Regression and Correlation

Topics Covered:

- Dependent and independent variables.
- Scatter diagram.
- Correlation coefficient.
- Linear Regression line.

by Dr.I.Namestnikova

Introduction

Regression analysis is used to model and analyse numerical data consisting of values of an **independent variable** X (the variable that we fix or choose deliberately) and **dependent variable** Y.

The main purpose of finding a relationship is that the knowledge of the relationship may enable events to be predicted and perhaps controlled.

Correlation coefficient

To measure the strength of the linear relationship between X and Y the **sample correlation coefficient** r is used.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{xy}}},$$

$$S_{xy} = n\sum xy - \sum x \sum y,$$

$$S_{xx} = n\sum x^2 - \left(\sum x\right)^2, \quad S_{yy} = n\sum y^2 - \left(\sum y\right)^2$$

Where x and y observed values of variables X and Y respectively.

Important notes

1. If the calculated r value is **positive** then the slope will **rise** from left to right on the graph. If the calculated value of r is **negative** the slope will **fall** from left to right.

2. The r value will **always** lie between -1 and $+1$. If you have an r value outside of this range you have made an error in the calculations.

3. Remember that a correlation does not necessarily demonstrate a causal relationship. A significant correlation only shows that two factors vary in a related way (positively or negatively).

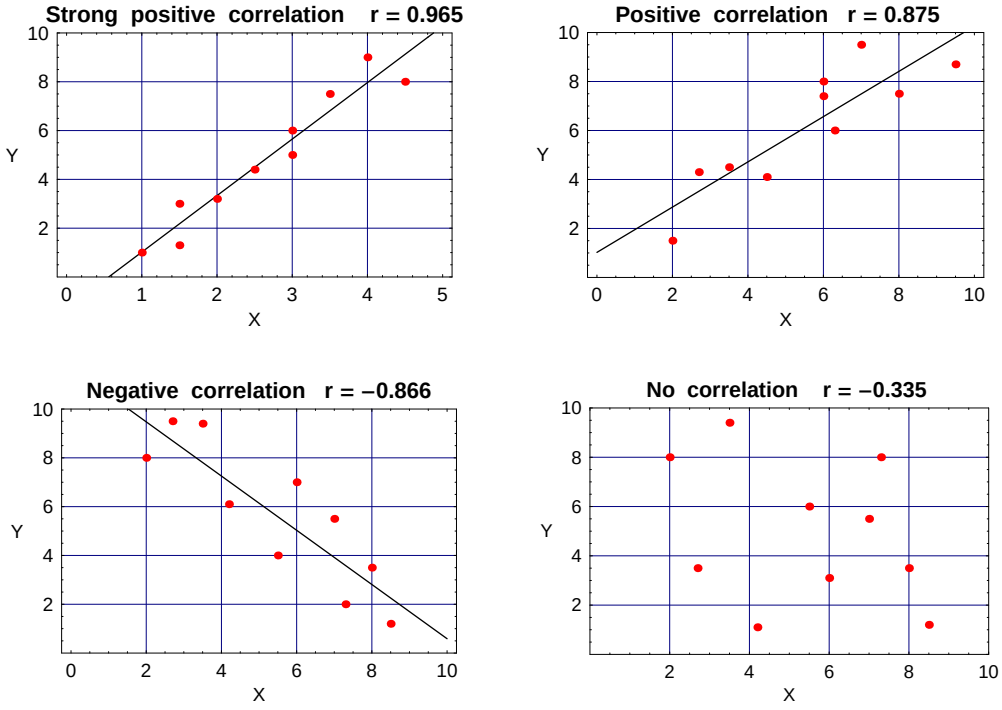
4. The formula above can be rewritten as

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \qquad \sigma_x = \sqrt{\frac{1}{n}\sum x^2 - \bar{x}^2}, \quad \sigma_y = \sqrt{\frac{1}{n}\sum y^2 - \bar{y}^2}$$

$$\sigma_{xy} = \frac{1}{n}\sum xy - \bar{x}\bar{y}, \quad \bar{x} = \frac{1}{n}\sum x, \quad \bar{y} = \frac{1}{n}\sum y$$

Scatter Diagrams

Scatter diagrams are used to graphically represent and compare two sets of data. The **independent variable** is usually plotted on the X axis. The **dependent variable** is plotted on the Y axis. By looking at a scatter diagram, we can see whether there is any connection (**correlation**) between the two sets of data. A scatter plot is a useful summary of a set of bivariate data (two variables), usually drawn before working out a linear correlation coefficient or fitting a regression line. It gives a good visual picture of the relationship between the two variables, and aids the interpretation of the correlation coefficient or regression model.



From plots one can see that if the more the points tend to cluster around a straight line and the higher the correlation (the stronger the **linear relationship** between the two variables). If there exists a random scatter of points, there is no relationship between the two variables (very low or zero correlation).

Very low or zero correlation could result from a non-linear relationship between the variables. If the relationship is in fact non-linear (points clustering around a curve, not a straight line), the correlation coefficient will not be a good measure of the strength.

A scatter plot will also show up a **non-linear relationship** between the two variables and whether or not there exist any outliers in the data.

Example 1

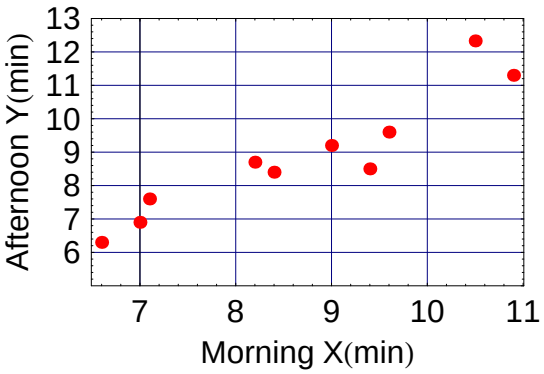
Determine on the basis of the following data whether there is a relationship between the time, in minutes, it takes a person to complete a task in the morning X and in the late afternoon Y.

Morning (x) (min)	8.2	9.6	7.0	9.4	10.9	7.1	9.0	6.6	8.4	10.5
Afternoon (y) (min)	8.7	9.6	6.9	8.5	11.3	7.6	9.2	6.3	8.4	12.33

Solution

The data set consists of $n = 10$ observations.

Step 1.



To construct the scatter diagram for the given data set to see any correlation between two sets of data.

From the scatter diagram we can conclude that it is likely that there is a linear relationship between two variables.

