Teaching Mathematics and Statistics in Sciences, IPA HU-SRB/0901/221/088 - 2011

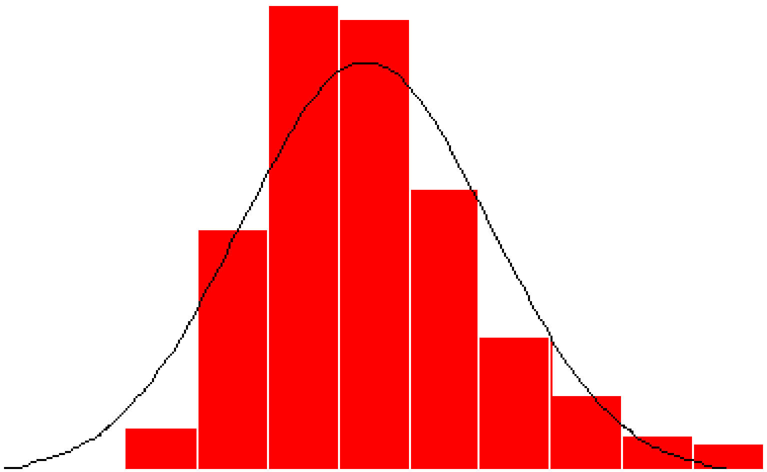
Biostatistics

Author: **Krisztina Boda** PhD

University of Szeged
Department of Medical Physics and Informatics

www.model.u-szeged.hu
www.szote.u-szeged.hu/dmi

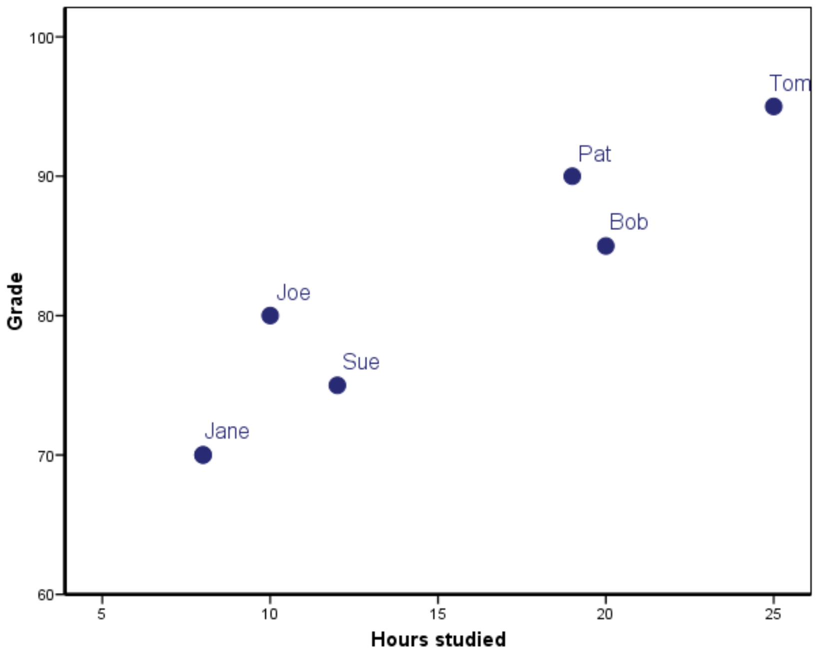
Correlation and regression II.
Regression using transformations



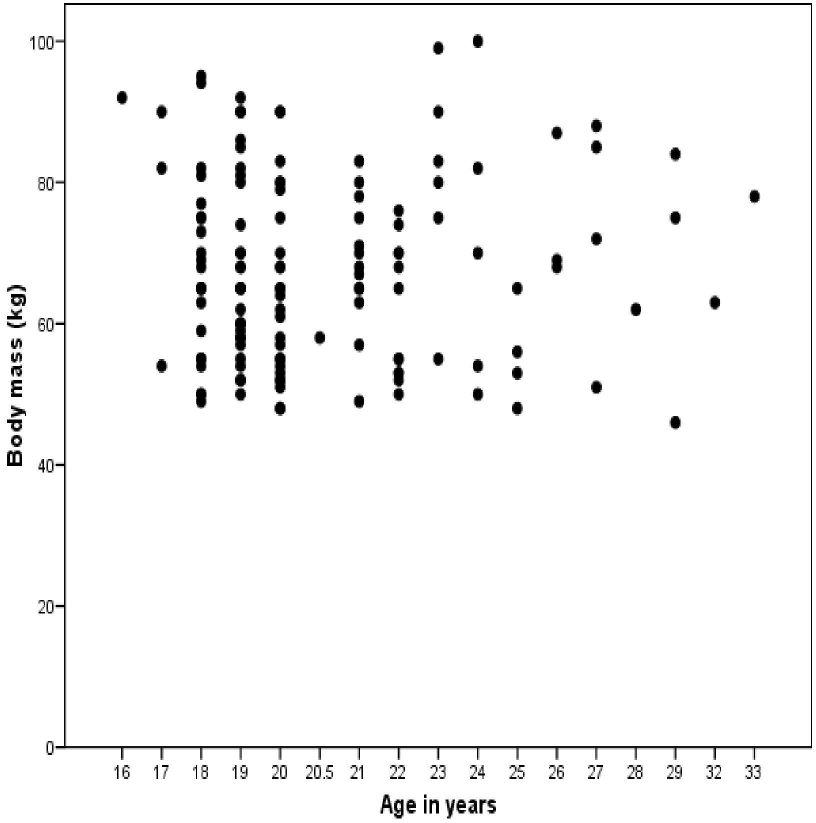
Scatterplot

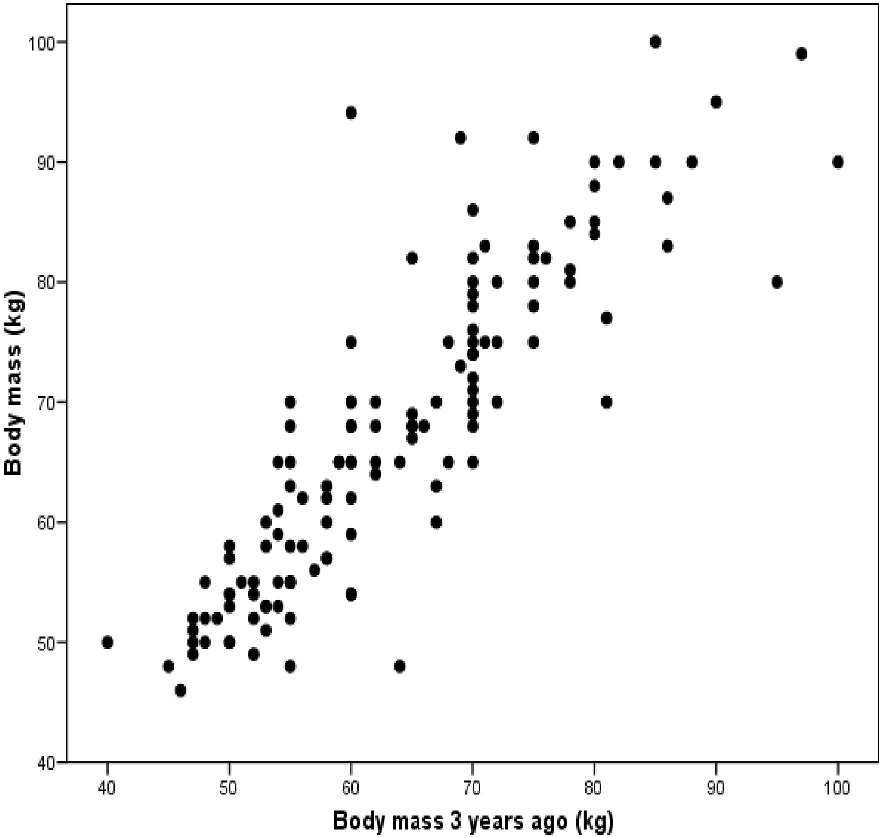
Relationship between two continuous variables

Student	Hours studied	Grade
Jane	8	70
Joe	10	80
Sue	12	75
Pat	19	90
Bob	20	85
Tom	25	95



Scatterplot
Other examples





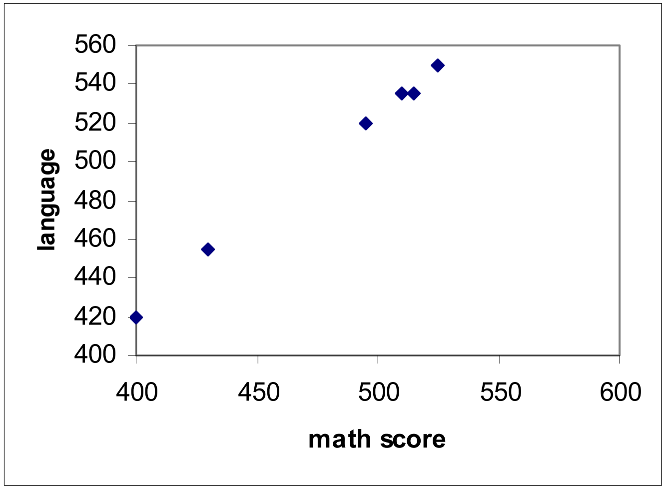
Example II.

- Imagine that 6 students are given a battery of tests by a vocational guidance counsellor with the results shown in the following table:

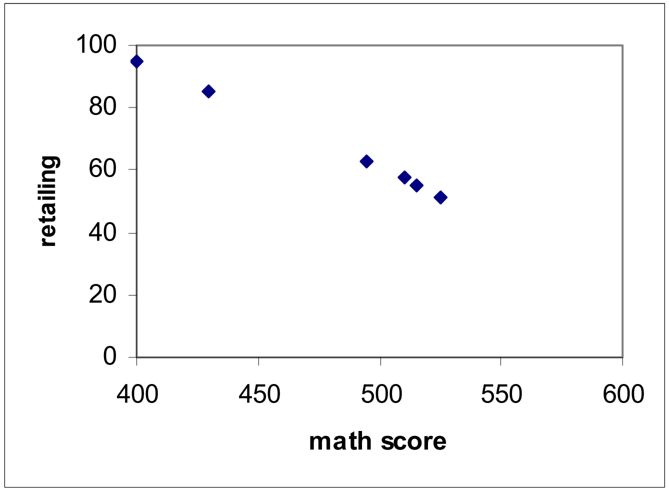
	STUDENT	RETAIL	THEATER	MATH	LANGUAGE
1	Pat	51.00	30.00	525.00	550.00
2	Sue	55.00	60.00	515.00	535.00
3	Inez	58.00	90.00	510.00	535.00
4	Arnie	63.00	50.00	495.00	520.00
5	Gene	85.00	30.00	430.00	455.00
6	Bob	95.00	90.00	400.00	420.00

- Variables measured on the same individuals are often related to each other.

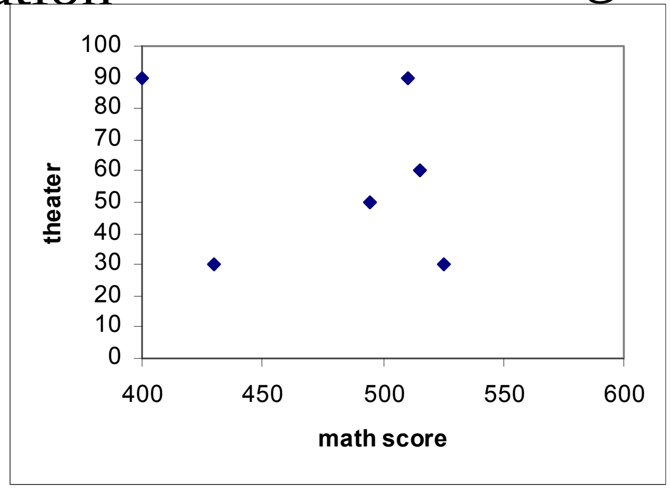
Possible relationships



positive correlation



negative correlation



no correlation

Describing <u>linear</u> relationship with number: the coefficient of correlation (r).

Also called **Pearson** coefficient of correlation

- Correlation is a numerical measure of the strength of a linear association.
- The formula for coefficient of correlation treats *x* and *y* identically. There is no distinction between explanatory and response variable.
- Let us denote the two samples by $x_1, x_2, \dots x_n$ and $y_1, y_2, \dots y_n$, the coefficient of correlation can be computed according to the following formula

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

Karl Pearson

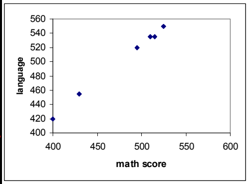
- Karl Pearson (27 March 1857 – 27 April 1936) established the discipline of mathematical statistics. http://en.wikipedia.org/wiki/Karl_Pearson

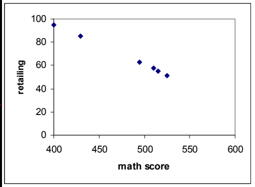
Properties of *r*

- Correlations are between -1 and +1; the value of *r* is always between -1 and 1, either extreme indicates a perfect linear association.

