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

Mathematical and Statistical Modelling in Medicine

Author: **Tibor Nyári** PhD

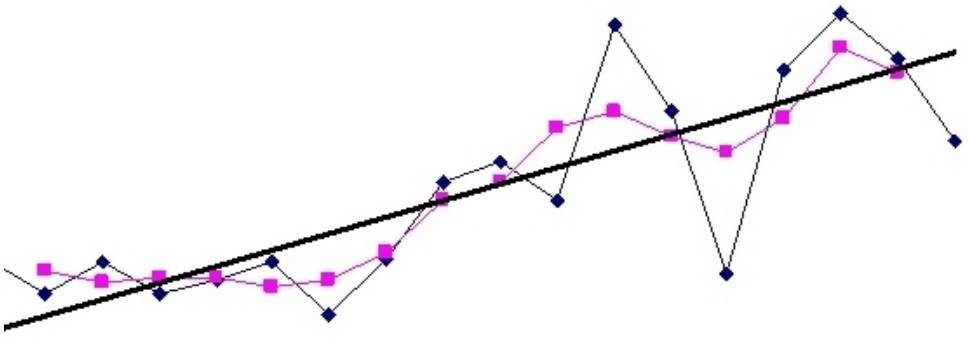
University of Szeged
Department of Medical Physics and Informatics

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

Correlation and regression

Linear regression

Multiple regression



Correlation and prediction

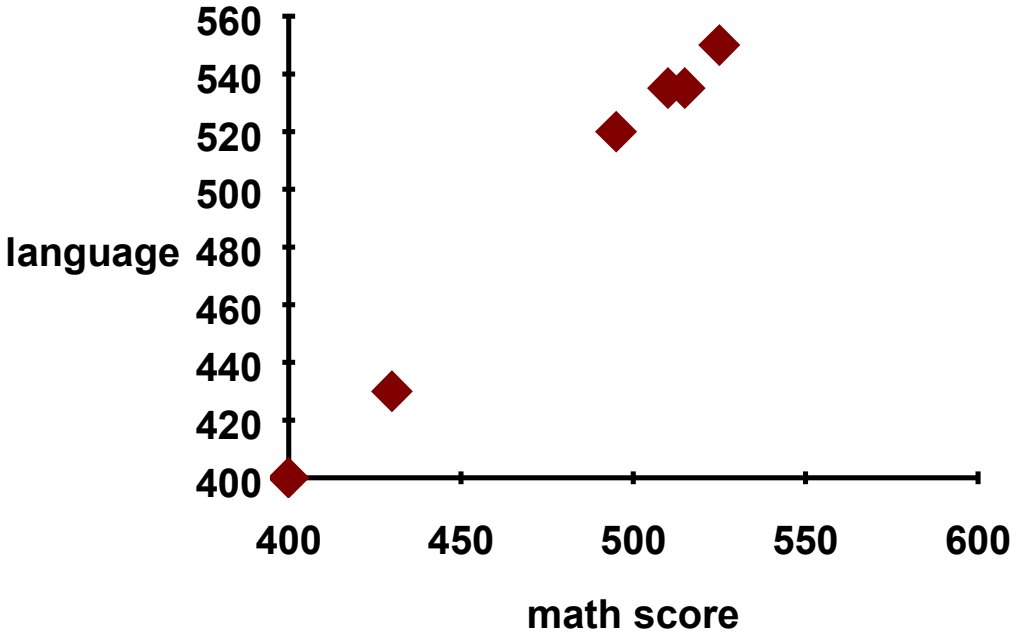
- Relationship between two variables
- It frequently happens that statisticians want to describe with a single number a relationship between two sets of scores.
- A number that measures a relationship between two sets of scores is called a correlation coefficient. There are several correlation coefficients for measuring various types of relationships between different kinds of measurements.
- We will illustrate the basic concepts of correlation by discussing only the Pearson correlation coefficient, which is one of the more widely used correlation coefficients.
  - The statistic is named for its inventor, Karl Pearson (1857-1936), one of the founders of modern statistics. It is denoted by r, and is used to measure what is called the linear relationship between two sets of measurements

