AdTrans project , 2023-1-PL01-KA220-HED-000158917

ADTRANS
ADVANCED DATA ANALYSIS FOR TRANSPORTATION AND MOBILITY ACADEMIC TEACHERS

Correlation and regression analysis

Joanna Wachnicka

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

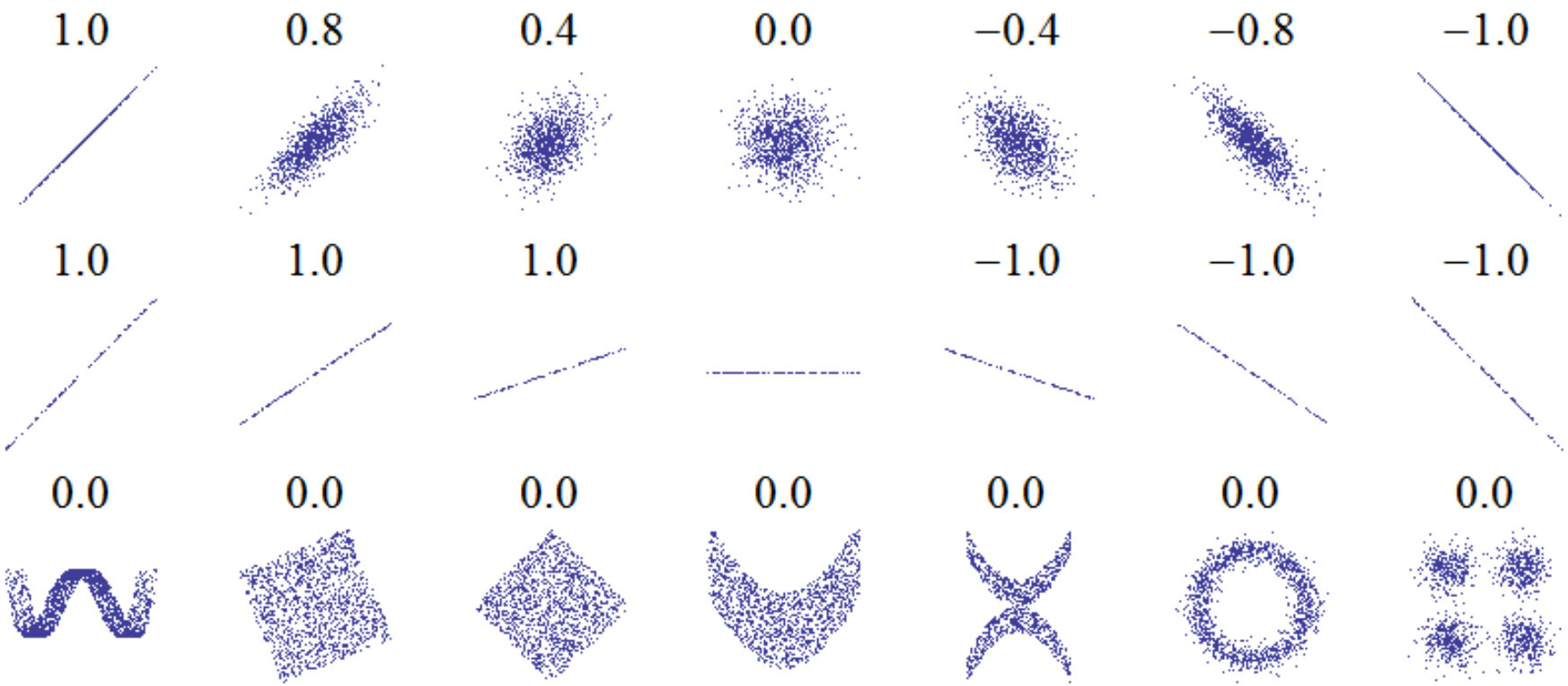
Correlation analysis

A linear correlation coefficient (Pearson's correlation coefficient) determines the degree to which variables are linearly related. The coefficient ranges from -1 to 1, and is interpreted as follows:

- $r_{xy} = 0$:negligible correlation
- $0.1 < r_{xy} \leq 0.3$:weak correlation
- $0.3 < r_{xy} \leq 0.5$:average correlation
- $0.5 < r_{xy} \leq 0.7$:high correlation
- $0.7 < r_{xy} \leq 0.9$:very high correlation
- $0.9 < r_{xy} \leq 1$:almost perfect correlation

- Positive correlation: an increase in the value of one trait corresponds to an increase in the average value of the other trait.
- Negative correlation: an increase in the value of one trait corresponds to a decrease in the average value of the other trait.