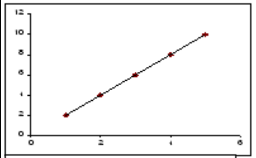
$$-1 \leq r \leq 1.$$

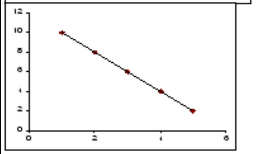
- a) If *r* is near +1 or -1 we say that we have high correlation.



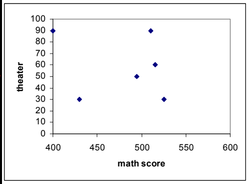


- b) If *r*=1, we say that there is perfect positive correlation.
  If *r*= -1, then we say that there is a perfect negative correlation.

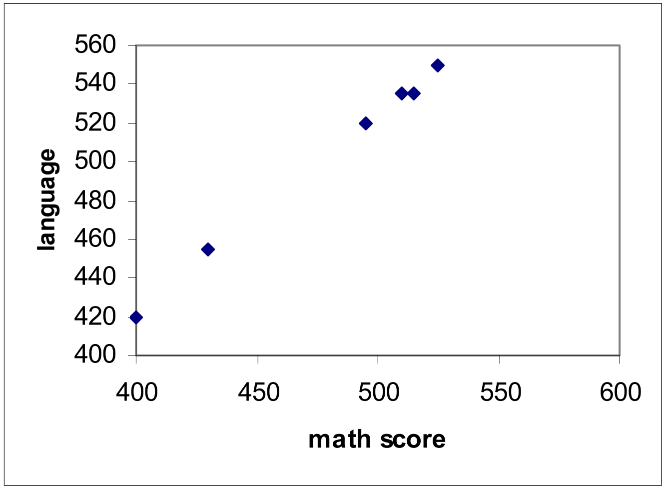




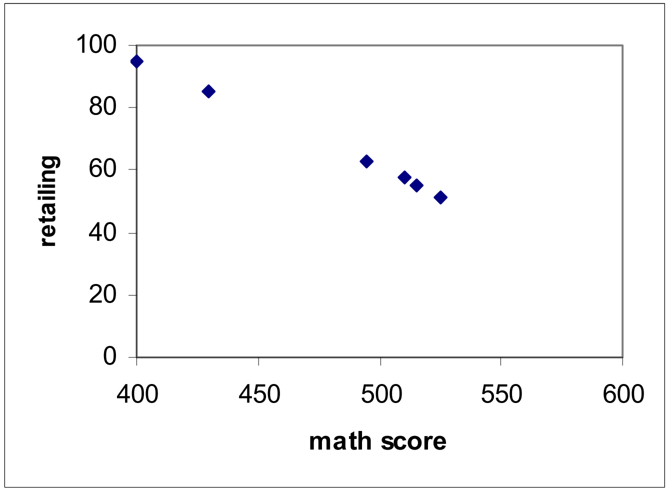
- c) A correlation of zero indicates the absence of linear association. When there is no tendency for the points to lie in a straight line, we say that there is no correlation (*r*=0) or we have low correlation (r is near 0).



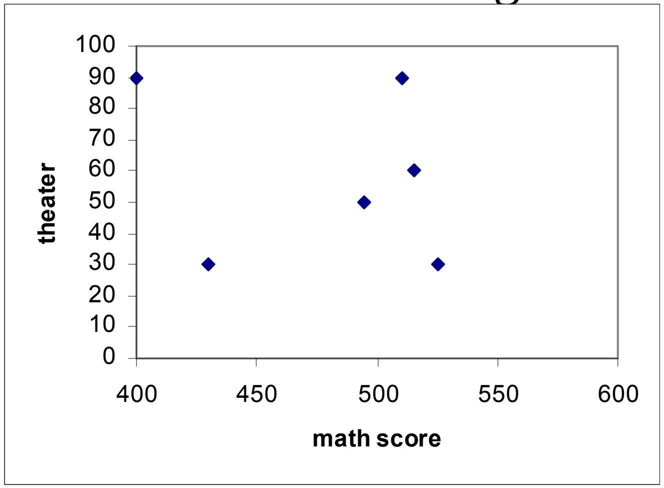
Calculated values of r



positive correlation, **r=0.9989**

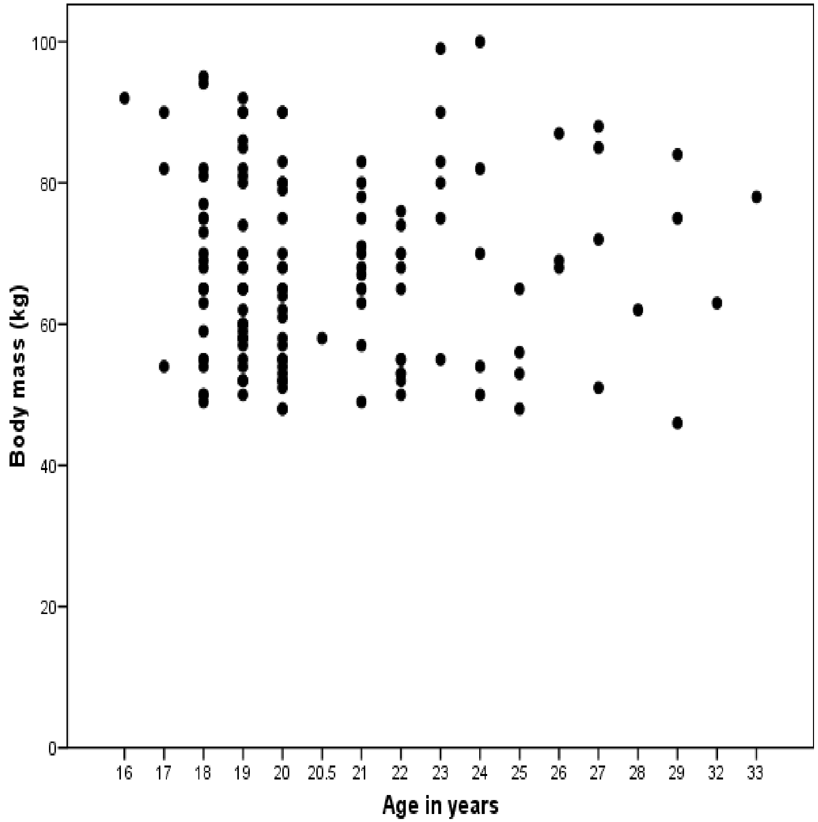


negative correlation, **r=-0.9993**

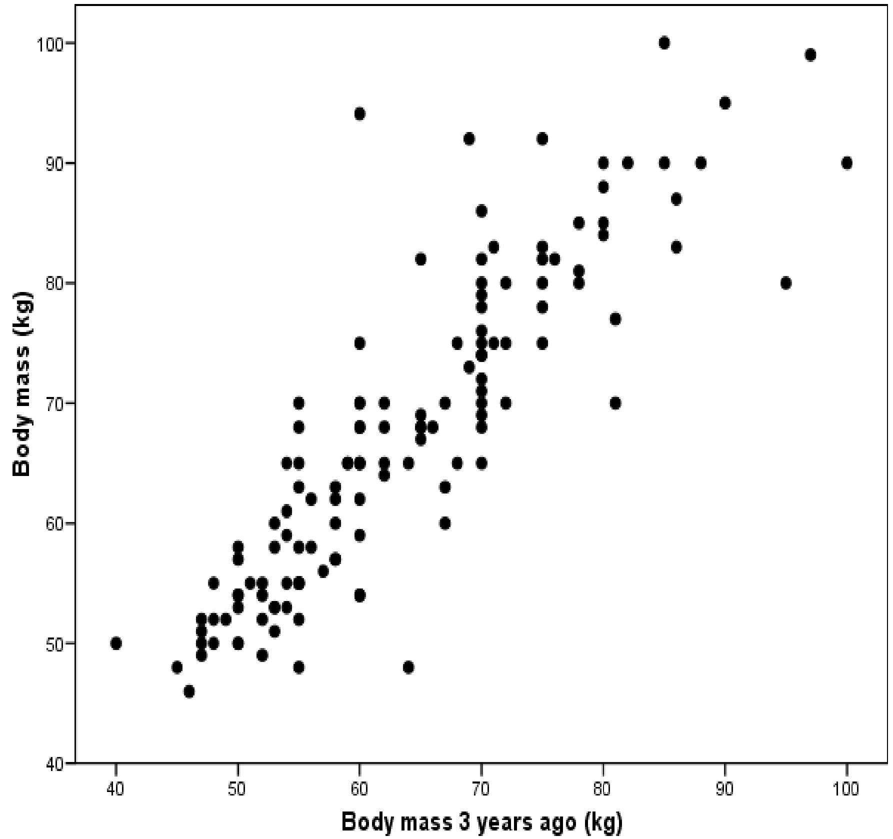


no correlation, **r=-0.2157**

Scatterplot
Other examples



r=0.018



r=0.873

Correlation and causation

- a correlation between two variables does not show that one causes the other.

When is a correlation „high”?

- What is considered to be high correlation varies with the field of application.
- The statistician must decide when a sample value of r is far enough from zero, that is, when it is sufficiently far from zero to reflect the correlation in the population.

Testing the significance of the coefficient of correlation

- The statistician must decide when a sample value of r is far enough from zero to be significant, that is, when it is sufficiently far from zero to reflect the correlation in the population.

Testing the significance of the coefficient of correlation

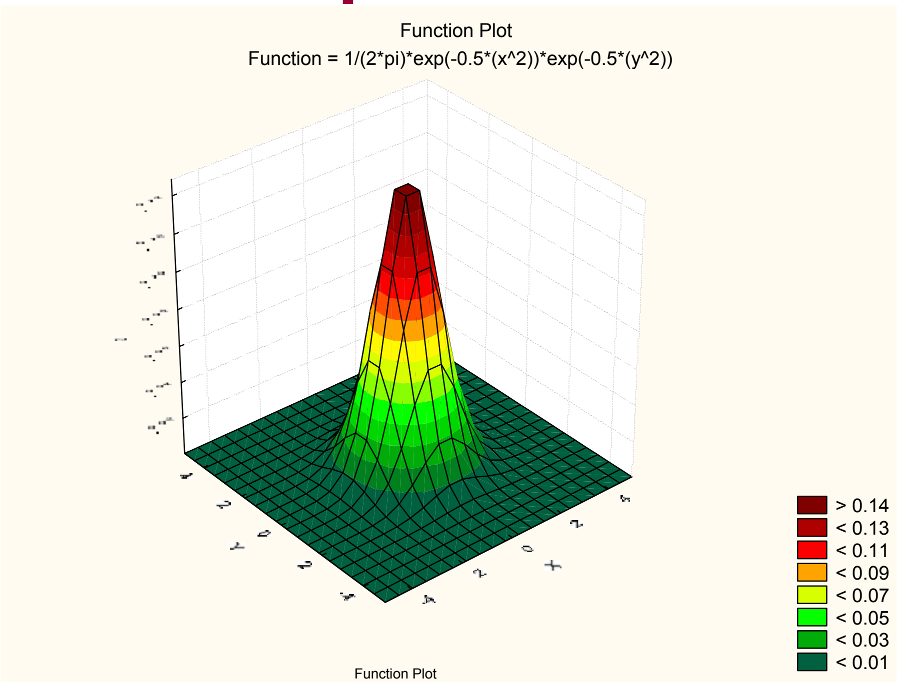
- The statistician must decide when a sample value of r is far enough from zero to be significant, that is, when it is sufficiently far from zero to reflect the correlation in the population. This test can be carried out by expressing the t statistic in terms of r.
- H_0: $\boldsymbol{\rho}=0$ (greek rho=0, correlation coefficient in population = 0)
- H_a: $\boldsymbol{\rho} \neq 0$ (correlation coefficient in population ≠ 0)
- Assumption: the two samples are drawn from a bivariate normal distribution
- If H0 is true, then the following t-statistic has n-2 degrees of freedom

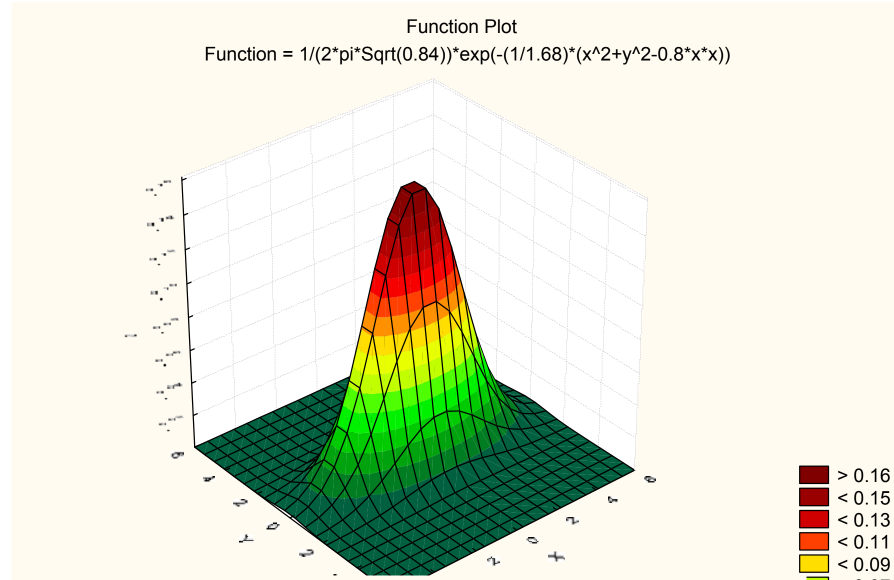
$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} = r \cdot \sqrt{\frac{n-2}{1-r^2}}$$

