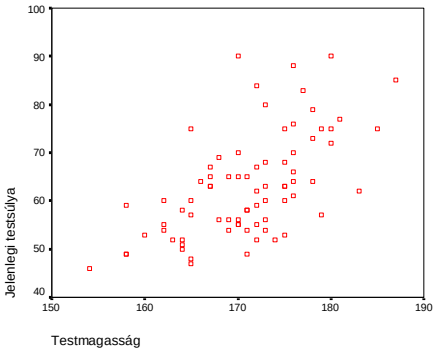
Practice, correlation, regression

Body weights and body heights were measured of a certain group of students. Examine the relationship between body weight in function of body height. Prepare a figure.

What is the appropriate figure to characterise to relationship between two continuous variables?.....



What is your opinion about the
a) direction ……. b) form…….c) strength……
of the relationship?

The coefficient of correlation was found r=0.595
What is expressed in the absolute value of the correlation?.....
Its sign?.....
What is your opinion about the relationship based only on the value of r?.....

Why is it important to examine whether the correlation is significant?.....

Significance of the correlation (n=77)

Null hypothesis: ρ=0 (Greek rho, correlation in the population). There is no correlation in the population between the two variables.
Alternative hypothesis: ρ≠0. There exist a nonzero correlation in the population between the two variables.

t-value of the correlation coefficient: $t = r \cdot \sqrt{\frac{n-2}{1-r^2}}$ ……

Degrees of freedom: df=n-2……..
Critical t-value in the table: …….
Significance……..
Consequences

The equation of the regression is y=0.978x-104.172.
Explain the meaning of the coefficients ……..
What is the "ideal" weight for a person with body = 160 based on the regression equation?
……
Find the above statistics in the output of the SPSS software.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.595[a]	.354	.345	8.71352

a. Predictors: (Constant), Testmagasság

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3119.952	1	3119.952	41.092	.000[a]
	Residual	5694.412	75	75.925		
	Total	8814.364	76			

a. Predictors: (Constant), Testmagasság
b. Dependent Variable: Jelenlegi testsúlya

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-104.172	26.112		-3.989	.000
	Testmagasság	.978	.153	.595	6.410	.000

a. Dependent Variable: Jelenlegi testsúlya

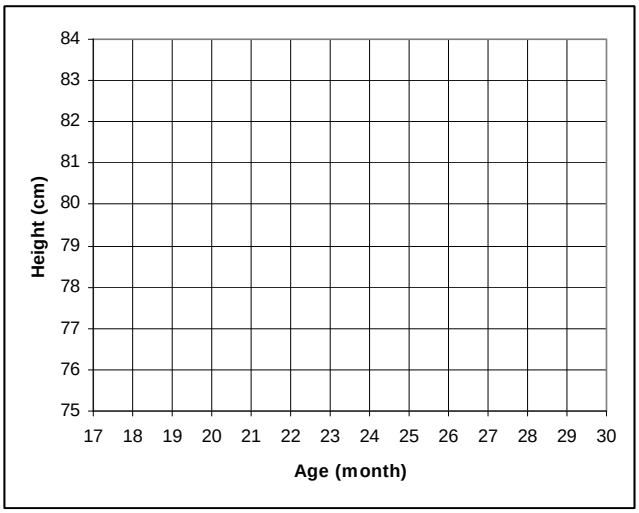
Homework, correlation, regression

Children's Heights

How do children grow? The pattern of growth varies from child to child, so we can best understand the general pattern by following the average height of a number of children. The table below presents the mean heights of a group of children in Kalama, an Egyptian village that is the site of a study of nutrition in developing countries. The data were obtained by measuring the heights of a sample of 161 children from the village each month from 18 to 30 months of age.

Prepare a scatterplot of the data. Age is the explanatory variable, which we plot on the horizontal scale.

Age (months)	Height (cm)
18	76.1
19	77
20	78.1
21	78.2
22	78.8
23	79.7
24	79.9
25	81.1
26	81.2
27	81.8
28	82.8
29	83.5



Following our usual strategy for examining data, we look first for an overall pattern and then for deviations from that pattern.
What is the overall pattern of growth?.....
Are there outliers? ……

The coefficient of correlation was found r=0.994
What is your opinion about the relationship based only on the value of r?.....