Step 2. Set out a table as follows and calculate all required values $\sum x,\ \sum y,\ \sum x^2,\ \sum y^2,\ \sum xy.$

Morning (x) (min)	Afternoon (y) (min)	x^2	y^2	xy
8.2	8.7	67.24	75.69	71.34
9.6	9.6	92.16	92.16	92.16
7.0	6.9	49.00	47.61	48.30
9.4	8.5	88.36	72.25	79.90
10.9	11.3	118.81	127.69	123.17
7.1	7.6	50.41	57.76	53.96
9	9.2	81.00	84.64	82.80
6.6	6.3	43.56	39.69	41.58
8.4	8.4	70.56	70.56	70.56
10.5	12.33	110.25	151.29	129.465
$\sum x = 86.7$	$\sum y = 88.8$	$\sum x^2 = 771.35$	$\sum y^2 = 819.34$	$\sum xy = 792.92$

Step 3.

Calculate

$$S_{xy} = n\sum xy - \sum x \sum y = 10 \times 792.92 - 86.7 \times 88.8 = 230.24$$

$$S_{xx} = n\sum x^2 - (\sum x)^2 = 10 \times 771.35 - (86.7)^2 = 196.61$$

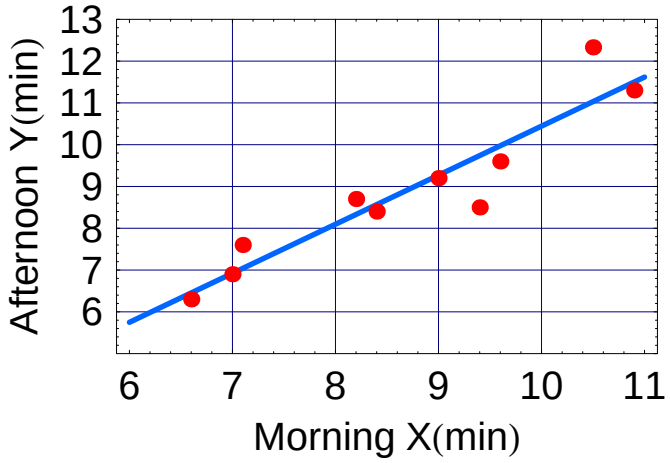
$$S_{yy} = n\sum y^2 - (\sum y)^2 = 10 \times 819.34 - 88.8^2 = 307.96$$

Step 4

Finally we obtain **correlation coefficient** r

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{xy}}} = \frac{230.24}{\sqrt{196.61 \times 307.96}} = 0.9357$$

The correlation coefficient is closed to 1 therefore the linear relationship exists between the two variables.



It would be tempting to try to fit a line to the data we have just analysed - producing an **equation** that shows the relationship, The method for this is called **linear regression**. By using linear regression method the line of best fit is

Regression equation: $y = 1.171x - 1.273$

This line is shown in blue on the above graph. How to find this equation one can see in the next section.

Linear regression analysis: fitting a regression line to the data

When a scatter plot indicates that there is a **strong linear relationship** between two variables (confirmed by **high correlation coefficient**), we can fit a straight line to this data which may be used to predict a value of the dependent variable, given the value of the independent variable.

Recall that the equation of a **regression line** (straight line) is

$$y = a + bx$$

$$b = \frac{S_{xy}}{S_{xx}} \qquad a = \bar{y} - b\bar{x} = \frac{\sum_i y_i - b\sum_i x_i}{n}$$

To illustrate the technic, let us consider the following data.

Example 2

Suppose that we had the following results from an experiment in which we measured the growth of a cell culture (as optical density) at different pH levels.

pH	3	4	4.5	5	5.5	6	6.5	7	7.5
Optical density	0.1	0.2	0.25	0.32	0.33	0.35	0.47	0.49	0.53

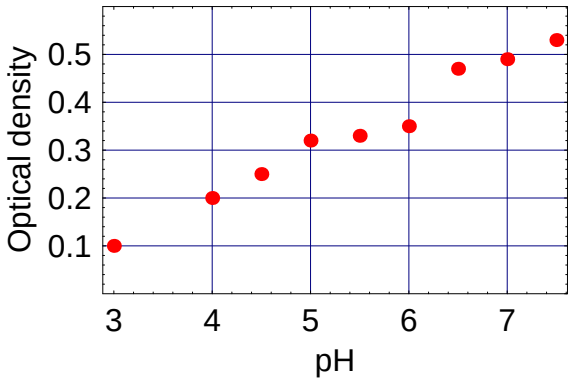
Find the equation to fit these data.

Solution

We can follow the same procedures for correlation, as before.

The data set consists of $n = 9$ observations.

Step 1. To construct the scatter diagram for the given data set to see any correlation between two sets of data.



These results suggest a linear relationship.

Step 2. Set out a table as follows and calculate all required values $\sum x,\ \sum y,\ \sum x^2,\ \sum y^2,\ \sum xy.$

pH (x)	Optical density(y)	x^2	y^2	xy
3	0.1	9	0.01	0.3
4	0.2	16	0.04	0.8
4.5	0.25	20.25	0.0625	1.125
5	0.32	25	0.1024	1.6
5.5	0.33	30.25	0.1089	1.815
6	0.35	36	0.1225	2.1
6.5	0.47	42.25	0.2209	3.055
7	0.49	49	0.240	3.43
7.5	0.53	56.25	0.281	3.975
$x = 49$	$y = 3.04$	$x^2 = 284$	$y^2 = 1.1882$	$xy = 18.2$
$\bar{x} = 5.444$	$\bar{y} = 0.3378$			

Step 3.

Calculate

$$S_{xy} = n\sum xy - \sum x \sum y = 9 \times 18.2 - 49 \times 3.04$$
$$= 163.8 - 148.96 = 14.84.$$

$$S_{xx} = n\sum x^2 - \left(\sum x\right)^2 = 2556 - 2401 = 155.$$

$$S_{yy} = n\sum y^2 - \left(\sum y\right)^2 = 10.696 - 9.242 = 1.454$$

Step 4.

Finally we obtain **correlation coefficient** r

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{xy}}} = \frac{14.84}{\sqrt{155 \times 1.454}} = 0.989$$

The correlation coefficient is closed to 1 therefore it is likely that the linear relationship exists between the two variables. To verify the correlation r we can run a hypothesis test.

Step 5. **A hypothesis test**

● **Hypothesis** about the **population** correlation coefficient ρ

1. The **null hypothesis** $H_0 : \rho = 0.$
2. The **alternative hypothesis** $H_A : \rho \neq 0.$

● **Distribution of test statistic.** When H_0 is true ($\rho = 0$) and the assumption are met, the appropriate test statistic is distributed as **Student's t distribution**

(the **test statistics** is $t = r\sqrt{\dfrac{n-2}{1-r^2}}$ with $n - 2$ degrees of freedom).

