1st "Training the Trainers" Workshop
(22. 11. – 03. 12. 2010)

Quantitative Methods in Political Science - A short introduction into descriptive and inferential statistics

Dr. Sebastian Jäckle

UNI FREIBURG

descriptive statistics - **tables**

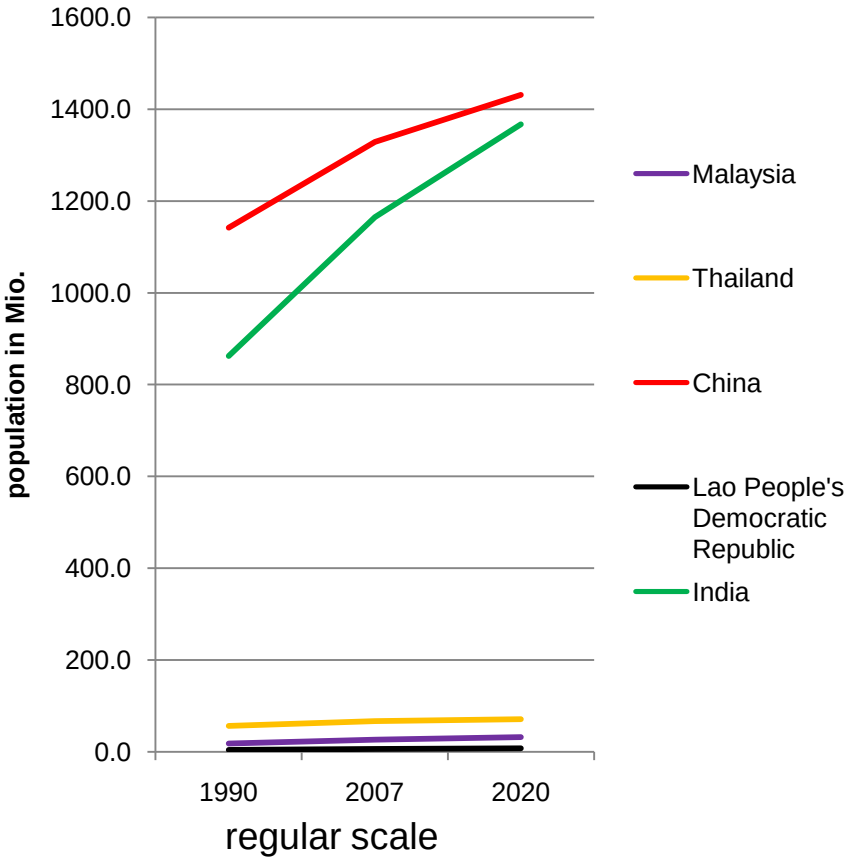
Example 1:

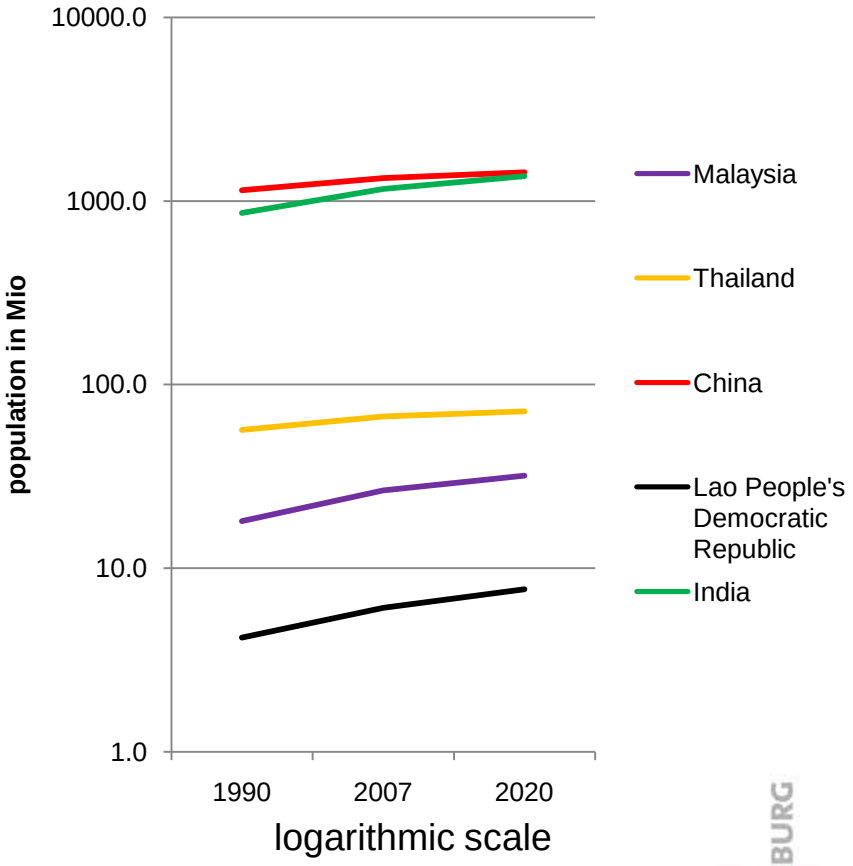
Transnational comparison of government expenditure ratios

country	gov. expenditure ratio	country	gov. expenditure ratio
Austria	49.6	Italy	53.0
Belgium	55.2	Japan	32.3
Canada	46.9	Luxemburg	50.0
Danmark	58.4	Netherlands	55.6
Finland	41.2	Norway	54.8
France	49.9	Portugal	42.9
Germany	46.0	Spain	42.7
Great Britain	42.1	Sweden	61.4
Greece	50.9	Switzerland	30.7
Ireland	43.1	USA	35.4

descriptive statistics – graphs
line graph

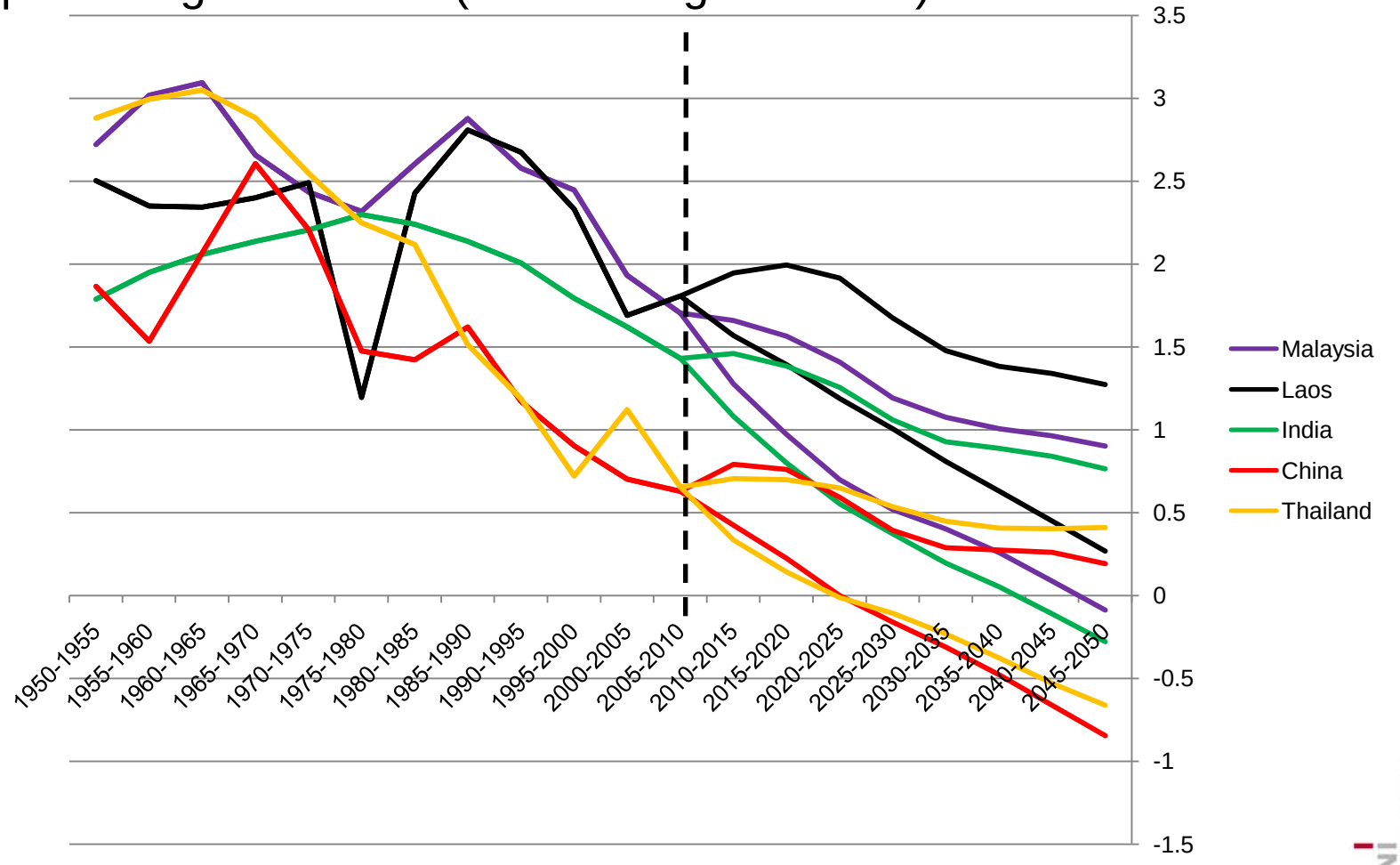
Population growth in Asian countries: different display options





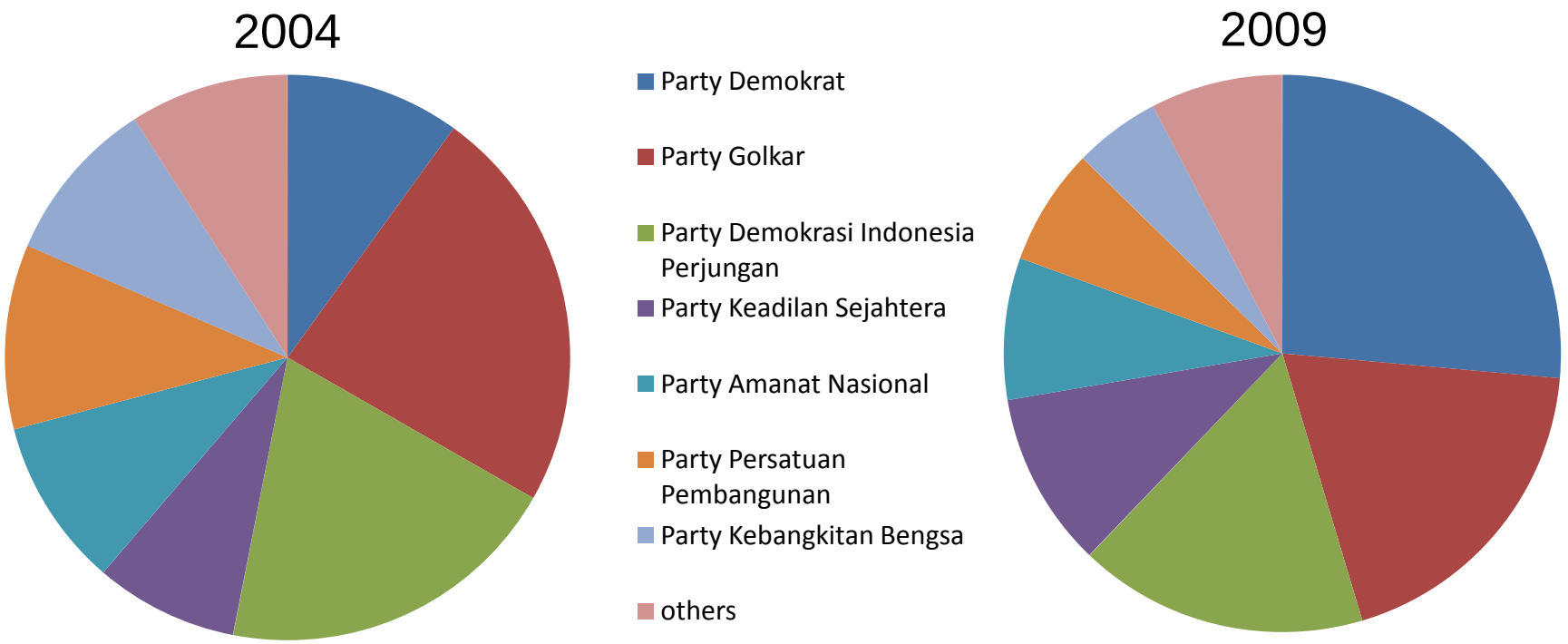
descriptive statistics – graphs
line graph

population growth rates (low and high variants)



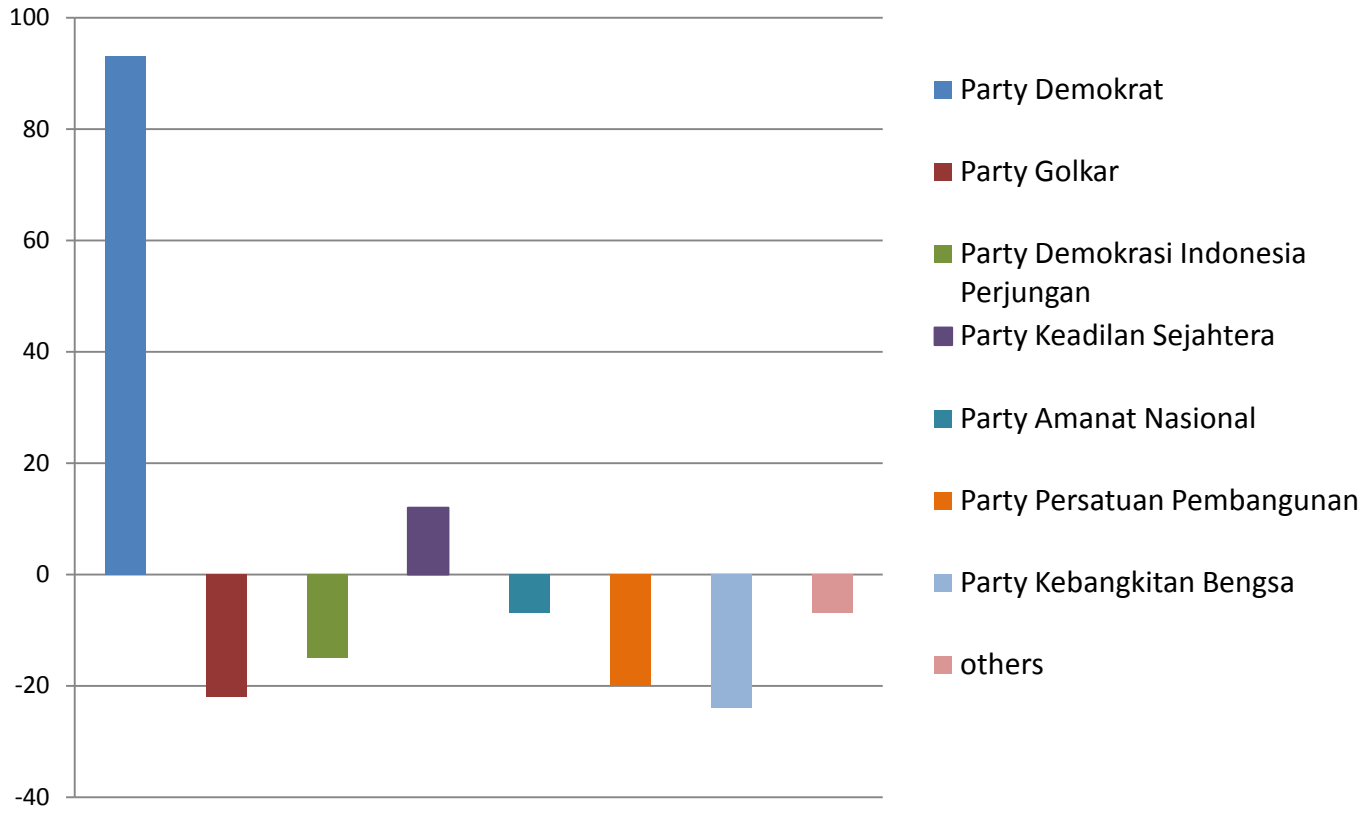
descriptive statistics – graphs
pie chart