Significance of the correlation
Null hypothesis: ……
Alternative hypothesis: ……..
n=……..

t-value of the correlation coefficient: $t = r \cdot \sqrt{\frac{n-2}{1-r^2}}$ =……

Degrees of freedom: df=n-2=……..
Critical t-value in the table: …….
Significance……..
Consequences

The equation of the regression is y = 0.635x + 64.928.
What is the "ideal" height for a child with at age = 10 based on the regression equation?
……

Practice, regression using transformations

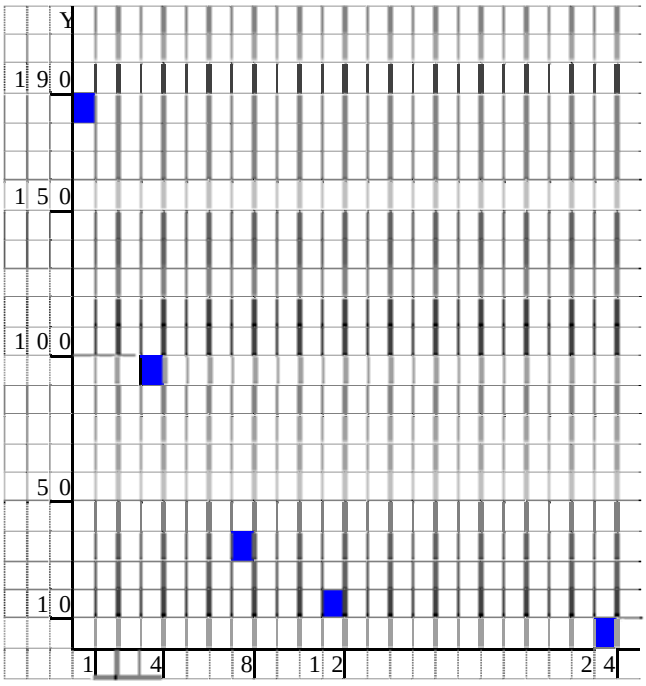
The following table gives the effect of a drug in time. X(time) is measured in hours, y is the effect of the drug blood level). The third column of the table contains the logarithm of y.

Let's prepare a scatterplot of the above data

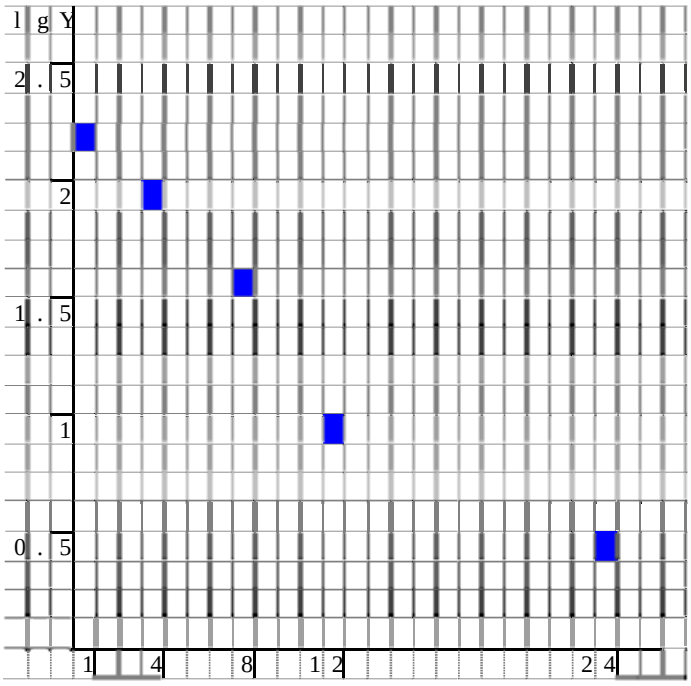
a) The *y* axis is scaled by the original measurements of y

b) The *y* axis is scaled by the logarithm measurements of y

x: time(hours)	*y*	$log_{10}\, y$
1	184.33	2.27
4	87.63	1.94
8	33.05	1.52
12	9.30	.97
24	2.80	.45



a)



b)

Describe the direction of the relationship. Are the variables positively or negatively associated?

.....*positively*..... ..... *positively*.....