To explain how *r* works and what is meant by a linear relationship, we will look at a few over simplified examples. It is unlikely that a real application of the correlation coefficient would be made with so few scores. Imagine that 6 students are given a battery of tests by a vocational guidance counselor with the results shown in the following table

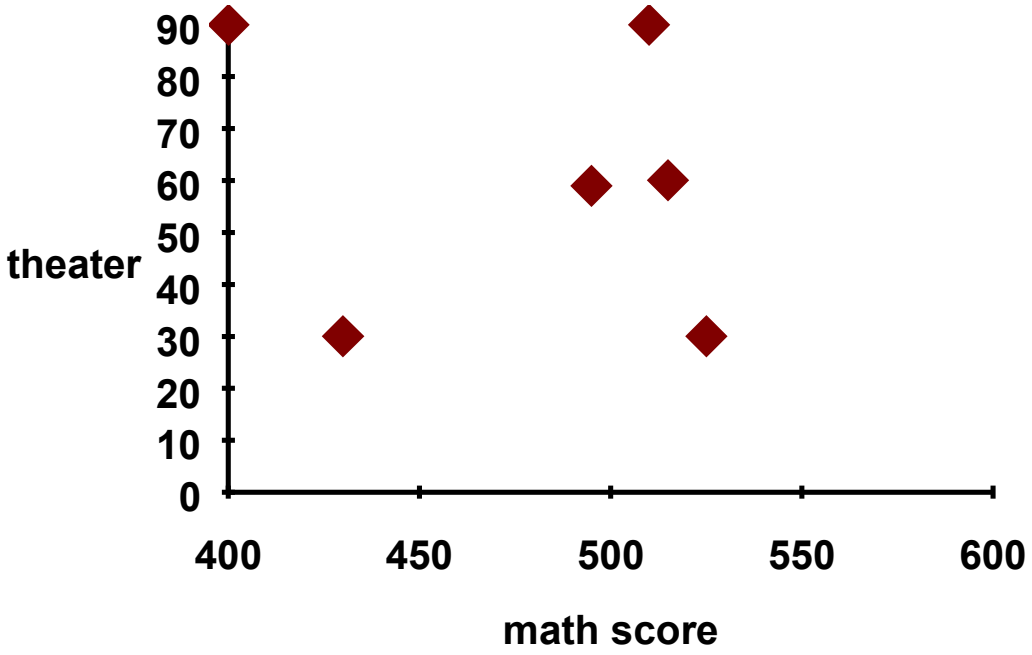
student	in retailing	in theater	math aptitude	language aptitude
Pat	51	30	525	550
Sue	55	60	515	535
Inez	58	90	510	535
Amie	63	50	495	520
Gene	85	30	430	455
Bob	95	90	400	420

- The counselor might want to see if there are any correlation among there set of marks. For example, between math and language.
- Let us draw a graph called scattergram to investigate this relationship.
- We put math scores on the horizontal axis, but that is not important. We could have put it on the vertical axis. After both axes are drawn and labeled, we use one dot for each person.
- You will notice there things about the scattergram.
- 1. There is one point for each pair of scores, 6 points in all.
- 2. The points are arranged approximately in a straight line. When this happens we say that there is a good linear correlation between the two variables.
- 3. The higher numbers in the math column of the table correspond to the higher numbers in the language column. This causes the line to slope up to right. This is called **positive correlation**.