Indonesian parliamentary election results (parliamentary seat shares)



descriptive statistics – graphs
bar chart

Seat gains and losses



descriptive statistics – measures
mode

Mode :
value that occurs most frequently

- possible from nominal scale level **and beyond;**
- not meaningful when the distribution is heavily skewed

Example datarow

1,1,2,2,3,4,5,6,6,6,7,7
→ Mode = 6

For continuous variables it makes sense to identify the modal class instead of a mostly non-existing mode.

descriptive statistics – measures
median

Median:

the value separating the upper half of a distribution from the lower half; above it and below it there are 50% of all values

- possible from ordinal scale level and beyond

For unclassed data: x_1, ..., x_n are the ordered values ($x_1 \leq$, ..., $\leq x_n$)

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{when } n \text{ is odd} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right) & \text{when } n \text{ is even and } x \text{ is a metric variable} \end{cases}$$

Example datarow: 1,1,2,2,3,4,5,6,6,6,7,7

→ $\tilde{x} = 4{,}5$

Remember that the data have to be ordered by size for calculating the median!

descriptive statistics – measures
arithmetic mean

Arithmetic mean:

- possible from interval scale level and beyond

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Example datarow: 1,1,2,2,3,4,5,6,6,6,7,7

$$\longrightarrow \quad \bar{x} = 4{,}17$$

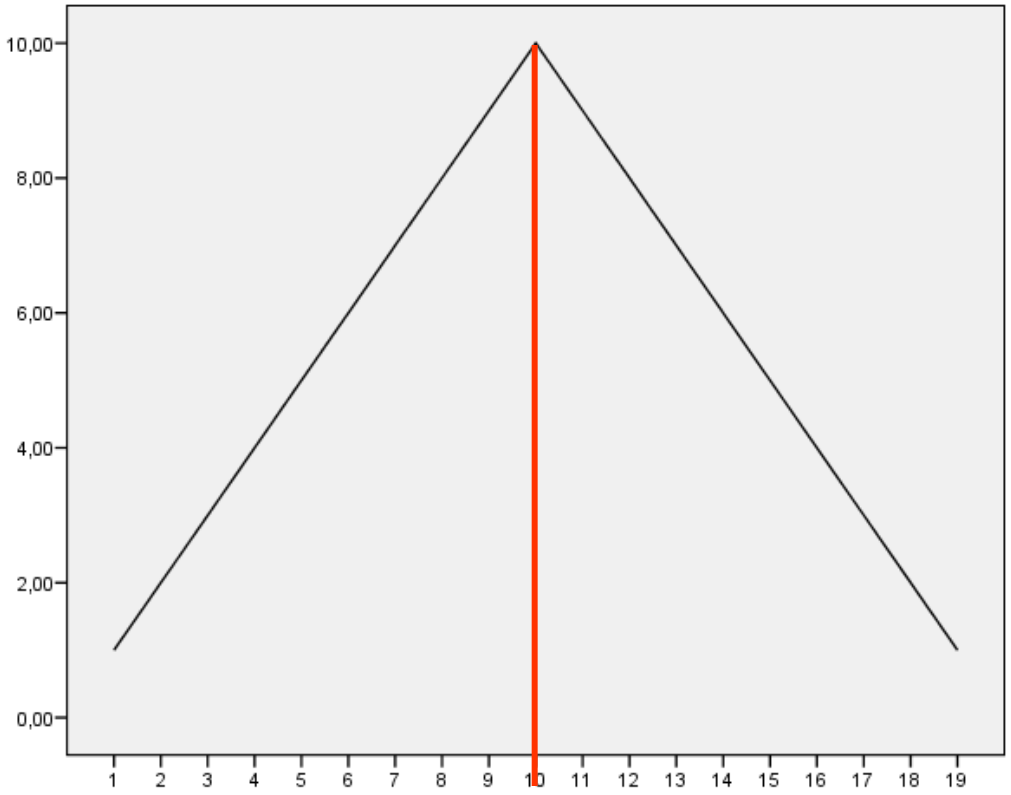
descriptive statistics – **shapes of distribution**

possible shapes of distribution

→ **symmetrical**

Fechners rule:

$$\overline{x} = \tilde{x} = x_{\text{mod}}$$



descriptive statistics – **shapes of distribution**

right-skewed

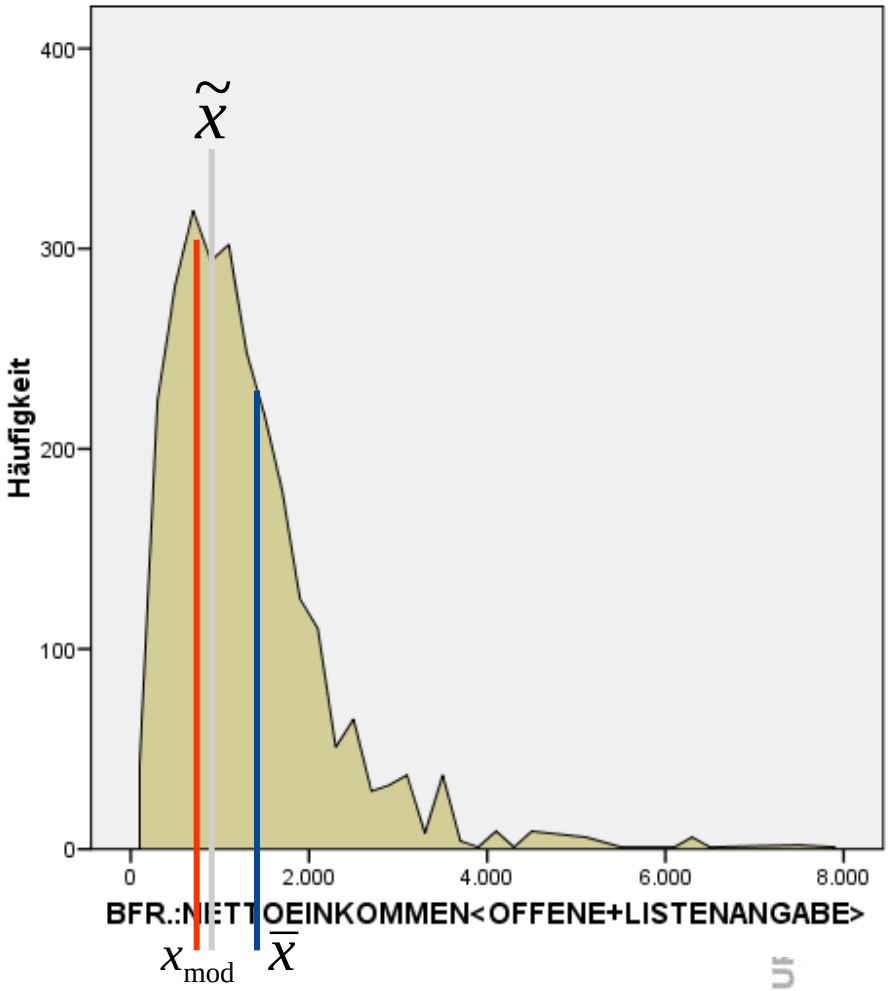
Example: Net income in Germany

Statistiken

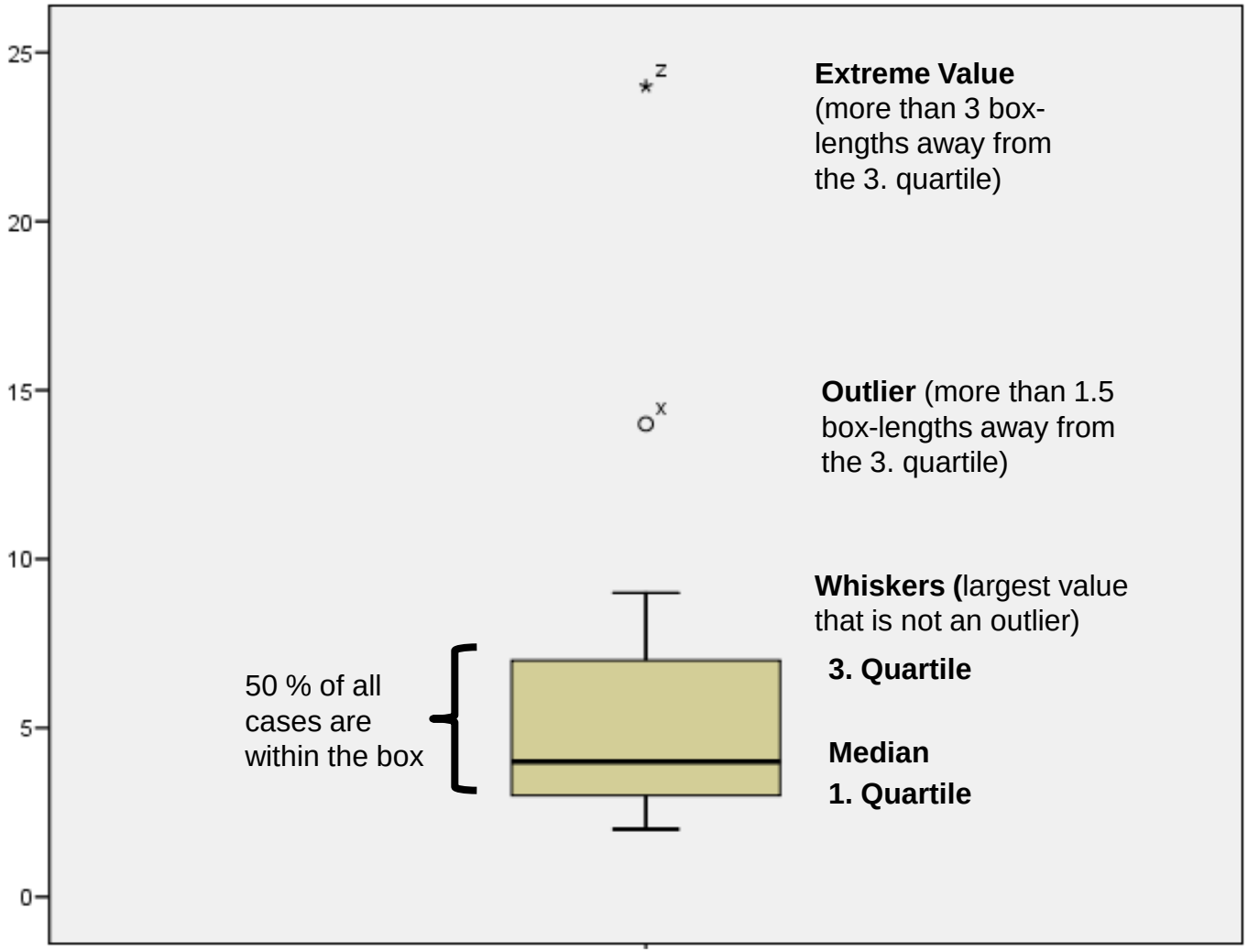
BFR.:NETTOEINKOMMEN<OFFENE+LISTENANGABE

N	Gültig	2644
	Fehlend	777
Mittelwert		1249,88
Median		1063,00
Modus		1000

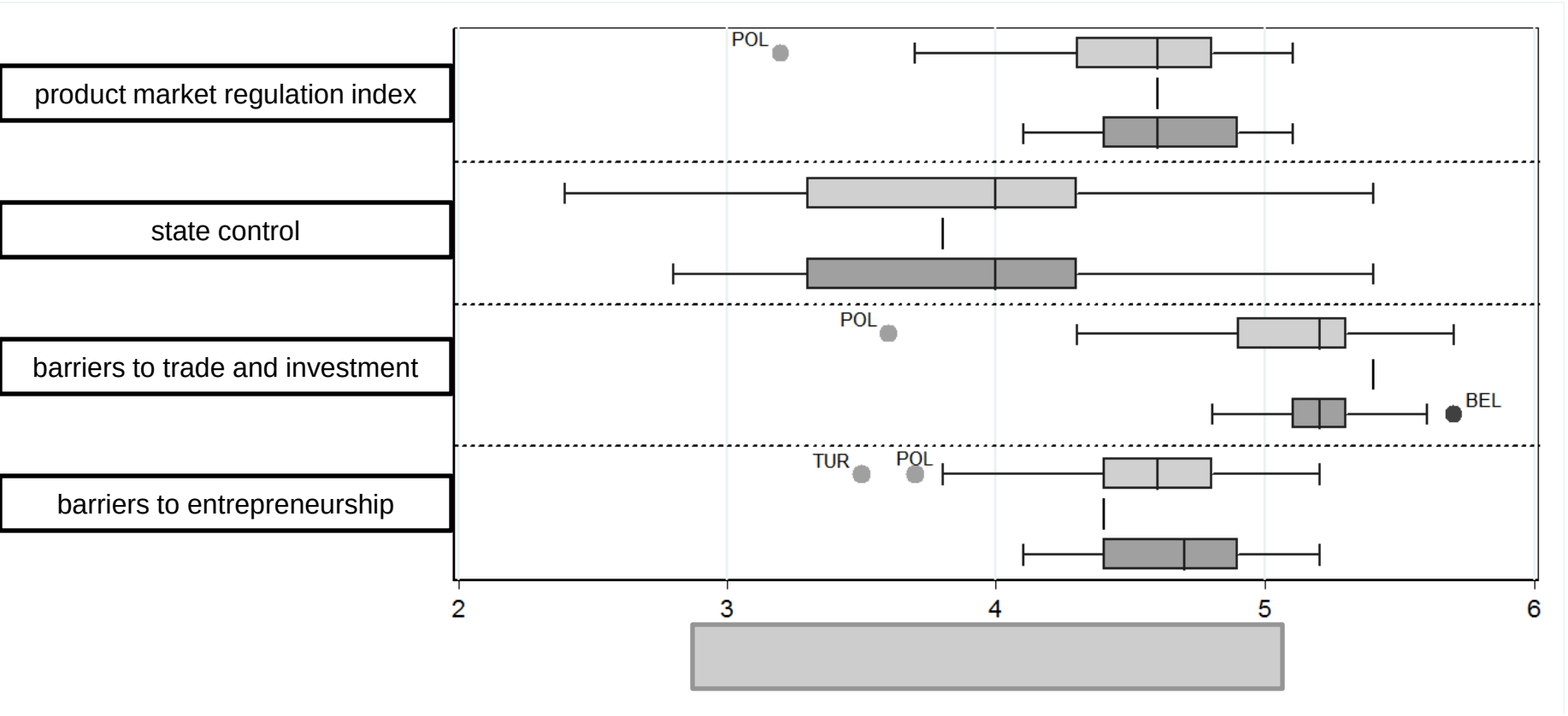
$$\bar{x} > \tilde{x} > x_{\text{mod}}$$



descriptive statistics – graphical display options for distributions **boxplot**



descriptive statistics – graphical display options for distributions **boxplot**



Own presentation based on the OECD-product regulation index from 2003. The blackline in between the two boxplots indicates the position of germany. The boxes include 50% of all cases, 25% are below, 25% are above. The line inside the box is the median. The whiskers indicate the last value that is no outlier/extreme value. Outliers are more than 1.5 box-length from the first or third quartile, extreme values more than three box length away.