The number of degrees of freedom is two less than the number of points on the graph ($9 - 2 \equiv 7$ degrees of freedom in our example because we have 9 points).

● **Decision rule.** If we let $\alpha = 0.025, \quad 2\alpha = 0.05$, the critical values of t in the present example are ± 2.365 (e.g. see John Murdoch, "Statistical tables for students of science, engineering, psychology, business, management, finance", 1998, Macmillan, 79 p., Table 7).

If, from our data, we compute a value of t that is either greater or equal to 2.365 or less than or equal to -2.365, we will reject the null hypothesis.

● **Calculation of test statistic.**

$$t = 0.989\sqrt{\frac{7}{1-0.989^2}} = 17.69$$

● **Statistical decision.** Since the computed value of the test statistic exceed the critical value of t, we **reject** the null hypothesis.

● **Conclusion.** We conclude that there is a **very highly significant positive correlation** between pH and growth as measured by optical density of the cell culture.

Step 6.

Now we want to use **regression analysis** to find the line of best fit to the data. We have done nearly all the work for this in the calculations above.

The **regression equation** for y on x is: $y = bx + a$ where b is the **slope** and a is the **intercept** (the point where the line crosses the $y-$axis)

We calculate b and a as:

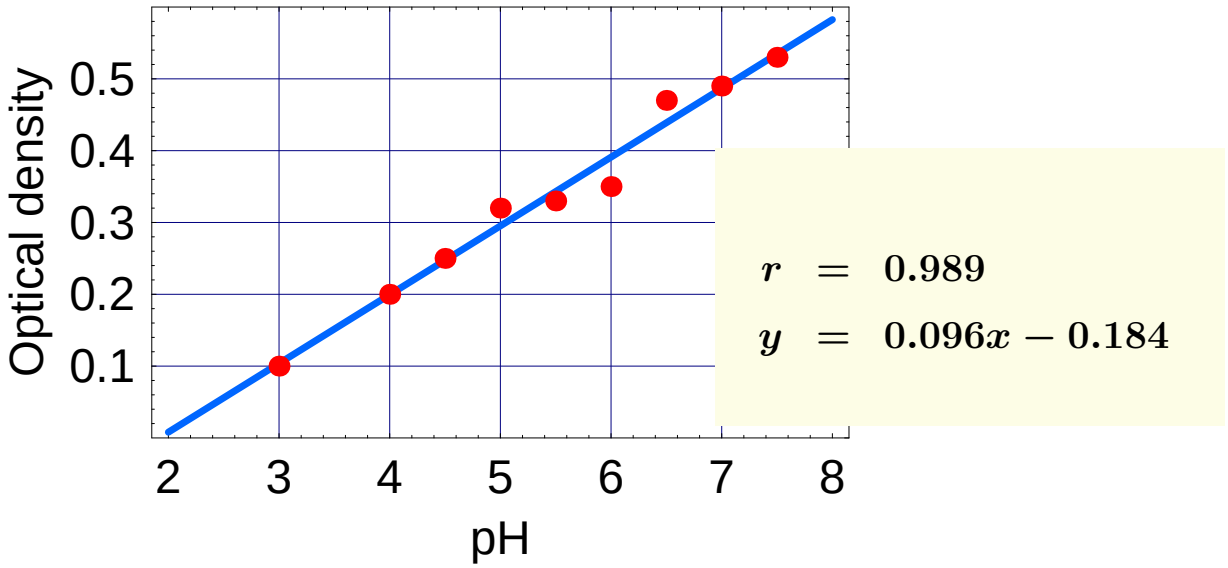
$$b = \frac{S_{xy}}{S_{xx}} = \frac{14.84}{155} = 0.096$$

$$\begin{aligned} a &= \bar{y} - b\bar{x} \\ &= 0.3378 - 0.096 \cdot 5.444 \\ &= 0.3378 - 0.5226 = -0.184 \end{aligned}$$

So the equation for the line of best fit is:

$$y = 0.096x - 0.184$$

(to 3 decimal places).



Example 3

The tensile strength of a cable for upper-limb prosthesis was investigated. Stainless steel cable is commonly available in three sizes (diameters): $\mathbf{1.19}$ mm, $\mathbf{1.59}$ mm and $\mathbf{2.38}$ mm. Four tests were performed for each diameter size and the results are given in the table below

Cable diameter (mm)	Cable cross area (mm^2)	Tensile strength (KN)
$\mathbf{1.19}$	$\mathbf{1.1122}$	$\mathbf{1.27}$
$\mathbf{1.19}$	$\mathbf{1.1122}$	$\mathbf{1.45}$
$\mathbf{1.19}$	$\mathbf{1.1122}$	$\mathbf{1.43}$
$\mathbf{1.19}$	$\mathbf{1.1122}$	$\mathbf{1.36}$
$\mathbf{1.59}$	$\mathbf{1.9856}$	$\mathbf{2.20}$
$\mathbf{1.59}$	$\mathbf{1.9856}$	$\mathbf{2.56}$
$\mathbf{1.59}$	$\mathbf{1.9856}$	$\mathbf{2.38}$
$\mathbf{1.59}$	$\mathbf{1.9856}$	$\mathbf{2.45}$
$\mathbf{2.38}$	$\mathbf{4.4488}$	$\mathbf{4.58}$
$\mathbf{2.38}$	$\mathbf{4.4488}$	$\mathbf{5.03}$
$\mathbf{2.38}$	$\mathbf{4.4488}$	$\mathbf{5.67}$
$\mathbf{2.38}$	$\mathbf{4.4488}$	$\mathbf{4.39}$

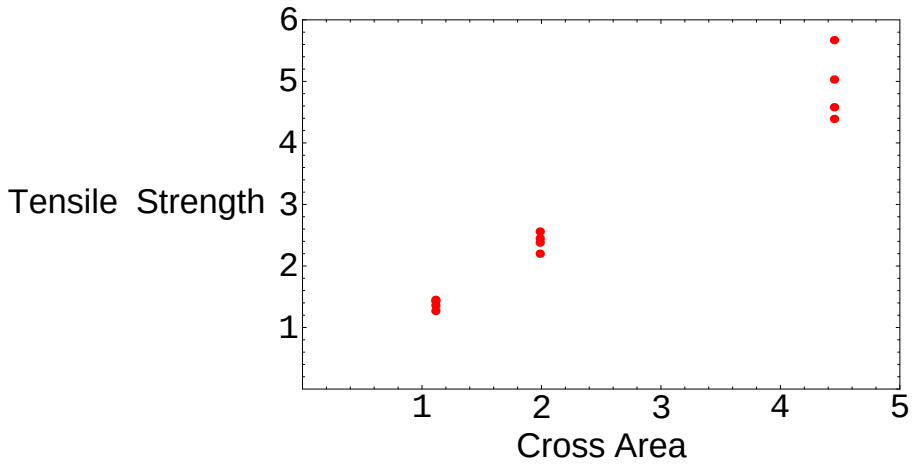
Let be $\boldsymbol{X}$ = Cable cross area ($\boldsymbol{mm^2}$), $\boldsymbol{Y}$ =Tensile strength (KN)

