HUMBOLDT-UNIVERSITÄT ZU BERLIN

HUMBOLDT-UNIVERSITÄT ZU BERLIN
Resource Economics

Faculty of Life Sciences | Thaer-Institute

Introduction online course on Mathematics and Statistics

Preparatory Course for M.Sc. Integrated Natural Resource Management

5 Correlation and Regression Analysis

Department for Agricultural Economics | Resource Economics Group

HUMBOLDT-UNIVERSITÄT ZU BERLIN

HUMBOLDT-UNIVERSITÄT ZU BERLIN
Resource Economics

Syllabus

5 Correlation and Regression Analysis

5.1 Introduction into correlation and regression analysis

Correlation is a statistical method used to determine whether a relationship between variables exists [1].

Regression is a statistical method used to describe the nature of the relationship between variables, that is, positive or negative, linear or nonlinear [1].

Example: A researcher wants to examine the relationship between GHG emissions of a country and its wealth. For this purpose he analysed the data on Gross Domestic Product (GDP) and Greenhouse Gas (GHG) emissions in 2012 in 33 OECD countries.

- Assumption: the amount of GHG emissions depends on economic activity of a country.
- GDP pro capita → an independent variable.
- The volume of GHG emissions pro capita → a dependent variable.

Independent variable → x variable.
Dependent variable → y variable.

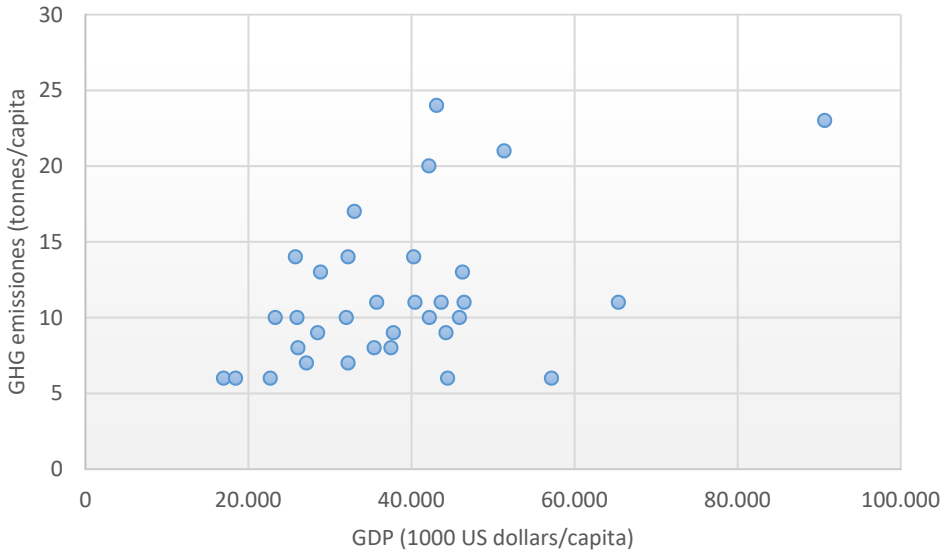
A ***scatter plot*** is a graph of the ordered pairs (x, y) of numbers consisting of the independent variable x and the dependent variable y [1].

HUMBOLDT-UNIVERSITÄT ZU BERLIN

HUMBOLDT-UNIVERSITÄT ZU BERLIN
Resource Economics

5.1 Introduction into correlation and regression analysis

Country	GDP (US 1000 dollars/capita)	GHG emissions intensity (tonnes/capita)
Australia	43.081	24
Austria	45.887,10	10
Belgium	42.212,30	10
Canada	42.144	20
Czech Republic	28.862,10	13
Denmark	44.250,90	9
Estonia	25.769,80	14
Finland	40.437,40	11
France	37.487,90	8
Germany	43.653,80	11
Greece	25.980	10
Hungary	22.701,50	6
Iceland	40.277,90	14
Ireland	46.269,30	13
Israel	32.007	10
Italy	35.424,20	8
Japan	35.738,30	11
Korea	32.222,90	14
Luxembourg	90.693,60	23
Mexico	16.957,70	6
Netherlands	46.457,10	11
New Zealand	32.991,30	17
Norway	65.394,30	11
Poland	23.310,20	10
Portugal	27.125,30	7
Slovak Republik	26.097,70	8
Slovenia	28.498,40	9
Spain	32.240,20	7
Sweden	44.433,60	6
Switzerland	57.174,90	6
Turkey	18.437,10	6
United Kingdom	37.788,30	9
United States	51.368,20	21