measures of dispersion – **range**

Range

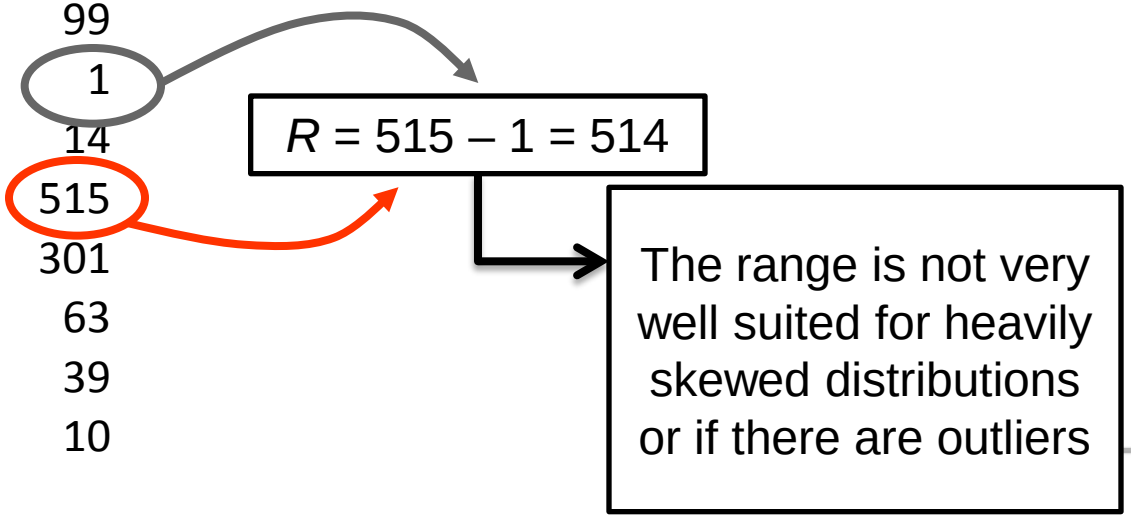
Difference between the largest and the smallest value of a given distribution

$$R = x_{i(\max)} - x_{i(\min)}$$

measures of dispersion – **range**

Per capita government expenditure on health at average exchange rate (US$)

Brunei Darussalam	413
Cambodia	7
China	31
India	7
Indonesia	12
Lao People's Democratic Republic	4
Malaysia	99
Myanmar	1
Philippines	14
Republic of Korea	515
Singapore	301
Thailand	63
Timor-Leste	39
Viet Nam	10



source: UNdata

measures of dispersion – **variance**

Variance:

Squared deviation from the arithmetic mean

$$S^2 = Q(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Problem: cannot be interpreted easily.

measures of dispersion – **standard deviation**

standard deviation:

square-root of the variance (for easier interpretations)

$$S = \sqrt{S^2}$$

rule of thumb:

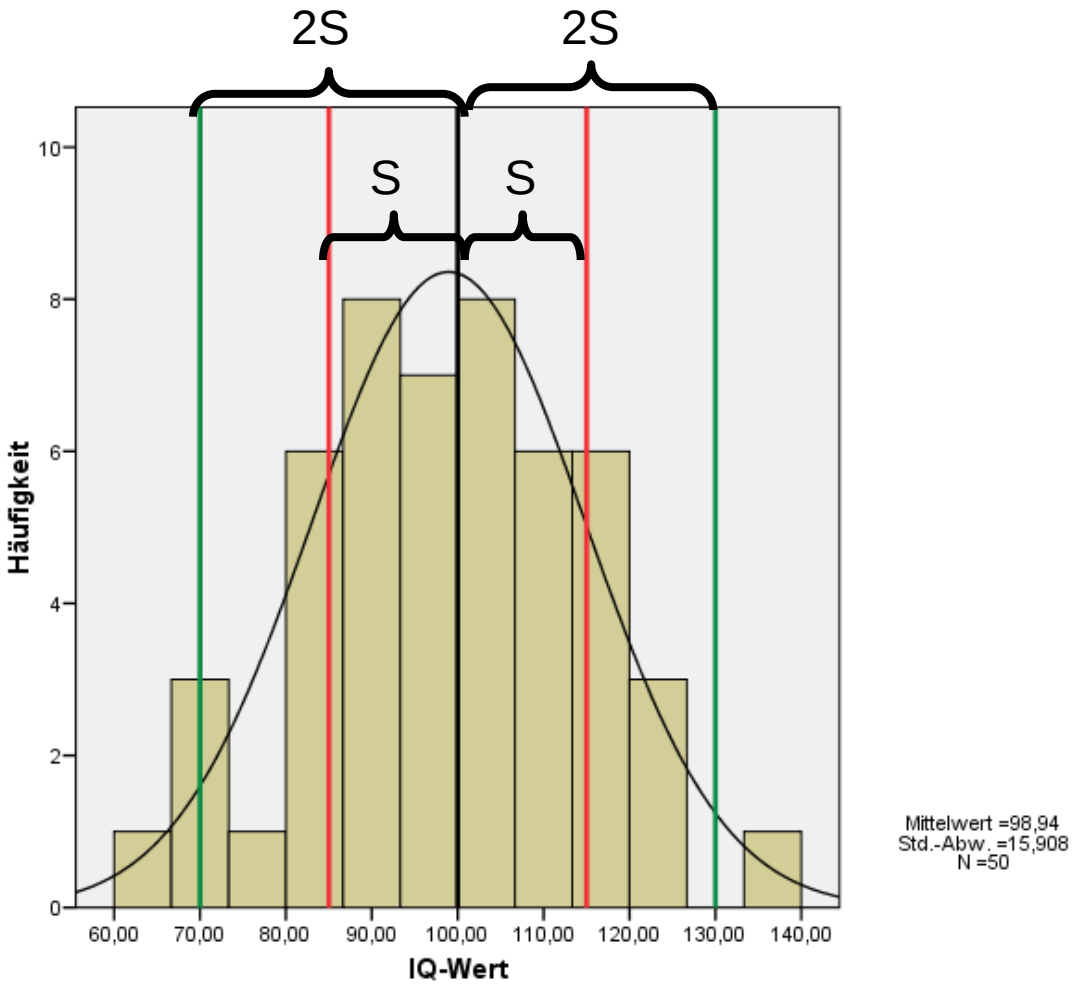
when the data is distributed normally, ca. 68% of all cases are in an inteval S, ca. 95% in an interval ± 2S and more than 99% in an interval ± 3S

measures of dispersion – **standard deviation**

Example:

Intelligence Quotient

The IQ is defined by the normal distribution of the sample under consideration. The value 100 is assigned to the mean of the sample with a standard deviation of 15.



measures of dispersion – **standard deviation**

Example:
government duration: Norway

government duration in days	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
901	901 – 816.7 = 84.3	7106.49
165	165 – 816.7 = -651.7	424712.89
1256	1256 – 816.7 = 439.3	192984.49
383	383 – 816.7 = -433.7	188095.69
1045	1045 – 816.7 = 228.3	52120.89
1137	1137 – 816.7 = 320.3	102592.09
355	355 – 816.7 = -461.7	213166.89
884	884 – 816.7 = 67.3	4529.29
582	582 – 816.7 = -234.7	55084.09
1459	1459 – 816.7 = 642.3	412549.29
Σ 8167	Σ 0	Σ 1652942.1

$$S^2 = \frac{1}{10} \cdot 1652942{,}1$$
$$= \underline{165294{,}21}$$

$$S = \sqrt{165294{,}21}$$
$$= \underline{406{,}56}$$

measures of association

measures of association

→ indicate the existence and the strength of a correlation between two variables

according to the level of measurement of the variables different measuerés of association can and have to be used:

- nominal scale:
  percentage difference d%, Phi Φ, Yules-Q, contingency coefficient C, Chi-sqare χ^2, Lambda λ

- ordinal scale:
  Spearmans Rho ρ, Goodman and Kruskals Gamma γ, Kendalls Tau-b and Tau-c

- Nominal in respect to interval scale:
  Eta-square η^2

- Interval scale:
  Pearsons r

measures of association (nominal scale)

Cross tabulations

Cross tabulations are the basis for the calculation of a number of different measures of association.

Variable 1	Variable 2		Sum
	B1	B2	
A1	a	b	∑ A1 (=a+b)
A2	c	d	∑ A2 (=c+d)
Sum	∑ B1 (=a+c)	∑ B2 (=b+d)	∑ ∑ (=a+b+c+d)

measures of association (nominal scale)
Phi coefficient

- Only possible for 2x2 tabulations (i.e. both variables are dichotomous)
- Two ways of calculating. Either:

$$\Phi = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

or using χ^2 (another measure which is not interpretable in its own right as a measure of association but forms the basis for a number of other measures)

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

Phi is **not** confined to a certain range of values and thus must be corrected to become interpretable/comparable.

measures of association (nominal scale)
Phi coefficient (adjustment alternative I)

Adjustment of the Phi coefficient is achieved by using the maximum/extreme Phi value:

$$\Phi_{kor} = \frac{\Phi}{\Phi_{extrem}}$$

Calculation of the extreme Phi value

The cell with the smallest value is set to zero and all other cells are changed in a way that keeps the marginal frequencies unchanged

→ The extreme Phi value can then be calculated using the same formula:

sex	employment status		sum
	employed	unemp loyed	
female	40	25	65
maile	80	5	85
sum	120	30	150

$$\Phi = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{40 \cdot 5 - 25 \cdot 80}{\sqrt{65 \cdot 85 \cdot 120 \cdot 30}} = -0{,}404.$$

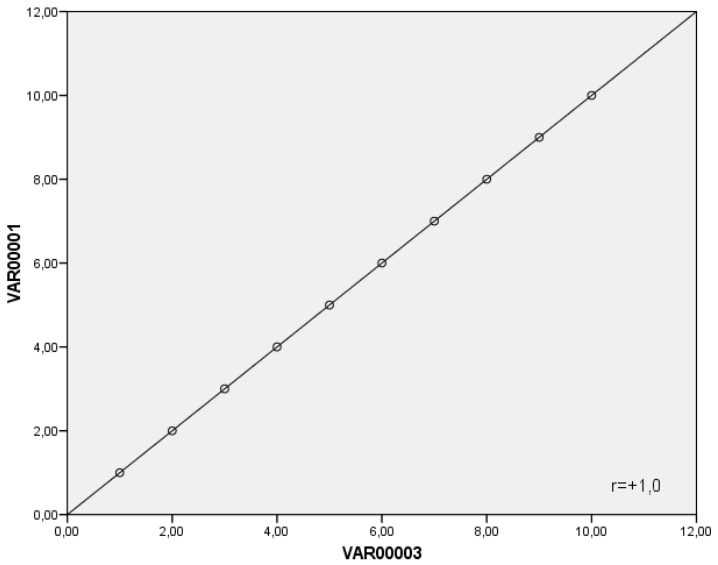
sex	employment status		sum
	employed	unempl oyed	
female	35	30	65
male	85	0	85
sum	120	30	150

