Describing Bivariate Data

Very often researchers are interested in more than just one variable that can be measured during their investigation. When two variables are measured on a single experimental unit, the resulting data are called bivariate data.

Methods for graphing bivariate data depend whether the variables are qualitative or quantitative.

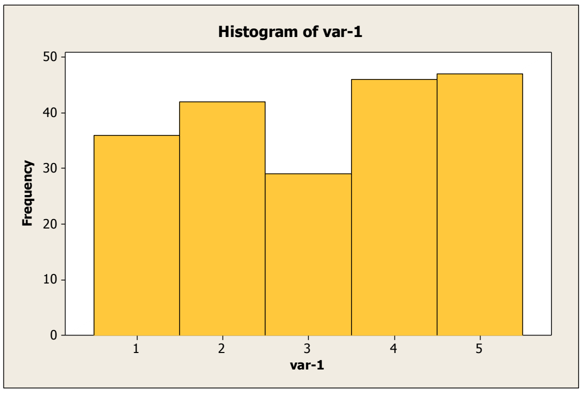
Graphs for Qualitative Variables

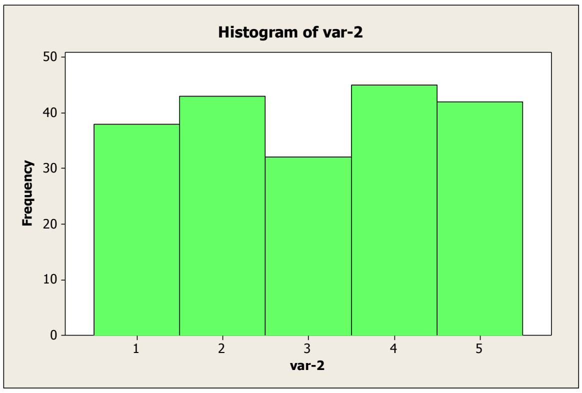
var-1	var-2
2	3
4	5
4	1
3	2
3	5
2	2
.	.

Rows: var-1 Columns: var-2

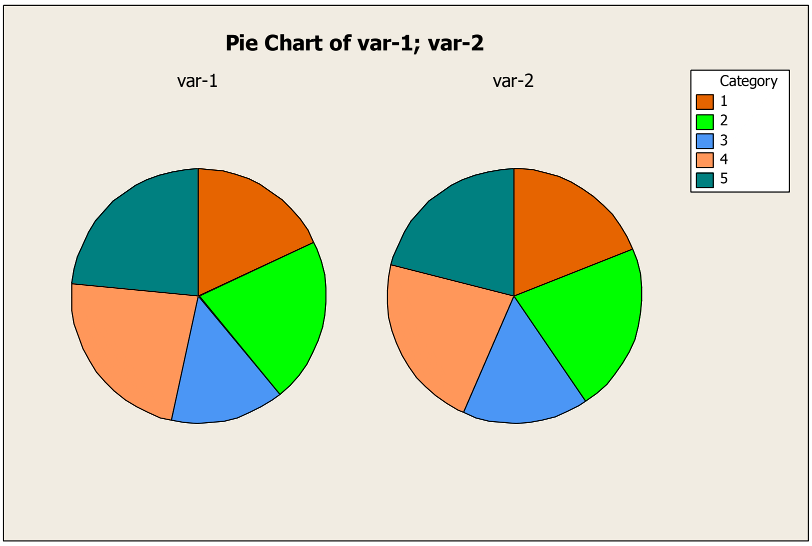
	1	2	3	4	5	All
1	9	7	8	7	5	36
2	8	10	5	11	8	42
3	6	7	3	4	9	29
4	6	9	8	11	12	46
5	9	10	8	12	8	47
All	38	43	32	45	42	200