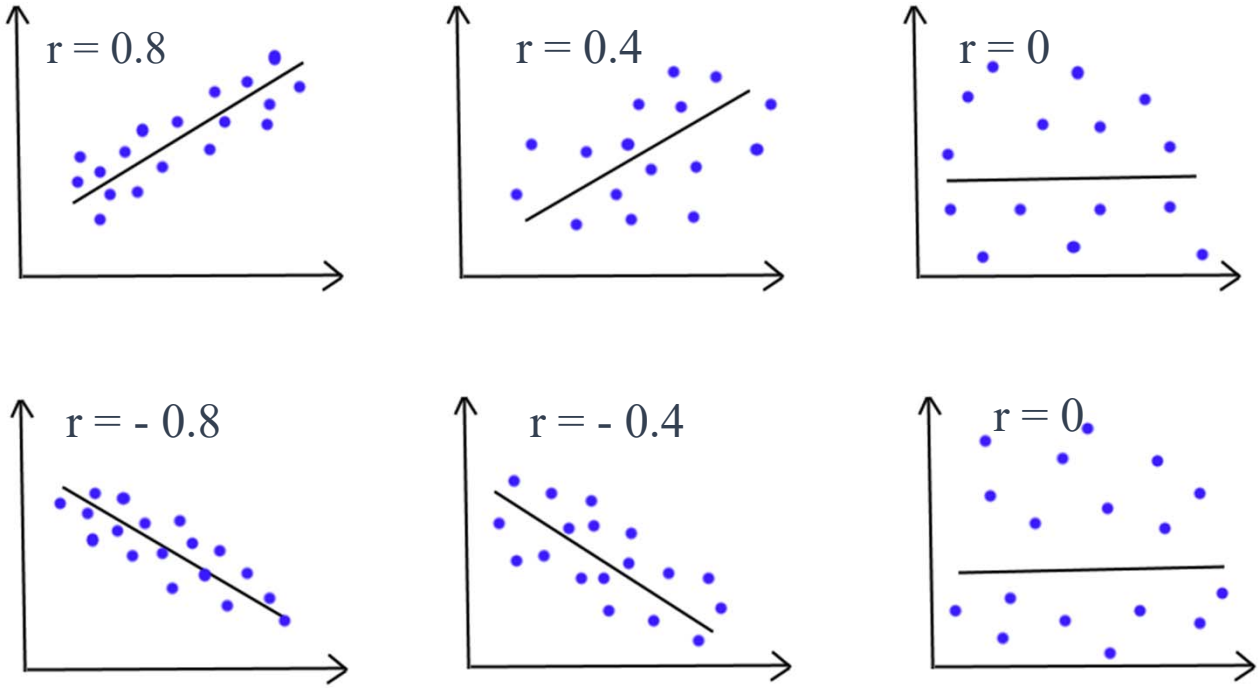


5.2.1 Sample correlation coefficient

Correlation coefficient computed from the sample data measures the strength and direction of a linear dependence between two variables and is denoted by *r* [1].

$-1 < r < +1$

- positive linear relationship between the variables → $r = 1$;
- negative linear relationship between the variables → $r = -1$ [1] .



5.2.1 Sample correlation coefficient

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

where n is the number of data pairs [1] .

Example: *Using the formula for computing the correlation coefficient, calculate its value for the data of European OECD countries* [1; 8].