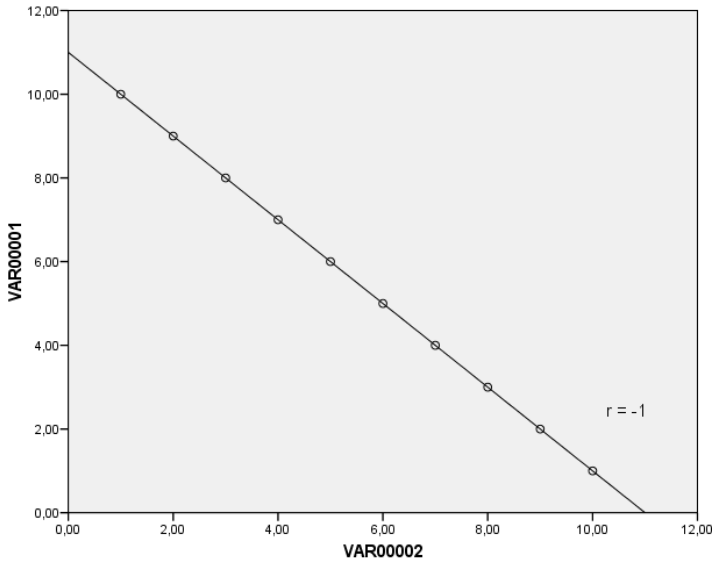
$$\Phi_{extrem} = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{35 \cdot 0 - 30 \cdot 85}{\sqrt{65 \cdot 85 \cdot 120 \cdot 30}} = -0{,}572.$$

$$\Phi_{kor} = \frac{\Phi}{\Phi_{extrem}} = \frac{-0{,}404}{-0{,}572} = 0{,}706.$$

measures of association (interval)
Pearson correlation coefficient r

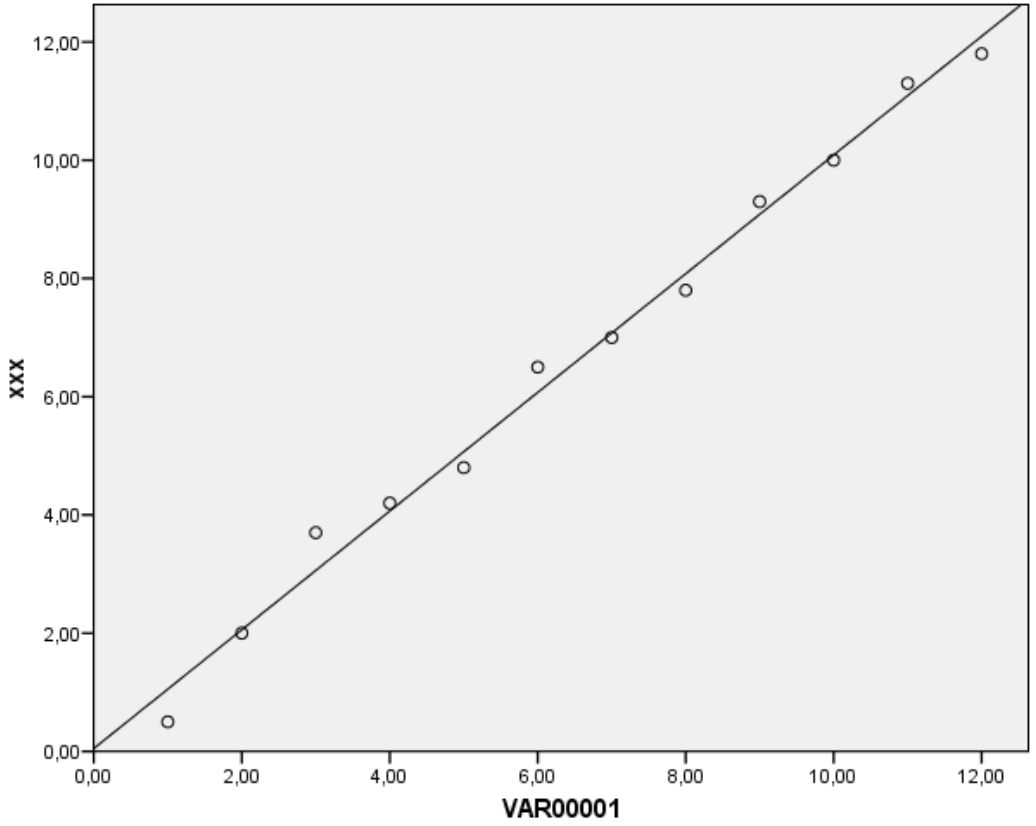
- Can be used for two metric variables
- Measures the linear correlation between two variables
- range: -1 (= perfect negative correlation) to +1 (perfect positive correlation)





Always inspect the scatter plot first!

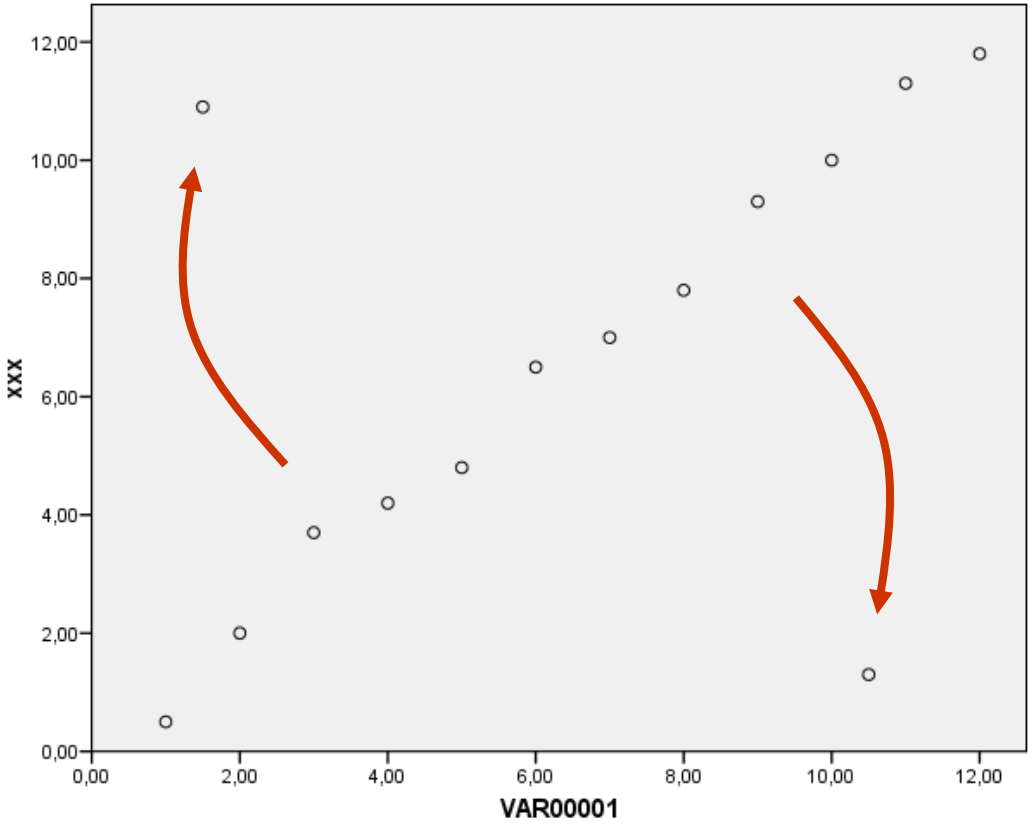
measures of association (interval)
Pearson correlation coefficient r



strong linear correlation

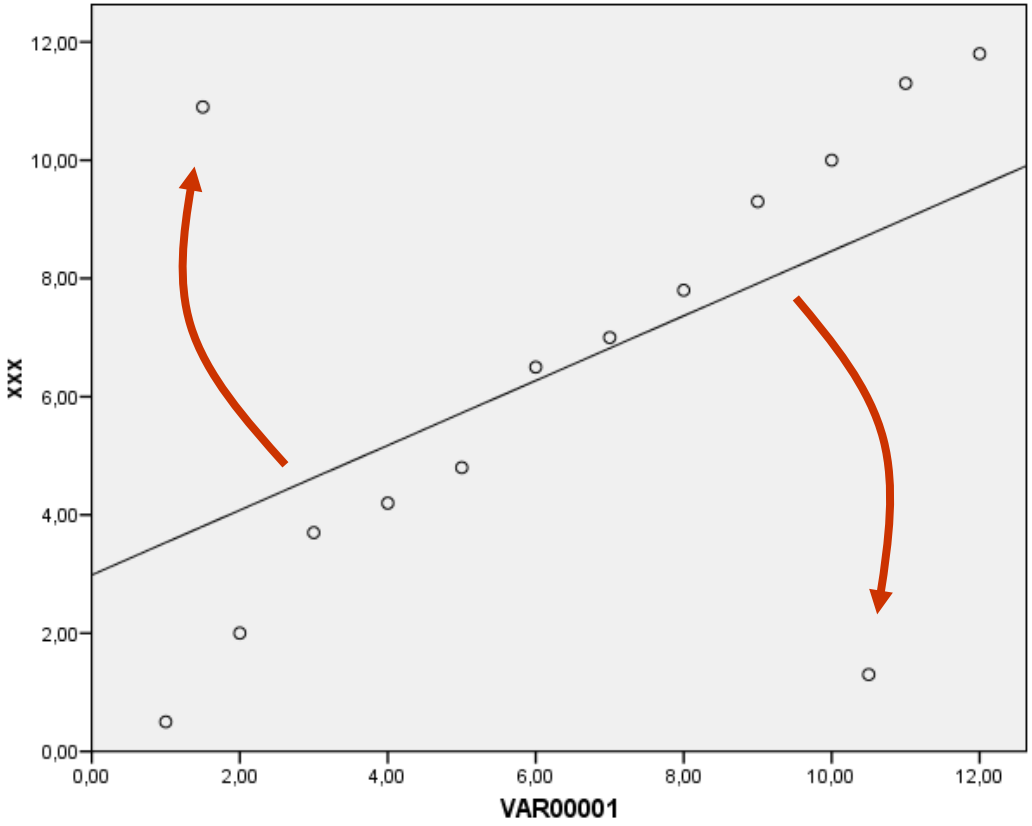
→ Pearsons r is meaningful here

measures of association (interval)
Pearson correlation coefficient r

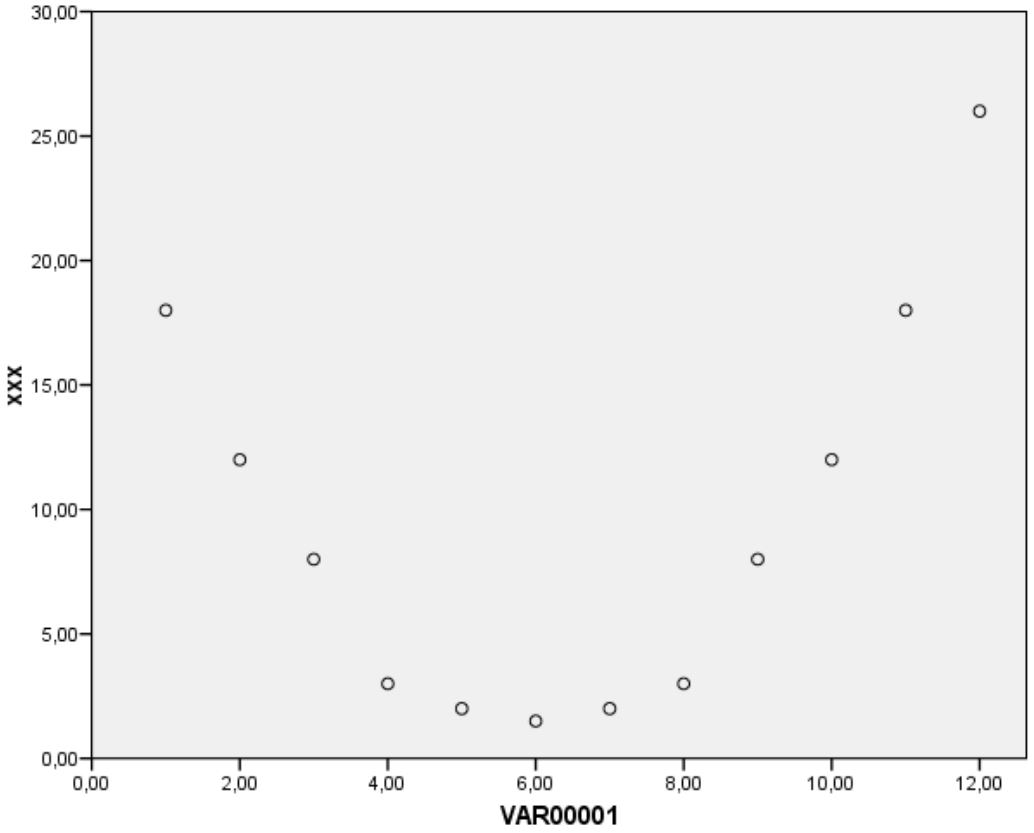


How does the inclusion of the two outliers change the linear correlation?

