
100,00
50,00
Quarterly Sales (1000s)
,00
5,00
10,00
15,00
20,00
25,00
30,00
Student Population (1000s)

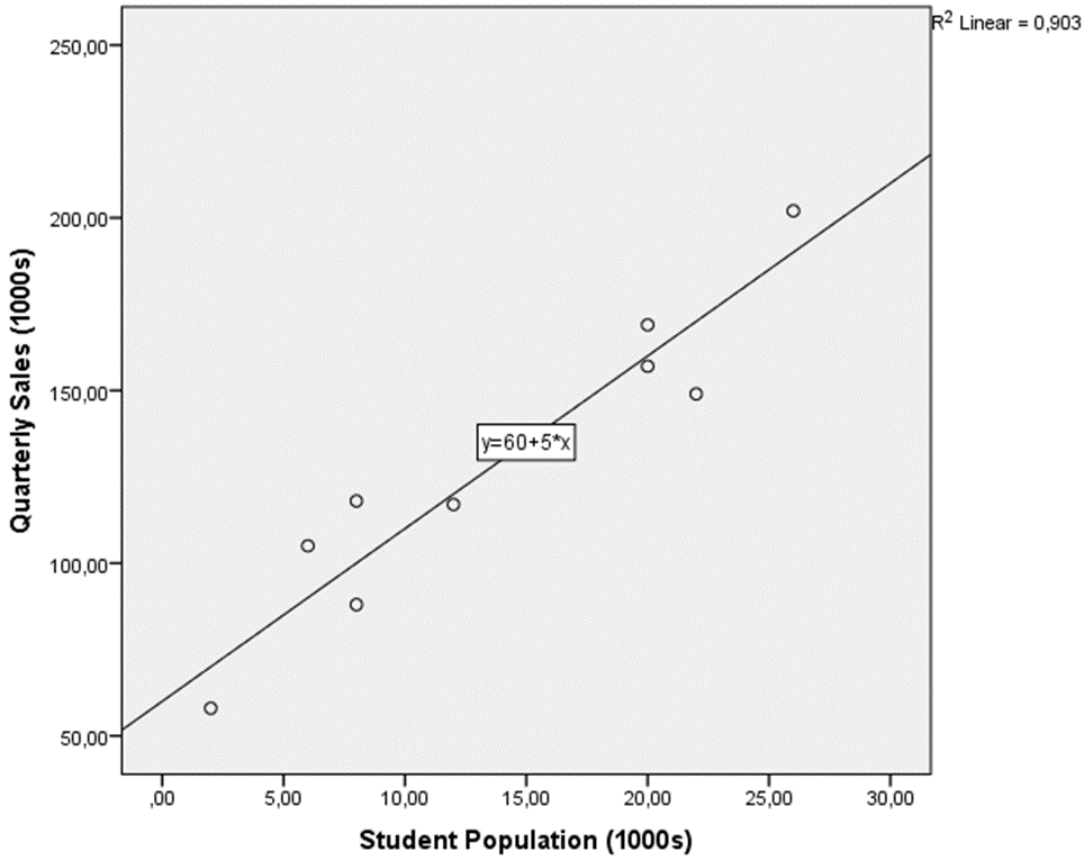
2.

Working Table

x_i	y_i	x_i^2	$x_i \cdot y_i$	y_i^2
2	58	4	116	3364
6	105	36	630	11025
8	88	64	704	7744
8	118	64	944	13924
12	117	144	1404	13689
16	137	256	2192	18769
20	157	400	3140	24649
20	169	400	3380	28561
22	149	484	3278	22201
26	202	676	5252	40804
140	1300	2528	21040	184730

$$\begin{cases} 10b_0 \quad 140b_1 = 1300 \\ 140b_0 + 2528b_1 = 21040 \end{cases} \Rightarrow \quad b_0 = 60, \quad b_1 = 5$$

$$y^* = 60 + 5x$$


R2 Linear = 0,903
250,00
200,00
150,00
100,00
50,00
Quarterly Sales (1000s)
y=60+5*x
,00
5,00
10,00
15,00
20,00
25,00
30,00
Student Population (1000s)

3.

Working Table

i	x_i	y_i	y_i^*	$y_i - y_i^*$	$(y_i - y_i^*)^2$	$(y_i - \bar{y})^2$
1	2	58	70	-12	144	5184
2	6	105	90	15	225	625
3	8	88	100	-12	144	1764
4	8	118	100	18	324	144
5	12	117	120	-3	9	169
6	16	137	140	-3	9	49
7	20	157	160	-3	9	729
8	20	169	160	9	81	1521
9	22	149	170	-21	441	361
10	26	202	190	12	144	5184
	140	1300	1300	0	1530	15730

$$SSR = SST - SSE = 15730 - 1530 = 14200$$

$$r^2 = \frac{14200}{15730} = 0.9027.$$

We can, therefore, conclude that 90.27% of the total sum of squares can be explained by using the regression equation

$$r = \sqrt{0.9027} = 0.9501.$$

Or alternatively:

$$r = \frac{10 \cdot 21040 - 140 \cdot 1300}{\sqrt{\left(10 \cdot 2528 - 140^2\right) \cdot \left(10 \cdot 184730 - 1300^2\right)}} \approx 0.9501.$$

4.

$$y^*(27) = 195 \text{ (1000s)}$$

Testing for Significance

To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero.