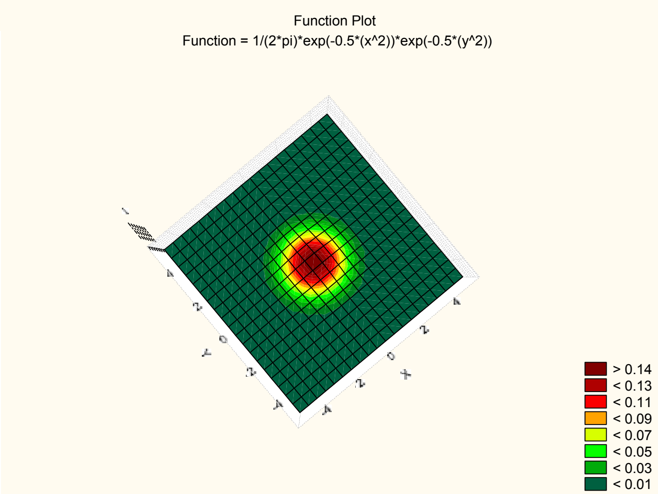
Bivariate normal distributions

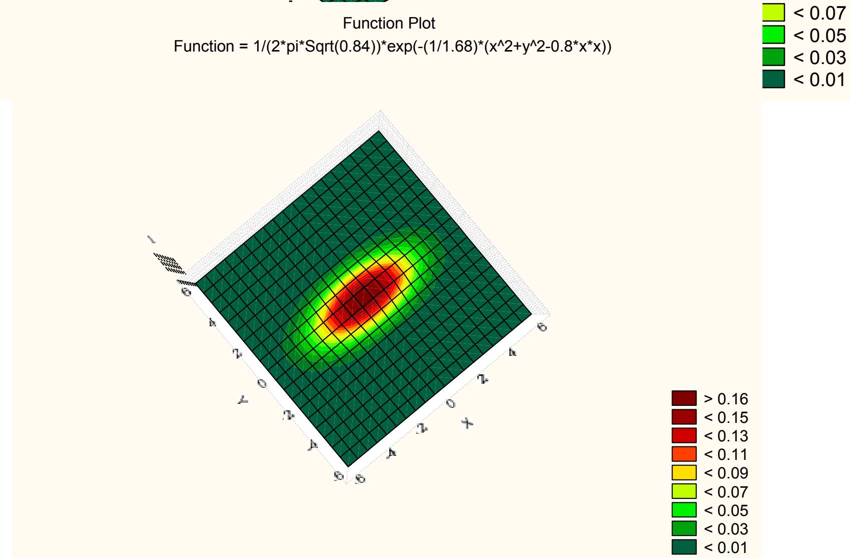
$\rho=0$

$\rho=0.4$









Testing the significance of the coefficient of correlation

- The following t-statistic has n-2 degrees of freedom

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} = r \cdot \sqrt{\frac{n-2}{1-r^2}}$$

- Decision using statistical table:
  - If $|t| > t_{\alpha,n-2}$, the difference is significant at α level, we reject H_0 and state that the population correlation coefficient is different from 0.
  - If $|t| < t_{\alpha,n-2}$, the difference is not significant at α level, we accept H_0 and state that the population correlation coefficient is not different from 0.

- Decision using p-value:
  - if $p < \alpha$ a we reject H_0 at α level and state that the population correlation coefficient is different from 0.

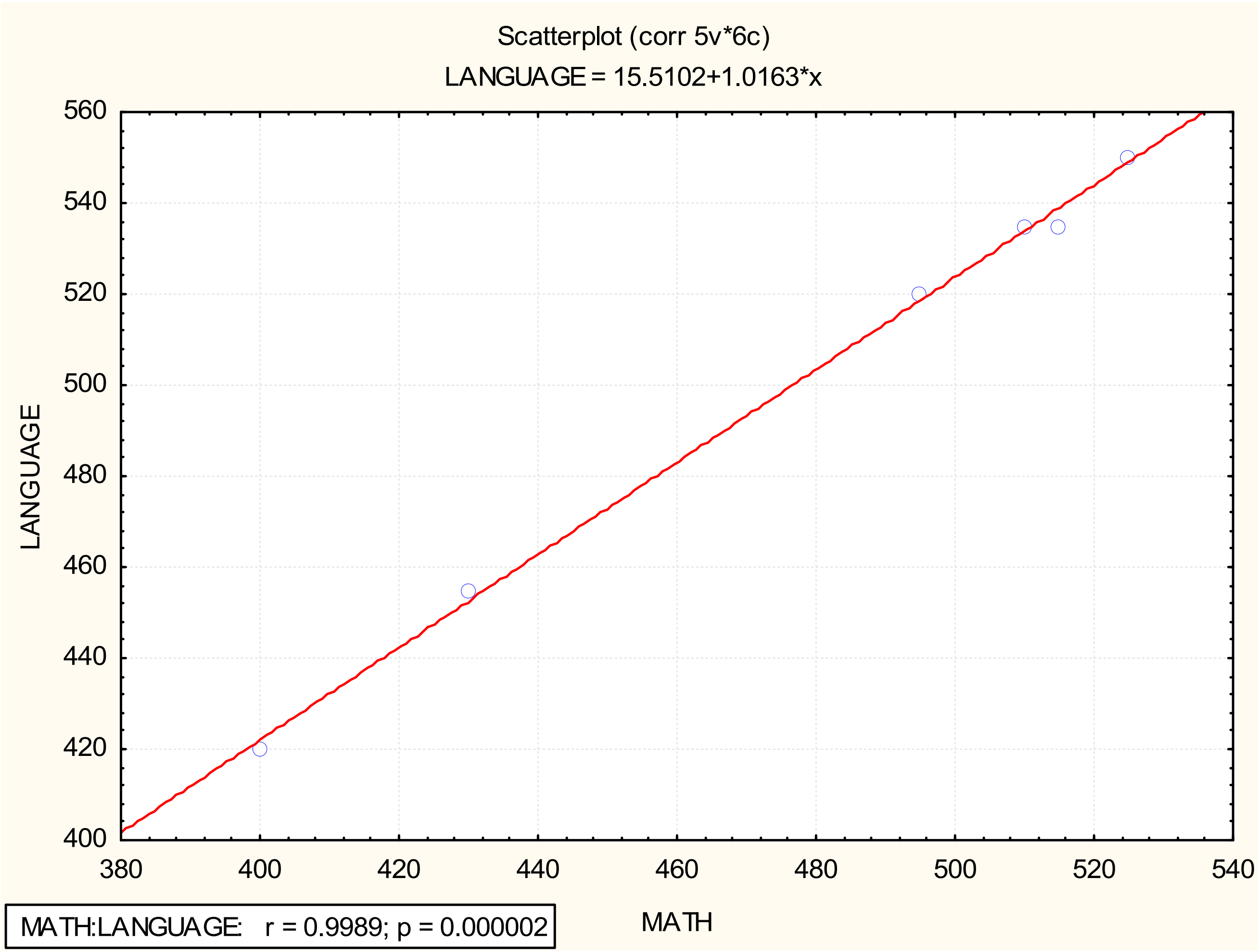
Example 1.

- The correlation coefficient between math skill and language skill was found *r*=0.9989. Is significantly different from 0?
- H_0: the correlation coefficient in population = 0, $\boldsymbol{\rho}$ =0.
- H_a: the correlation coefficient in population is different from 0.
- Let's compute the test statistic:

$$t=\frac{0.9989\cdot\sqrt{6-2}}{\sqrt{1-0.9989^2}}=0.9989\cdot\sqrt{\frac{4}{1-0.9989^2}}=42.6$$

- Degrees of freedom: df=6-2=4
- The critical value in the table is $t_{0.05,4}$ = 2.776.
- Because 42.6 > 2.776, we reject H_0 and claim that there is a significant linear correlation between the two variables at 5 % level.

Example 1, cont.



p<0.05, the correlation is significantly different from 0 at 5% level

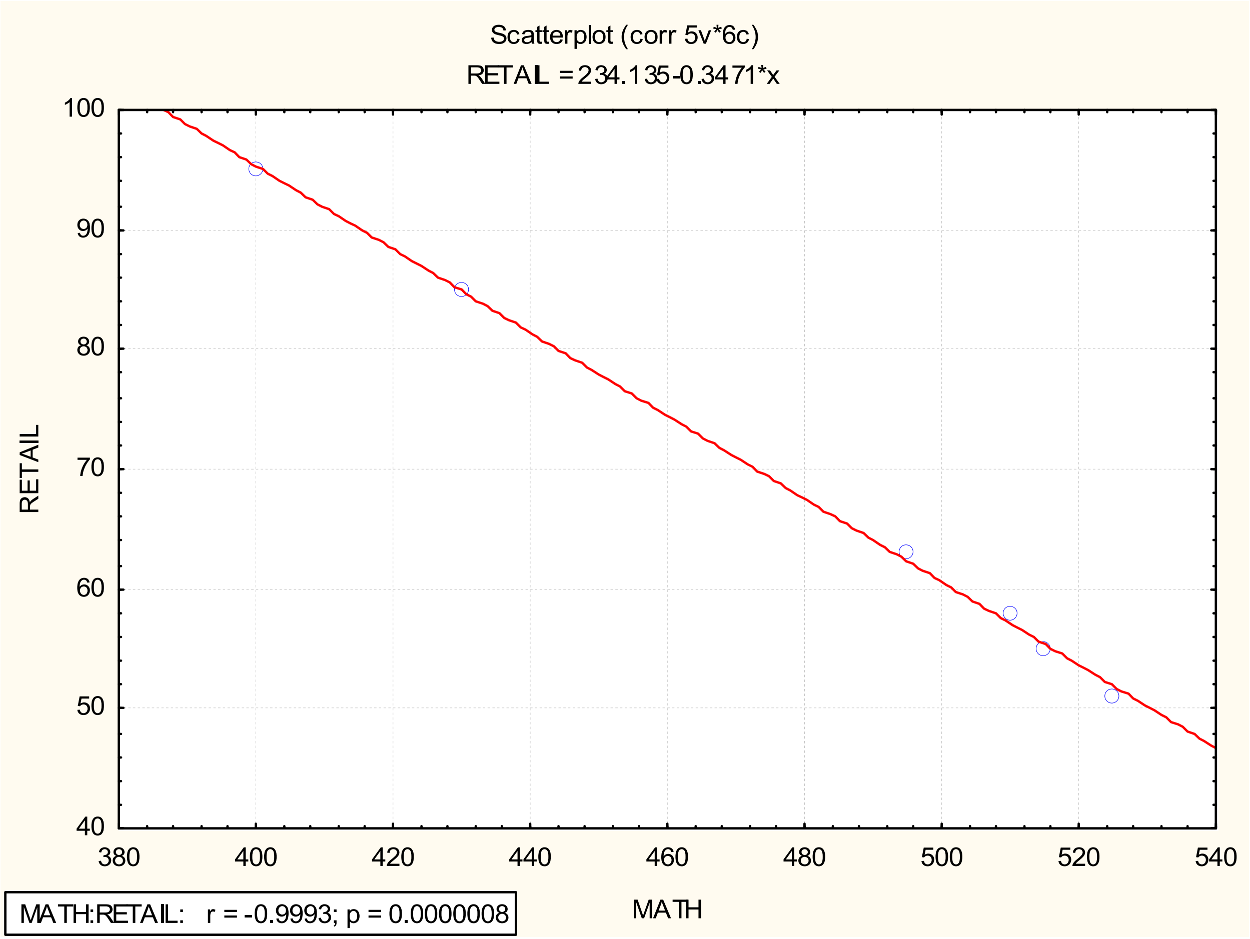
Example 2.

- The correlation coefficient between math skill and retailing skill was found r= -0.9993. Is significantly different from 0?
- H_0: the correlation coefficient in population = 0, $\boldsymbol{\rho}$ =0.
- H_a: the correlation coefficient in population is different from 0.
- Let's compute the test statistic:

$$t = \frac{-0.9993 \cdot \sqrt{6-2}}{\sqrt{1-0.9993^2}} = -0.9993 \cdot \sqrt{\frac{4}{1-0.9986}} = -53.42$$

- Degrees of freedom: df=6-2=4
- The critical value in the table is $t_{0.05,4}$ = 2.776.
- Because |-53.42|=53.42 > 2.776, we reject H_0 and claim that there is a significant linear correlation between the two variables at 5 % level.

Example 2., cont.