Correlations between two variables



https://pl.wikipedia.org/ Pearson correlation coefficient

Correlation analysis

- Linear correlation coefficient from a sample:

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

- When the sample is very large, the normal distribution can be used to construct a confidence interval for the correlation coefficient or to test the hypothesis about the value of the correlation coefficient.

Regression analysis

- The regression function is used to study the relationships between variables. It is an analytical expression assigning the mean values of the dependent variable to specific values of the independent variable.
  - Dependent variable (explained) — the one assumed to depend on another variable.
  - Independent variable (explaining) — its value is treated as given and not explained; it is assumed that independent variables determine the dependent variable or affect it.
- The simplest form is a linear function: $y = ax + b$
- The equation parameters (a, b) are usually determined by the least squares method.

Regression analysis

$$\hat{y} = ax + b$$

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

Regression analysis

The coefficient of determination R^2 is a measure of the goodness of fit of the regression function. It indicates what proportion of the variable y has been explained by linear regression relative to variable x:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where:

- $\hat{y}_t$:value of the explained variable from the model (regression equation)
- y_t :actual value of the explained variable
- $\bar{y}$:arithmetic mean of the empirical values of the explained variable

Examples

10.1. We would like to examine how temperature changes depending on the depth below the earth's surface. Using data from the table, calculate the correlation coefficient r, determine the equation of the regression line that allows estimation of temperature based on the depth data, calculate the coefficient of determination R^2 ,present the results in the form of a graph, and interpret the results of the calculations.

Depth	Temperature
200	10
400	15
600	23
800	26
1000	33
1200	37

Excel spreadsheet using the formulas shown in the presentation:

	Depth	Temperature	xi-xsr	y-ysr	Numerator	(xi-xsr)^2	(y-ysr)^2
	200	10	-500	-14	7000	250000	196
	400	15	-300	-9	2700	90000	81
	600	23	-100	-1	100	10000	1
	800	26	100	2	200	10000	4
	1000	33	300	9	2700	90000	81
	1200	37	500	13	6500	250000	169
Average	700	24		sum	19200	700000	532

Quotient		Iloraz	372400000
Square Root		pierwiastka	19297,66825
r=		r=	0,994938857

Explanation of calculations performed in table, part 1:

200 – 700 = - 500

10 – 24= - 14

(-500) x (-14) = - 7000

	Depth	Temperature	xi-xsr	y-ysr	Numerator	(xi-xsr)^2	(y-ysr)^2
	200	10	-500	-14	7000	250000	196
	400	15	-300	-9	2700	90000	81
	600	23	-100	-1	100	10000	1
	800	26	100	2	200	10000	4
	1000	33	300	9	2700	90000	81
	1200	37	500	13	6500	250000	169
Average	700	24		sum	19200	700000	532

Quotient	Iloraz		372400000
Square Root	pierwiastka		19297,66825
r=	r=		0,994938857

Explanation of calculations performed in table, part 2:

	Depth	Temperature	xi-xsr	y-ysr	Numerator	(xi-xsr)^2	(y-ysr)^2
	200	10	-500	-14	7000	250000	196
	400	15	-300	-9	2700	90000	81
	600	23	-100	-1	100	10000	1
	800	26	100	2	200	10000	4
	1000	33	300	9	2700	90000	81
	1200	37	500	13	6500	250000	169
Average	700	24		sum	19200	a 700000	b 532