measures of association (interval)
Pearson correlation coefficient r

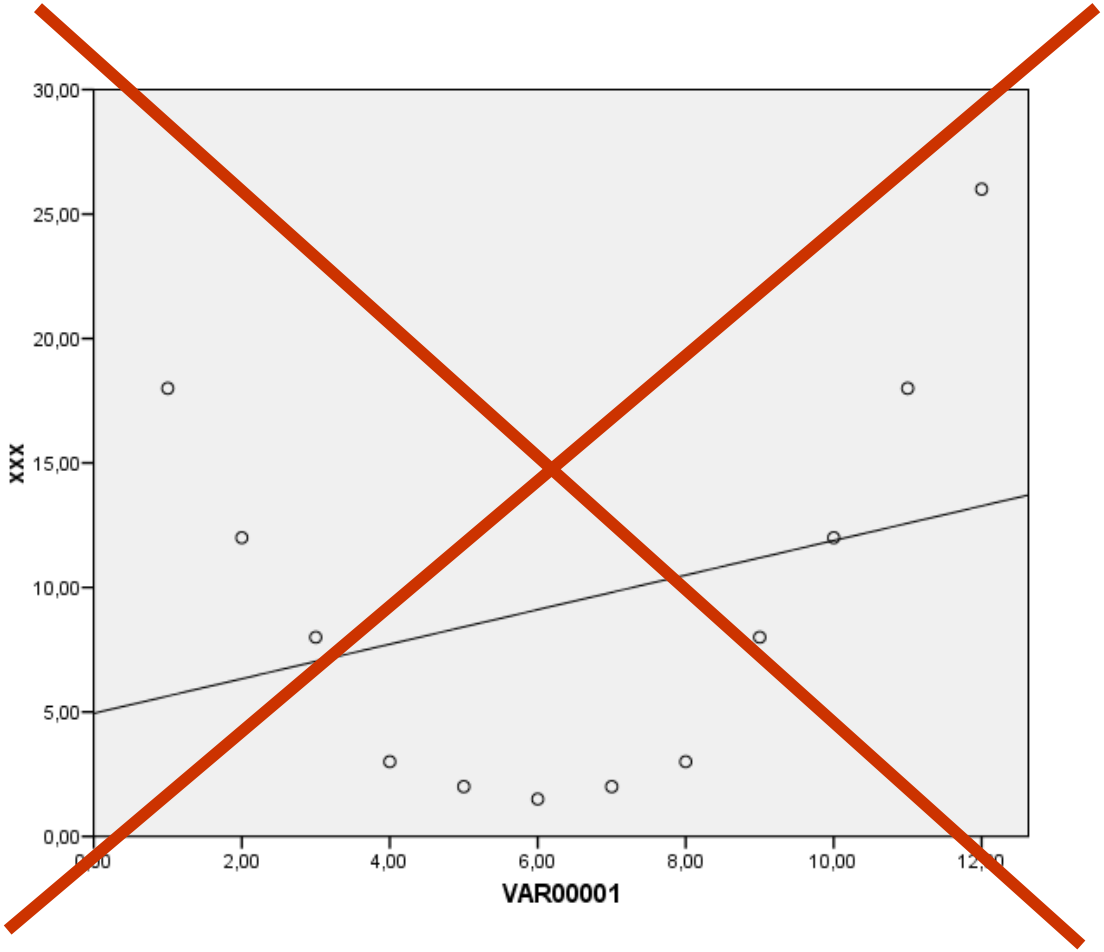


measures of association (interval)
Pearson correlation coefficient r



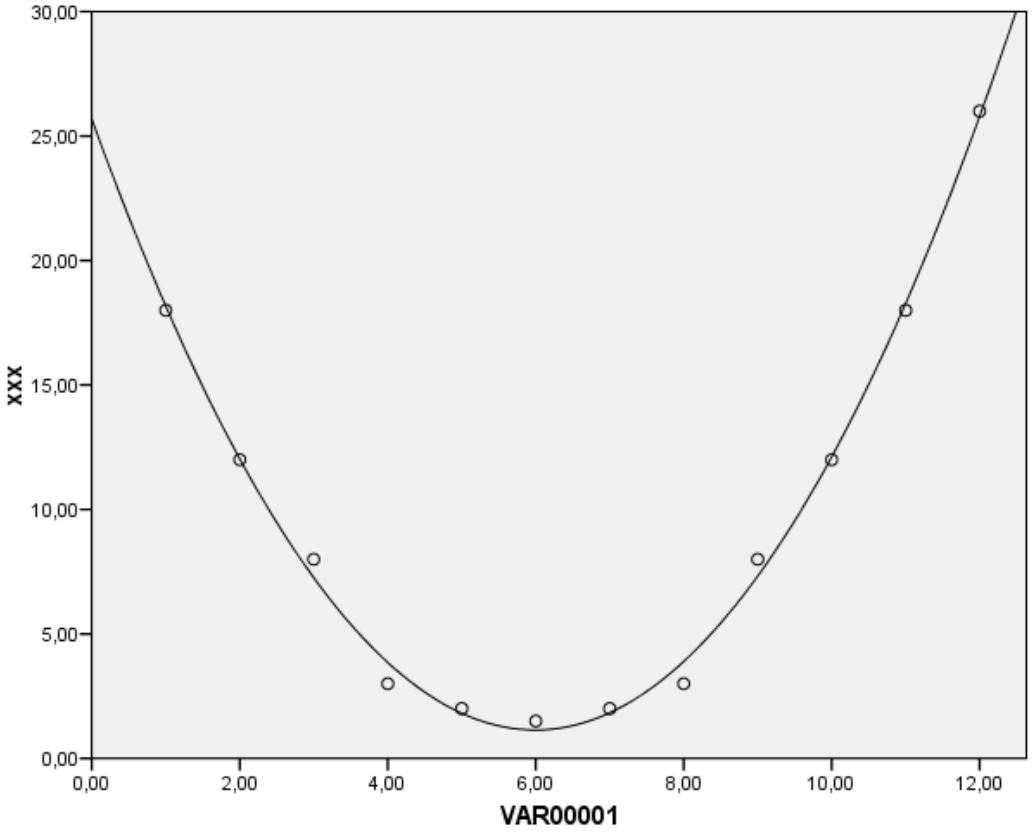
Is Pearsons r meaningful in this case as well?

measures of association (interval)
Pearson correlation coefficient r



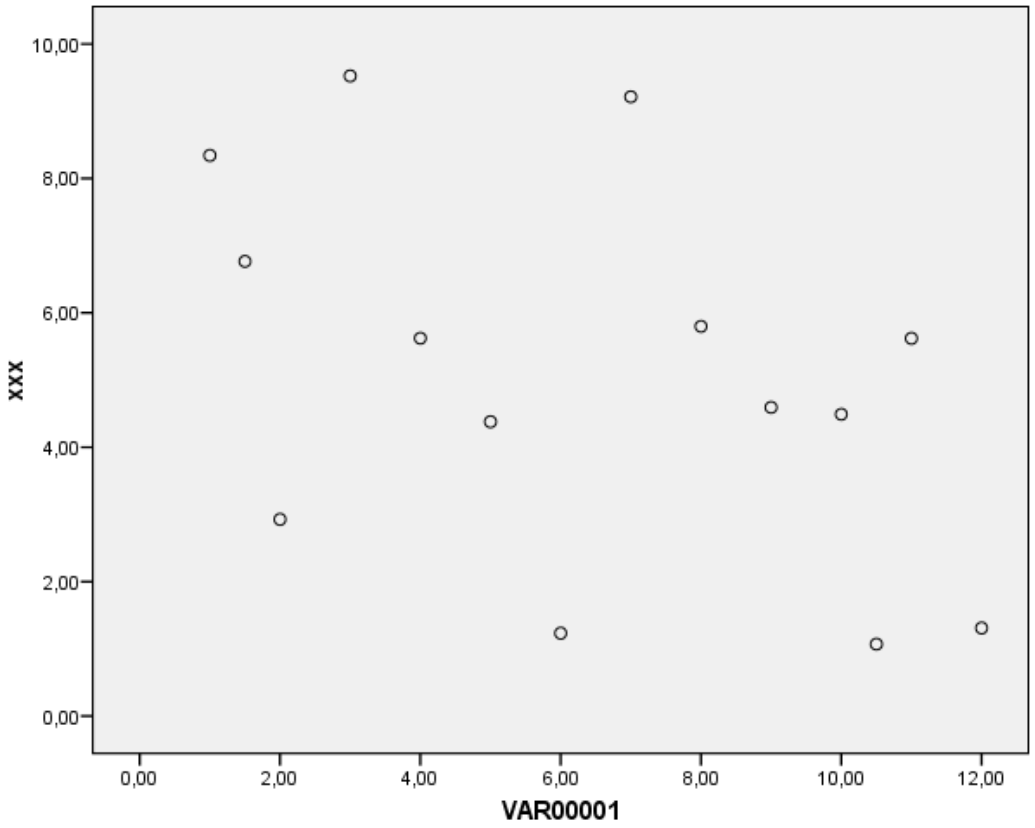
obviously there is no linear relationship → It does not make any sense to calculate Pearsons r as a measure of association

measures of association (interval)
Pearson correlation coefficient r



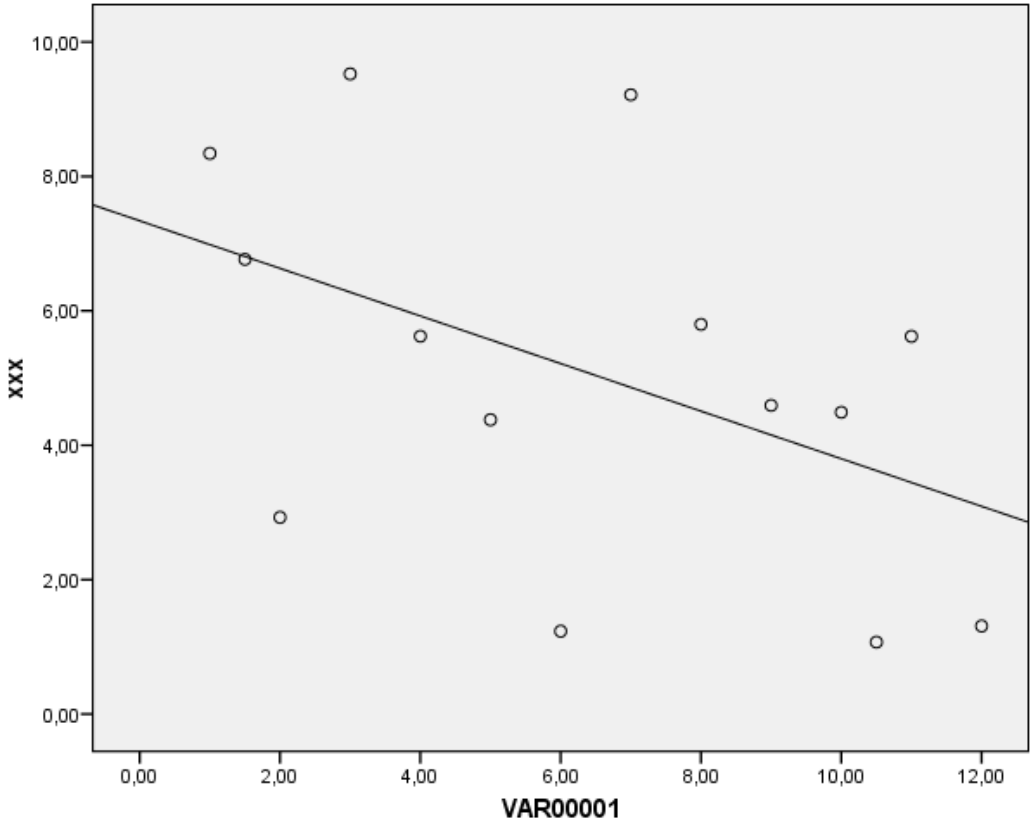
A quadratic or cubic function would be suited much better for this relationsship

measures of association (interval)
Pearson correlation coefficient r



What kind of relationship do we see here??

measures of association (interval)
Pearson correlation coefficient r

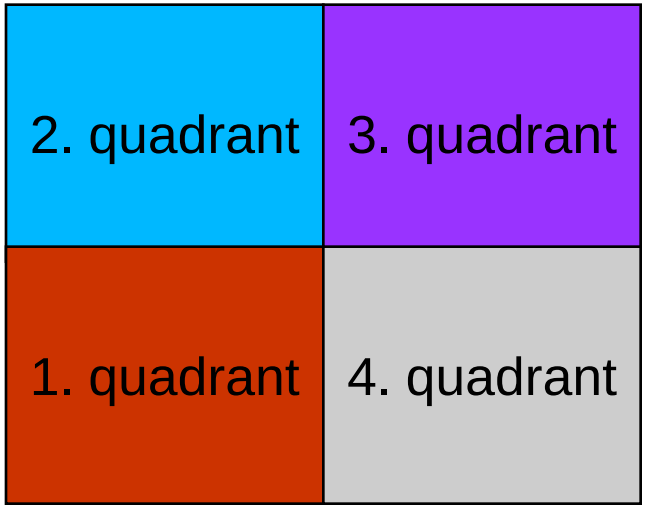


An inspection of the scatter plot reveals that there is, if any, only a very weak linear correlation between the two variables.

measures of association (interval)
Pearson correlation coefficient r

Calculation of Pearsons r

Concept of Covariance
→ at first graphical
The whole scatter plot is divided by the arithmetic means of the two variables into four quadrants.



1) When all the value-pairs are situated in the 1. and 3. quadrant
   → positive covariance
2) When al the value-pairs are situated in the 2. and 4. quadrant
   → negative covariance
3) When all the value-pairs are equally distributed among all four quadrants
   → no common covariance, i.e. the two variables are unconnected

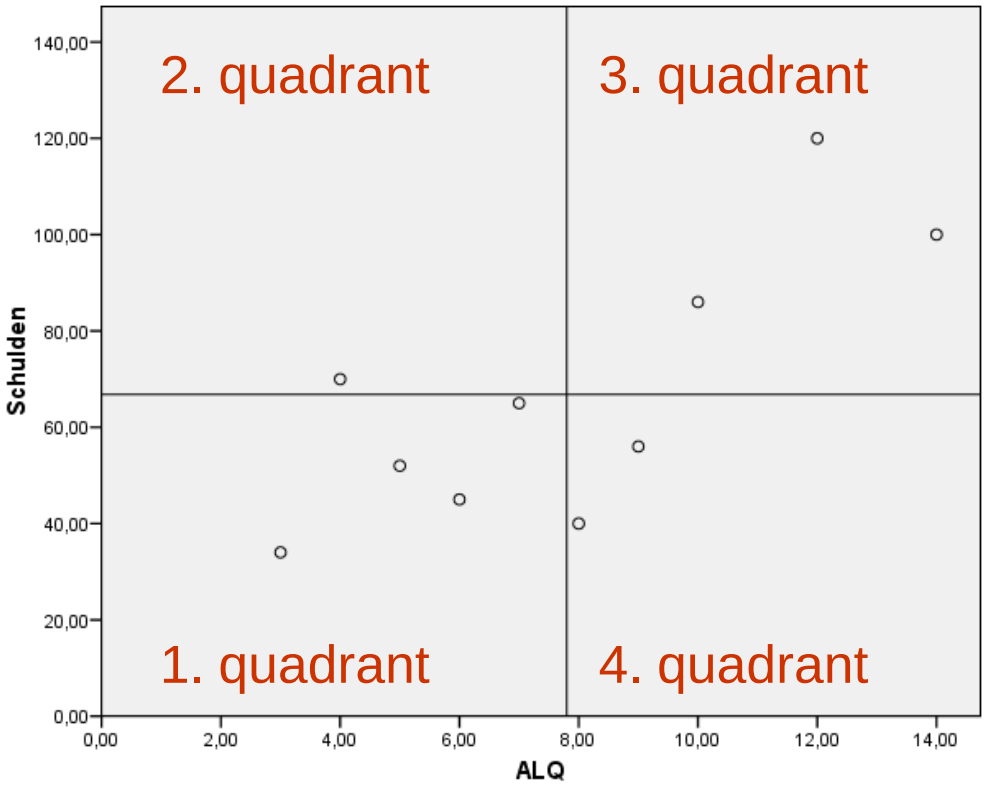
measures of association (interval)
Pearson correlation coefficient r

Example

Relationship between unemployment ratio and public debt ratio

Most value-pairs are situated in the 1. and 3. quadrant

→ there is a positive covariance



measures of association (interval)
Pearson correlation coefficient r

Calculating the covariance

$$COV_{(xy)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Pearsons r is defined as follows:

$$r = \frac{COV_{(xy)}}{S_x S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

measures of association (interval)
Pearson correlation coefficient r

country	unemploy x_i	debt y_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
A	12	120	4,2	17,64	53,2	2830,24	223,44
B	6	45	-1,8	3,24	-21,8	475,24	39,24
C	7	65	-0,8	0,64	-1,8	3,24	1,44
D	5	52	-2,8	7,84	-14,8	219,04	41,44
E	3	34	-4,8	23,04	-32,8	1075,84	157,44
F	10	86	2,2	4,84	19,2	368,64	42,24
G	4	70	-3,8	14,44	3,2	10,24	-12,16
H	8	40	0,2	0,04	-26,8	718,24	-5,36
I	9	56	1,2	1,44	-10,8	116,64	-12,96
J	14	100	6,2	38,44	33,2	1102,24	205,84
	Σ 78	Σ 668	Σ 0,0	Σ 111,60 S^2=11,16 S_x=3,34	Σ 0,0	Σ 6919,60 S^2=691,96 S_y=26,31	Σ680.60 COV=68,06

measures of association (interval)
Pearson correlation coefficient r

$$r = \frac{COV_{(xy)}}{S_x S_y} = \frac{68.06}{3.34 \cdot 26.31} = 0.774.$$

→Pearsons r indicates a relatively strong positive linear relationship between the unemployment rate and the public debt ratio.