- *Make a table and fill it with corresponding values.*
- *Substitute in the formula and compute r:*

$$r = \frac{(24)(10464928{,}7) - (958427{,}8)(244)}{\sqrt{[(24)(43640381673) - (958427{,}8)^2][(24)(2784) - (244)^2]}}$$

$$= \frac{251158288{,}8 - 233856383{,}2}{\sqrt{[1047369160152 - 918583847812{,}84][66816 - 59536]}}$$

$$= \frac{17301905{,}6}{\sqrt{(128785312339{,}16)(7280)}} \approx 0{,}565$$

Country	GDP x (US dollars/capita)	GHG emissions intensity y (tonnes/capita)	xy	x^2	y^2
Austria	45.887,10	10	458871	2105625946	100
Belgium	42.212,30	10	422123	1781878271	100
Czech Republic	28.862,10	13	375207,3	833020816,4	169
Denmark	44.250,90	9	398258,1	1958142151	81
Estonia	25.769,80	14	360777,2	664082592	196
Finland	40.437,40	11	444811,4	1635183319	121
France	37.487,90	8	299903,2	1405342646	64
Germany	43.653,80	11	480191,8	1905654254	121
Greece	25.980	10	259800	674960400	100
Hungary	22.701,50	6	136209	515358102,3	36
Iceland	40.277,90	14	563890,6	1622309228	196
Ireland	46.269,30	13	601500,9	2140848122	169
Italy	35.424,20	8	283393,6	1254873946	64
Luxembourg	90.693,60	23	2085952,8	8225329081	529
Netherlands	46.457,10	11	511028,1	2158262140	121
Norway	65.394,30	11	719337,3	4276414472	121
Poland	23.310,20	10	233102	543365424	100
Portugal	27.125,30	7	189877,1	735781900,1	49
Slovak Republik	26.097,70	8	208781,6	681089945,3	64
Slovenia	28.498,40	9	256485,6	812158802,6	81
Spain	32.240,20	7	225681,4	1039430496	49
Sweden	44.433,60	6	266601,6	1974344809	36
Switzerland	57.174,90	6	343049,4	3268969190	36
United Kingdom	37.788,30	9	340094,7	1427955617	81
∑	958427,8	244	10464928,7	43640381673	2784

5.2.2 Population correlation coefficient

The sample correlation coefficient: $r \neq 0$ if the value of r is high enough to observe a quite strong linear dependence between the variables or it happened due to chance → hypothesis testing [1].

The ***population correlation coefficient ρ*** is the correlation computed by using all possible pairs of data values (x, y) taken from a population [1].

H_0 : $\rho = 0$ → no correlation between the x and y variables of the population.

H_1 : $\rho \neq 0$ → there is significant correlation between the variables of the population.

- If H_0 is rejected → there is a significant difference between the value of r and 0.
- If H_0 is not rejected → The value of r does not significantly differ from 0 and is probably due to chance.

To test the significance of the r value, use the table of Critical Values for Pearson Correlation Coefficient (see next slide). The table shows the values of the correlation coefficient that are significant for a specific α level and a specific number of degrees of freedom (the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary) [4] .

➢ For 5 degrees of freedom and at $\alpha = 0.01$ → a critical value is 0.874. Thus, any value of $r > +0.874$ or $r < -0.874$ will be significant and the null hypothesis will be rejected [1] .

Example: Let us test the significance of the computed correlation coefficient of the sample data from the example at $\alpha = 0.05$.

- H_0: $\rho = 0$ H_1: $\rho \neq 0$
- 24 -2 = 22 degrees of freedom
- According to the table the critical value for 22 degrees of freedom and $\alpha = 0.05$ is $\rho = 0.404$ → which means that any value of $r > +0.404$ or $r < -0.404$ will be significant. The correlation coefficient for the data of European OECD countries in the example equals to 0.565.
- $r > \rho$ (0.565 > 0.404) → the null hypothesis has to be rejected → there is enough evidence to claim that there is a significant linear dependence between the variables [1;8] .

5.2.2 Population correlation coefficient

Critical Values for Pearson Correlation Coefficient [2]

df = n – 2 n = # of pairs of data	Level of significance for two-tailed test			
	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
11	.476	.553	.634	.684
12	.458	.532	.612	.661
13	.441	.514	.592	.641
14	.426	.497	.574	.628
15	.412	.482	.558	.606
16	.400	.468	.542	.590
17	.389	.456	.528	.575
18	.378	.444	.516	.561
19	.369	.433	.503	.549
20	.360	.423	.492	.537
21	.352	.413	.482	.526
22	.344	.404	.472	.515
23	.337	.396	.462	.505
24	.330	.388	.453	.495
25	.323	.381	.445	.487
26	.317	.374	.437	.479
27	.311	.367	.430	.471
28	.306	.361	.423	.463
29	.301	.355	.416	.456
30	.296	.349	.409	.449
35	.275	.325	.381	.418
40	.257	.304	.358	.393
45	.243	.288	.338	.372
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.302
80	.183	.217	.256	.284
90	.173	.205	.242	.267
100	.164	.195	.230	.254

5.2.2 Population correlation coefficient

The strong association between the variables exists:

→ there is a direct cause-and-effect relationship between the variables (x causes y);

→ there is a reverse cause-and-effect relationship (y causes x);

→ the relationship of the variables might be caused from another variable;

→ there might be complex interrelationships among multiple variables;

→ the obtained relationship is due to chance [1] .