Separate bar charts

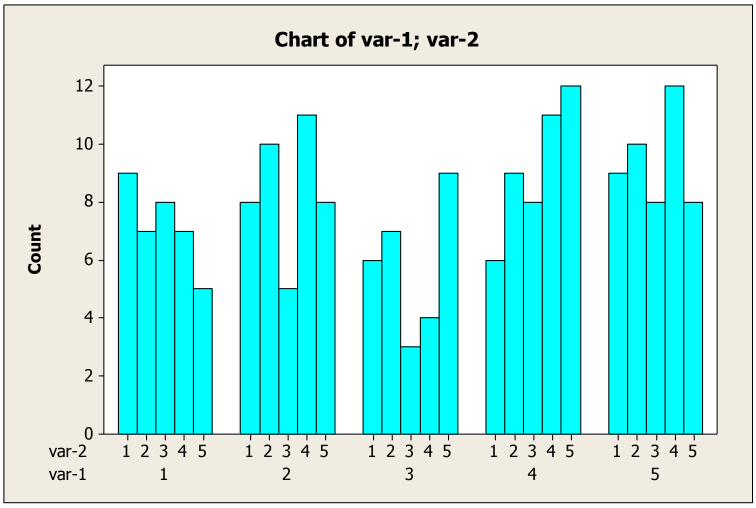




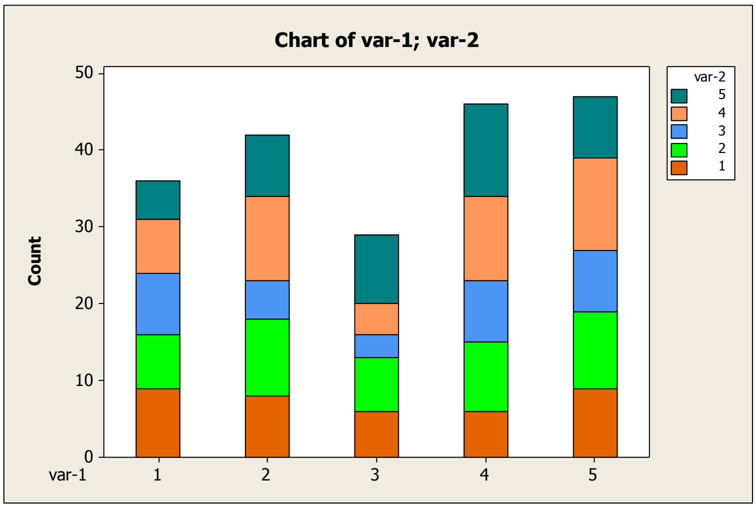
MTB > hist c1 MTB > hist c2



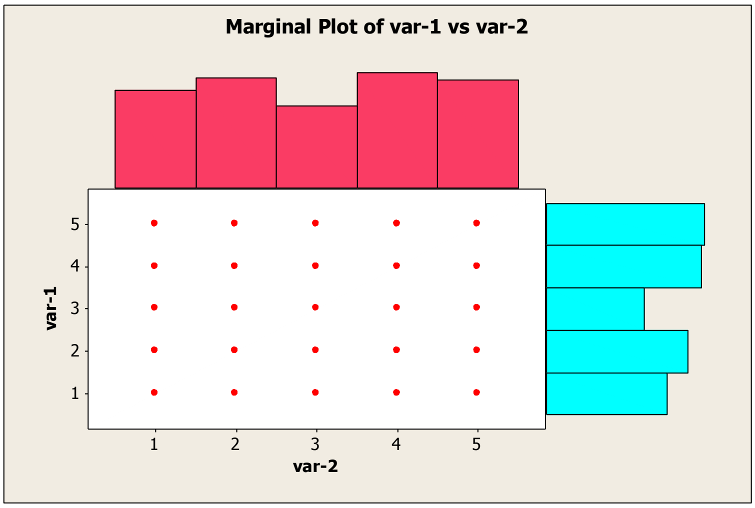
Comparative bar charts




```
MTB > Chart 'var-1';
SUBC>   Group 'var-2';
SUBC>   Bar.
```