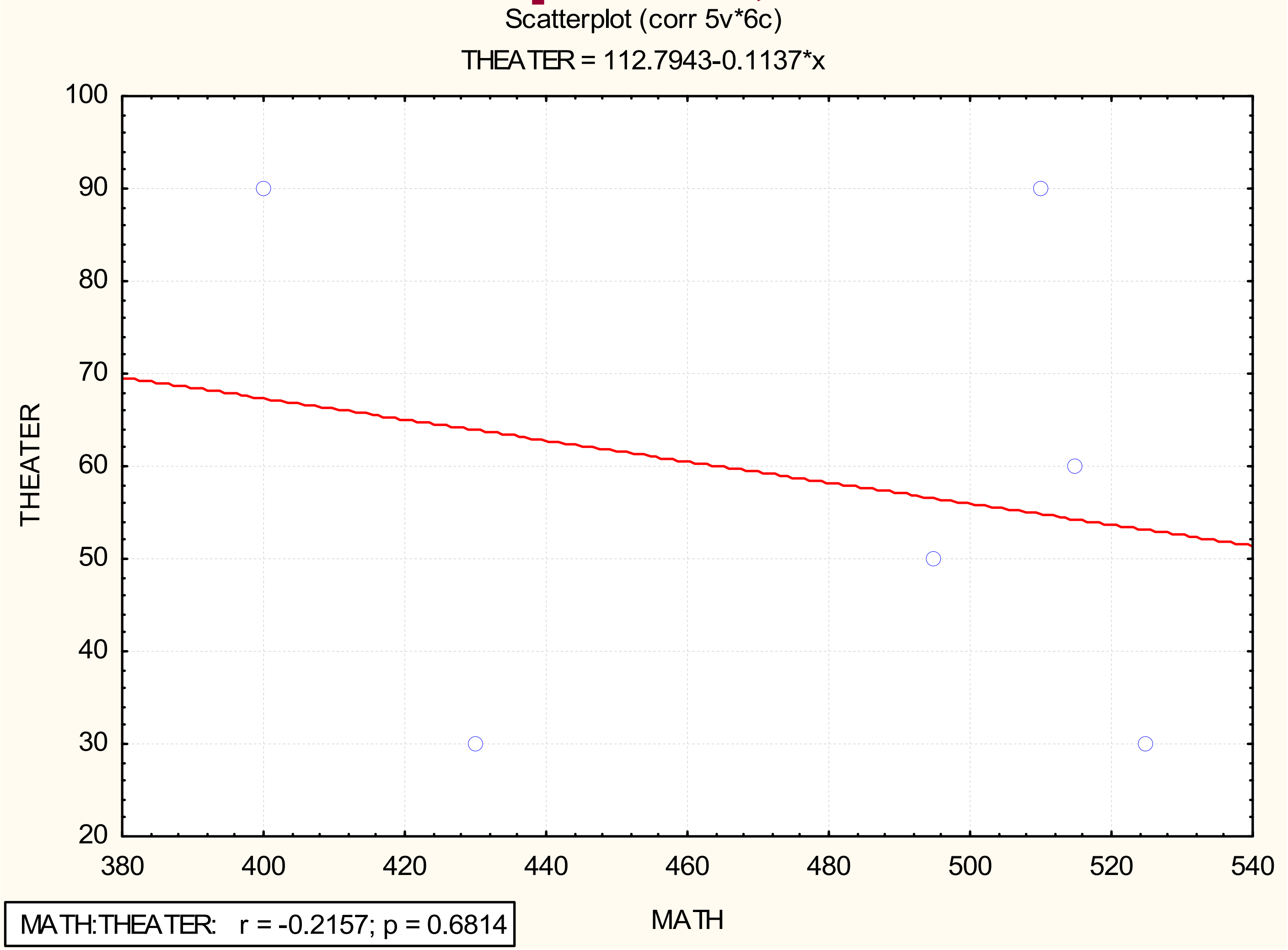
Example 3.

- The correlation coefficient between math skill and theater skill was found *r*= -0.2157. Is significantly different from 0?
- H_0: the correlation coefficient in population = 0, $\boldsymbol{\rho}$ =0.
- H_a: the correlation coefficient in population is different from 0.
- Let's compute the test statistic:

$$t = \frac{-0.2157 \cdot \sqrt{6-2}}{\sqrt{1-0.2157^2}} = -0.2157 \cdot \sqrt{\frac{4}{1-0.04653}} = -0.4418$$

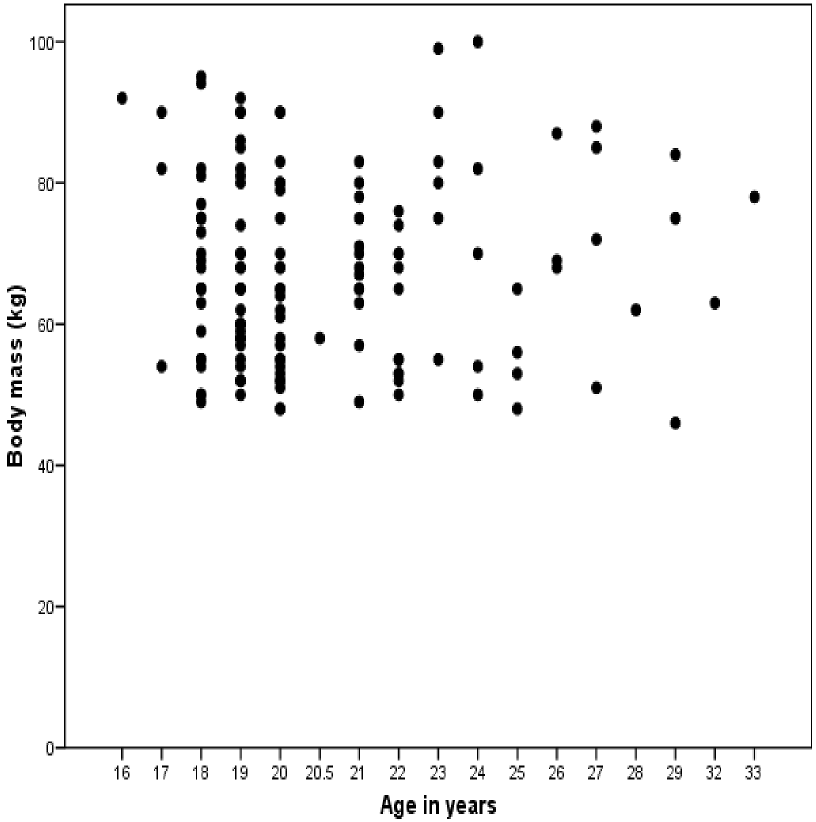
- Degrees of freedom: df=6-2=4
- The critical value in the table is $t_{0.05,4}$ = 2.776.
- Because |-0.4418|=0.4418 < 2.776, we do not reject H_0 and claim that there is no a significant linear correlation between the two variables at 5 % level.

Example 3., cont.