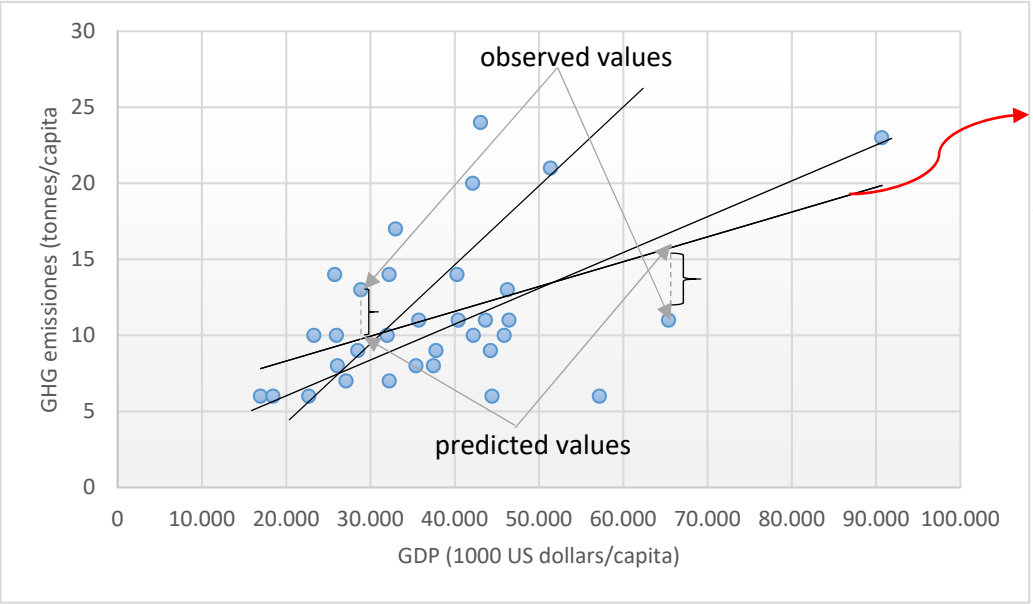
Exercise #1: Using the data below [5;6] and the correlation analysis, try to answer the following research question: whether income may explain water use pattern ?

	GDP/cap in PPS (Index EU28=100) x	Water withdrawal m3/yr/cap y
Belgium	121	590
Bulgaria	45	817
Czech Republic	83	165
Denmark	129	119
Germany	120	391
Estonia	65	1337
Greece	85	841
Spain	96	705
France	108	512
Croatia	59	143
Italy	104	790
Cyprus	100	167

Latvia	53	176
Lithuania	60	704
Luxembourg	256	136
Hungary	64	557
Malta	84	134
Netherlands	134	639
Austria	126	452
Poland	62	313
Portugal	82	812
Romania	52	320
Slovenia	83	464
Finland	116	309

5.3.1. Regression analysis

Regression analysis is a statistical process for estimating the relationships among variables [1].



Regression line (or the data´s line of best fit) → the shorter overall distance from the points to this line is - the better the prediction is [10;8].

- Algebra: equation of the line is y = mx + b, where *m* defines the slope of the line and *b* is the *y* intercept.
- Statistics: equation of the regression line is y´ = a + bx, where *a* is the y´ intercept and *b* is the slope of the line [1] .

5.3.2. Regression line equation

Formulas for the regression line is $y' = a + bx$ [1]:

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Example: *Using the formulas for the regression line, find its equation for the data of european OECD countries* [8] .