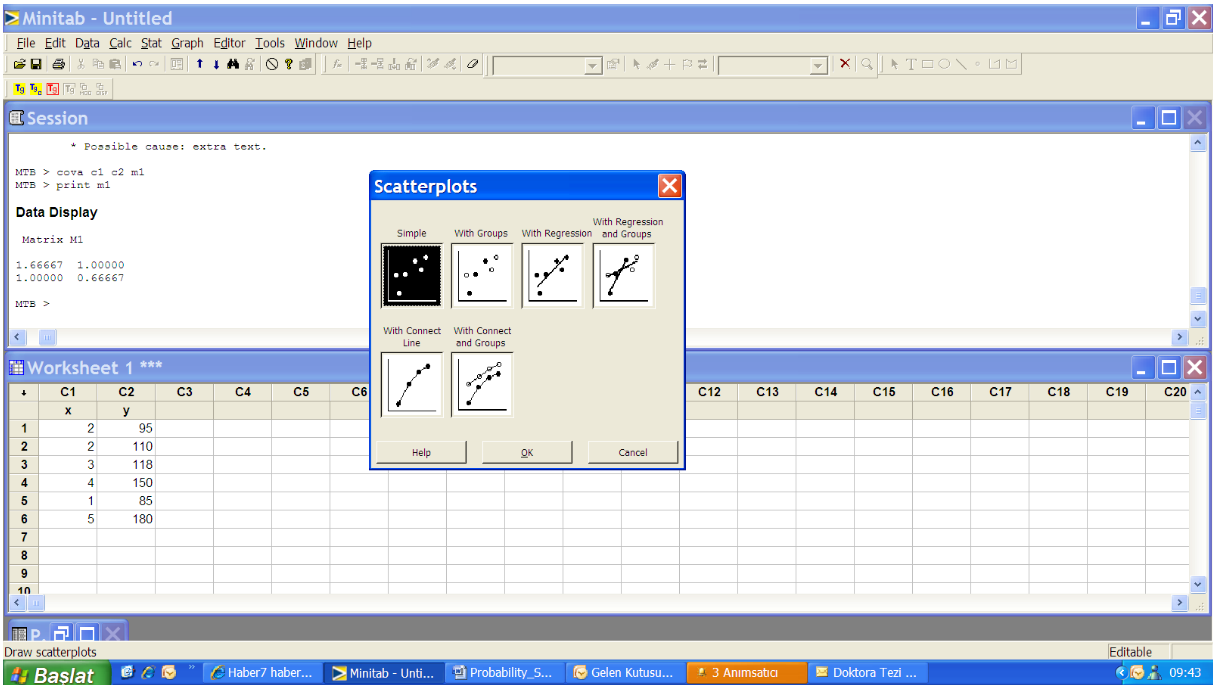

```
MTB > Chart 'var-1';
SUBC>   Group 'var-2';
SUBC>   Stack;
SUBC>   Bar.
```