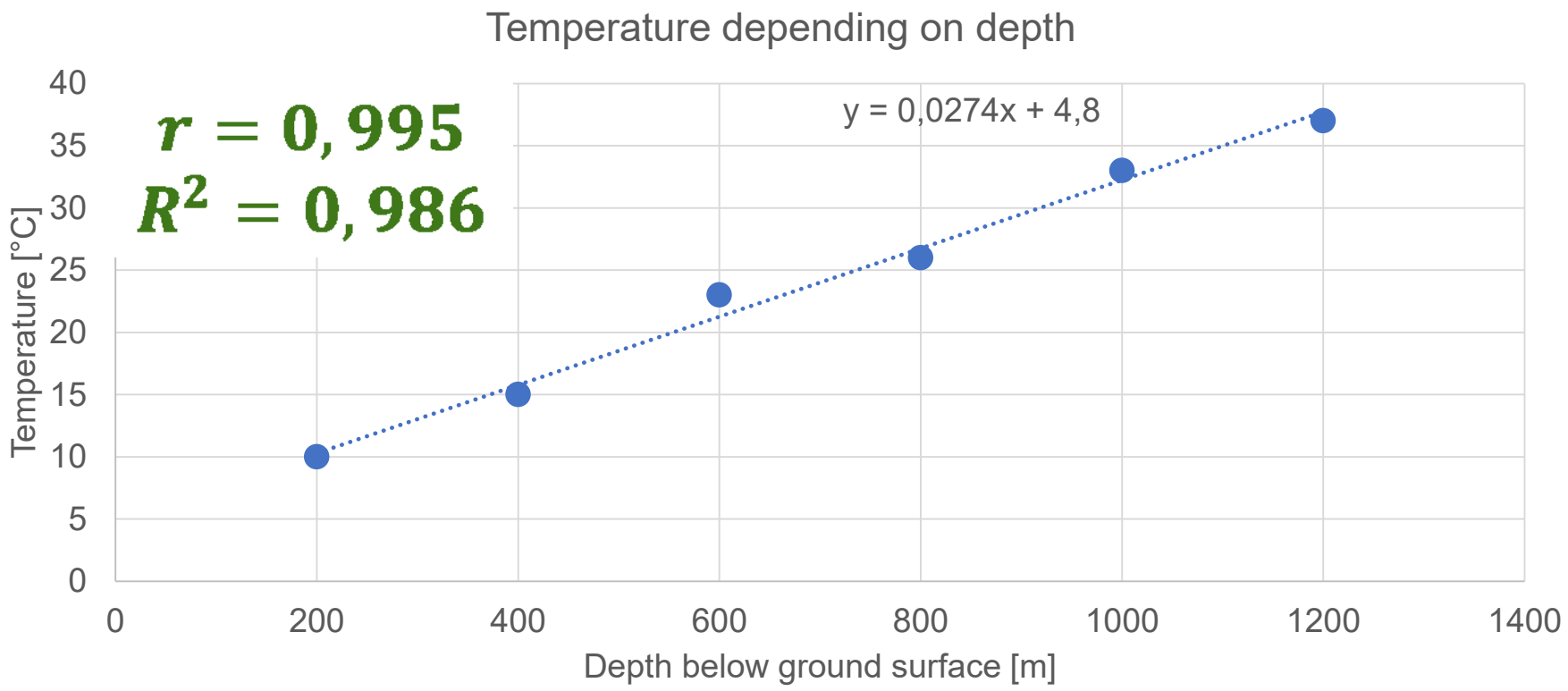
Iloraz 372400000 | a x b

pierwiastka 19297,66825 | $\sqrt{axb}$

r= 0,994938857

19200/19297,7

Solution


Temperature depending on depth
$r = 0,995$
$R^2 = 0,986$
y = 0,0274x + 4,8
Temperature [°C]
40
35
30
25
20
15
10
5
0
0
200
400
600
800
1000
1200
1400
Depth below ground surface [m]

Examples

We would like to examine whether there is a relationship between age and blood pressure in patients with hypertension. Using the data in the table, calculate the correlation coefficient r, determine the equation of the regression line to estimate blood pressure depending on the age of the patient with hypertension, calculate the coefficient of determination R^2 .Present the results as a graph. Interpret the results of the calculations."