- *Make a table and fill it in with corresponding values* → *Use the table from slide 6.*
- *Substitute in the formula and compute* a *and* b*:*

$$a = \frac{(244)(43640381673) - (958427{,}8)(10464928{,}7)}{(24)(43640381673) - (958427{,}8)^2} = 4.802$$

$$b = \frac{(24)(10464928{,}7) - (958427{,}8)(244)}{(24)(43640381673) - (958427{,}8)^2} = 1.343$$

- *Insert the values into the equation:*

  $y' = 4.802 + 1.343x$

5.3.2. Regression line equation

Predictions for the dependent variable [1]:

- when the regression line shows a reasonable good fit with the points;
- when the correlation coefficient shows significant interrelation of the variables;
- when the data does not go much beyond the scope of the available sample data .

Example: *Use the regression line to predict GHG emissions for a country which GDP might be equal to 50 000 $/capita* [8].

- *As the x values are in 1000s ($/capita) → 50 000/1000 = 50.*
- *Insert this value into equation: y´= 4.802 + 1.343(50) = 71.952*
- *Hence, when a country´s GDP is equal to 50 000 $/capita, ist GHG emissions will be approcimately 71.962 tonnes/capita.*

Key assumptions of linear regression models [1]**:**

- Variables are normally distributed.
- The standard deviation of dependent and independent variables are equal .

A ***marginal change*** is the amount of the change in one variable when the other variable changes by exactly 1 unit [1].

Influential points are outliers that might affect the regression line equation [1].

5.4.1 Variation for the regression model

Examine the presented hypothetical regression model [1;11]*:*

x	2	4	6	8	10	12	14	16
y	4	8	10	6	16	20	18	24

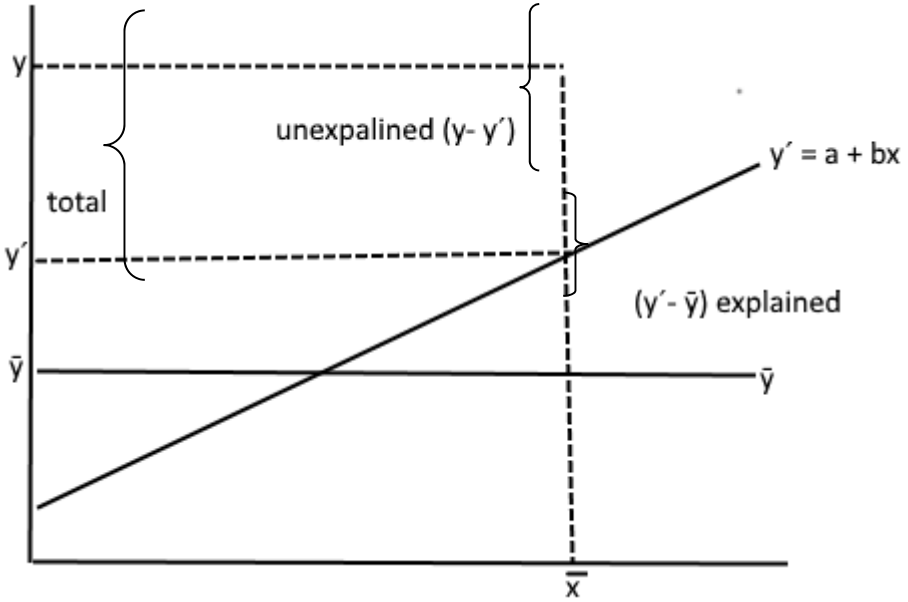
- *The regression line equation: $y' = -2.068 + 2.202x$*
- *The correlation coefficient: $r = 0,96$*
- *$x = 2 \rightarrow y = 4$ (according to the data) and $y' = -2.068 + (2.202)(2) = 2.336$*

The ***total variation*** $\sum(y - \bar{y})^2$ represents the sum of the squares about the mean [1].

$$\sum(y - \bar{y})^2 = \sum(y' - \bar{y})^2 + \sum(y - y')^2$$

The ***explained variation*** $\sum(y' - \bar{y})^2$ represents the deviation of the predicted value of each observation from the mean [1].

The ***unexplained variation*** $\sum(y - y')^2$ is the variation due to chance and represents the deviation of each observation from its predicted value [1].