1. Construct a scatter diagram to illustrate these results.
2. Calculate the correlation coefficient for the data and comment on the result.
3. Obtain the least squares estimates for the sample regression equation of ”Tensile strength ” on ”Cable cross area”.
4. Estimate the tensile strength for cable with cross area $\mathbf{3}$.
5. Comment on the suitability of using the sample regression equation to estimate the tensile strength for cable with cross area $\mathbf{5}$.

Solution

We can follow the same procedure, as before. The data set consists of $n = 12$ observations.

1. To construct the scatter diagram for the given data set to see any correlation between two sets of data.



These results may suggest a linear relationship.

2. Set out a table as follows and calculate all required values $\sum x, \quad \sum y, \quad \sum x^2, \quad \sum y^2, \quad \sum xy.$

Cable cross area (x)	Tensile strength (y)	x^2	y^2	xy
1.1122	1.27	1.237	1.613	1.413
1.1122	1.45	1.237	2.103	1.613
1.1122	1.43	1.237	2.045	1.591
1.1122	1.36	1.237	1.850	1.513
1.9856	2.20	3.943	4.840	4.368
1.9856	2.56	3.943	6.554	5.083
1.9856	2.38	3.943	5.664	4.726
1.9856	2.45	3.943	6.003	4.867
4.4488	4.58	19.792	20.976	20.376
4.4488	5.03	19.792	25.301	22.378
4.4488	5.67	19.792	32.149	25.225
4.4488	4.39	19.792	19.272	19.530
$x = 30.19$	$y = 34.77$	$x^2 = 99.89$	$y^2 = 128.37$	$xy = 112.68$
$\bar{x} = 2.5158$	$\bar{y} = 2.8975$			

Calculate

$$S_{xy} = n\sum xy - \sum x \sum y = 12 \times 112.68 - 30.19 \times 34.77$$
$$= 1352.16 - 1049.71 = 302.454$$

$$S_{xx} = n\sum x^2 - \left(\sum x\right)^2 = 1198.68 - 911.436 = 287.244$$

$$S_{yy} = n\sum y^2 - \left(\sum y\right)^2 = 1540.44 - 1208.95 = 331.487$$

Finally we obtain **correlation coefficient** r

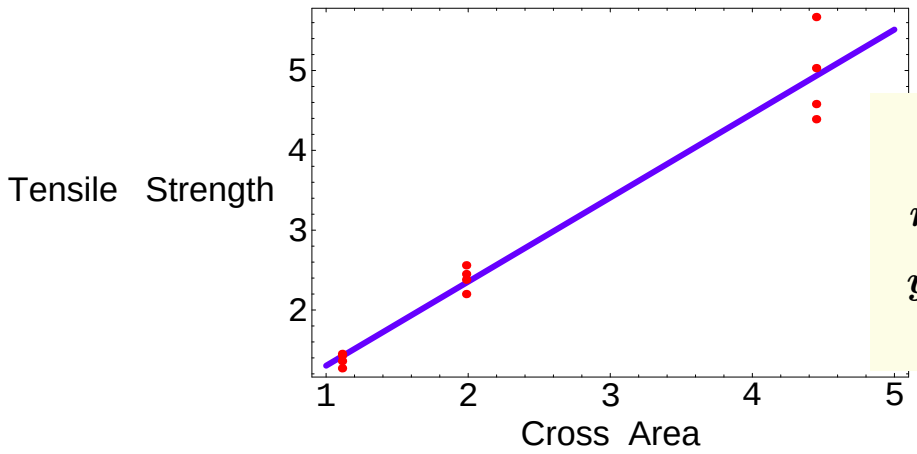
$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{xy}}} = \frac{302.454}{\sqrt{287.244 \times 331.487}} = 0.9802$$

The correlation coefficient is closed to 1 therefore it is likely that the linear relationship exists between the two variables. To verify the correlation we can run a hypothesis test.

3. We calculate b and a as:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{302.454}{287.244} = 1.053$$
$$a = \bar{y} - b\bar{x} = 2.8975 - 1.053 \cdot 2.5158 = 0.249$$

So the equation for the line of best fit is: $y = 1.053x + 0.249$ (to 3 decimal places).



$$r = 0.9802$$
$$y = 1.053x + 0.249$$

4. To estimate the tensile strength for cable with cross area 3 we need to substitute $x = 3$ into the regression equation $y(3) = 1.053 \times 3 + 0.249 = 3.408$

5. The sample regression equation is not suitable to estimate the tensile strength for cable with cross area 5 because this value is outside the test range $(1.1 \leqslant x \leqslant 4.45)$.

Example 4

To make a prosthesis we must know the force that acts in it as the person moves. This force depends on the adjacent musculature. Records of the variation with time of the force in hip joints during level walking show two maximum values in the stance phase of each cycle. A total of 16 subjects took part in the study. An indication of the variation in average of the maxima hip joint force with body weight W and the ratio of stride length L to height H are given in the table below.