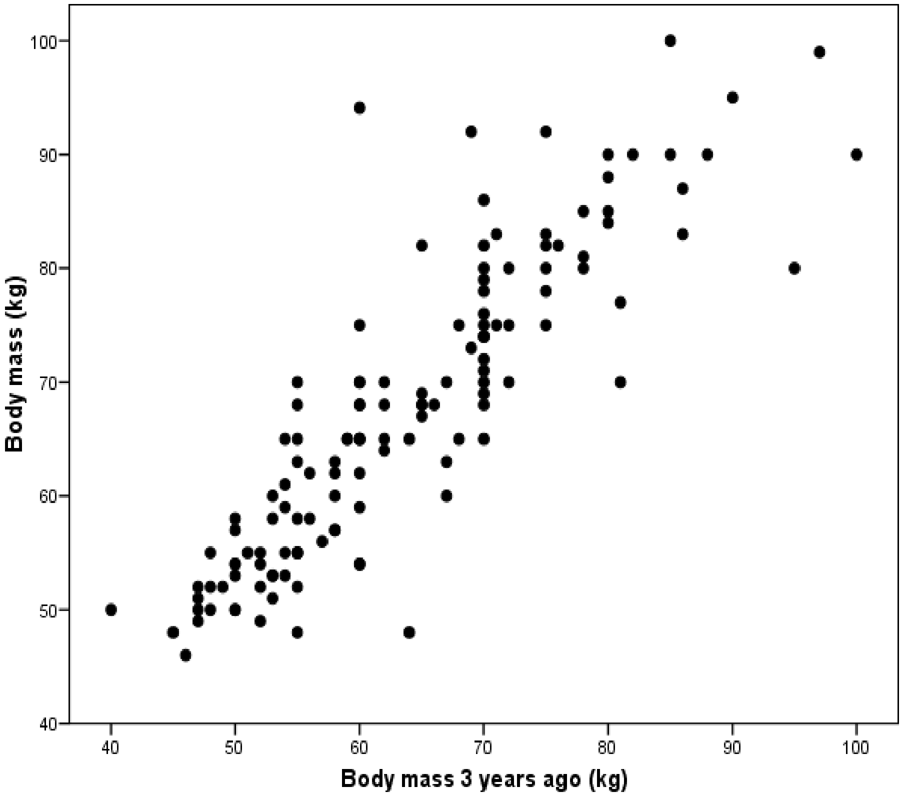


Significance of the correlation

Other examples



r=0.018, p=0.833

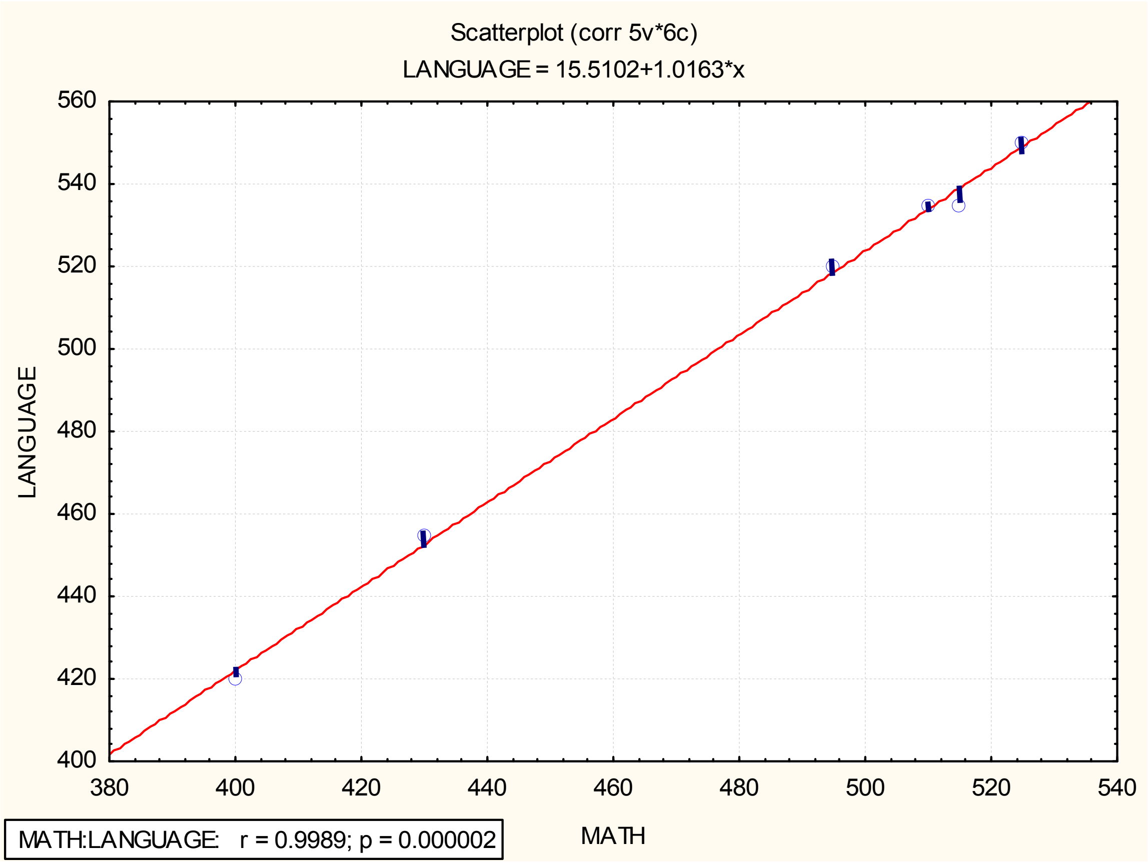


r=0.873, p<0.0001

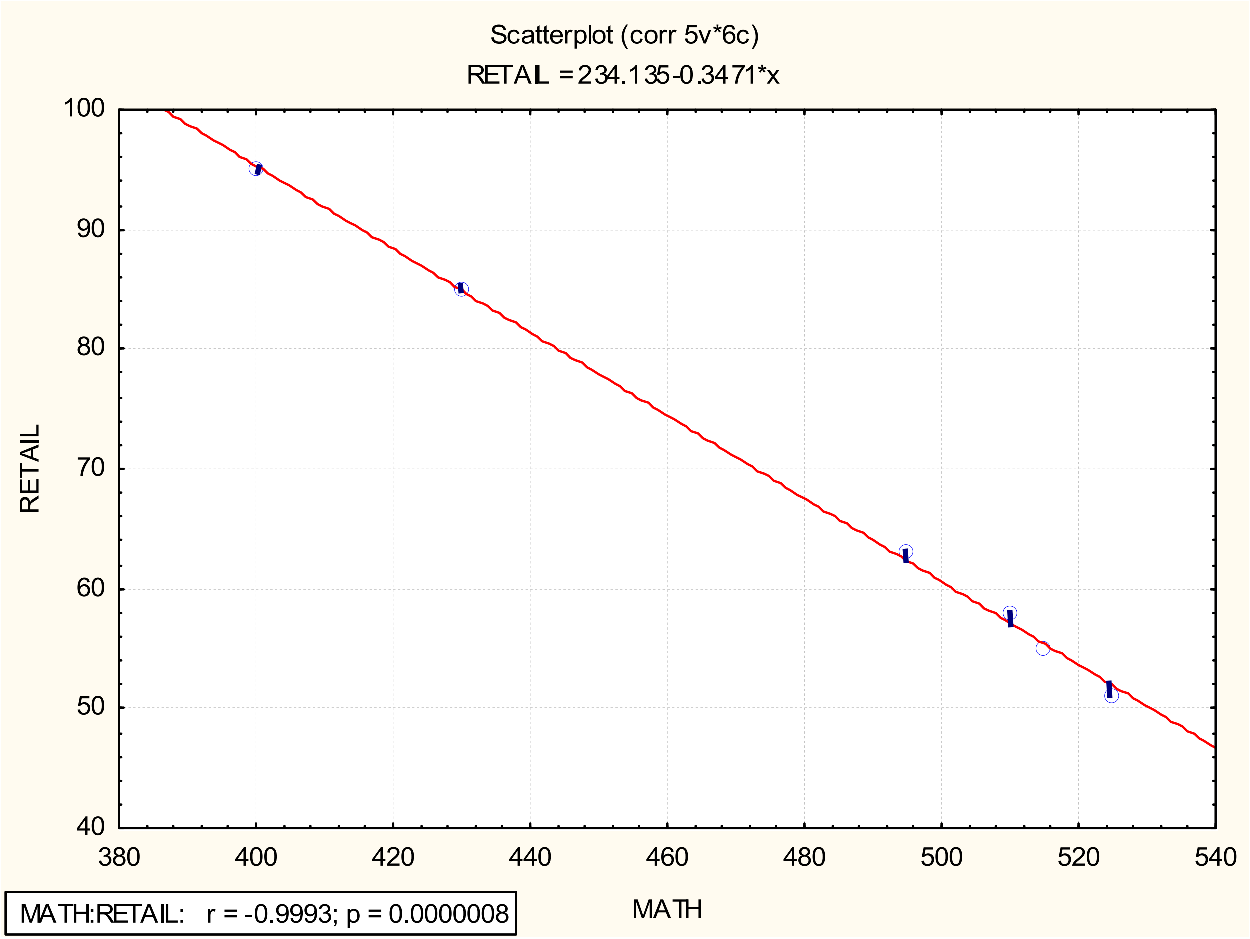
Prediction based on linear correlation: the linear regression

- When the form of the relationship in a scatterplot is linear, we usually want to describe that linear form more precisely with numbers.
- We can rarely hope to find data values lined up perfectly, so we fit lines to scatterplots with a method that compromises among the data values. This method is called the method of least squares.
- The key to finding, understanding, and using least squares lines is an understanding of their failures to fit the data; the residuals.

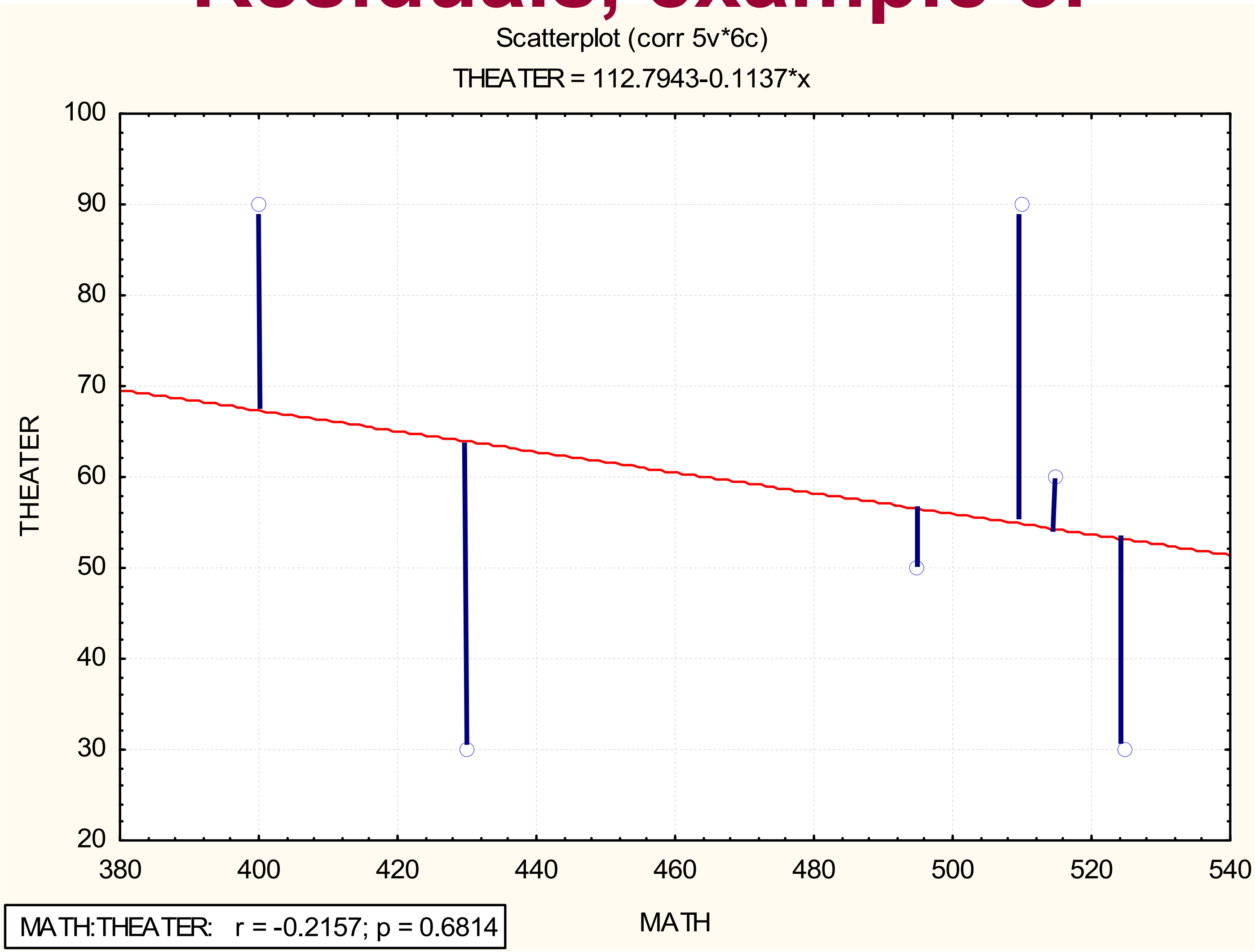
Residuals, example 1.



Residuals, example 2.



Residuals, example 3.



Prediction based on linear correlation: the linear regression

- A straight line that best fits the data:

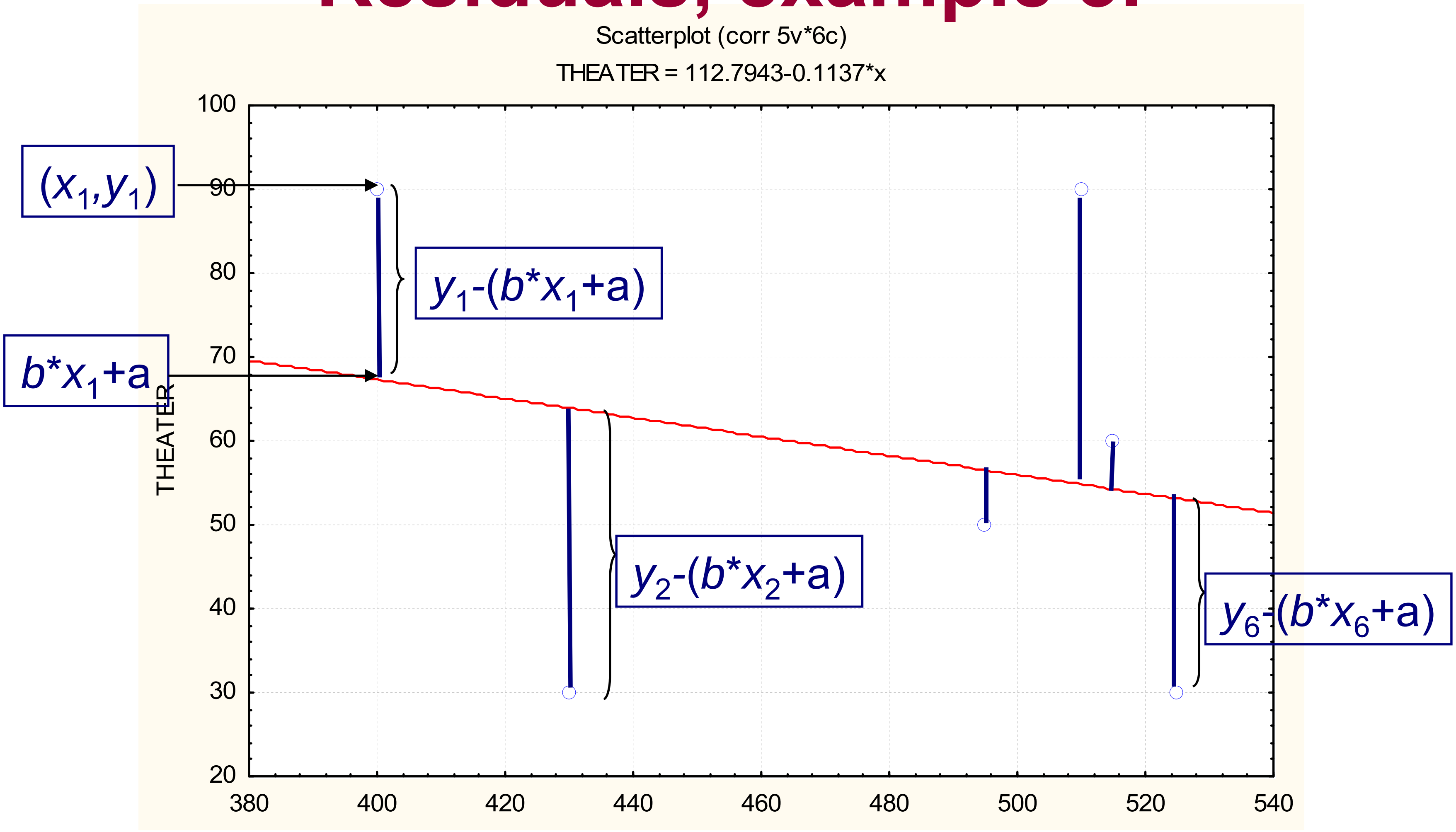
$$y=bx + a \quad or \quad y= a + bx$$

  is called regression line
- **Geometrical meaning of *a* and *b*.**
- *b*: is called regression coefficient, slope of the best-fitting line or regression line;
- *a*: *y*-intercept of the regression line.

- The principle of finding the values ***a*** and ***b*,** given $x_1,x_2,\ldots x_n$ and $y_1,y_2,\ldots y_n$.
- Minimizing the sum of squared residuals, i.e.

$$\Sigma(y_i-(a+bx_i))^2 \rightarrow \min$$

Residuals, example 3.



The general equation of a line is y = a + b x. We would like to find the values of *a* and *b* in such a way that the resulting line be the best fitting line. Let's suppose we have *n* pairs of (x_i, y_i) measurements. We would like to approximate y_i by values of a line . If x_i is the independent variable, the value of the line is $a + b x_i$.

We will approximate y_i by the value of the line at x_i, that is, by $a + b x_i$. The approximation is good if the differences $y_i - (a + b \cdot x_i)$ are small. These differences can be positive or negative, so let's take its square and summarize:

$$\sum_{i=1}^{n} (y_i - (a + b \cdot x_i))^2 = S(a,b)$$

This is a function of the unknown parameters *a* and *b*, called also the sum of squared residuals. To determine *a* and *b*: we have to find the minimum of S(*a,b*). In order to find the minimum, we have to find the derivatives of S, and solve the equations

$$\frac{\partial S}{\partial a} = 0, \quad \frac{\partial S}{\partial b} = 0$$

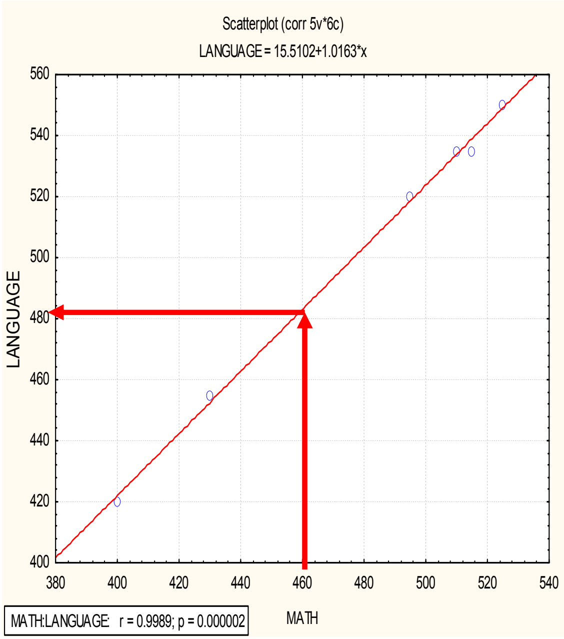
The solution of the equation-system gives the formulas for *b* and *a*:

$$b = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \text{ and } a = \overline{y} - b \cdot \overline{x}$$

It can be shown, using the 2nd derivatives, that these are really minimum places.