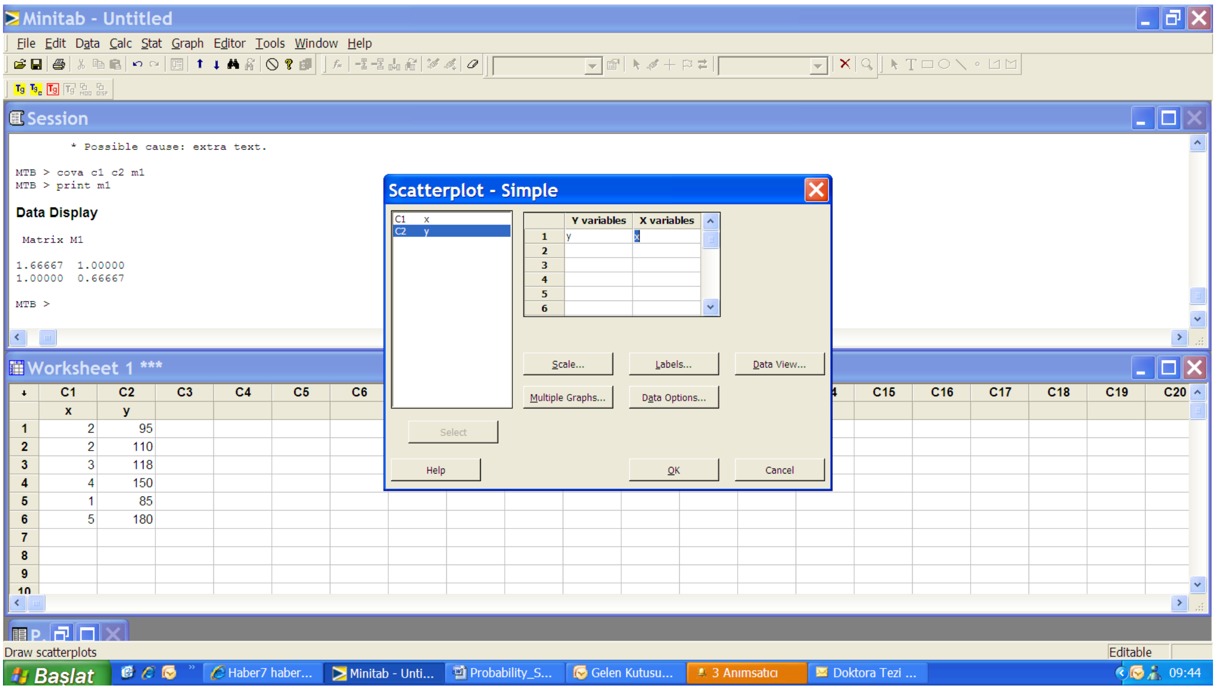


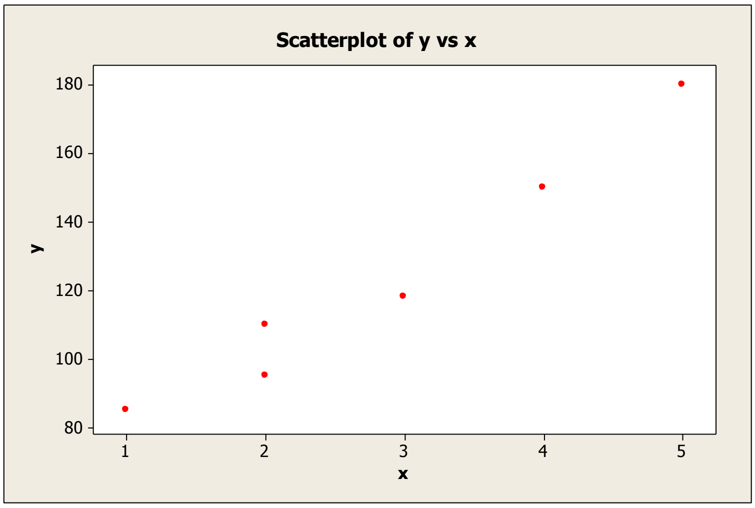

```
Margplot  'var-1'* 'var-2'.
```

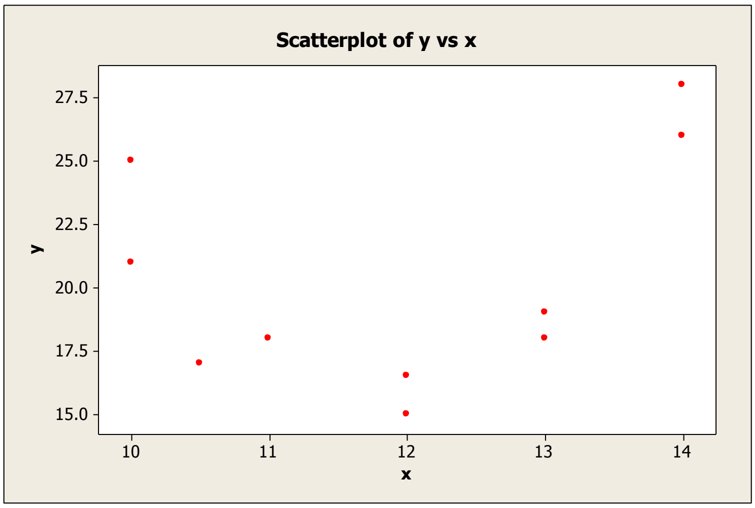
Scatter plots for two Quantitative Variables

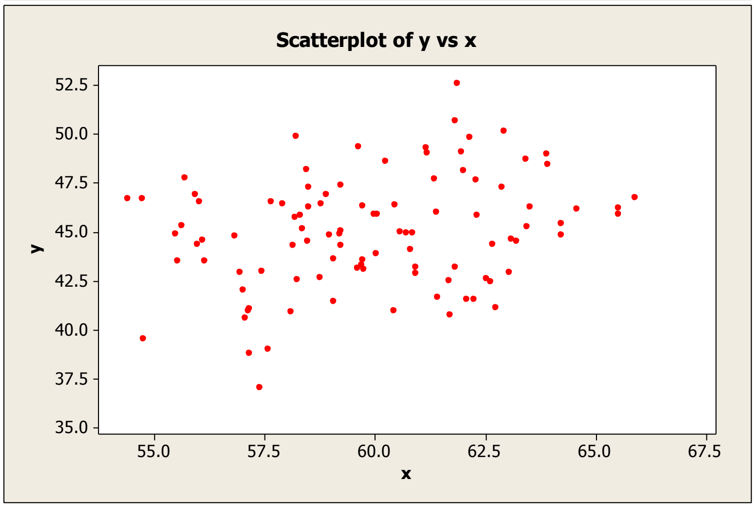
When both variables to be displayed on a graph are **quantitative**, one variable is plotted along the horizontal axis and the second along the vertical axis. The first variable is often called x and the second is called y, so that the graph takes the form of a plot on the (x, y) axis. Each pair of data values is plotted as a point on this two-dimensional graph, called a **scatterplot.**