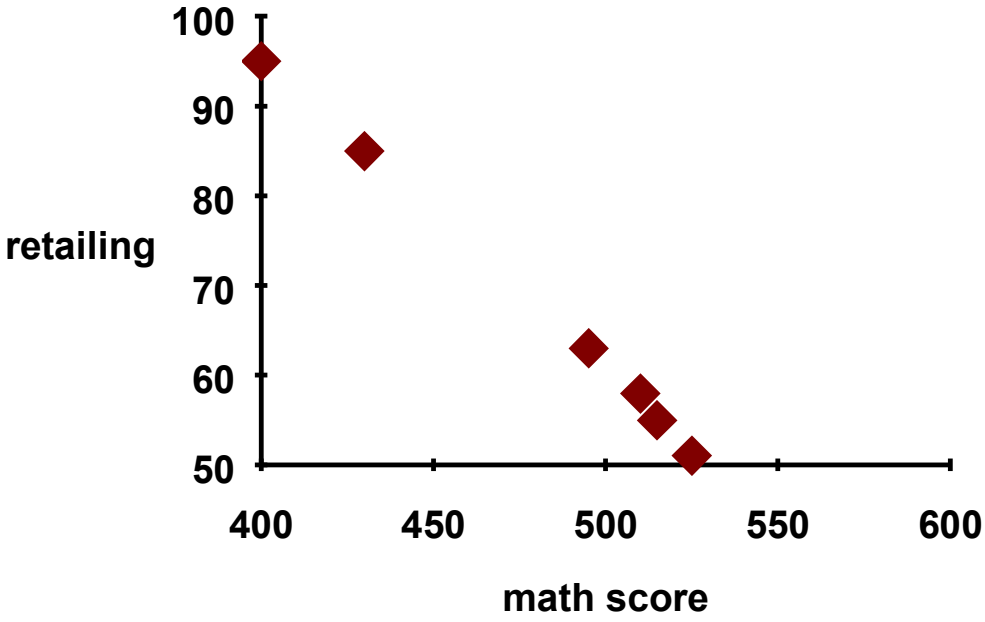
Math aptitude vs Language aptitude



Math aptitude vs interest in theather



Math aptitude vs interest in retailing



Correlation

- You will notice that there is no special tendency for the points to appear in a straight line. We say that there is a little or no correlation between the math scores and the theater-interest scores.
- Also note that it is not necessary for both variables to be scored on the same scale, since the correlation coefficient describes the pattern of the scores, not the actual values.
- Relationship between math scores and retailing-interest scores: there is a tendency for the points to lie in a line that slopes down to the right. This is called negative correlation. (The higher scores in the column for math correspond to the low scores in the column for retailing interest)

Computation of r
(r denotes the correlation coefficient)

- Let us denote the two samples by : $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$.

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

Properties of r

- The value of r is always between **-1 and 1**.
- 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).
- If r is near +1 or -1 we say that we have high correlation. If **r=1**, we say that there is perfect positive correlation. If **r=-1**, then we say that there is a perfect negative correlation

Testing the significance of r

- Suppose that we examined an entire population and computed the correlation coefficient for two variables.
- If this coefficient equaled zero, we would say that there is no correlation between these two variables in this population. Consequently, when we examine a random sample taken from a population, then a sample value of r near zero is interpreted as reflecting no correlation between the variables in the population.
- A sample value of r far from zero (near 1 or -1) indicates that there is some correlation in the population. 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.

The t-test

- H_0: correlation coefficient in population = 0, in notation: $\rho = 0$
- H_a: $\rho \neq 0$
- This test can be carried out by expressing the t statistic in terms of r. It can be proven that the statistic has t-distribution with 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_{table} denotes the value of the table corresponding to n-2 degrees of freedom and probability,
  - if $|t| > t_{table}$, we reject H_0 and state that the population correlation coefficient,ρ is different from 0.
- **Decision using p-value**: if $p < \alpha$ (=0.05) we reject H_0 and state that the population correlation coefficient, ρ is different from 0

Example

- The correlation coefficient between math skill and language skill was found r=0.9989
- H_0: correlation coefficient in population = 0, in notation: $\rho = 0$
- H_a: $\rho \neq 0$
- Let's compute the test statistic