$\frac{WL}{H}$ (kg)	33.6	41.4	43.3	44.1	45.6	46.0	49.8	53.2	53.8	54.7	55.2	58.3	59.7	62.2	66.3	72.1
Mean hip joint force F (kN)	1.400	1.300	1.050	1.320	1.200	1.107	1.560	2.070	2.200	1.730	1.870	2.520	2.370	2.640	2.380	2.850

Let be $X = \frac{WL}{H}$ and mean hip joint force Y

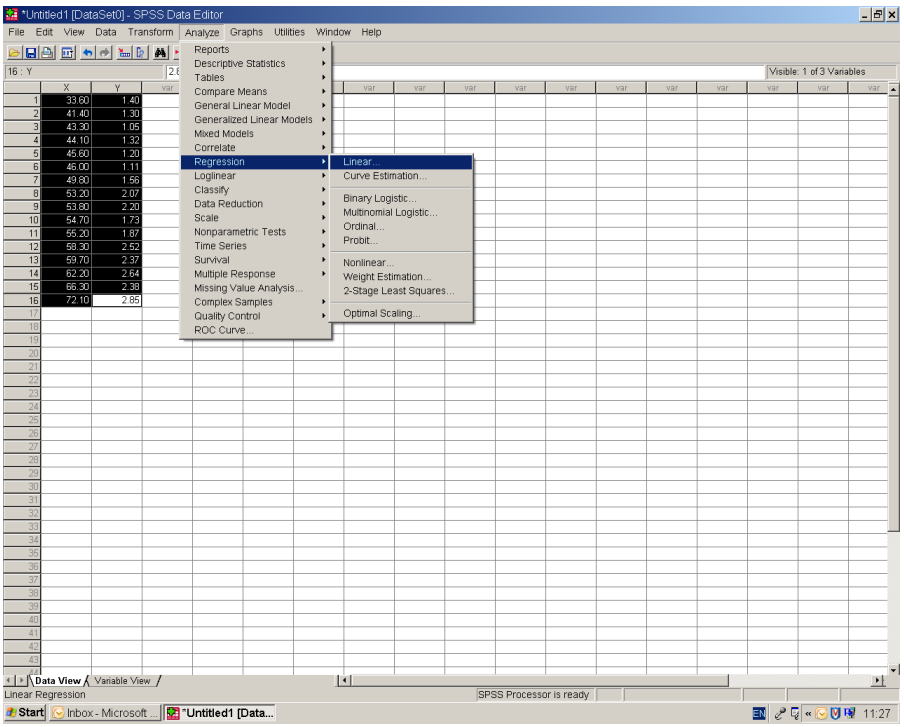
- Construct a scatter diagram to illustrate these results.
- Calculate the correlation coefficient for the data and comment on the result.
- Obtain the least squares estimates for the sample regression equation of "$\frac{WL}{H}$ " on "Mean hip joint force".

Solution

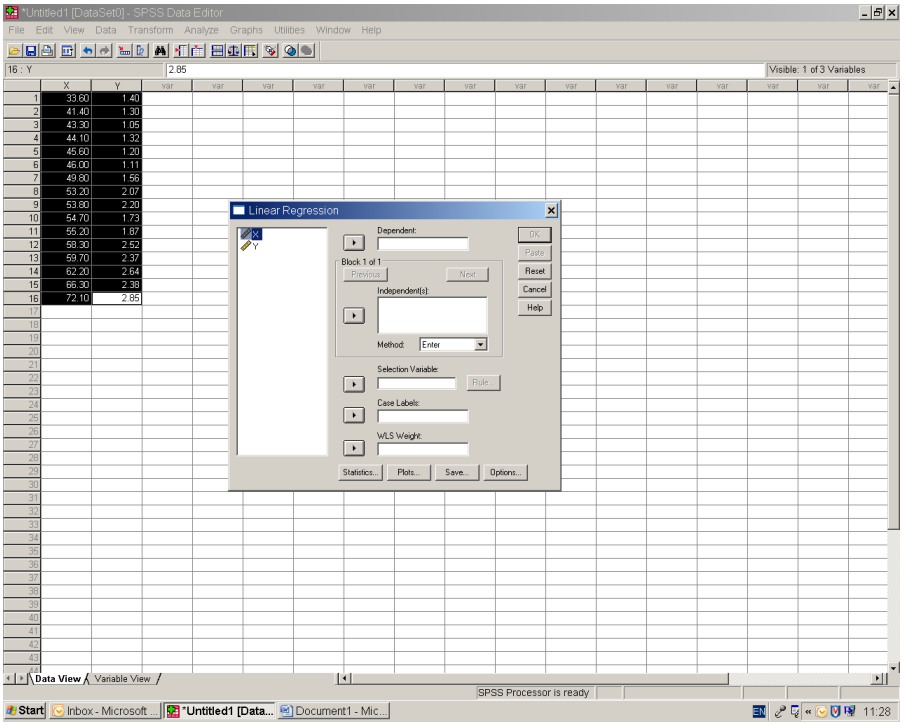
We can follow the same procedure, as before. Another option is to use SPSS or Excel.

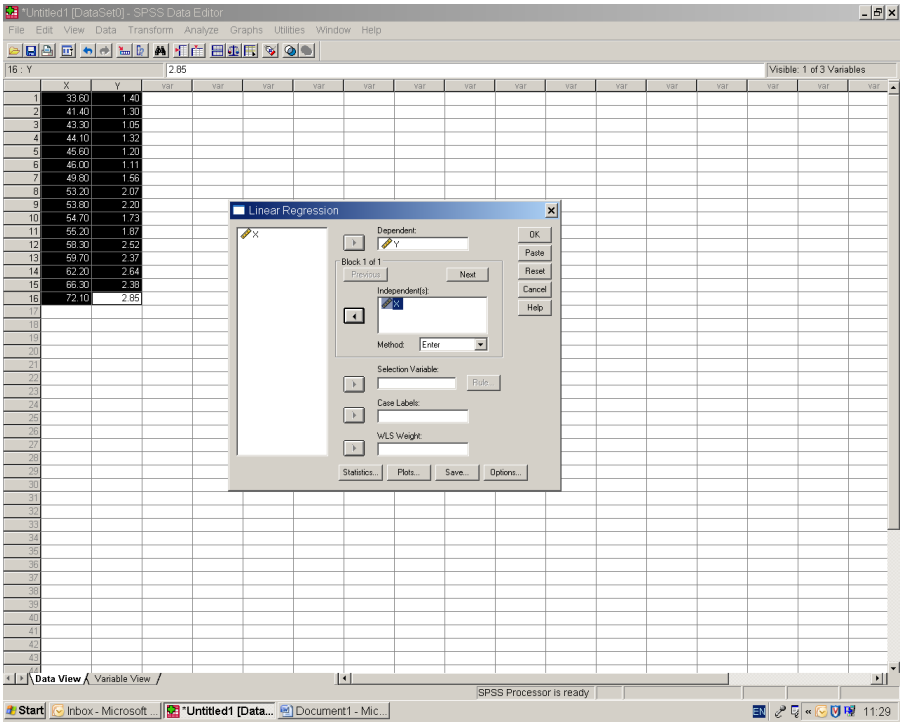
Example of using SPSS for Regression Analysis

Open **Data.sav**. Click **Analyze**, click **Regression** and click **Linear**.