5.4.1 Variation for the regression model

- Find the predicted y' values:
- $x = 2 \rightarrow y' = -2.068 + (2.202)(2) = 2.336$
- $x = 4 \rightarrow y' = -2.068 + (2.202)(4) = 6.74$
- $x = 6 \rightarrow y' = -2.068 + (2.202)(6) = 11.144$
- $x = 8 \rightarrow y' = -2.068 + (2.202)(8) = 15.548$
- $x = 10 \rightarrow y' = -2.068 + (2.202)(10) = 19.952$
- $x = 12 \rightarrow y' = -2.068 + (2.202)(12) = 24.356$
- $x = 14 \rightarrow y' = -2.068 + (2.202)(14) = 28.76$
- $x = 16 \rightarrow y' = -2.068 + (2.202)(16) = 33.164$
- Find the mean of the y values:

$$\bar{y} = \frac{4 + 8 + 14 + 10 + 16 + 24 + 30 + 36}{8} = 17.75$$

- Find the total variation:
- $(4 - 17.75)^2 = 189.06$
- $(8 - 17.75)^2 = 95.06$
- $(14 - 17.75)^2 = 14.06$
- $(10 - 17.75)^2 = 60.06$
- $(16 - 17.75)^2 = 3.06$
- $(24 - 17.75)^2 = 39.06$
- $(30 - 17.75)^2 = 150.06$
- $(36 - 17.75)^2 = 333.06$
- $\sum(y - \bar{y})^2 = 883.48$

The values of $(y - y')$ are called residuals.

- Find the explained variation:
- $(2.336 - 17.75)^2 = 237.59$
- $(6.74 - 17.75)^2 = 121.22$
- $(11.144 - 17.75)^2 = 43.64$
- $(15.548 - 17.75)^2 = 4.85$
- $(19.952 - 17.75)^2 = 4.85$
- $(24.356 - 17.75)^2 = 43.64$
- $(28.76 - 17.75)^2 = 121.22$
- $(33.164 - 17.75)^2 = 237,59$
- $\sum(y' - \bar{y})^2 = 814.6$

- Find the unexplained variation:
- $(4 - 2.336)^2 = 2.77$
- $(8 - 6.74)^2 = 1.59$
- $(14 - 11.144)^2 = 8.16$
- $(10 - 15.548)^2 = 30.78$
- $(16 - 19.952)^2 = 15.62$
- $(24 - 24.356)^2 = 0.13$
- $(30 - 28.76)^2 = 1.54$
- $(36 - 33.164)^2 = 8.04$
- $\sum(y - y')^2 = 68.63$

A ***residual*** is the difference between the actual value of y and the predicted value y for a given x value [1].

5.4.2 Coefficient of determination

The ***coefficient of determination***, r^2, is a measure of the variation of the dependent variable that is explained by the regression line and the independent variable [1].

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

→ From our previous example, $r^2 = 814.6/883.48 = 0.922$ → ~ 92%.

- Coefficient of nondetermination: $1 - r^2$
- The coefficient of determination can be also computed by squaring the value of the correlation coefficient. →r =0.96 and r^2 ~ 0,922.
- $0 \leq r^2 \leq 1$:
  - The more the value of the coefficient is close to 1, the more it is close to a perfect fit → the more reliable is the model for predictions.
  - The closer is the value of the coefficient to 0 → the more unreliable is the model for predictions [1;12] .

5.4.3 Standard error of the estimate

The ***standard error of the estimate***, s_{est}, is the standard deviation of the observed y values about the predicted y´values [1].

$$s_{est} = \frac{\sum(y - y')^2}{n - 2}$$

$$s_{est} = \frac{68.63}{8 - 2} \sim 11.44$$

→ The standard deviation of observed values about the predicted values is 11.44.

$$s_{est} = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n - 2}}$$

Example: *compute the standard error of the estimate by using the data from the hypothetical regression model (see slide 13)* [1] .

- *Compute the product of x and y values.*
- *Find the square of each y value.*
- *Calculate the sum of y values, of the product of x and y values and of the squares of y values.*
- *Insert the values into the formula:*

$$s_{est} = \sqrt{\frac{1772 - (-2.068)(106) - (2.202)(1184)}{8 - 2}} \sim 10.13$$

x	y	xy	y^2
2	4	8	16
4	8	32	64
6	10	60	100
8	6	48	36
10	16	160	256
12	20	240	400
14	18	252	324
16	24	384	576
	$\sum y = 106$	$\sum xy = 1184$	$\sum y^2 = 1772$

5.5 Multiple regression equation

Multiple regression is a statistical process, which allows to explain the relationship between a dependent variable and several explanatory (independent) variables [1].