Equation of regression line for the data of Example 1.

- $y = 1.016 \cdot x + 15.5$
  the slope of the line is 1.016
- Prediction based on the equation: what is the predicted score for language for a student having 400 points in math?
- $y_{predicted} = 1.016 \cdot 400 + 15.5 = 421.9$



Computation of the correlation coefficient from the regression coefficient.

- There is a relationship between the correlation and the regression coefficient:

$$r = b \cdot \frac{s_x}{s_y}$$

- where s_x, s_y are the standard deviations of the samples .
- From this relationship it can be seen that the sign of r and b is the same: if there exist a negative correlation between variables, the slope of the regression line is also negative .

Hypothesis tests for the parameters of the regression line

- **Does y really depend on x? (not only in the sample, but also in the population?)**
- Assumption: the two samples are drawn from a bivariate normal distribution
- ***One possible method:***
- ***t*-test for the slope of the regression line**
- **H$_0$**: $b_{pop}=0$ the slope of the line is 0 (horizontal line)
- **Ha**: $b_{pop}\neq 0$
- If H0 is true, then the statistic t= b/SE(b) follows t-distribution with n-2 degrees of freedom

Hypothesis tests for the parameters of the regression line

- **Does y really depend on x? (not only in the sample, but also in the population?)**
- *Another possible method (equivalent)*
- **F-test for the regression – analysis of variance for the regression**
- ***Denote the estimated value*** $\hat{y}_i = bx_i + a$
- ***The following decomposition is true:***

Total variation of y =	Variation doe to its dependence on x +	Residual
SStot	SSx	SSe

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

Analysis of variance for the regression

Source of variation	Sum of Squares	Degrees of freedom	Variancie	F
Regression	SSr	1	SSr	$F = \frac{SSr}{SSh/(n-2)}$
Residual	SSh	n-2	SSh/n-2	
Total	SStot	n-1		

F has two degrees of freedom: 1 and n-2.
This test is equivalent to the t-test of the slope of the line, and to the t-test of the coefficient of correlation (they have the same p-value).

SPSS ouput for the relationship between age and body mass

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.018	.000	-.007	13.297

The independent variable is Age Age in years.

Coefficient of correlation, r=0.018

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	7.866	1	7.866	.044	.833
Residual	23515.068	133	176.805		
Total	23522.934	134			

The independent variable is Age in years.

Significance of the regression

(=significance of correlation), p=0.833

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Age Age in years	.078	.372	.018	.211	.833
(Constant)	66.040	7.834		8.430	.000

Significance of the slope =significance of the correlation, p=0.833

Equation of the regression line: $y=0.078x+66.040$

Significance of the constant term, p<0.0001

SPSS output for the relationship between body mass 3 years ago and at present

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.873	.763	.761	5.873

The independent variable is Mass Body mass (kg).

Coefficient of correlation, r=0.873

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	17420.202	1	17420.202	418.494	.000
Residual	5411.365	130	41.626		
Total	22831.568	131			

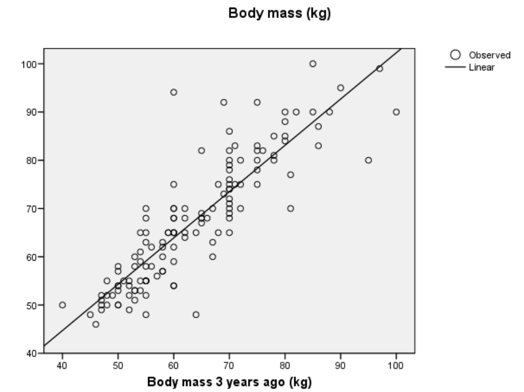
The independent variable is Body mass 3 years ago (kg).

Significance of the regression (=significance of correlation), $p<0.0001$

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Body mass 3 years ago (kg)	.960	.047	.873	20.457	.000
(Constant)	6.333	3.039		2.084	.039

Equation of the regression line:
$y=0.96x+6.333$



Coefficient of determination

- The square of the correlation coefficient multiplied by 100 is called the coefficient of determination.
- It shows the percentages of the total variation explained by the linear regression.
- **Example.**
- The correlation between math aptitude and language aptitude was found r =0,9989.
  The coefficient of determination, $r^2 = 0.9978$.
  So 91.7% of the total variation of Y is caused by its linear relationship with X .

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.9989	.9978	.997	2.729

The independent variable is Matematika.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	13707.704	1	13707.704	1840.212	.000
Residual	29.796	4	7.449		
Total	13737.500	5			

The independent variable is Matematika.

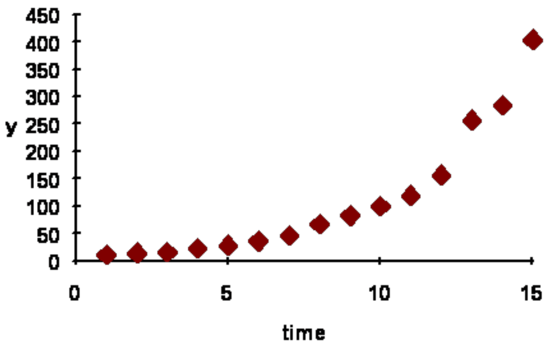
$r2$ from the ANOVA table:
$r2$ = Regression SS/Total SS=
=13707.704/13737.5= 0.9978 .

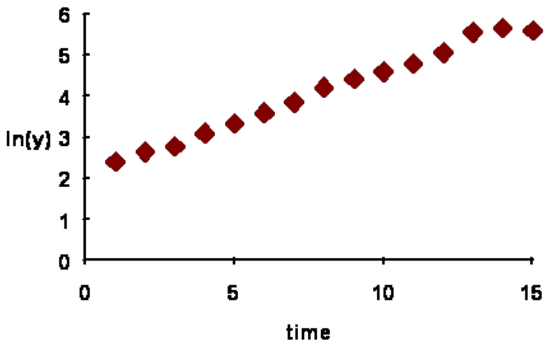
Regression using transformations

- Sometimes, useful models are not linear in parameters. Examining the scatterplot of the data shows a functional, but not linear relationship between data.

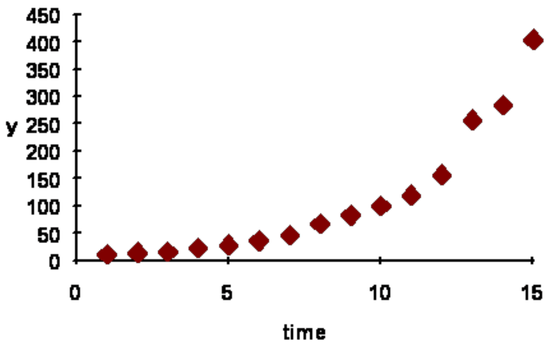
Example

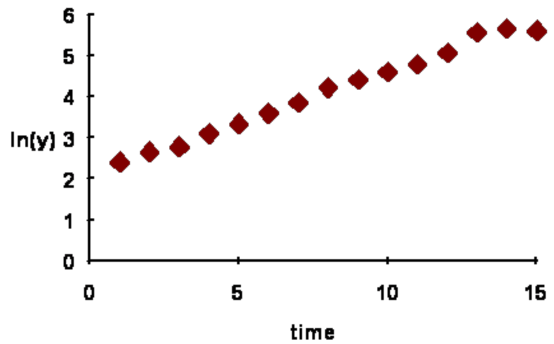
- A fast food chain opened in 1974. Each year from 1974 to 1988 the number of steakhouses in operation is recorded.
- The scatterplot of the original data suggests an exponential relationship between x (year) and y (number of Steakhouses) (first plot)
- Taking the logarithm of y, we get linear relationship (plot at the bottom)





- Performing the linear regression procedure to x and log (y) we get the equation
- $\log y = 2.327 + 0.2569\,x$
- that is
- $y = e^{2.327 + 0.2569\,x} = e^{2.327}e^{0.2569x} = 1.293e^{0.2569x}$ is the equation of the best fitting curve to the original data.





$y = 1.293e^{0.2569x}$

$\log y = 2.327 + 0.2569\, x$

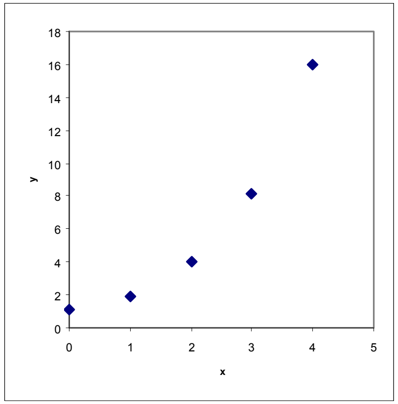
Types of transformations

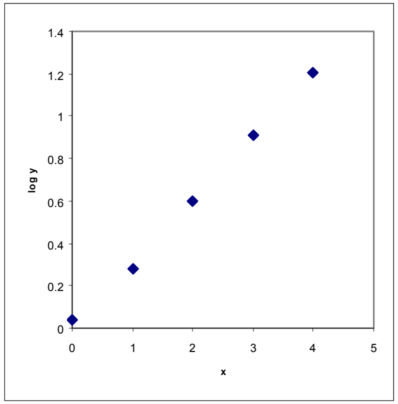
- Some non-linear models can be transformed into a linear model by taking the logarithms on either or both sides. Either 10 base logarithm (denoted log) or natural (base e) logarithm (denoted ln) can be used. If a>0 and b>0, applying a logarithmic transformation to the model

Exponential relationship ->take log y

x	y	lg y
0	1.1	0.041393
1	1.9	0.278754
2	4	0.60206
3	8.1	0.908485
4	16	1.20412

- Model: $y=a*10^{bx}$
- Take the logarithm of both sides:
- $\lg y = \lg a + bx$
- so lg y is linear in x

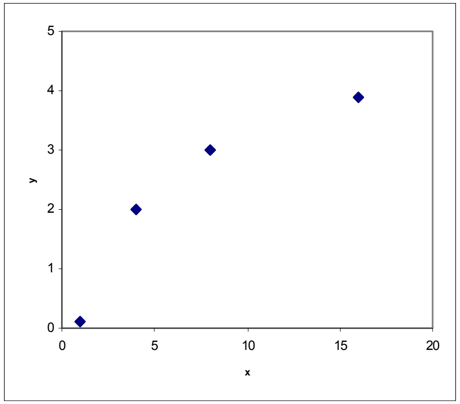


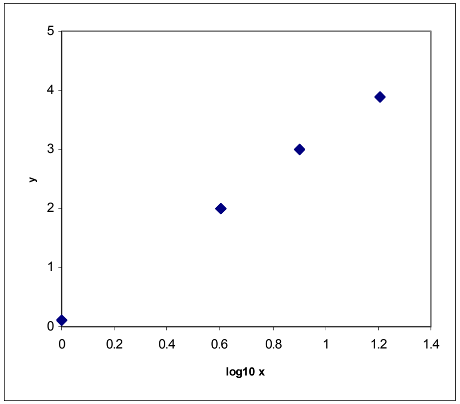


Logarithm relationship ->take log x

x	y	log x
1	0.1	0
4	2	0.60206
8	3.01	0.90309
16	3.9	1.20412

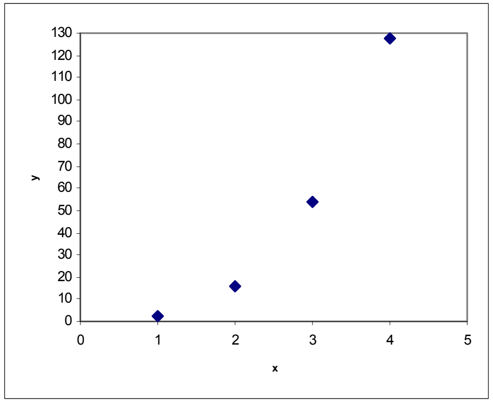
- Model: *y*=*a*+lgx
- so *y* is linear in lg *x*