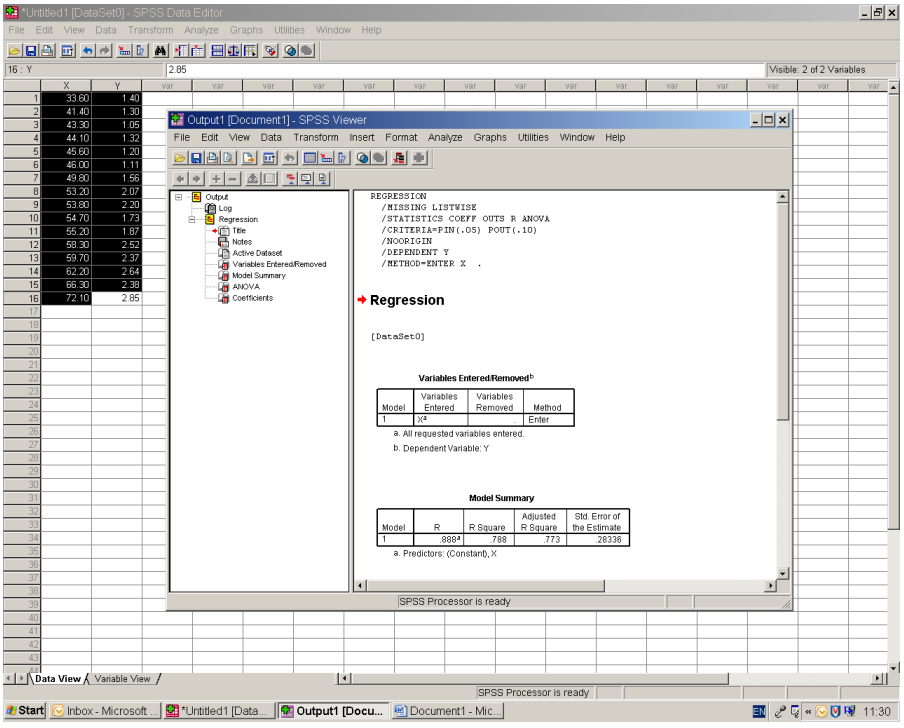


In the **Linear Regression** dialog box click ***X*** and click an arrow to move ***X*** into **Independent Variable** list. In the **Linear Regression** dialog box click ***Y*** and click an arrow to move ***Y*** into **Dependent Variable** list.