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

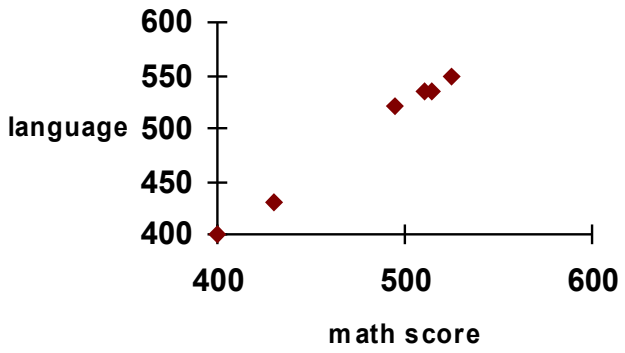
- The critical value in the table is $t_{0.05,4} = 2.776$.
- Because $42.6 > 2.776$, we reject H_0 and claim that there is a significant linear correlation between the two variables at 95 % level.

Prediction based on linear correlation: the linear regression

- If the statistician determines that there is high linear correlation between two variables, we can try to represent the correspondence by an ideal line - a line that best represents the linear correspondence.
- We can then write the formula which determines this line, and use this formula which determines this line, and use this formula to predict, for instance, which value of the Y variable corresponds ideally to any given value of the X variable.

Example

- Let us suppose that math aptitude and language aptitude have a high positive correlation.
- Suppose we have found a formula which predicts language aptitude from scores of math. aptitude.
- Given that value of math aptitude 410 scores, the formula predicts 432.2 scores of language
- language = 1.016 * math + 15.5
- r =0.9989,
- r^2 = 91.7 %)



How to get the formula for the line which is used to get the best point estimates?

- The general equation of a line is y = a + b x.
- We are going 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 estimate 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 xi, that is, by a + b x_i. The approximation is good if the differences are small. These differences can be positive or negative, so let's take its square and summarize

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

Least squares method of fit

- 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

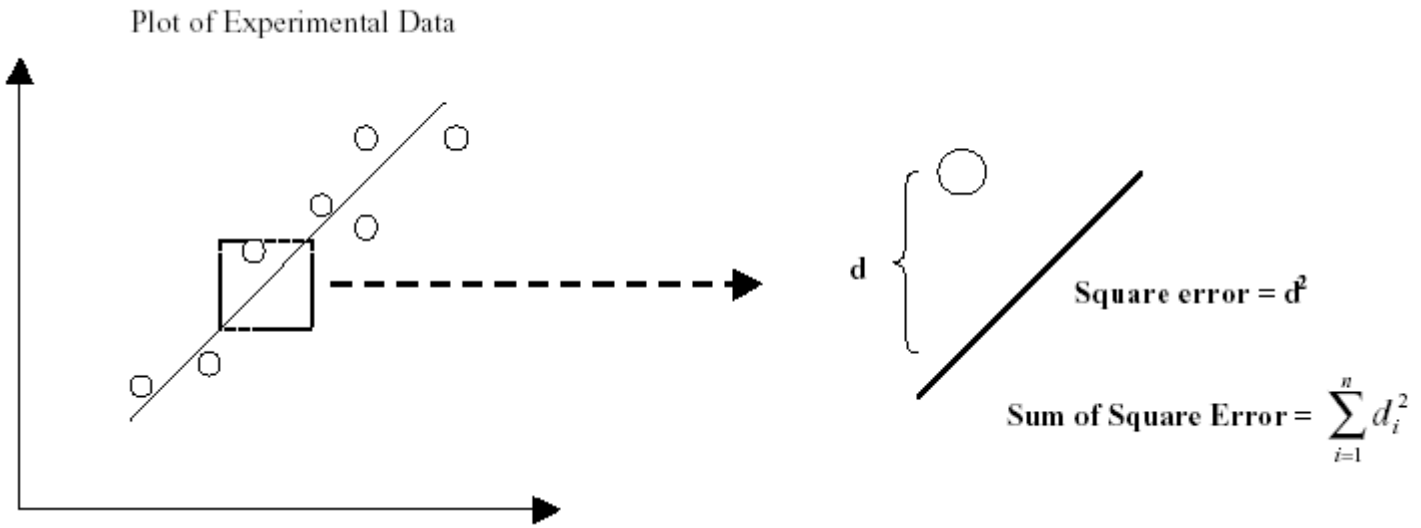
$$\frac{\partial S}{\partial a} = 0, \quad \frac{\partial S}{\partial b} = 0$$