Describe the form of the relationship. Is it linear?

.....*no*..... .....*yes*.....

The coefficient of correlation *r*= - 0.789 The coefficient of correlation *r*= -0.970

The *t*-value of corr. *t*= - 2.222 The *t*-value of corr. *t*= - 6.894

Compare the absolute value of computed *t*-values to the *t* in the table ($t_{3,0.05}$=3.182). Is the correlation significant at 5% level?

.....*no*..... .....*yes*.....

Is the correlation significant by the *p*-value computed by SPSS?

p=0.113, significance … *no*..... *p*=0.006, significance … *yes* .....

Explain again the relationship based on computations.

The relationship is not linear *The relationship is linear*

because the correlation not significant *because the correlation is significant*

The equation of the regression line on the logarithmic scale is: $log_{10}\, y = 2.206 - 0.0794 \cdot x$.

Re-transforming the relationship we get the equation of the exponential curve:

$$y = 10^{\,2.206 - 0.0794 \cdot x} = 160.69 \cdot 10^{-0.0794 \cdot x}$$

Homework, regression using transformations

NAME..…….....DATE:.....

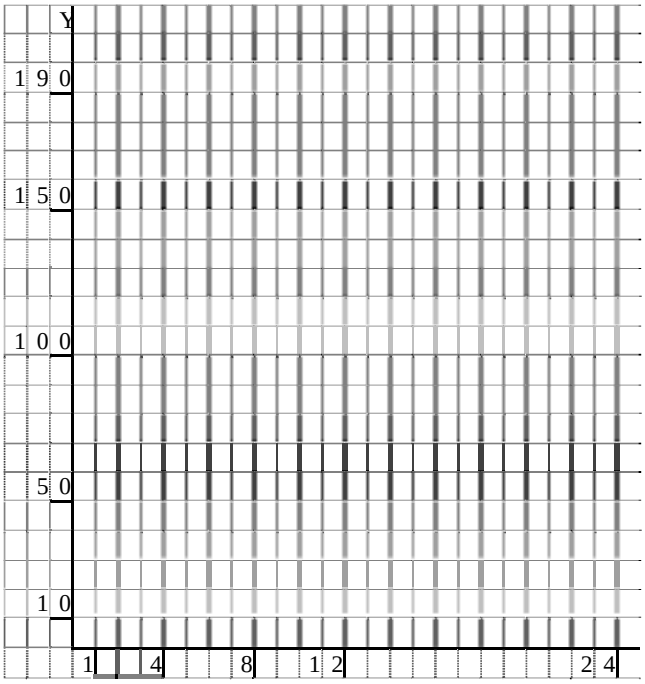
The following table gives the effect of a drug in time. *x*(time) is measured in hours, *y* is the effect of the drug blood level). The third column of the table contains the logarithm of y.

Let's prepare a scatterplot of the above data

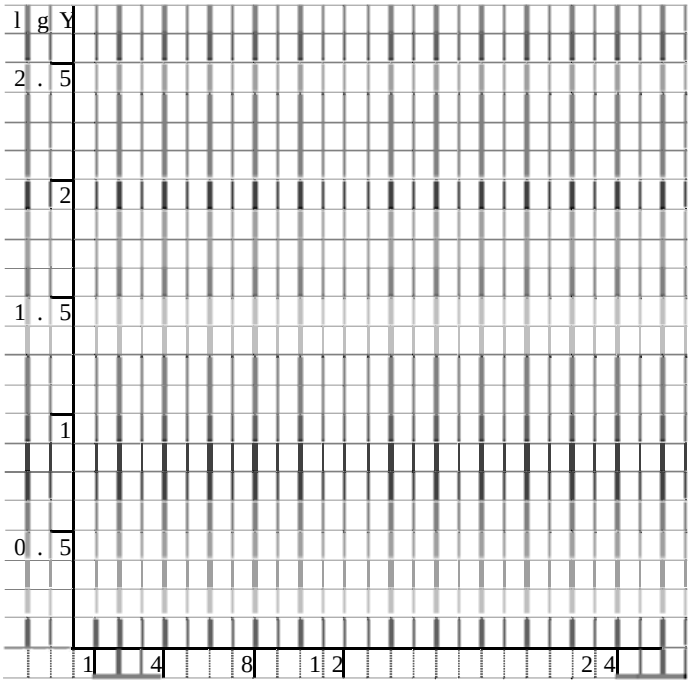
c) The *y* axis is scaled by the original measurements of *y*

d) The *y* axis is scaled by the logarithm measurements of *y*

x: time(hours)	*y*	$log_{10}\, y$
1	184.33	2.27
4	87.63	1.94
8	33.05	1.52
12	9.30	.97
24	2.80	.45



a)



b)

Describe the direction of the relationship. Are the variables positively or negatively associated?