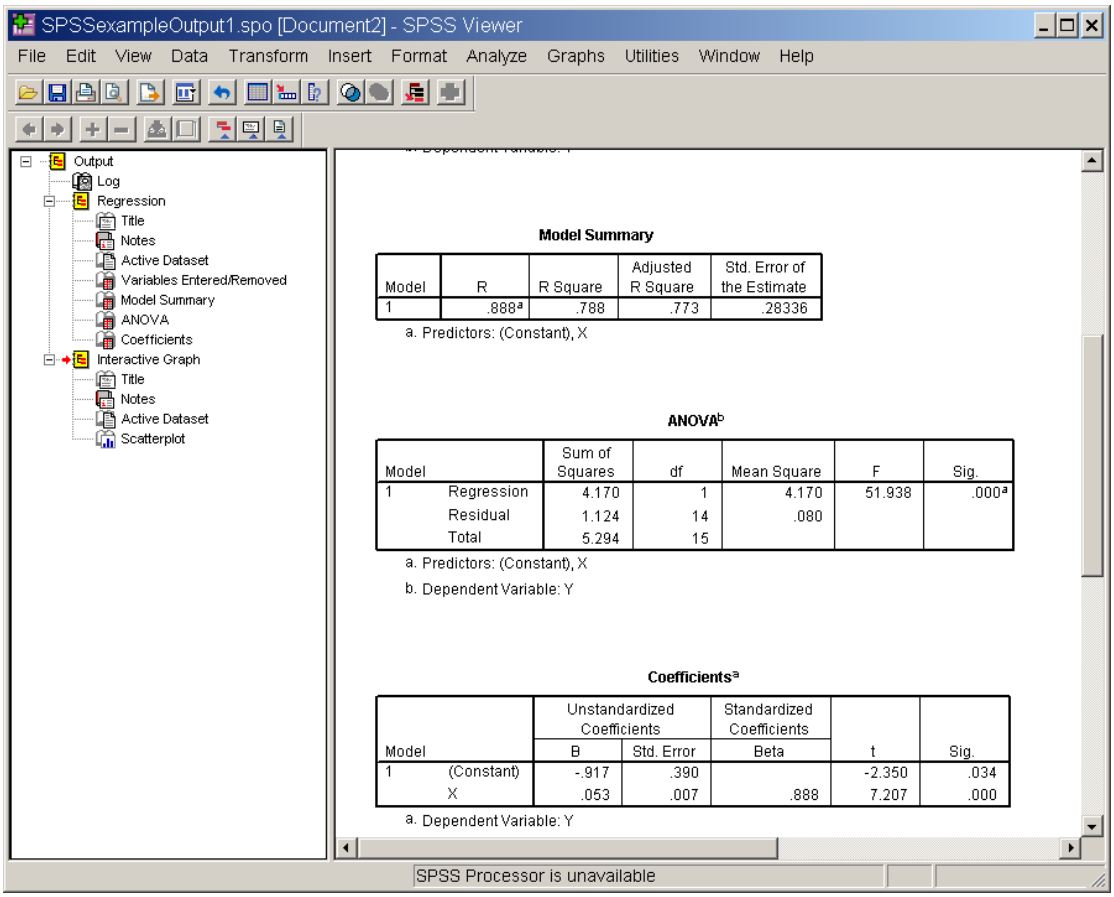




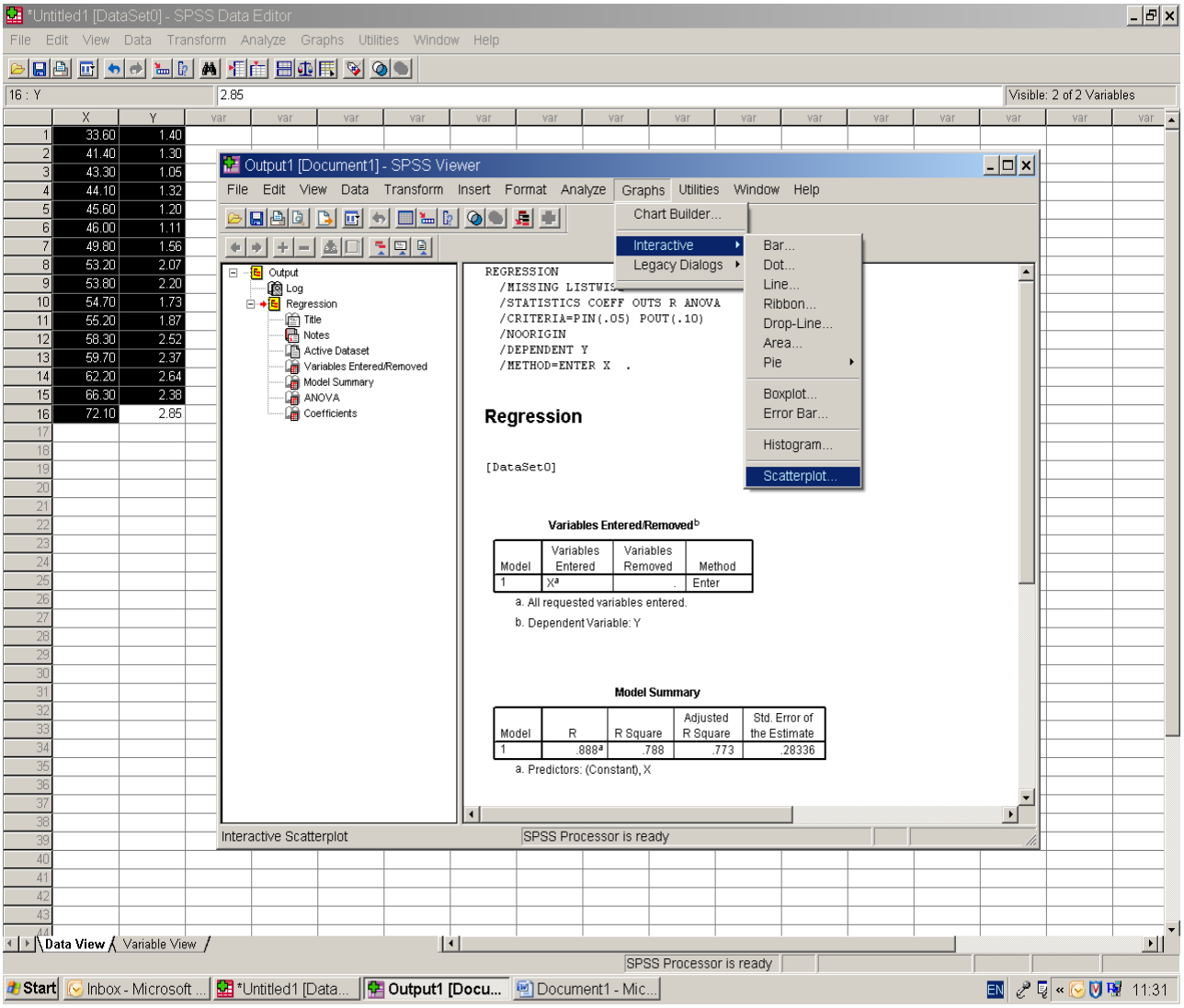
Then click **OK**

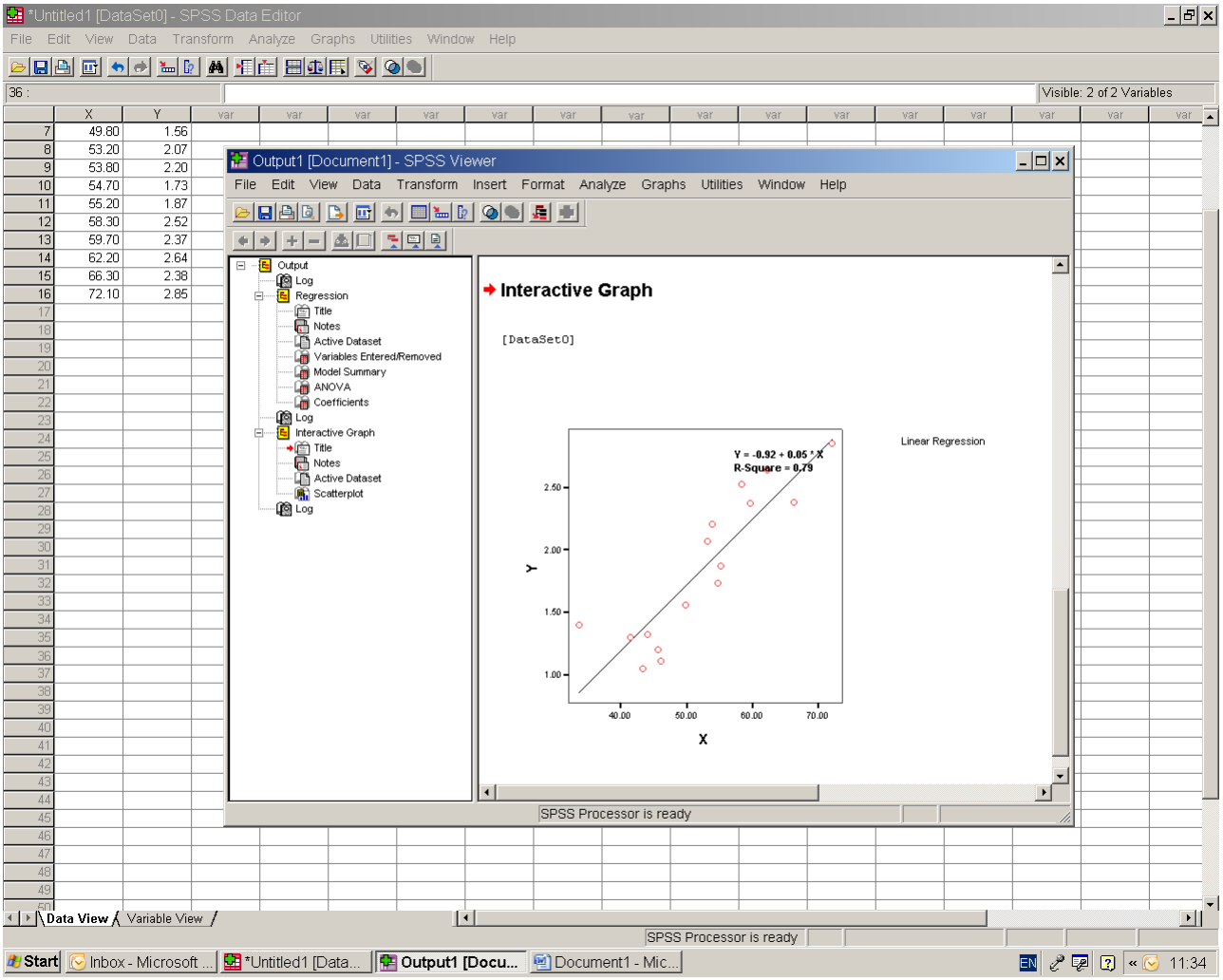


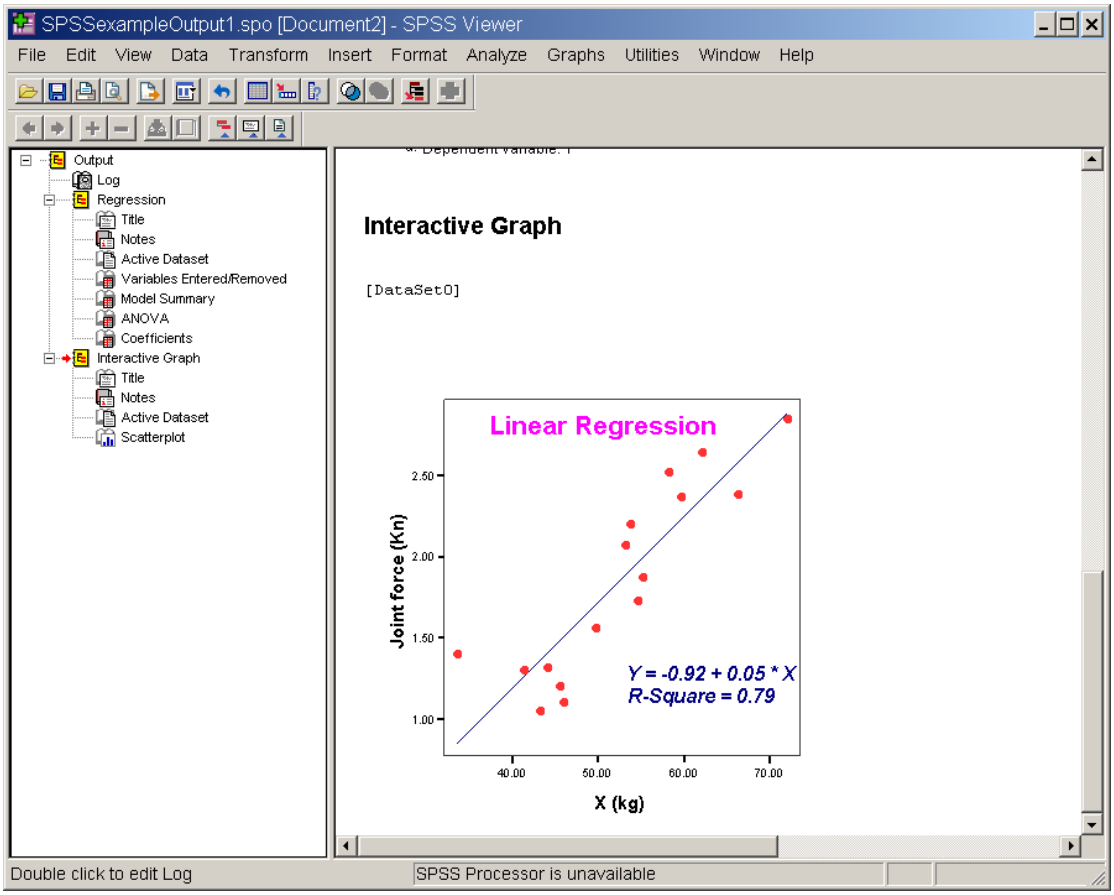
You can use the **Output Viewer** to browse resultes.



You can add a scatter plot. Click **Graphs**, click **Iteractive,** than click **Scatterplot**. The scatter plot can be edited, just double click on it.


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
16 : Y
2.85
Visible: 2 of 2 Variables
X Y
1 33.60 1.40
2 41.40 1.30
3 43.30 1.05
4 44.10 1.32
5 45.60 1.20
6 46.00 1.11
7 49.80 1.56
8 53.20 2.07
9 53.80 2.20
10 54.70 1.73
11 55.20 1.87
12 58.30 2.52
13 59.70 2.37
14 62.20 2.64
15 66.30 2.38
16 72.10 2.85
Output1 [Document1] - SPSS Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help
Chart Builder...
Interactive
Legacy Dialogs
Bar...
Dot...
Line...
Ribbon...
Drop-Line...
Area...
Pie
Boxplot...
Error Bar...
Histogram...
Scatterplot...
Output
Log
Regression
Title
Notes
Active Dataset
Variables Entered/Removed
Model Summary
ANOVA
Coefficients
REGRESSION
/MISSING LISTWIS
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT Y
/METHOD=ENTER X .
Regression
[DataSet0]
Variables Entered/Removed[b]
Model | Variables Entered | Variables Removed | Method
1 | X[a] | . | Enter
a. All requested variables entered.
b. Dependent Variable: Y
Model Summary
Model | R | R Square | Adjusted R Square | Std. Error of the Estimate
1 | .888[a] | .788 | .773 | .28336