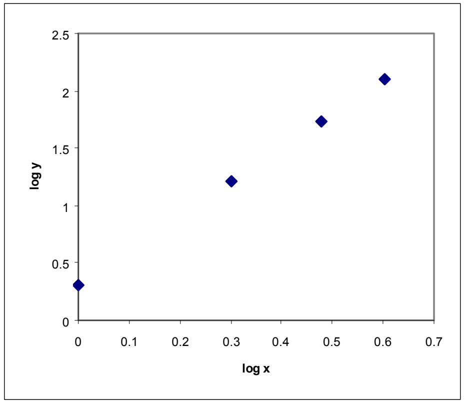




Power relationship ->take log x and log y

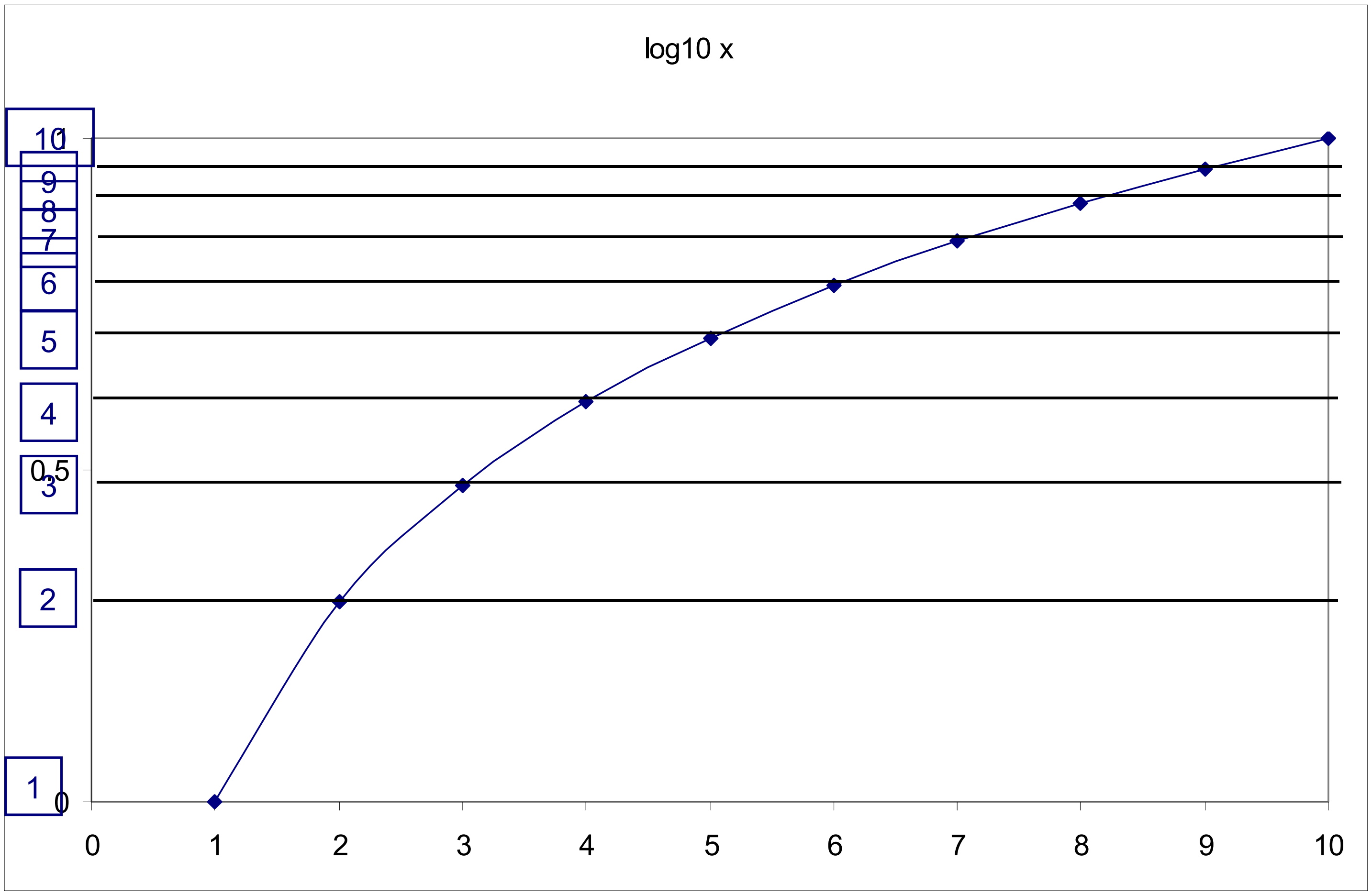
x	y	log x	log y
1	2	0	0.30103
2	16	0.30103	1.20412
3	54	0.477121	1.732394
4	128	0.60206	2.10721



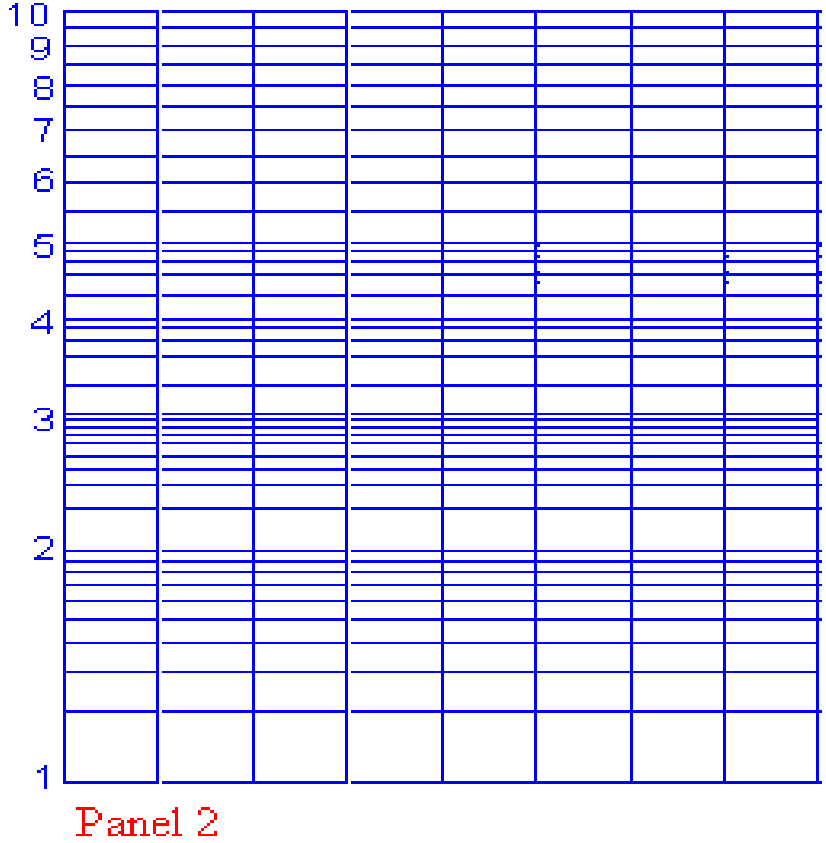


- Model: $y=ax^b$
- Take the logarithm of both sides:
- lg y =lg$a$+b lg$x$
- so lg$y$ is linear in lg x

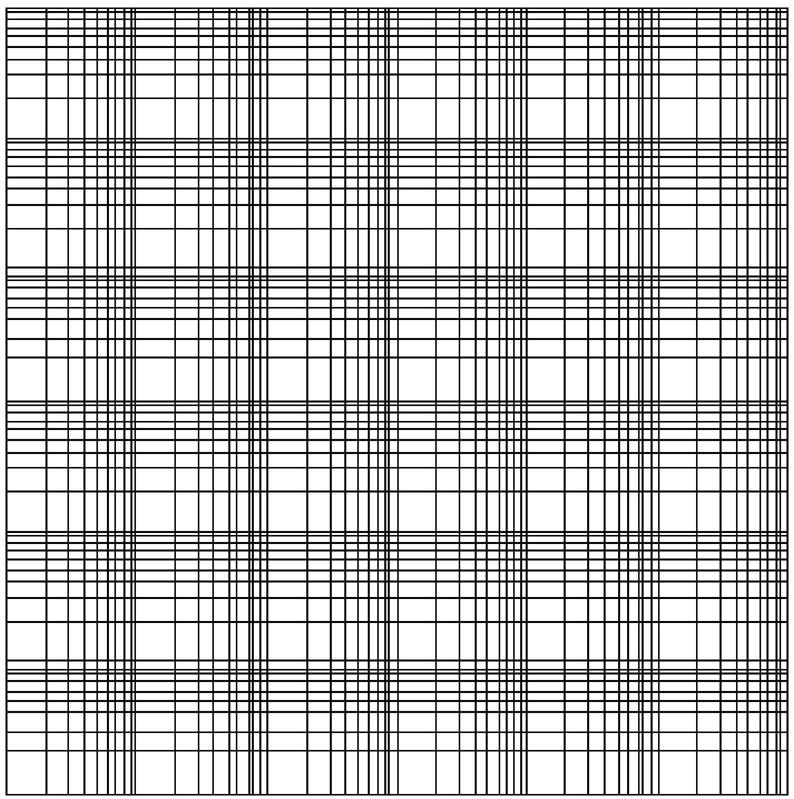
Log10 base logarithmic scale



Logarithmic papers



Semilogarithmic paper

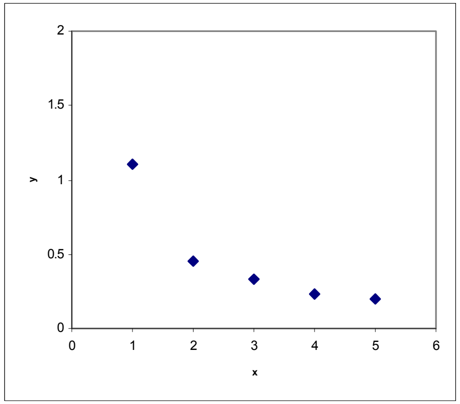


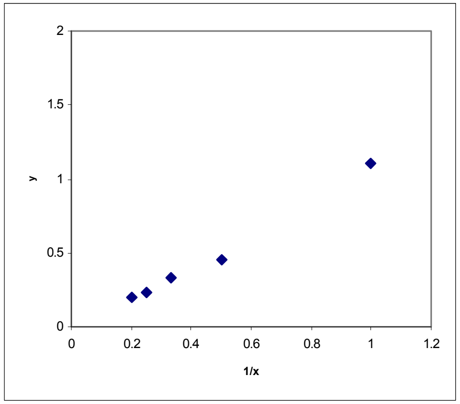
log-log paper

Reciprocal relationship ->take reciprocal of x

x	y	1/x
1	1.1	1
2	0.45	0.5
3	0.333	0.333333
4	0.23	0.25
5	0.1999	0.2

- Model: $y=a+b/x$
- $y=a+b*1/x$
- so y is linear in $1/x$





Example from the literature

Circulation
JOURNAL OF THE AMERICAN HEART ASSOCIATION

American Heart Association®
Learn and Live℠

Correlation between echocardiographic endocardial surface mapping of abnormal wall motion and pathologic infarct size in autopsied hearts

GT Wilkins, JF Southern, CY Choong, JD Thomas, JT Fallon, DE Guyer and AE Weyman

Circulation 1988;77;978-987
Circulation is published by the American Heart Association. 7272 Greenville Avenue, Dallas, TX 72514
 Print ISSN: 0009-7322. Online ISSN: 1524-4539