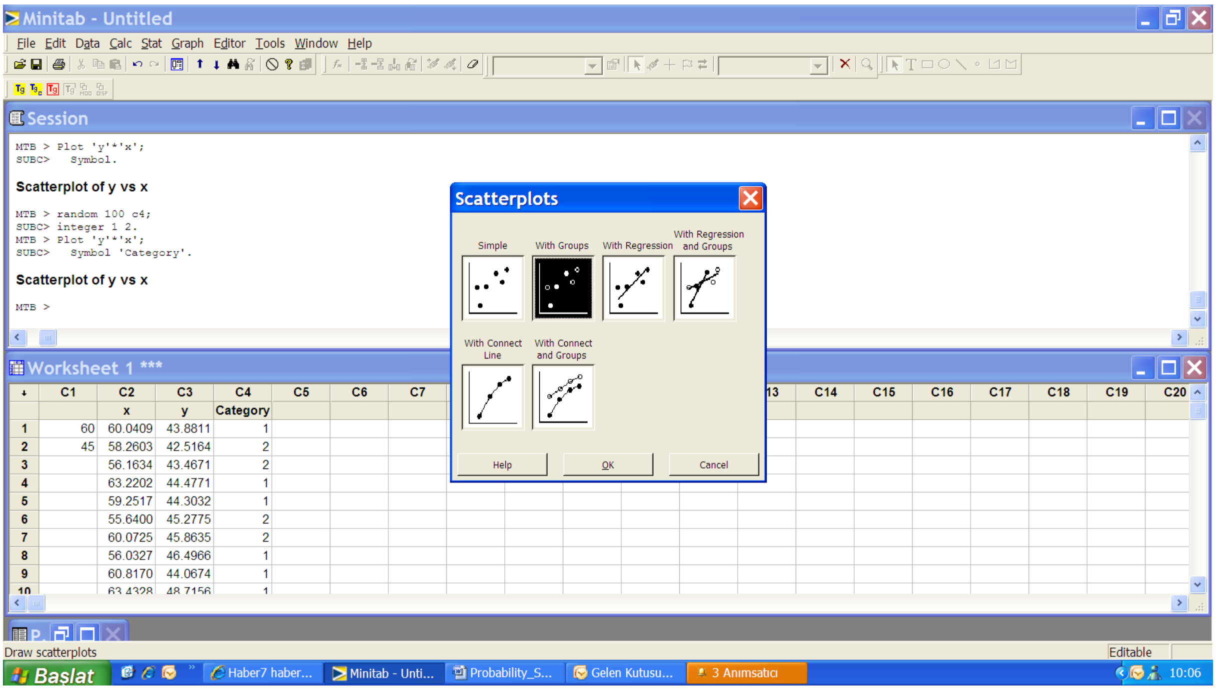


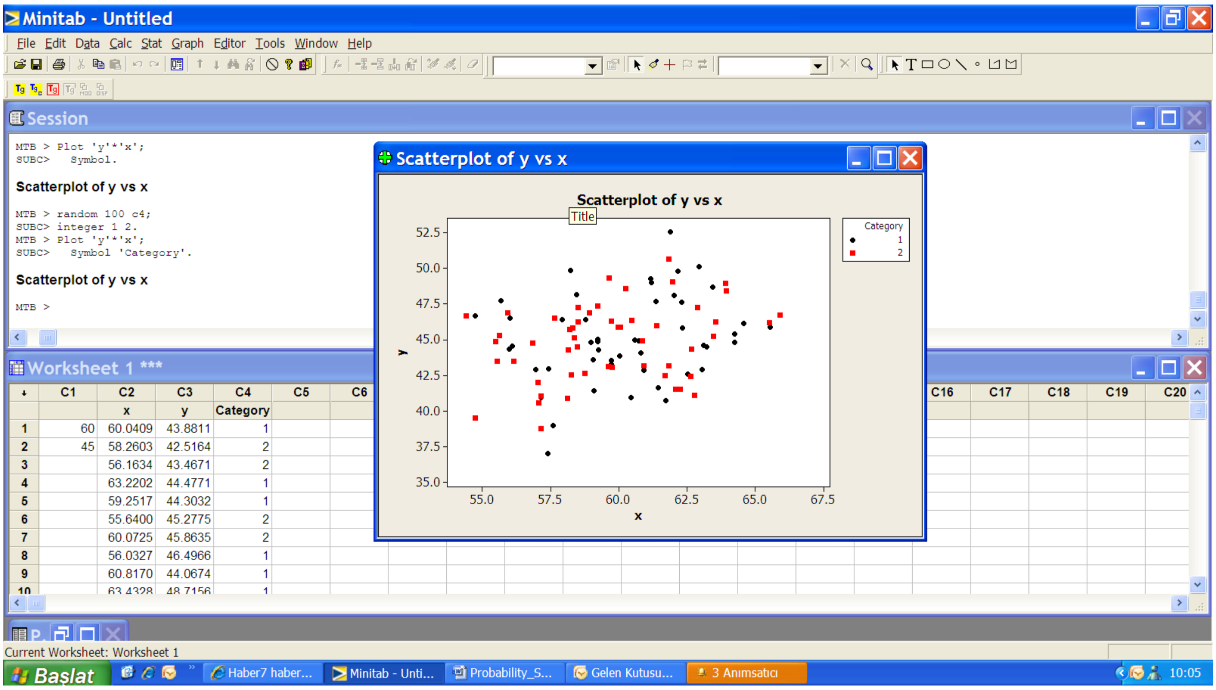




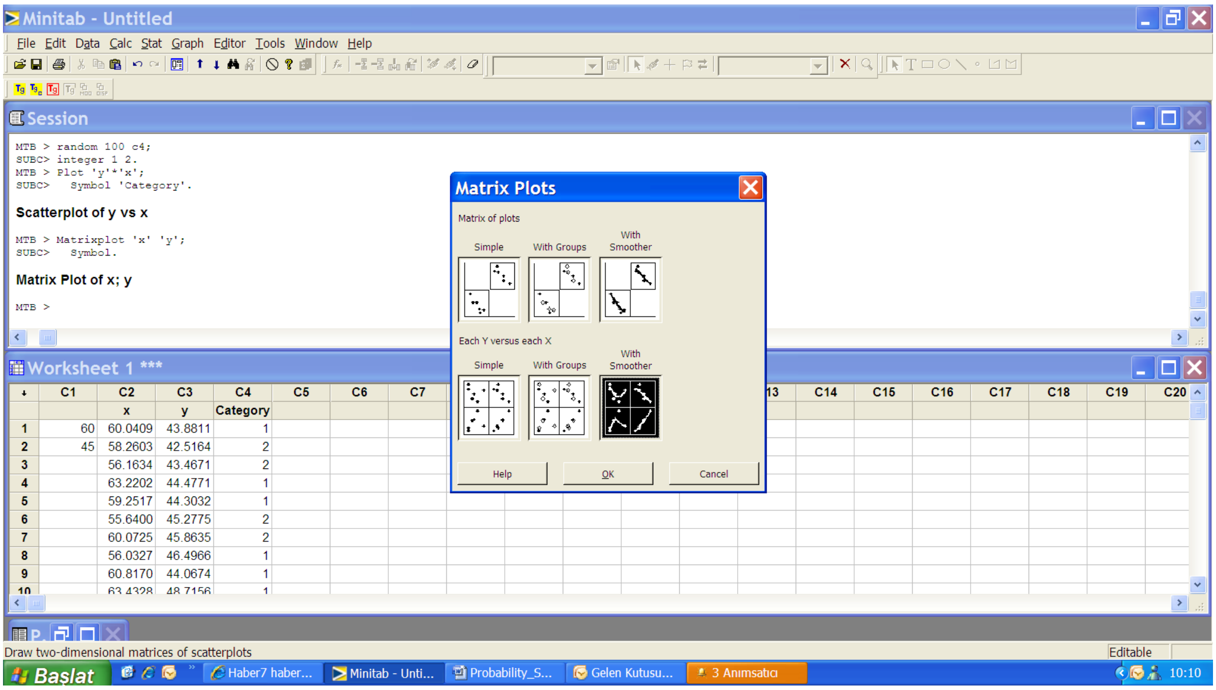
- **What type of pattern do you see?**
- **How strong is the pattern?**
- **Are there any unusual observations?**