BUT: -

-it‘s only a bivariate analysis and other factors could play a roll as well

-and: **correlation ≠ causality**

OLS regression – **basic concept**

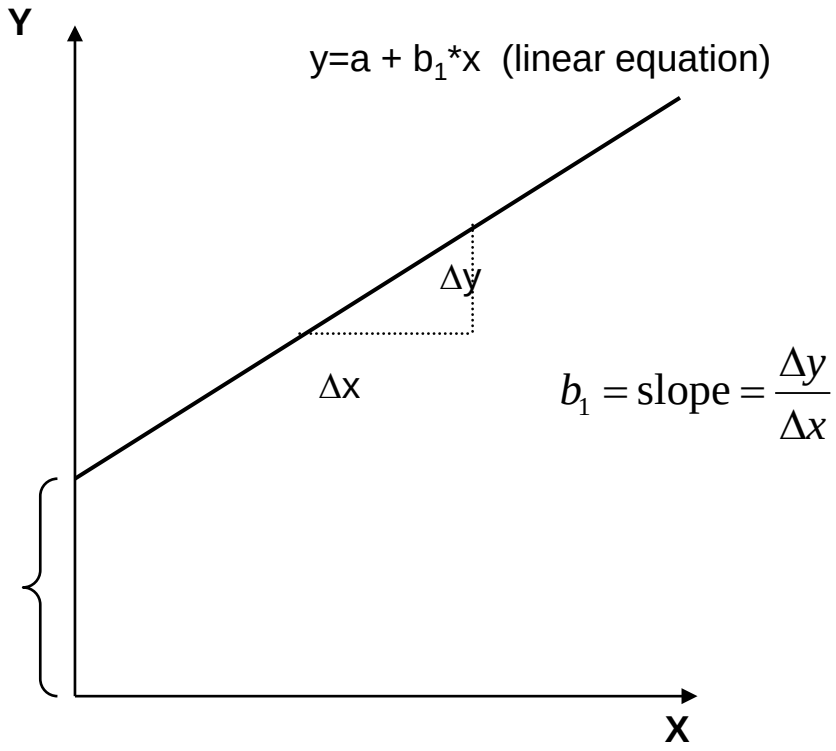
The bivariate regression is described by the following function

$$Y = a + b_1 \cdot X_1.$$

With a being the constant term (intercept) and b_1 being slope parameter

- Both the dependent and the independent variable have to be measured on a metric level.
- The assumption of a linear relationship is fundamental!

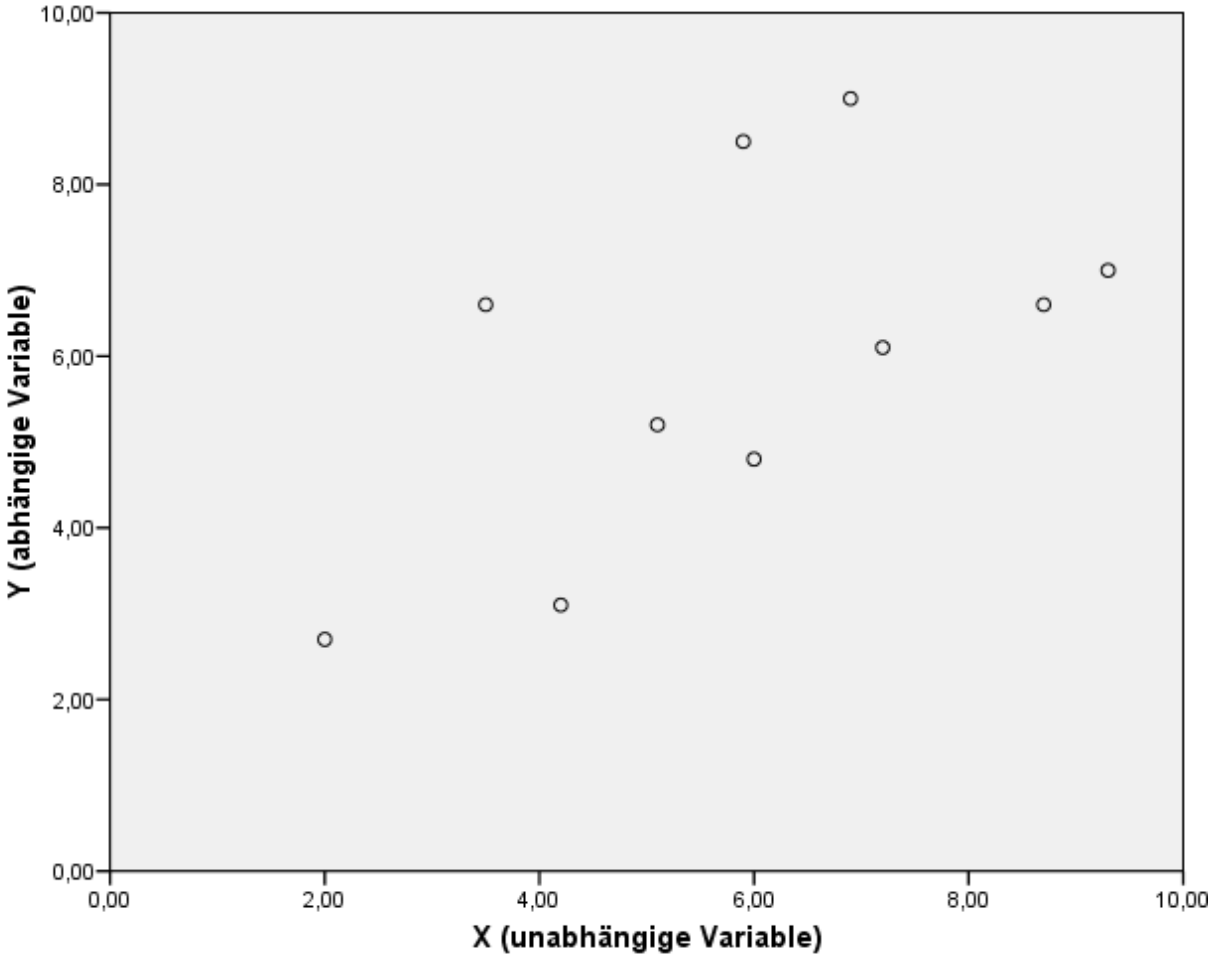
OLS regression – **regression equation**



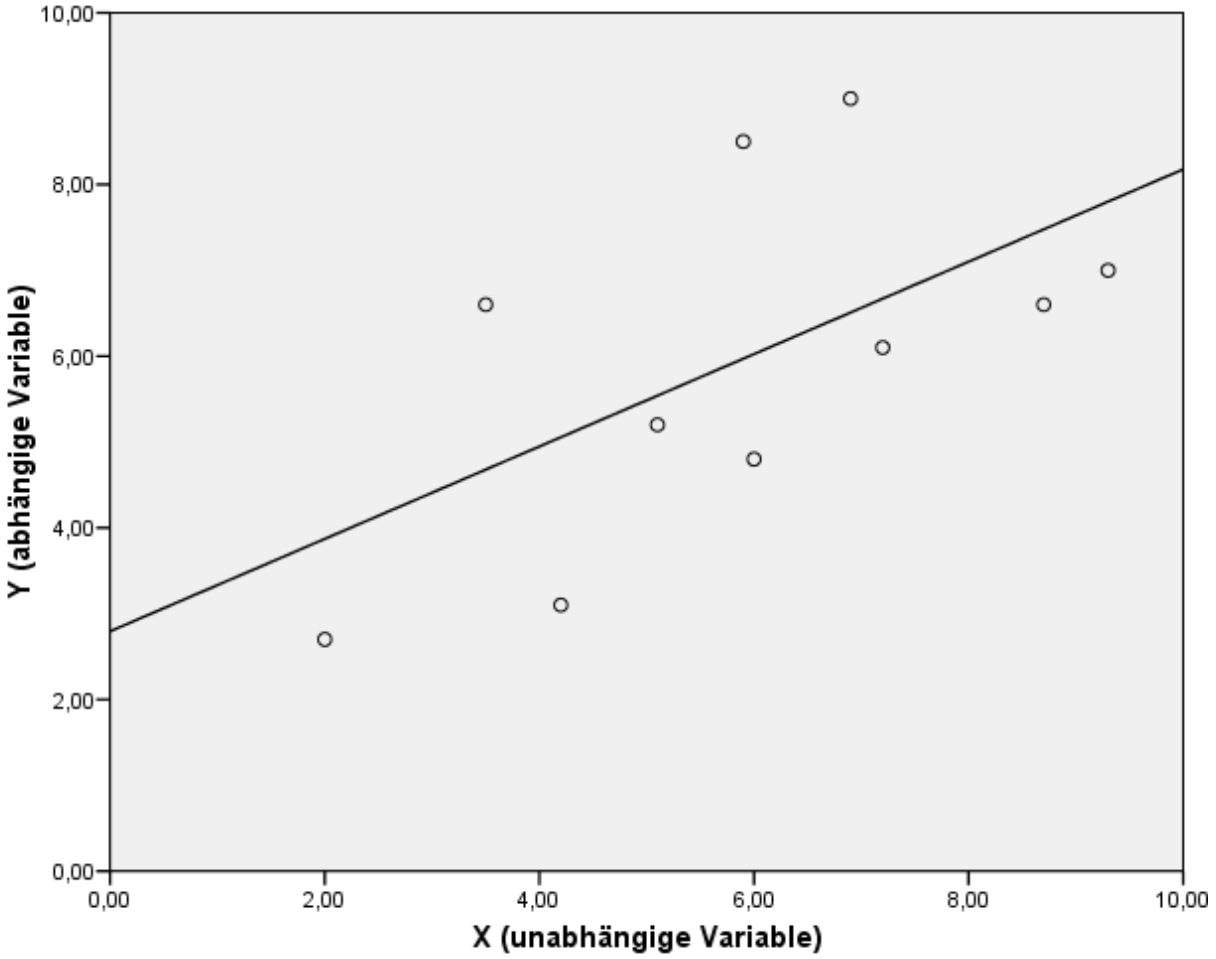
$$b_1 = \text{slope} = \frac{\Delta y}{\Delta x}$$

a = intercept

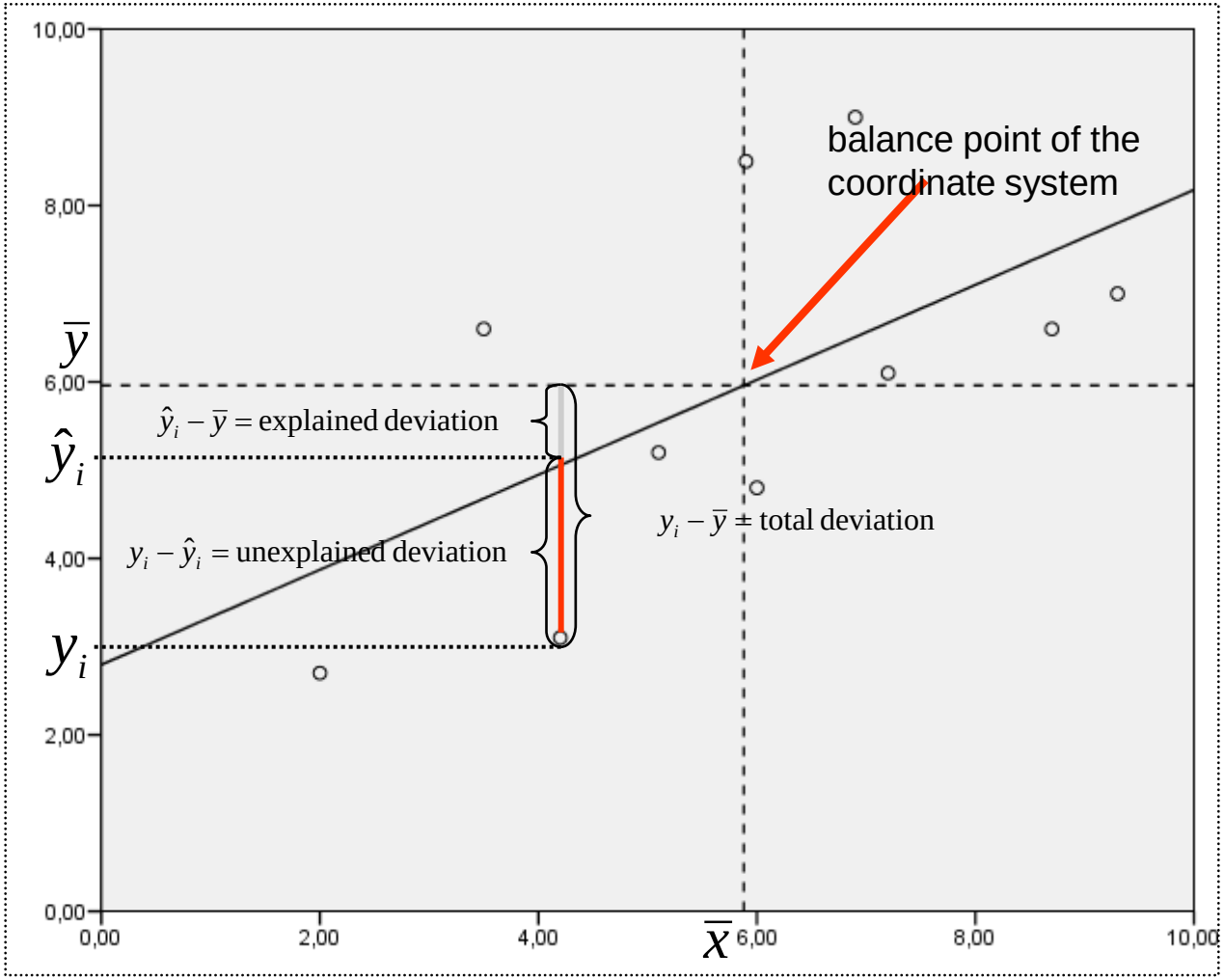
OLS regression – **regression equation**



OLS regression – **regression equation**



OLS regression – **regression equation**



OLS regression – **regression equation**

The goal of regression analysis is to minimize the unexplained deviations

→ Because the sum of the unexplained deviations is zero (points above and below the regression line cancel each other out) the sum of the squared unexplained deviations is minimized

(= **OLS regression; Ordinary Least Squares**)