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

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

$$a = \bar{y} - b \cdot \bar{x}$$

Least squares linear regression



Geometrical meaning of *a* and *b*

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

Coefficient of determination and coefficient of correlation

- It can be shown that the ratio of the explained and the total variation is the square of the correlation coefficient

$$r^2 = \frac{Explained}{Total} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\left[\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

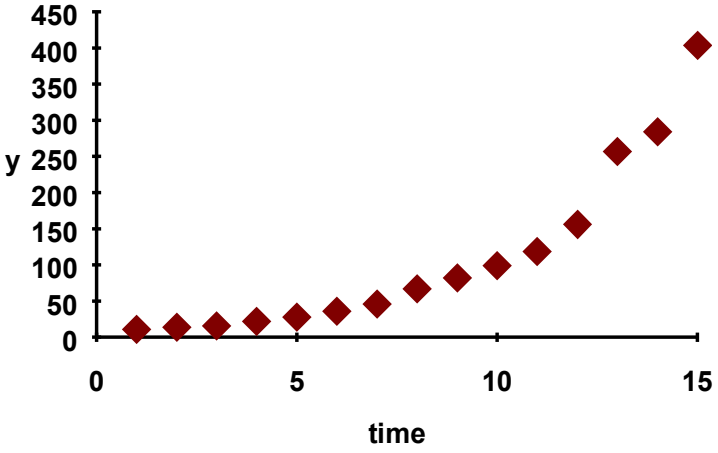
- This is called coefficient of determination. Generally it is multiplied by 100. The square of the correlation coefficient shows the percentages of the total variation explained by the linear regression.
- Model goodness-of-fit statistics

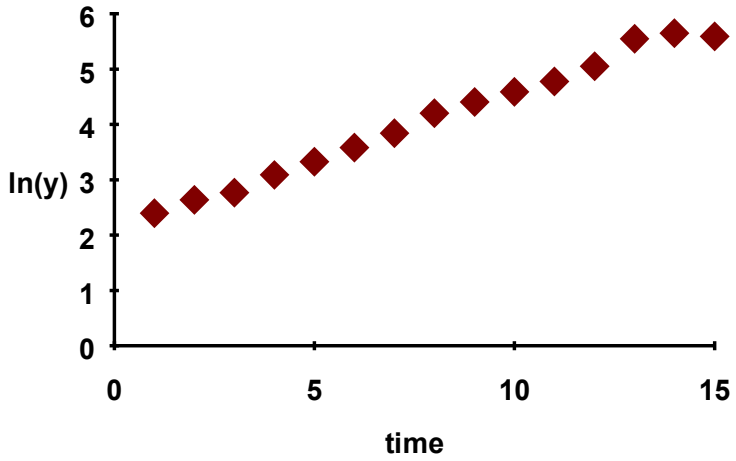
Regression using transformations

- Up to this point, we have suited linear models, when the relationship between x and y had the form
- **y=a +b x**. This model is linear in parameters
- Sometimes, however, useful models are not linear in parameters. Examining the scatterplot of the data shows a functional, but not linear relationship between data. In special cases we are able to find the bets fitting curve to the data.
- For instance, the model
- $y=a\ (b^x)$
- is not linear in parameters. Here the independent variable x enters as an exponent. To apply the technique of estimation and prediction of linear regression, we must transform such a nonlinear model into a linear model that is linear in parameters.
- Some non-linear models can be transformed into a linear model by taking the logarithms on 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
- $y=a\ (b^x)$

  resulted $\log y = \log a + x \log b$
- If we let Y=log y and A = log a and B = log b, the transformed version of the model becomes
- Y=A + B x
- Thus we see that the model with dependent variable log y is linear in the parameters A and B.