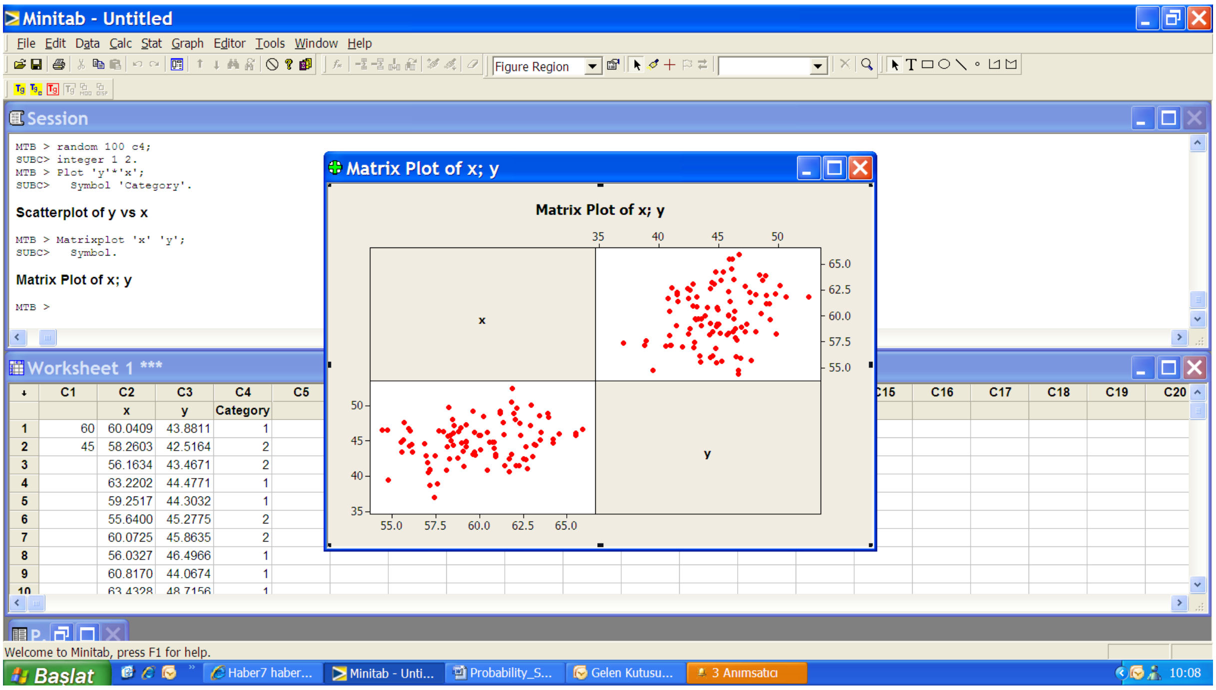
Scatter Plot with groups





Matrix Plot





Numerical Measures for Quantitative Bivariate Data

A constant rate of increase or decrease is perhaps the most common pattern found in bivariate scatterplots. A simple measure that serves this purpose is called the **correlation coefficient**, denoted by r, and is defined as

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

or

$$r = \frac{s_{xy}}{s_x s_y}$$

The quantities S_x and S_y are the standard deviations for the variables x and y, respectively. **The new quantity S_{xy} is called the** **covariance** **between x and y and is defined as**

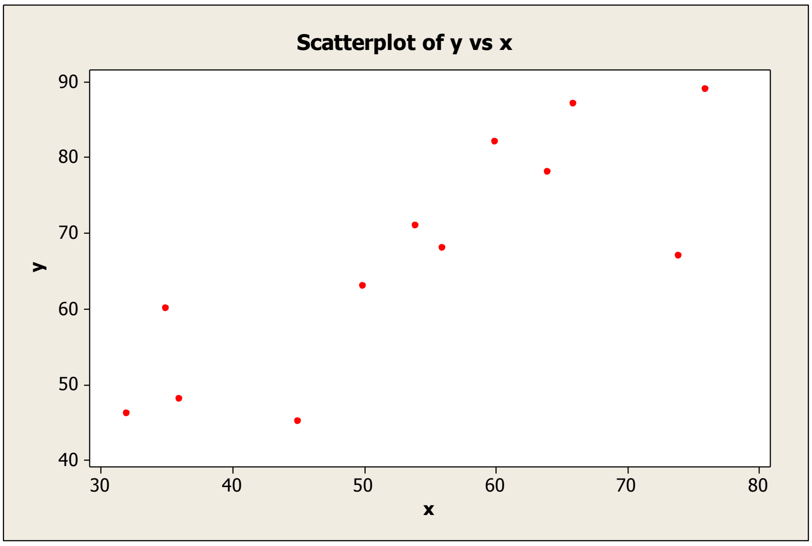
$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{n-1}$$