- The unexplained deviations are called **residuals**

OLS regression – **regression equation**

Calculating the equation for the regression line

1) slope parameter

$$b_1 = \frac{Cov(xy)}{S_x^{\,2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

2) intercept

$$a = \bar{y} - b_1 \cdot \bar{x}.$$

OLS regression – **determining the regression parameters**

Example: Influence of the unemployment rate on public debt

country	unemploy x_i	debt y_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
A	12	120	4,2	17,64	53,2	2830,24	223,44
B	6	45	-1,8	3,24	-21,8	475,24	39,24
C	7	65	-0,8	0,64	-1,8	3,24	1,44
D	5	52	-2,8	7,84	-14,8	219,04	41,44
E	3	34	-4,8	23,04	-32,8	1075,84	157,44
F	10	86	2,2	4,84	19,2	368,64	42,24
G	4	70	-3,8	14,44	3,2	10,24	-12,16
H	8	40	0,2	0,04	-26,8	718,24	-5,36
I	9	56	1,2	1,44	-10,8	116,64	-12,96
J	14	100	6,2	38,44	33,2	1102,24	205,84
	Σ 78	Σ 668	Σ 0,0	Σ 111,60 S^2=11,16 S_x=3,34	Σ 0,0	Σ 6919,60 S^2=691,96 S_y=26,31	Σ680.60 COV=68,06

OLS regression – **determining the regression parameters**

$$b_1 = \frac{Cov(xy)}{S_x^{\ 2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{68{,}06}{11{,}16} = 6{,}099.$$

$$a = \bar{y} - b_1 \cdot \bar{x} = 66{,}8 - 6{,}099 \cdot 7{,}8 = 19{,}22.$$

→ regression function:

$$\text{public debt} = 19{,}22 + 6{,}1 \cdot \text{unemployment}$$

OLS regression – **coefficient of determination R^2**

- Measure that indicates the overall fit of the estimated regression function
- Perfect correlation, i.e. all points lie exactly on the regression line
  → $R^2 = 1$
- No linear correlation, i.e. the total deviations equal the unexplained deviations
  → $R^2 = 0$
- Multiplied with 100 R^2 can be interpreted as the percentage of variance that is explained by the regression equation

R^2 can be calculated using the Pearson correlation coefficient:

$R^2 = r^2$

Compare slide 39: $r = 0{,}774$ → $R^2 = 0{,}60$.

→ The unemployment rate thus explains 60% of the variance of the public debt ratio.

Exercise – regression by hand

OLS regression –
Interpreting the SPSS output

Modellzusammenfassung[b]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Durbin-Watson-Statistik
1	,774[a]	,600	,550	18,604	1,969

a. Einflußvariablen : (Konstante), ALQ

b. Abhängige Variable: DEBT

R = Pearsons r(correlation coefficient)

R^2 = coefficient of determination

OLS regression –
Interpreting the SPSS output

Modellzusammenfassung[b]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Durbin-Watson-Statistik
1	,774[a]	,600	,550	18,604	1,969

a. Einflußvariablen : (Konstante), ALQ

b. Abhängige Variable: DEBT

Adjusted coefficient of determination R^2_{adj} :

-Adjusts on the number of independent variables, because every additional explaining variable within a multivariate regression increases the R^2.

- Calculation with this formula: n = number of cases

$$R_{adj}^{\ 2} = R^2 - \frac{p-1}{n-p} \cdot (1 - R^2).$$

p = number of regressors (=independent variables) + intersept

OLS regression –
Interpreting the SPSS output

Modellzusammenfassung[b]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Durbin-Watson-Statistik
1	,774[a]	,600	,550	18,604	1,969

a. Einflußvariablen : (Konstante), ALQ

b. Abhängige Variable: DEBT

Standard error of the estimate:

With normally distributed data, 95% of all cases fall into an interval of ±2 standard deviations → Construction of the confidence intervals for the estimation

Calculation of the standard error for the estimation: $\hat{y} \pm 2 \cdot s_e$

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}} = \sqrt{\frac{2768{,}92}{8}} = 18{,}60.$$

Country A has an estimated value of 92.41 → if it was a random sample 95% of the cases from the basic population would be in the interval [43.01; 117.42]

OLS regression –
Interpreting the SPSS output

Modellzusammenfassung[b]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Durbin-Watson-Statistik
1	,774[a]	,600	,550	18,604	1,969

a. Einflußvariablen : (Konstante), ALQ

b. Abhängige Variable: DEBT

Durbin-Watson tests for autocorrelation(= correlation with values from the previous period) → not meaningful/necessary for cross sectional analyses

$0 < d < d_l$: positive autocorrelation

$d_l < d < d_u$: value is in the domain of uncertainty

$d_u < d < 4-d_u$ no autocorrelation

$4-d_u < d < 4d_l$: value is in the domain of uncertainty

$4-d_l < d < 4$: negative autocorrelation

→ The calculated empirical Durbin-Watson value d has to be compared with a theoretical value from a Durbin-Watson-table

OLS regression – **Interpreting the SPSS output**

Lower limit (d_l) and upper limit (d_u) of the critical values of the Durbin-Watson-test;

Level of significance ($\alpha = .05$).

T number of observations, and
K is the number of explaining variables

In our example:
→ T=10, K =1

There is no autocorrelation when d is in the interval [1.32; 2.68]

d = 1.98
→ **no autocorrelation**

	$K=1$		$K=2$		$K=3$		$K=4$		$K=5$	
T	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u
10	0.88	1.32	0.70	1.64	0.52	2.02	0.38	2.41	0.24	2.82
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20										
21										
22										
23										
24										
25										
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81

Rule of thumb:
The closer the empirical Durbin-Watson value comes to 2, autocorrelation becomes less likely
If it is close to 0 or 4, autocorrelation is very likely

OLS regression –
Interpreting the SPSS output

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	4150,684	1	4150,684	11,992	,009[a]
	Residuen	2768,916	8	346,114		
	Gesamt	6919,600	9			

a. Einflußvariablen : (Konstante), ALQ

b. Abhängige Variable: DEBT

Explained Sum of Squares (= explained variation)

Unexplained Sum of Squares (= unexplained variation)

Total Sum of Squares (= total variation)

OLS regression –
Interpreting the SPSS output

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	4150,684	1	4150,684	11,992	,009[a]
	Residuen	2768,916	8	346,114		
	Gesamt	6919,600	9			

a. Einflußvariablen : (Konstante), ALQ

b. Abhängige Variable: DEBT

df = degrees of freedom)

- One degree of freedom for the regression equation, one for the independent variable and 8 for the residuals

- More degrees of freedom (with the residuals) „lead“ to a more stable the regression function

→ Rule of thumb: There should be at least 10 degrees of freedom left for the residuals.

OLS regression – **Interpreting the SPSS output**

What are degrees of freedom?

Example

n=3

Arithmetic mean = 3

$x_1 = 2, x_2 = 3$ → $x_3 = 4.$

$x_1 = 1, x_2 = 10$ → $x_3 = -2$

→ In this example we have two degrees of freedom because we can always choose two elements freely but then have to adjust the third element, so that the arithmetic mean is again 3.

OLS regression –
Interpreting the SPSS output

ANOVA[b]

Modell		Quadrats umme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	4150,684	1	4150,684	11,992	,009[a]
	Residuen	2768,916	8	346,114		
	Gesamt	6919,600	9			

a. Einflußvariablen : (Konstante), ALQ

b. Abhängige Variable: DEBT

Mean square: (= Sum of squares/df)

- not relevant for direct interpretation

-However the mean square can be used to calculate another important test statistic, the F-value

$$F = \frac{\text{mean square of the regression}}{\text{mean square of the residuals}} = \frac{4150.68}{346.11} = 11.99.$$

OLS regression –
Interpreting the SPSS output

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	4150,684	1	4150,684	11,992	,009[a]
	Residuen	2768,916	8	346,114		
	Gesamt	6919,600	9			

a. Einflußvariablen : (Konstante), ALQ

b. Abhängige Variable: DEBT

F-statistics:

- Tests the significance of the overall regression, i.e. the significance of the coefficient of determination

- When the level of significance is lower than .05 a linear relationship can be assumed

OLS regression – **Interpreting the SPSS output**

General procedure for the F-test:

1) Put forward the null-hypothesis → no relationship between independent and dependent variable
2) Setting the level of significance (probability that a rejection of the null-hypothesis is not happening at random); generally 95% or 99%
3) Calculation of the empirical F-value:

$$F = \frac{R^2(n-2)}{1-R^2} = \frac{0.600 \cdot (10-2)}{1-0.600} = 12.00.$$

4) Comparison between the empirical F-value and a theoretical one from a F-table

OLS regression –
Interpreting the SPSS output

Für 99% Sicherheit:

degrees of freedom for the numerator

	f_1	1	2	3	4	5	6	7	8	9	10
f_2											
1		4052	4999	5403	5625	5764	5859	5929	5981	6023	6056
2		98, 49	99, 00	99, 17	99, 25	99, 30	99, 33	99, 35	99, 36	99, 38	99, 40
3		34, 12	30, 81	29, 46	28, 71	28, 24	27, 91	27, 67	27, 49	27, 34	27, 23
4		21, 20	18, 00	16, 69	15, 98	15, 52	15, 21	14, 98	14, 80	14, 66	14, 54
5		16, 26	13, 27	12, 06	11, 39	10, 97	10, 67	10, 44	10, 27	10, 14	10, 04
6		13, 47	10, 92	9, 78	9, 15	8, 75	8, 47	8, 26	8, 10	7, 98	7, 87
7		12, 25	9, 55	8, 45	7, 85	7, 46	7, 19	6, 99	6, 84	6, 72	6, 62
8		11, 26	8, 65	7, 59	7, 01	6, 63	6, 37	6, 18	6, 03	5, 91	5, 81
9		10, 56	8, 02	6, 99	6, 42	6, 06	5, 80	5, 61	5, 47	5, 35	5, 26
10		10, 04	7, 56	6, 55							4, 85
11		9, 65	7, 21	6, 22							4, 54
12		9, 33	6, 93	5, 95							4, 30
13		9, 07	6, 70	5, 74							4, 10
14		8, 86	6, 51	5, 56							4, 94
15		8, 68	6, 36	5, 42	4, 89	4, 56	4, 32	4, 11	4, 00	3, 89	3, 80

degrees of freedom for the denominator

$F_{emp} = 12.00 > F_{theo} = 11.26$

→ null-hypothesis can be rejected on the 99% confidence probability level

OLS regression –
Interpreting the SPSS output

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	4150,684	1	4150,684	11,992	,009[a]
	Residuen	2768,916	8	346,114		
	Gesamt	6919,600	9			

a. Einflußvariablen : (Konstante), ALQ

b. Abhängige Variable: DEBT

SPSS calculates the exact level of significance

→The probability that the null-hypothesis has not been wrongly rejected is thus:

P = 1 – 0.009 = 0,991.

→ confidence probability = 99%

→ Null-hypothesis has to be rejected and the regression equation is significant!

OLS regression –
Interpreting the SPSS output

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	19,231	14,943		1,287	,234
	ALQ	6,099	1,761	,774	3,463	,009

a. Abhängige Variable: DEBT

intercept

slope parameter

standard deviation of the regression coefficient

$$S_b = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot (n-2)}} = \sqrt{\frac{2768.92}{111.6 \cdot 8}} = 1.761.$$

OLS regression –
Interpreting the SPSS output

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	T	Signifikanz
1	(Konstante)	19,231	14,943		1,287	,234
	ALQ	6,099	1,761	,774	3,463	,009

a. Abhängige Variable: DEBT

The betas are standardized regressioon coefficients. They are the value of regression coefficients when the regression is calculated with z-transformed values.

The size of the betas indicates the importance of an independent variable in comparison with another independent one.

→ Of only one independent variable exists, like in this case, the standardization with betas is not necessary!

OLS regression –
Interpreting the SPSS output

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	T	Signifikanz
1	(Konstante)	19,231	14,943		1,287	,234
	ALQ	6,099	1,761	,774	3,463	,009

a. Abhängige Variable: DEBT

The t-test tests whether the slope parameter is significantly different from zero.

T is calculated by dividing the regression coefficient by the standard error of the coefficient

Comparison of the calculated empirical t-value with a theoretical one out of the student-t-distribution (e.g. in Wagschal: 369)

SPSS calculates the exact level of significance as well = 0,09 → The slope parameter is higly significant

OLS regression – **the problem of multicollinearity**

- multicollinearity = high correlation among the independent variables (of course only possible with multivariate regressions)
- there is perfect multicollinearity when one of the independent variables is completely correlated with another one ($r = \pm 1$)

→the regression equation can not be estimated in this case

OLS regression – **the problem of multicollinearity**

Effects of multicollinearity:
-Statistical significance of the regression coefficients is reduced
-The t-statistics is reduced, because the standard deviations of the regression coefficients increases
-The estimated OLS regression coefficients still provide the best and most efficient estimators
-The prediction is not influenced by the coefficients
-The explanatory power of the models (and the explaining variables) is reduced

OLS regression – **the problem of multicollinearity**

Possibilities for the identification of multicollinearity

1) correlation matrix of the independent variables

2) Klein-test:
   Calculate a regression for every independent variable with all other independent variables as predictors in the regression equatíon

→ Tolerance value = 1- the so obtained coefficient of determination

→ Variance inflation factor (VIF) = reciprocal of the tolerance value

SPSS reports the tolerance as well as the VIF

→ There is multicollinearity if the correlation among the independet variables is high, the tolerance value is near zero or the VIF is clearly larger than one.

OLS regression – **the problem of multicollinearity**

Example for multicollinearity:

Predicting the GDP with internet connections, fax-machines and higher education.

First possibility

→ correlation matrix

Korrelationen

		internet connections per 10,000 people, 1997	fax machines per 1000 people (1996 or 1997)	higher education enrollment % total enrollment per age group, 1995,1996, 1997
internet connections per 10,000 people, 1997	Korrelation nach Pearson	1	,629**	,656**
	Signifikanz (2-seitig)		,000	,000
	N	174	134	149
fax machines per 1000 people (1996 or 1997)	Korrelation nach Pearson	,629**	1	,621**
	Signifikanz (2-seitig)	,000		,000
	N	134	139	115
higher education enrollment % total enrollment per age group, 1995,1996, 1997	Korrelation nach Pearson	,656**	,621**	1
	Signifikanz (2-seitig)	,000	,000	
	N	149	115	151

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

What is the treshold of correlation strenght for a problematic multicollinearity?