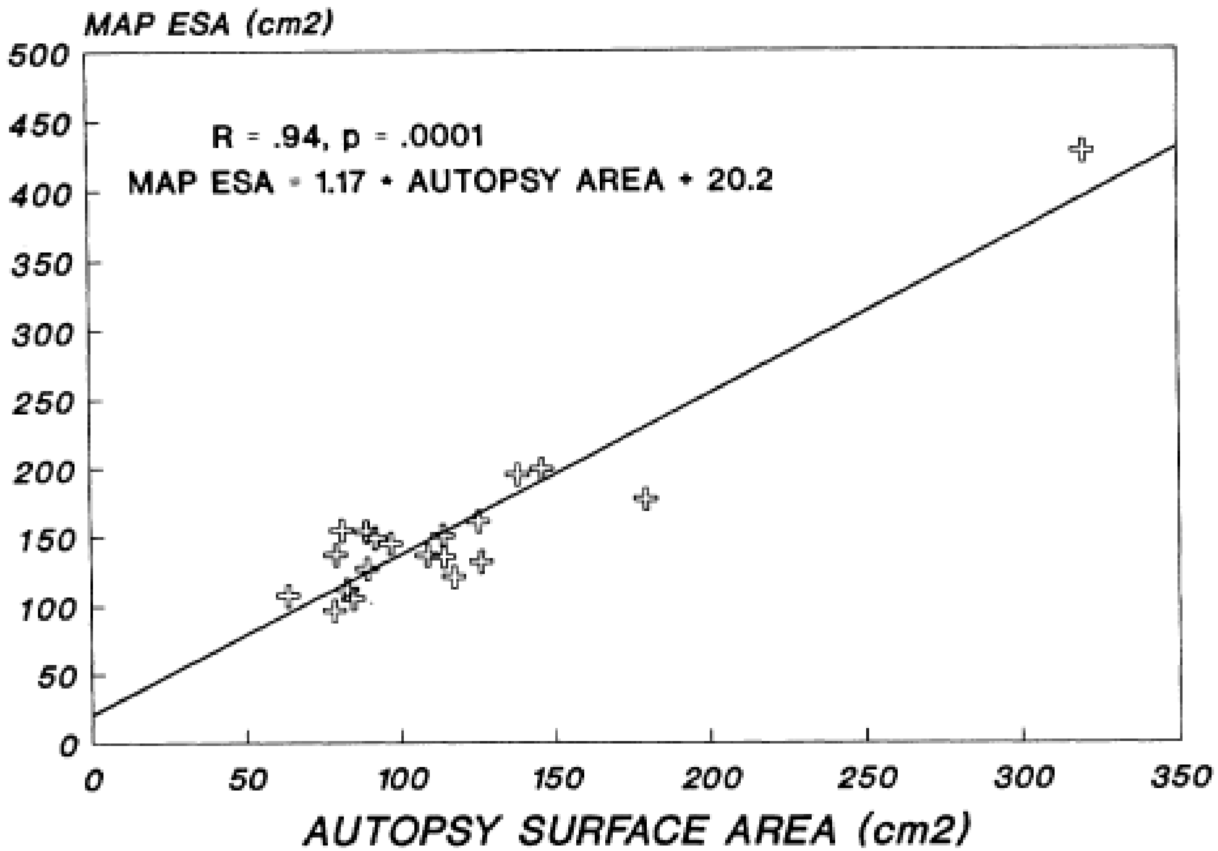
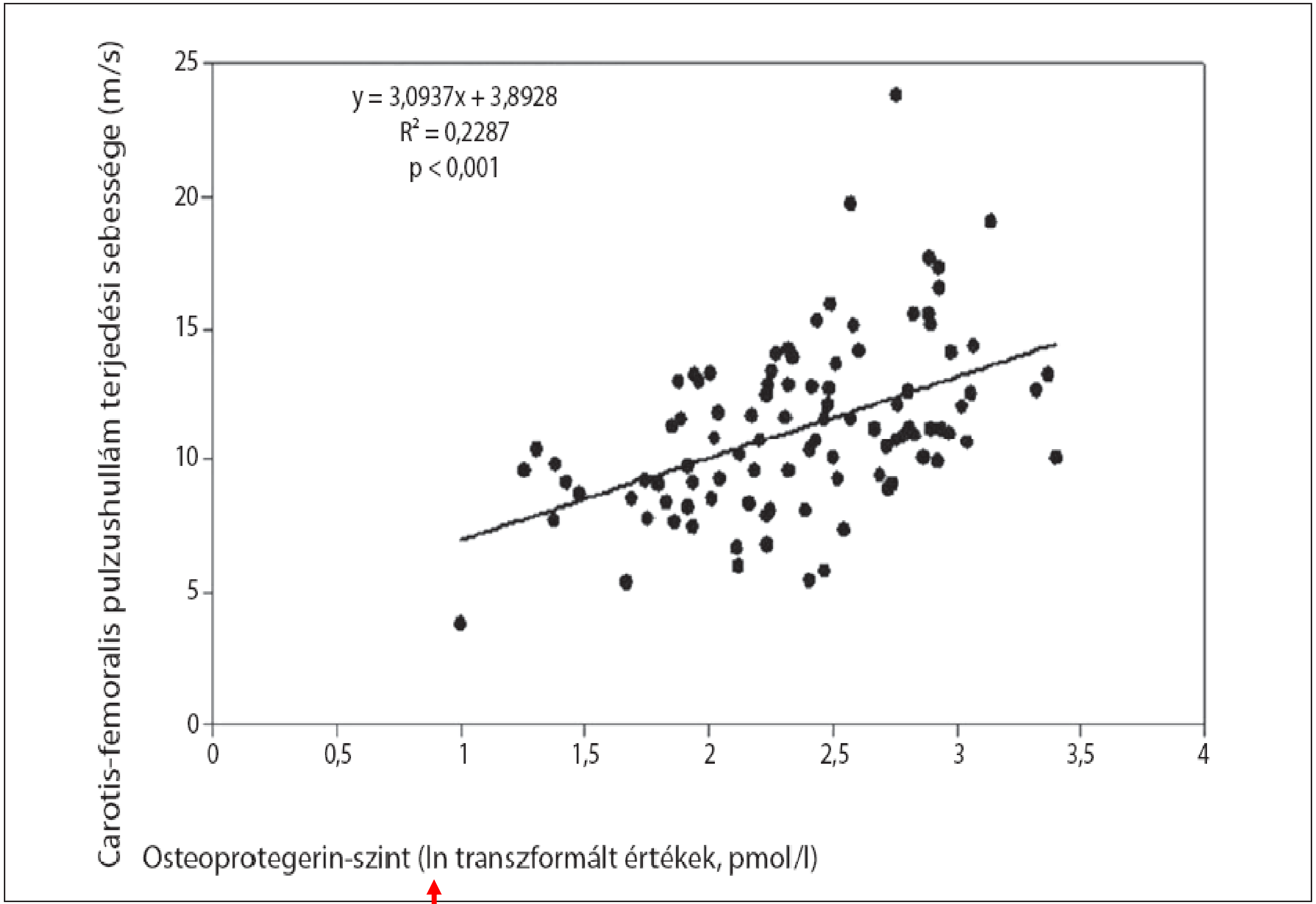


FIGURE 4. Correlation of the left ventricular endocardial surface area measured at autopsy (Autopsy Surface Area) with the endocardial surface area derived from the echocardiographic map (MAP ESA).

Example 2. EL HADJ OTHMANE TAHA és mtsai: Osteoprotegerin: a regulátor, a protektor és a marker. Orvosi Hetilap 2008 ■ 149. évfolyam, 42. szám ■ 1971–1980.



3. ábra | A carotis-femoralis PWV és az osteoprotegerin szérumszintje közötti lineáris regresszió

Useful WEB pages

- http://davidmlane.com/hyperstat/desc_biv.html
- http://onlinestatbook.com/stat_sim/reg_by_eye/index.html
- http://www.youtube.com/watch?v=CSYTZWFnVpg&feature=related
- http://www.statsoft.com/textbook/basic-statistics/#Correlationsb
- http://people.revoledu.com/kardi/tutorial/Regression/NonLinear/LogarithmicCurve.htm
- http://www.physics.uoguelph.ca/tutorials/GLP/

Review questions and problems

- Graphical examination of the relationship between two continuous variables (scatterplot)
- The meaning and properties of the coefficient of correlation
- Coefficient of correlation and linearity
- The significance of correlation: null hypothesis, t-value, degrees of freedom, decision
- The coefficient of determination
- The meaning of the regression line and its coefficients
- The principle of finding the equation of the regression line
- Hypothesis test for the regression
- Regression using transformations.

Problems

- Based on n=5 pairs of observations, the coefficient of correlation was calculated, its value r=0.7. Is the correlation signficnat at 5% level?
  - Null and alternative hypothesis:…………………
  - t-value of the correlation:........................ Degrees of freedom:.................................
  - Decision (t-value in the table t3,0.05=3.182) …………………..
- On the physics practicals the waist circumference was measured. The measurement was repeated three times. The relationship of the first two measurements were examined by linear regression. Interpret the results below (coefficient of correlation, coefficient of determination, the significance of correlation . Null hypothesis, t-value, p-value, the equation of the regression line)

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.980	.960	.960	2.267

The independent variable is Waist circumference 1.

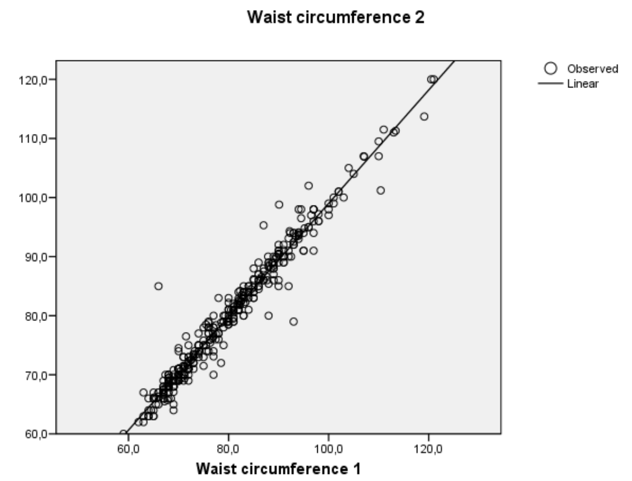
ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	44733.495	1	44733.495	8707.197	.000
Residual	1849.511	360	5.138		
Total	46583.007	361			

The independent variable is Waist circumference 1.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Waist circumference 1	.960	.010	.980	93.312	.000
(Constant)	3.061	.832		3.678	.000



The origin of the word „regression”. Galton: Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute 1886 Vol.15, 246-63

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

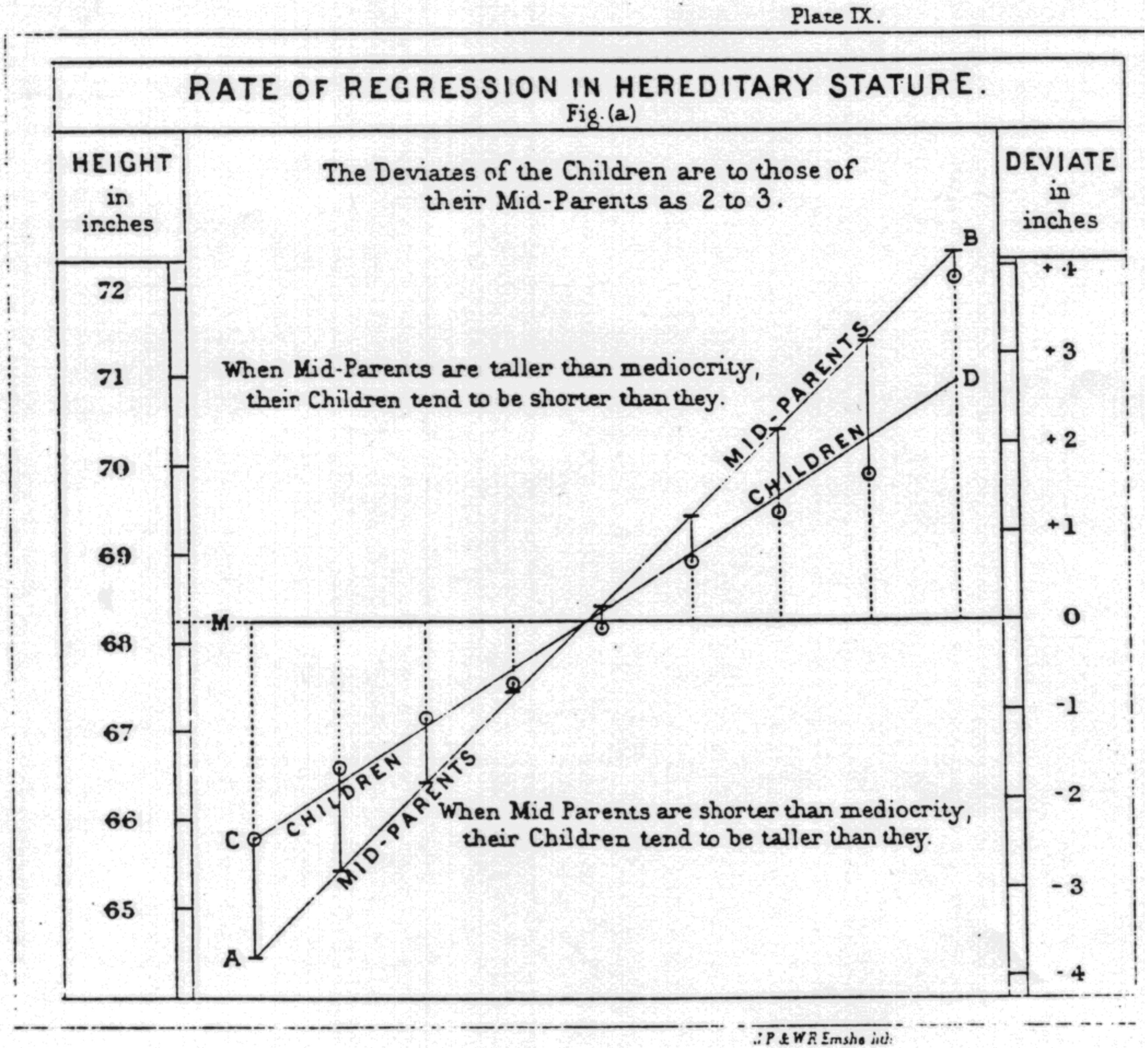


Plate IX.
RATE OF REGRESSION IN HEREDITARY STATURE.
Fig. (a)
HEIGHT in inches
The Deviates of the Children are to those of their Mid-Parents as 2 to 3.
DEVIATE in inches
72
71
70
69
68
67
66
65
+4
+3
+2
+1
0
-1
-2
-3
-4
B
D
M
C
A
When Mid-Parents are taller than mediocrity, their Children tend to be shorter than they.
MID-PARENTS
CHILDREN
When Mid Parents are shorter than mediocrity, their Children tend to be taller than they.
J P & W R Emslie lith