The multiple regression equation [1]:

$$y' = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

How to compute *a* [1]:

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$

How to compute *b* [7]:

$$b_1 = \left[\frac{r_{y,x1} - r_{y,x2}r_{x1,x2}}{1 - (r_{x1,x2})^2}\right]\left[\frac{SD_y}{SD_{x1}}\right]$$

$$b_2 = \left[\frac{r_{y,x2} - r_{y,x1}r_{x1,x2}}{1 - (r_{x1,x2})^2}\right]\left[\frac{SD_y}{SD_{x2}}\right]$$

Key assumptions for multiple regression are similar to those for a simple regression [1] :

- Variables are normally distributed.
- The standard deviations of dependent variables are the same for each value of the independent variable.
- There is a linear relationship between the outcome variable and the independent variables.
- The independent variables are not highly correlated with each other.

where SD_y – a standard deviation of a dependent variable,
SD_{x1} – a standard deviation of the first independent variable,
SD_{x2} – a standard deviation of the second independent variable.

5.5 Multiple correlation coefficient

The formula for computing the multiple correlation coefficient [1] :

$$R = \sqrt{\frac{r_{yx1}^2 + r_{yx2}^2 - 2r_{yx1}\, r_{yx2} r_{x1x2}}{1 - r_{x1x2}^2}}$$

Example: *Assume you collected data on sheet and rill erosion in some US States and want to analyse the causes of it. Based on the data on average annual precipitation, cultivated land and estimates of erosion in the US States, find the value of R* [3;9].

- *Finding correlation coefficient of each independent variable:*
  $r_{yx1} = 0.293$
  $r_{yx2} = 0.209$
  $r_{x1x2} = -0.506$

- *Calculating the multiple correlation coefficient:*

$$R = \frac{(0{,}293)^2 + (0{,}209)^2 - 2 \cdot 0.293 \cdot 0.209 \cdot (-0.506)}{1 - (-0{,}506)^2} =$$

$= 0.507$

→ *The two independent variables have a medium relation to sheet and rill erosion and therefore, it is quite difficult to estimate the erosion based only on these variables.*

State	Average annual precipitation (mm)	Cultivated land (1000 acres)	Estimated sheet and rill erosion (t/acre/year)
Connecticut	1279	77,9	2,25
Delaware	1160	480,1	1,5
Idaho	481	4571,7	2,24
Illinois	996	23621,5	4,11
Indiana	1060	12804,4	2,56
Iowa	864	24127,4	5,05
Kentucky	1242	3521,7	3,45
Maine	1072	163,8	1,49
Maryland	1131	1371,1	4,54
Massachusetts	1211	54	1,11
Michigan	833	6528,8	1,85
Minnesota	693	19665,1	1,95
Montana	390	12588,2	0,97
Nebraska	599	18024,4	2,65
New Hampshire	1103	20,7	0,73
New Jersey	1196	425,9	4,97
New York	1062	2705,4	1,98
North Dakota	452	22871,3	0,96
Ohio	993	10229,9	2,4
Oregon	695	2676,5	2,09
Pennsylvania	1089	3632	3,4
Rhode Island	1218	4,8	1,12
South Dakota	511	14437,5	1,26
Vermont	1085	138,1	1,3
Virginia	1125	1669,8	2,83
Washington	976	5483,6	4,98
West Virginia	1147	160,8	1,49
Wisconsin	829	8753,9	3,31
Wyoming	328	977,7	0,49

HUMBOLDT-UNIVERSITÄT ZU BERLIN

HUMBOLDT-UNIVERSITÄT ZU BERLIN
Resource Economics

5.5 Multiple correlation coefficient

Exercise #2: Using the data on sheet and rill erosion in another US States examine whether the average annual precipitation and land area subjected to cultivation have a strong relation to sheet and rill erosion there. How would the multiple regression equation for this set of data look like?