Two tests are commonly used:

1. t Test
2. F Test

T Test for Significance in Simple Linear Regression

Step 1:
Formulate the test hypotheses:

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$$

Step 2:
Calculate the test statistic:

$$t_{stat} = \frac{b_1}{s_{b_1}}$$

where

$$s_{b_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}},$$

$$s = \sqrt{\frac{SSE}{n-2}}$$

Step 3:
Conclusion:

$$\text{Reject } H_0 \text{ if} \begin{cases} p-value \leq \alpha & p-value \text{ approach} \\ t \leq -t_{\alpha/2} \text{ or } t \geq t_{\alpha/2} & \text{critical value approach} \end{cases}$$

Example

Let the significance level be equal to $\alpha = 0.01$.

Step 1:

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$$

Step 2:

$$s = \sqrt{\frac{1530}{10-2}} = 13.829$$

$$s_{b_1} = \frac{13.829}{\sqrt{568}} = 0.580$$

$$t_{stat} = \frac{5}{0.5803} = 8.62$$

Step 3:

p-value approach: $p-value < 0.0001 < 0.01 = \alpha \quad \Rightarrow$ reject H_0

critical value approach: $t = 8.62 \geq 3.355 = t_{8;0.01} \quad \Rightarrow$ reject H_0

F Test for Significance in Simple Linear Regression

Step 1:
Formulate the test hypotheses:

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$$

Step 2:
Calculate the test statistic:

$$F = \frac{MSR}{MSE}$$

where

$$MSR = \frac{SSR}{\text{Number of independent variables}}$$

$$MSE = s^2$$

Step 3:
Conclusion:

$$\text{Reject } H_0 \text{ if} \begin{cases} p-value \leq \alpha & p-value \text{ approach} \\ F \geq F_\alpha & \text{critical value approach} \end{cases}$$

(F_α is based on an F distribution with 1 degree of freedom in the numerator and $n-2$ degree of freedom in the denominator.)

Example

Let the significance level be equal to $\alpha = 0.01$.

Step 1:

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$$

Step 2:

$$F = \frac{14200}{191.25} = 74.25$$

Step 3:

p-value approach: $p - value < 0.0001 < 0.01 = \alpha \quad \Rightarrow \quad$ reject H_0

critical value approach: $F = 74.25 \geq 11.26 = F_{0.01} \quad \Rightarrow \quad$ reject H_0

Test of Significance for Correlation

A test of significance for a linear relationship between x and y can also be performed by using the sample correlation r. With ρ denoting the population correlation coefficient, the hypotheses are as follows:

$$H_0: \rho = 0, \quad H_1: \rho \neq 0.$$

A significant relationship can be concluded if H_0 is rejected. However, the T and F tests give the same result as the test for significance using the correlation coefficient.

Confidence Interval for β_1

$$\beta_1 \in \left[b_1 - t_{\alpha/2} s_{b_1},\ b_1 + t_{\alpha/2} s_{b_1} \right]$$

Example

$$\beta_1 \in \left[5 - 3.355 \cdot 0.5803,\ 5 + 3.355 \cdot 3.5803 \right] = \left[3.05,\ 6.95 \right].$$

Using the Estimated Regression Equation for Estimation and Prediction

Here we have the following possibilities:

- Point estimation
- Interval estimation for the mean value of y
- Interval prediction for an individual value of y

- **Point Estimate**

  By substituting a certain value for x in the regression function, we obtain a point estimate.

Example

Predict the sales for a university with 10000 students.

$y^*(10) = 60 + 5 \cdot 10 = 110$ (or \$110000).

- **Confidence Interval for the Mean of y**

$$E(y_p) \in \left[y_p^* - t_{\alpha/2} \cdot s_{y_p^*},\ y_p^* + t_{\alpha/2} \cdot s_{y_p^*} \right]$$

where

x_p: the particular given value for x

y_p: the value of y corresponding to the given x_p

$E(y_p)$: the expected value of y

$y_p^* = b_0 + b_1 x_p$: the point estimate of $E(y_p)$, when $x = x_p$

$$s_{y_p^*} = s \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Example

With $x_p = 10$ and $\alpha = 0.05$ we obtain:

$$s_{y_p^*} = 13.829 \cdot \sqrt{\frac{1}{10} + \frac{(10-14)^2}{568}} = 4.95$$

$$E(y_p) \in \left[110 - 2.306 \cdot 4.95,\ 110 + 2.306 \cdot 4.95 \right] = \left[98.585,\ 121.415 \right].$$

- **Confidence Interval for an Individual Value of y**

$$y_p \in \left[y_p^* - t_{\alpha/2} \cdot s_{ind},\ y_p^* + t_{\alpha/2} \cdot s_{ind} \right]$$

where

$$s_{ind} = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

Example

With $x_p = 10$ and $\alpha = 0.05$ we obtain:

$$s_{ind} = 13.829 \cdot \sqrt{1 + \frac{1}{10} + \frac{(10-14)^2}{568}} = 14.69$$

$$y_p \in \left[110 - 2.306 \cdot 14.69,\ 110 + 2.306 \cdot 14.69 \right] = \left[76.125,\ 143.875 \right].$$