Age	Blood pressure
67	150
73	162
64	154
74	168
54	135
61	148
65	158
46	128
72	165

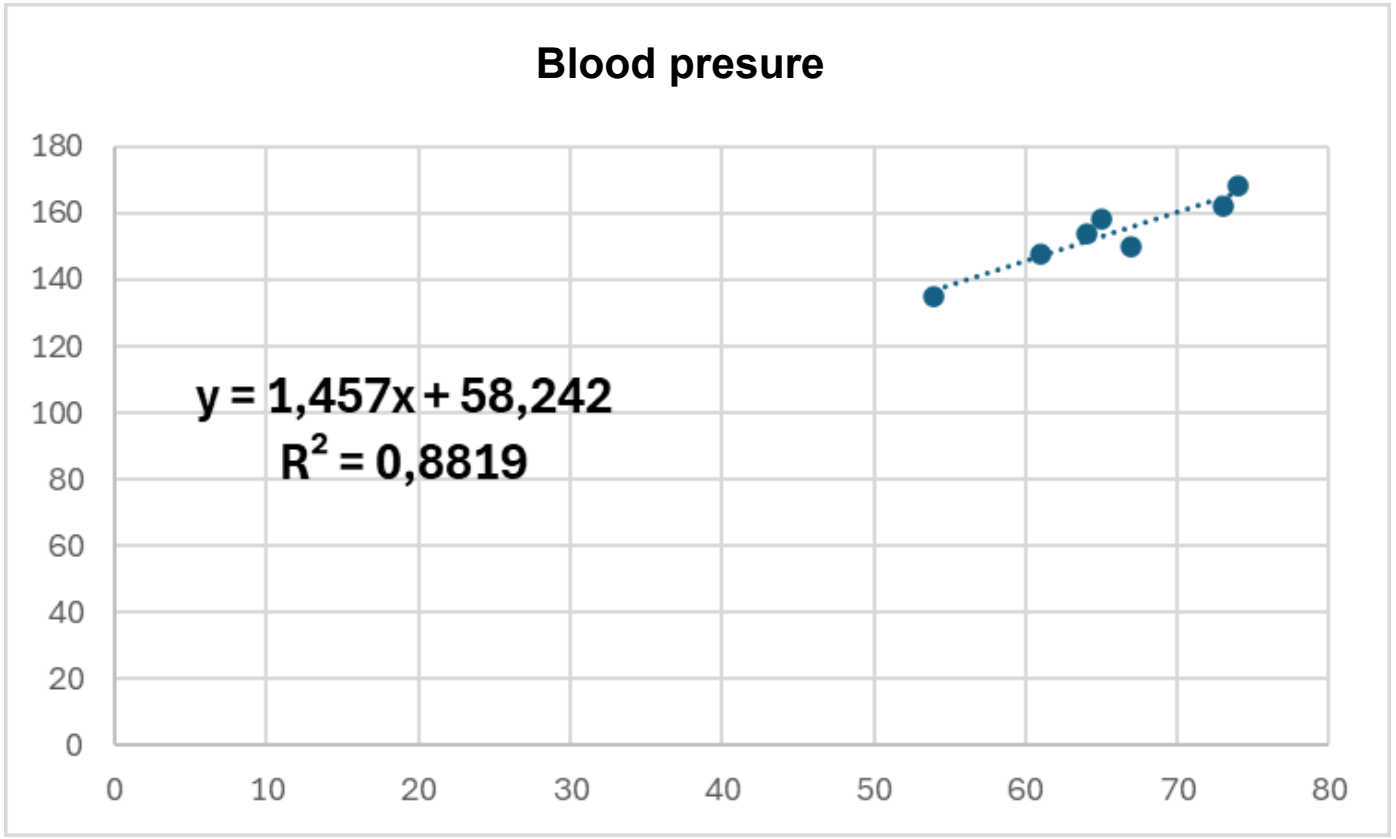
	age	blood pressure	xi-xsr	y-ysr	Numerator	(xi-xsr)^2	(y-ysr)^2
	67	150	1,571429	-3,5714286	-5,612244898	2,469387755	12,75510204
	73	162	7,571429	8,42857143	63,81632653	57,32653061	71,04081633
	64	154	-1,42857	0,42857143	-0,612244898	2,040816327	0,183673469
	74	168	8,571429	14,4285714	123,6734694	73,46938776	208,1836735
	54	135	-11,4286	-18,571429	212,244898	130,6122449	344,8979592
	61	148	-4,42857	-5,5714286	24,67346939	19,6122449	31,04081633
	65	158	-0,42857	4,42857143	-1,897959184	0,183673469	19,6122449
Mean	65,42857143	153,5714286		sum	416,2857143	285,7142857	687,7142857