a. Predictors: (Constant), X
Interactive Scatterplot
SPSS Processor is ready
Data View
Variable View
SPSS Processor is ready
Start
Inbox - Microsoft ...
*Untitled1 [Data...
Output1 [Docu...
Document1 - Mic...
EN
11:31


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Visible: 2 of 2 Variables
X
Y
49.80 1.56
53.20 2.07
53.80 2.20
54.70 1.73
55.20 1.87
58.30 2.52
59.70 2.37
62.20 2.64
66.30 2.38
72.10 2.85
Output1 [Document1] - SPSS Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help
Output
Log
Regression
Title
Notes
Active Dataset
Variables Entered/Removed
Model Summary
ANOVA
Coefficients
Log
Interactive Graph
Title
Notes
Active Dataset
Scatterplot
Log
Interactive Graph
[DataSet0]
Linear Regression
Y = -0.92 + 0.05 * X
R-Square = 0.79
2.50
2.00
1.50
1.00
Y
40.00
50.00
60.00
70.00
X
SPSS Processor is ready
Data View
Variable View
SPSS Processor is ready
Start
Inbox - Microsoft ...
*Untitled1 [Data...
Output1 [Docu...
Document1 - Mic...
EN
11:34


SPSSexampleOutput1.spo [Document2] - SPSS Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help
Output
Log
Regression
Title
Notes
Active Dataset
Variables Entered/Removed
Model Summary
ANOVA
Coefficients
Interactive Graph
Title
Notes
Active Dataset
Scatterplot
Interactive Graph
[DataSet0]
Linear Regression
Y = -0.92 + 0.05 * X
R-Square = 0.79
Joint force (Kn)
2.50
2.00
1.50
1.00
40.00
50.00
60.00
70.00
X (kg)
Double click to edit Log
SPSS Processor is unavailable