State	Average annual precipitation (mm)	Cultivated land (1000 acres)	Estimated sheet and rill erosion (t/acre/year)
Alabama	1480	2562	4,45
Arkansas	1284	7328,6	3,13
Connecticut	1279	77,9	2,25
Delaware	1160	480,1	1,5
Florida	1385	1485,1	1,15
Georgia	1287	4176,2	4,52
Hawaii	1618	183,8	2,86
Illinois	996	23621,5	4,11
Indiana	1060	12804,4	2,56
Iowa	864	24127,4	5,05
Kentucky	1242	3521,7	3,45
Louisiana	1528	5325,9	2,9
Maine	1072	163,8	1,49
Maryland	1131	1371,1	4,54
Massachusetts	1211	54	1,11
Michigan	833	6528,8	1,85
Minnesota	693	19665,1	1,95
Mississippi	1499	4947	4,23
Missouri	1071	10442	4,11
New Hampshire	1103	20,7	0,73
New Jersey	1196	425,9	4,97
New York	1062	2705,4	1,98
North Carolina	1279	5117	4,73
Ohio	993	10229,9	2,4
Pennsylvania	1089	3632	3,4
Rhode Island	1218	4,8	1,12
South Carolina	1264	2279	2,44
Tennessee	1376	3196,9	5,01
Vermont	1085	138,1	1,3
Virginia	1125	1669,8	2,83
West Virginia	1147	160,8	1,49
Wisconsin	829	8753,9	3,31

Questions of repetition

- Which types of interrelation between two variables exist? How can it be measured?
- What is meant by dependent and independent variables?
- What is the difference between correlation and regression analysis?
- What for is the correlation coefficient used? What is the range of it´s values?
- If two variables are correlated, does it always mean that one causes another?
- What does the regression line show? Which difference does exist between the equation of the line in mathematics and statistics?
- What kinds of variation of the regression model exist? How do they differ?
- What does the coefficient of determination show?
- What is the difference between simple and linear regression?
- What are the assumptions of the multiple regression?

Answers to exercises

Exercise #1:

r = - 0.255, which means that strong linear dependence between the variables. Therefore, according to the results, income may not explain water use patterns.

Exercise #2:

$r_{yx1} = 0.167$

$r_{yx2} = 0.344$

$r_{x1x2} = - 0.542$

R = 0.54 → The two independent variables have a medium relation to sheet and rill erosion.

$y' = - 6531.86 + 0.003x_1 + 1.25x_2$

HUMBOLDT-UNIVERSITÄT ZU BERLIN

HUMBOLDT-UNIVERSITÄT ZU BERLIN
Resource Economics

References:

[1] Allan G. Bluman, 2009. *Elementary Statistics: A Brief Version*, 7th Edition, New York: McGraw-Hill.

[2] Critical Values for Pearson Correlation Coefficient: http://www.statisticshowto.com/tables/ppmc-critical-values/

[3] Current results. Weather and science facts. Retrieved from https://www.currentresults.com/Weather/US/average-annual-state-precipitation.php

[4] Degrees of freedom: https://en.wikipedia.org/wiki/Degrees_of_freedom

[5]Eurostat:http://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tec00114&plugin=1

[6] FAO STATISTICAL YEARBOOK (2014) Europe and Central Asia, Food and Agriculture Food and Agriculture Organization of the United Nations Regional, Office for Europe and Central Asia Budapest: http://www.fao.org/3/a-i3621e.pdf

[7] Higgins, J. (2006): The Radical Statistician: A Beginners Guide to Unleashing the Power of Applied Statistics in The Real World (5th Ed.) Jim Higgins Publishing.

[8] OECD (2015), Environment at a Glance 2015: OECD Indicators, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264235199-en OECD Data. Retrieved from: https://data.oecd.org/gdp/gross-domestic-product-gdp.htm#indicator-chart

[9] U.S. Department of Agriculture. 2015. Summary Report: 2012 National Resources Inventory, Natural Resources Conservation Service, Washington, DC, and Center for Survey Statistics and Methodology, Iowa State University, Ames, Iowa. http://www.nrcs.usda.gov/technical/nri/12summary

[10] https://en.wikipedia.org/wiki/Regression_analysis

[11] https://www.armstrong.edu/academic-departments/mathematics

[12] https://mathbits.com/MathBits/TISection/Statistics2/correlation.htm