196489,7959
443,2716954

r= 0,939120902
R2= 0,881948068

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{416.28}{285.71} = 1.457$$

$$b = \bar{y} - a\bar{x} = 153.57 - 1.457 \cdot 65.43 = 58.242$$



- r=0.969r=0.969
- R2=0.931$R2$=0.931

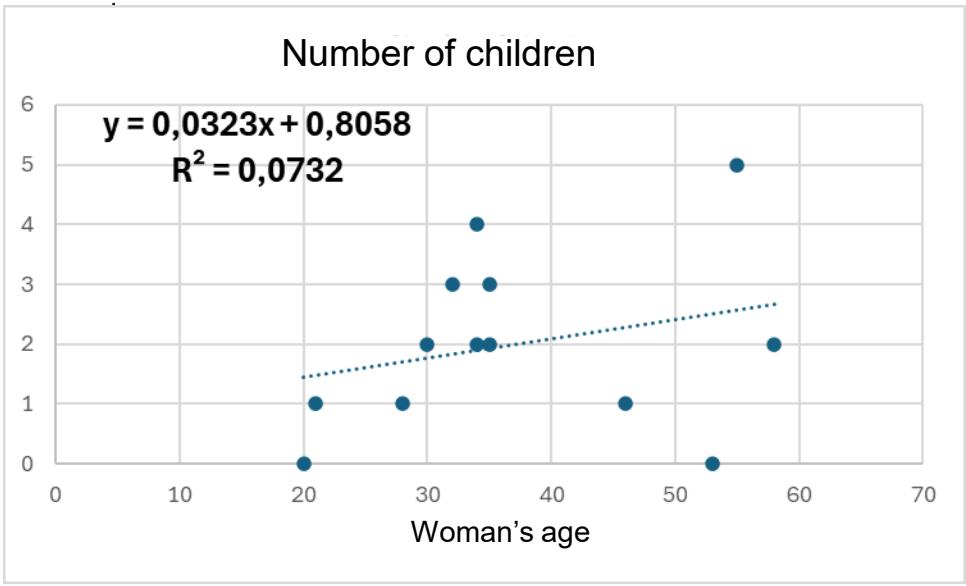
Examples

10.3. Investigate whether there is a relationship between a woman's age and the number of children she has. Calculate the value of the linear correlation coefficient. Present the data in the form of a scatterplot.

woman's age	number of children
55	5
21	1
35	2
58	2
28	1
30	2
32	3
20	0
35	3
46	1
34	2
53	0
34	4

Solution:

	age	No. of children	xi-xsr	y-ysr	numerator	(xi-xsr)^2	(y-ysr)^2
	55	5	18	3	54	324	9
	21	1	-16	-1	16	256	1
	35	2	-2	0	0	4	0
	58	2	21	0	0	441	0
	28	1	-9	-1	9	81	1
	30	2	-7	0	0	49	0
	32	3	-5	1	-5	25	1
	20	0	-17	-2	34	289	4
	35	3	-2	1	-2	4	1
	46	1	9	-1	-9	81	1
	34	2	-3	0	0	9	0
	53	0	16	-2	-32	256	4
	34	4	-3	2	-6	9	4
mean	37	2		sum	59	1828	26



	47528
	218,0091741
r=	0,270630813
R2=	0,073241037
a=	0,032275711
b=	0,805798687

Sources:

- Jerzy Greń: Mathematical Statistics. Models and Tasks. Fourth Edition, supplemented. State Scientific Publishing House. Warsaw, 1974.
- Kot, SM, Jakubowski, J., Sokołowski, A.: Statistics. Second revised edition. Difin Publishing House . Warsaw, 2011.
- Stanisz, A.: Accessible statistics course using STATISTICA PL on examples from medicine. Volume 1. Basic statistics. StatSoft Polska Publishing House. Cracow, 2006.