Example


450
400
350
300
y 250
200
150
100
50
0
0
5
10
15
time


6
5
4
ln(y) 3
2
1
0
0
5
10
15
time

Multiple linear regression

- The data on next slide show responses, percentages of total calories obtained from complex carbohydrates, for twenty male insulin-dependent diabetics who had been on a high-carbohydrate diet for six months. Compliance with the regime was thought to be related to age (in years), boddy weights (relative to ‘ideal’ weight for height) and other components of the diet, such as the perentage of calories as protein. These other variables are treated as explanatory variables

Carbohydrate	Age	Weight	Protein
33	33	100	14
40	47	92	15
37	49	135	18
27	35	144	12
30	46	140	15
43	52	101	15
34	62	95	14
48	23	101	17
30	32	98	15
38	42	105	14
50	31	108	17
51	61	85	19
30	63	130	19
36	40	127	20
41	50	109	15
42	64	107	16
46	56	117	18
24	61	100	13
35	48	118	18
37	28	102	14

Linear regression model between carbohydrate (Y) and age (X) variables

Regression Statistics	
Multiple R	0,059107
R Square	0,003494
Adjusted R Square	-0,05187
Standard Error	7,778111
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3,817852	3,817852	0,063106	0,804498
Residual	18	1088,982	60,49901		
Total	19	1092,8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	39,21943	6,677034	5,87378	1,46E-05
Age	-0,03509	0,139687	-0,25121	**0,804498**

Linear regression model between carbohydrate (Y) and weight (X) variables

Regression Statistics	
Multiple R	0,4074
R Square	0,165975
Adjusted R Square	0,11964
Standard Error	7,115798
Observations	20

$$Y = -0.186X + 58.164$$

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	181,3776	181,3776	3,58209
Residual	18	911,4224	50,63458	
Total	19	1092,8		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	58,16381	10,98103	5,296754	4,91E-05
Weight	-0,18576	0,098149	-1,89264	**0,074601**

Linear regression model between carbohydrate (Y) and protein (X) variables

Regression Statistics	
Multiple R	0,462889
R Square	0,214266
Adjusted R Square	0,170614
Standard Error	6,906721
Observations	20

$$Y = 1.579X + 12.478$$

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	234,1497	234,1497	4,908511
Residual	18	858,6503	47,7028	
Total	19	1092,8		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	12,47868	11,44351	1,090459	0,289894
Protein	1,579957	0,713133	2,215516	0,039855

Multiple linear regression **model between carbohydrate (Y), weight (X1) and protein (X2) variables**

Regression Statistics	
Multiple R	**0,667414**
R Square	0,445441
Adjusted R Square	0,380199
Standard Error	5,970624
Observations	20

$$Y = -0.22X_1 + 1.824X_2 + 33.13$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	486,7781	243,389	6,827498	**0,006661**
Residual	17	606,0219	35,64835		
Total	19	1092,8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	33,13032	12,57155	2,635341	0,017361
Weight	-0,22165	0,083262	-2,66208	**0,016423**

Multiple linear regression **model between carbohydrate (Y), age (X1), weight (X2) and protein (X3) variables**

Regression Statistics	
Multiple R	**0,693211925**
R Square	0,480542773
Adjusted R Square	0,383144543
Standard Error	5,956419107
Observations	20

$$Y = -0.228X_2 + 1.958X_3 + 36.96$$

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	525,1371	175,0457	4,933794	**0,012971**
Residual	16	567,6629	35,47893		
Total	19	1092,8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	36,96005591	13,07128	2,827577	0,012131
Age	-0,113676356	0,109325	-1,0398	0,313893
Weight	-0,228017362	0,083289	-2,73767	**0,014599**

Model selection

- AGE was not correlated with carbohydrate neither in simple nor in multiple linear regression models.
- Thus choose model of carbohydrate (Y), weight (X1) and protein (X2) variables as
- Y=-0.22*X1+1.82*X2