....

Describe the form of the relationship. Is it linear?

....

The coefficient of correlation *r*= - 0.789 The coefficient of correlation *r*= -0.970

The *t*-value of corr. *t*= - 2.222 The *t*-value of corr. *t*= - 6.894

Compare the absolute value of computed *t*-values to the *t* in the table ($t_{3,0.05}$=3.182). Is the correlation significant at 5% level?

....

Is the correlation significant by the *p*-value computed by SPSS?

p=0.113, significance …… *p*=0.006, significance ……

Explain again the relationship based on computations.

....

....

The equation of the regression line on the logarithmic scale is: $log_{10}\, y = 2.206 - 0.0794 \cdot x$.

Re-transforming the relationship we get the equation of the exponential curve: ……

Practice, chi-squared test

2x2 contingency tables.

Example.

Two medicines are being compared regarding a particular side effect; 60 similar patients are split randomly into two groups, one of each group. The results (***observed frequencies***) are presented in the following table (called ***contingency table***):

	Side effects	No side effects	Sum
Drug A	10	20	
Drug B	5	25	
Sum			

We can test the hypothesis that drug type and side effects are independent.

The name of the test:……..

H0: ……
HA:…….

Find the expected frequencies and check the assumption of the test. Formula for expected frequencies: row total*column total/grand total

	Side effects	No side effects	Sum
Drug A			
Drug B			
Sum			

Assumptions:
Number of cells with expected frequencies **<5** must be **less than 20%** of the number of cells!

Assumptions violated?.....

Find the value of the test statistic $X^2 = \sum \frac{(O_i - E_i)^2}{E_i} =$ ……

Find the degrees of freedom…*(rows-1)(column-1)=*…….

The critical value in the table with α=0.05 is χ2=3.84.

Result of the test:…….
Conclusion:……..

Let's denote the frequencies in a 2x2 table by a,b,c,d:

	B1	B2	Sum
A1	a	b	
A2	c	d	
Sum			

Then the formula for the chi-square test can be calculated directly from these frequencies as well:

$\chi 2 = \frac{(ad - bc)^2 (a + b + c + d)}{(a + b)(c + d)(a + c)(b + d)}$ with 1 degrees of freedom.

Homework

Solve the following problems and fill in this sheet by hand. Terminus: next week.
Name:

Two medicines are being compared regarding a particular side effect; 100 similar patients are split randomly into two groups, one of each group. In the group of drug A, 10 side effects were observed while in the group of drug B only 2 side effects were observed. Test whether drug and side effects are independent

	Side effects	No side effects	Sum
A	a=	b=	50
B	c=	d=	50
Sum			100

Null hypothesis:…….

Alternative hypothesis:……..

Find the expected frequencies and check the assumption of the test.

Find the value of the test statistic $\chi 2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}$ ……

Find the degrees of freedom…….

The critical value in the table with α=0.05 is χ2=3.84.

Significance:…….

Conclusion:……..

Homework

Solve the following problems and fill in this sheet by hand. Terminus: next week.
Name:

Two medicines are being compared regarding a particular side effect; 100 similar patients are split randomly into two groups, one of each group. In the group of drug A, 10 side effects were observed while in the group of drug B only 5 side effects were observed. Test whether drug and side effects are independent

	Side effects	No side effects	Sum
A	a=	b=	50
B	c=	d=	50
Sum			100

Null hypothesis:…….

Alternative hypothesis:……..

Find the expected frequencies and check the assumption of the test

Find the value of the test statistic $\chi 2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}$ ……

Find the degrees of freedom…….

The critical value in the table with α=0.05 is χ2=3.84.

Significance:…….

Conclusion:……..