→As a rule of thumb from .7 onwards it can be problematic

→In this case there is probably no problem with multicollinearity

OLS regression – **the problem of multicollinearity**

Second possibility:

Klein-test

Regression of the independent variables on each other

1- R^2 = Varianz

1/Varianz = VIF

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,731[a]	,534	,525	14,00938

a. Einflußvariablen : (Konstante), fax machines per 1000 people (1996 or 1997), internet connections per 10,000 people, 1997

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	15,239	1,497		10,178	,000
	internet connections per 10,000 people, 1997	,092	,016	,499	5,929	,000
	fax machines per 1000 people (1996 or 1997)	,323	,090	,304	3,605	,000

a. Abhängige Variable: higher education enrollment % total enrollment per age group, 1995,1996, 1997

1 - 0,534 = 0,466

1/ 0,466 = 2,146

OLS regression – **the problem of multicollinearity**

Klein-test in SPSS:

analyse→ regression → linear → statistics → collinearity diagnosis

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	Kollinearitätsstatistik	
		B	Standardfehler	Beta			Toleranz	VIF
1	(Konstante)	1724,620	526,793		3,274	,001		
	fax machines per 1000 people (1996 or 1997)	197,262	23,077	,514	8,548	,000	,537	1,861
	internet connections per 10,000 people, 1997	1,689	4,362	,025	,387	,699	,453	2,208
	higher education enrollment % total enrollment per age group, 1995,1996, 1997	166,810	23,577	,457	7,075	,000	,466	2,145

a. Abhängige Variable: gdp per capita, 1995,1996, or 1997

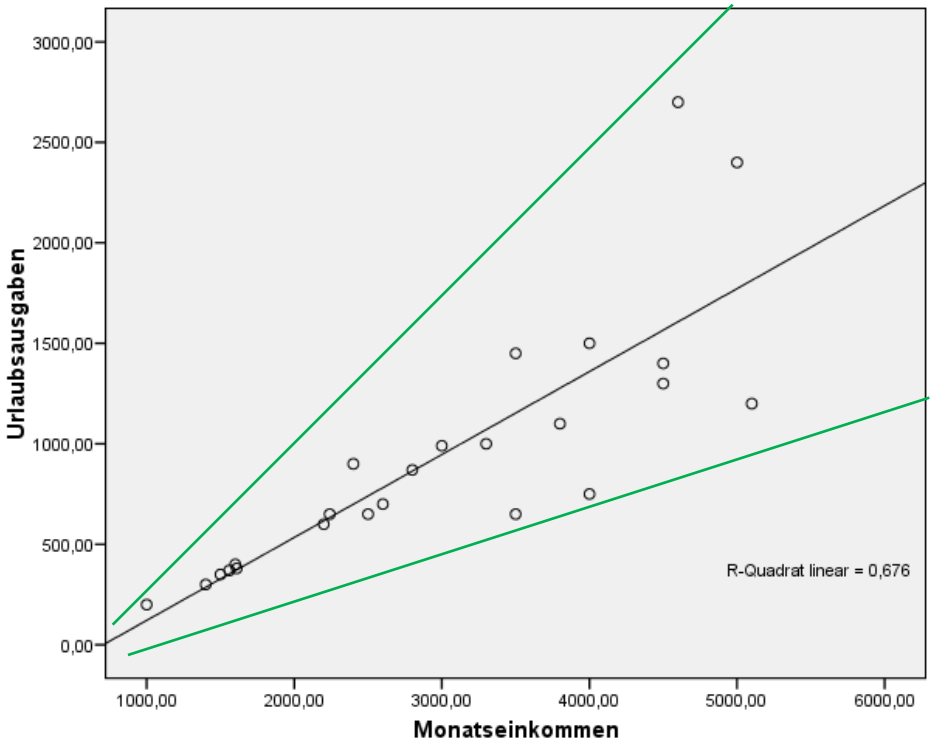
Tolerance as well as VIF do not indicate a notable level of multicollinearity

OLS regression – **the problem of heteroscedasticity**

Heteroscedasticity = heterogeneity of variance

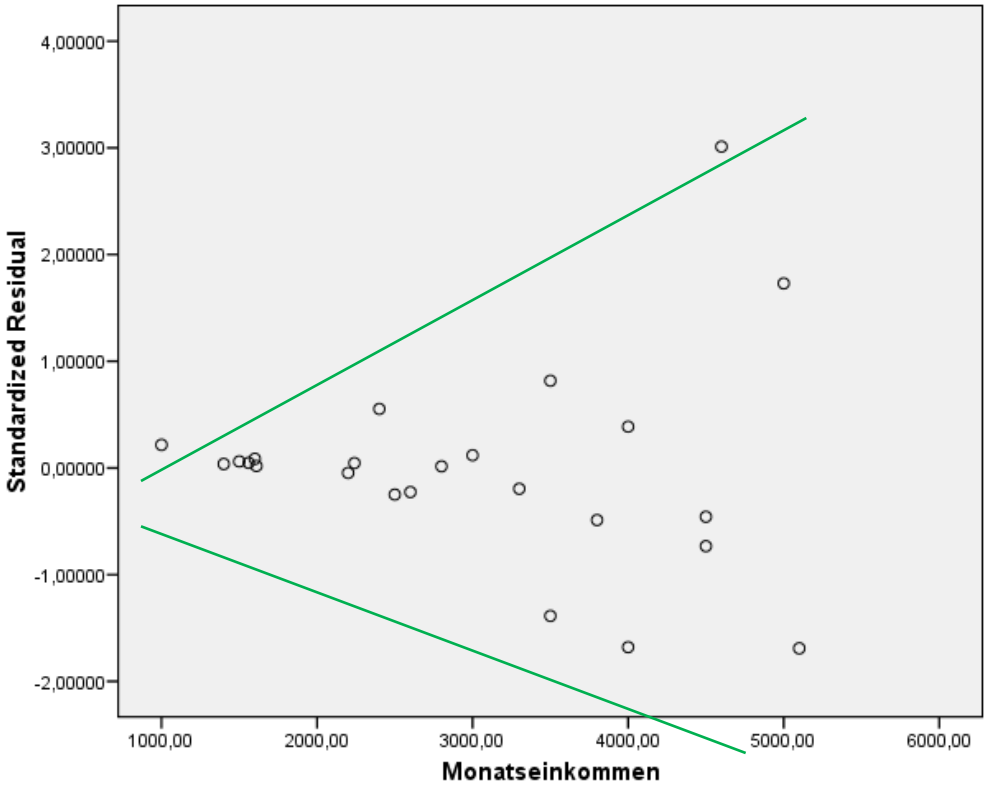
→ typically a wedge-shaped variation of the residuals

→ Identification via scatter plot (dependent vs. independent variable)



OLS regression – **the problem of heteroscedasticity**

Or even better: Scatter-plot of the residuals (y-axes) and the independent variable (x-axes)



OLS regression – **the problem of heteroscedasticity**

Effects of heteroscedasticity:

- Heteroscedasticity is a violation of the basic assumptions of the linear regression model (the constant variance within the residuals = homoscedasticity)

- In case of heteroscedasticity the standard error can be underestimated and the t-statistics can increase → the estimates are no longer efficient in that case

OLS regression – **the problem of heteroscedasticity**

The Goldfeld-Quandt test on heteroscedsticity

- Divide the observations into two groups
- Under the assumption of homoscedasticity the variances in both groups must be equal
- Order the values for the independent variable in ascending order → build two groups
- Estimate the regression equation separately for both groups
- Calculate the ratio of squared residuals between both groups

- The null hypothesis is homoscedasticity and from this follows a F-distribution with n_1-k (degrees of freedom for the numerator) and n_2-k (degrees of freedom for the denominator).
- n_1 and n_2 are the respective sizes of the groups and k is the number of regressors plus the intercept
- Large F-values result in a rejection of the null-hypothesis (→ i.e. there is significant heteroscedasticity)

OLS regression – **the problem of heteroscedasticity**

Goldfeld-Quandt test for the example of holiday expenditure and monthly income:

1) Sort the independent variable (monthly income) in ascending order
2) Make two groups of equal size (one high, one low income)
3) Calculate the regression for both groups separately
4) Report the residuals for both groups separately
5) Square the residuals for both groups separately
6) Sum the squared residuals for both groups separately
7) F_{emp} = squared Residuals (high income) / squared residuals (low income)
8) Compare the empirical F-value with the theoretical one from a F-table.

OLS regression – **the problem of heteroscedasticity**

Für 99% Sicherheit:

degrees of freedon for the numerator

degrees of freedom for the denominator

f_2 \ f_1	1	2	3	4	5	6	7	8	9	10
1							5929	5981	6023	6056
2							99, 35	99, 36	99, 38	99, 40
3							27, 67	27, 49	27, 34	27, 23
4							14, 98	14, 80	14, 66	14, 54
5							10, 44	10, 27	10, 14	10, 04
6	13, 47	10, 92	9, 78	9, 15	8, 75	8, 47	8, 26	8, 10	7, 98	7, 87
7	12, 25	9, 55	8, 45	7, 85	7, 46	7, 19	6, 99	6, 84	6, 72	6, 62
8	11, 26	8, 65	7, 59	7, 01	6, 63	6, 37	6, 18	6, 03	5, 91	5, 81
9	10, 56	8, 02	6, 99	6, 42	6, 06	5, 80	5, 61	5, 47	5, 35	5, 26
10	10, 04	7, 56	6, 55	5, 99	5, 64	5, 39	5, 20	5, 06	4, 94	4, 85
11						5, 07	4, 89	4, 74	4, 63	4, 54
12						4, 82	4, 64	4, 50	4, 39	4, 30
13						4, 62	4, 44	4, 30	4, 19	4, 10
14	8, 86	6, 51	5, 56	5, 04	4, 70	4, 46	4, 28	4, 14	4, 03	4, 94
15	8, 68	6, 36	5, 42	4, 89	4, 56	4, 32	4, 11	4, 00	3, 89	3, 80

There are 24 variables→

n_1 and n_2 = 12 →

10 numerator- and 10 denominator degrees of freedom

If F_{emp} > 4,85 → heteroscedasticity

OLS regression – **the problem of outliers**

Graphical identification:

- Identification via scatter plot
- Observed values that are distant from the regression line can be interpreted as outliers
- Identification possible as well via boxplot or stem and leaf diagram

OLS regression – **the problem of outliers**

Statistical identification of outliers via **leverages:**

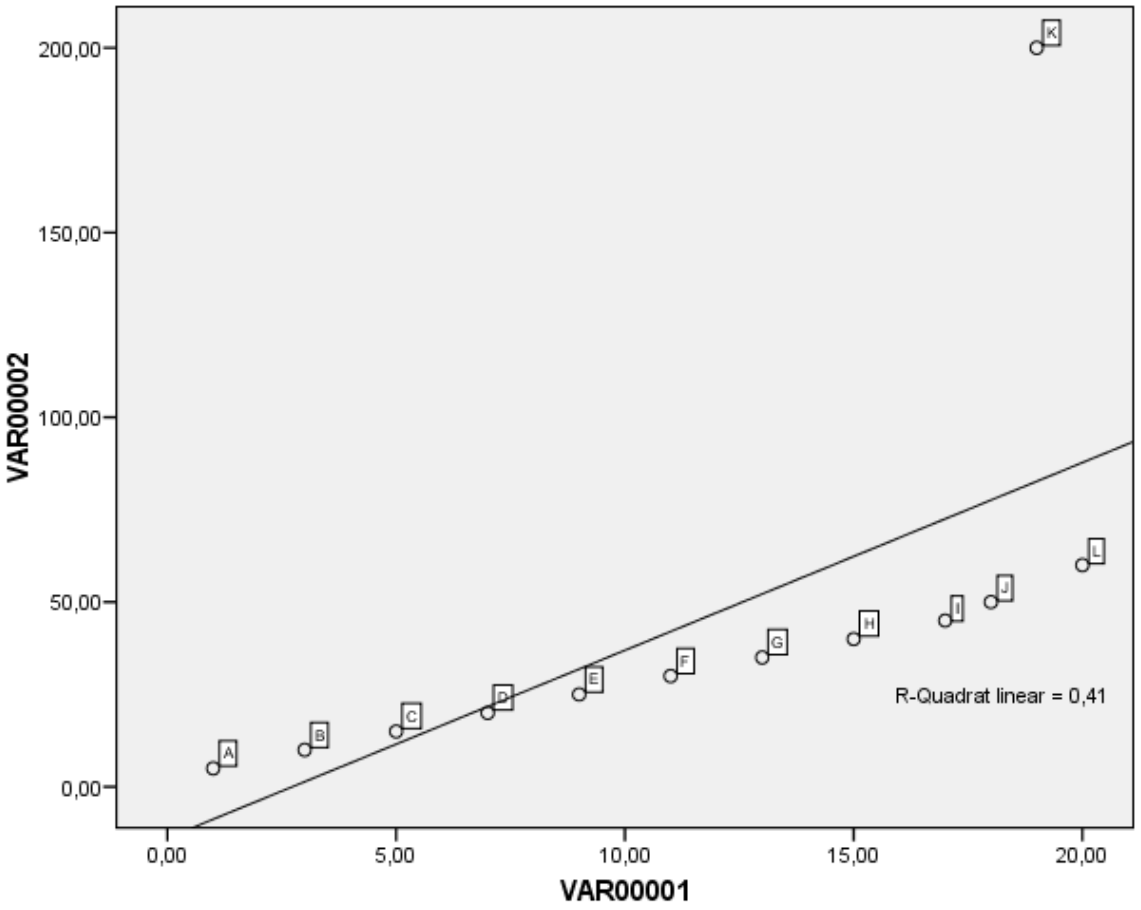
- Leverages indicate the influence of single cases on the assumed linear relationship
- Leverages can be between zero and (n-1)/n, at which k/n (k= number of regressors, n = number of cases) is expected on average.
- Leverages below .2 are not problematic, between .2 and .5 you have to be cautious and values above 2k/n or above .5 are always critical
- Cases with extremely high leverages should be excluded from the statistical analysis
- Closely related to the leverages is the Mahalanobis-distance which indicates how far the values of the independent variables of one case deviate from the mean of all cases

OLS regression – **the problem of outliers**

Problem:

Leverages are not meaningful if the outlier is only at the dependent variable.

→Case K is clearly an outlier, but its leverage (.12) as well as its Mahalanobis-distance (1.32) indicate no outlier tendency.



OLS regression – **the problem of outliers**

Solution: Cook-value (calculated by SPSS as well)

- The Cook-value is a measure assessing how much the residuals of all cases would change, if a certain case is excluded from the calculation of the regression coefficients

→ The Cook value is 1.28 for the case K and hence a lot larger than the values of the other cases → K is an outlier

Thus for the identification of outliers generally leverages, Mahalanobis-distance and the Cook-value should be considered.