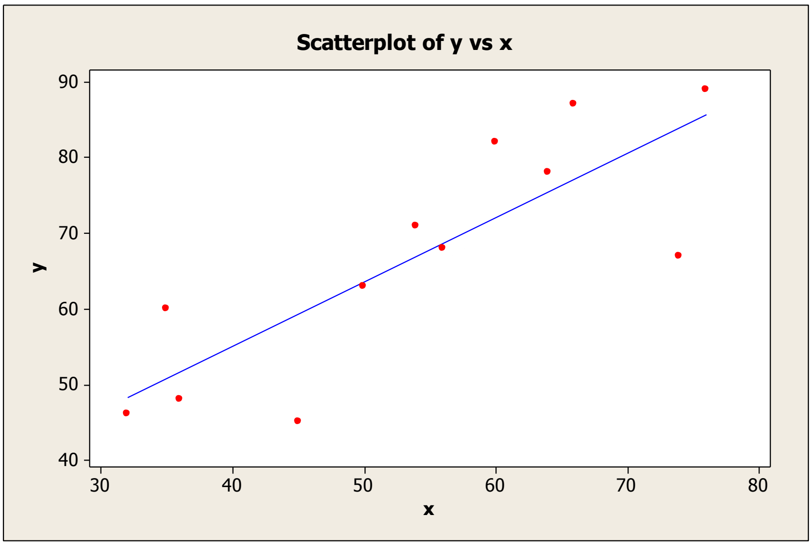
Covariance

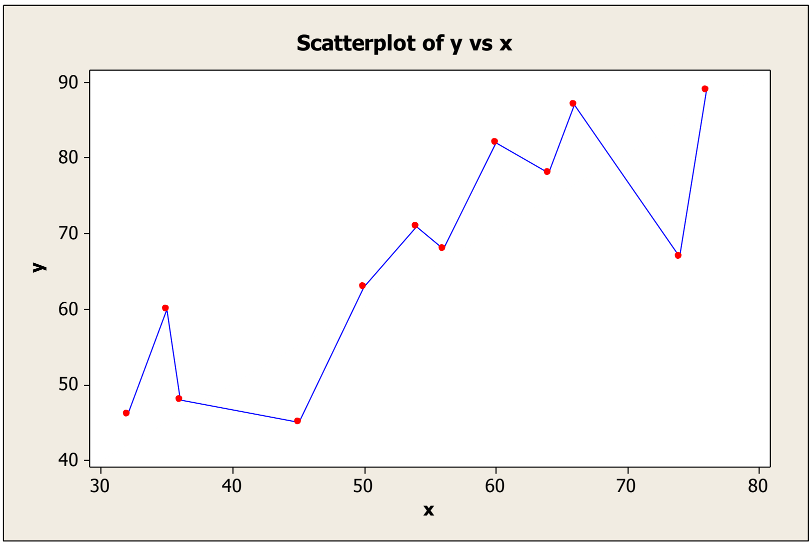
Covariance is a measure of the linear relationship between two variables. <u>Covariance is not standardized, unlike the correlation coefficient</u>. Therefore, covariance values can range from <u>negative infinity to positive infinity</u>. A correlation coefficient is the covariance divided by the product of each variable's standard deviation.

Example:

x	y	x^2	y^2	xy
60	82	3600	6724	4920
74	67	5476	4489	4958
36	48	1296	2304	1728
64	78	4096	6084	4992
45	45	2025	2025	2025
66	87	4356	7569	5742
76	89	5776	7921	6764
56	68	3136	4624	3808
32	46	1024	2116	1472
35	60	1225	3600	2100
50	63	2500	3969	3150
54	71	2916	5041	3834
648	**804**	**37426**	**56466**	**45493**


Scatterplot of y vs x
y
x
90
80
70
60
50
40
30
40
50
60
70
80


Scatterplot of y vs x
y
x
90
80
70
60
50
40
30
40
50
60
70
80


Scatterplot of y vs x
y
x
90
80
70
60
50
40
30
40
50
60
70
80

$$s_x = \left(\frac{\sum x^2 - \left(\sum x\right)^2 / n}{n-1} \right)^{1/2} = \left(\frac{37426 - 648^2 / 12}{11} \right)^{1/2} = 14.875$$

$$s_y = \left(\frac{\sum y^2 - \left(\sum y\right)^2 / n}{n-1} \right)^{1/2} = \left(\frac{56466 - 804^2 / 12}{11} \right)^{1/2} = 15.368$$

$$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{n-1} = \frac{45493 - \frac{(648)(804)}{12}}{11} = 188.818$$

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
x	12	0	54.00	4.29	14.88	32.00	38.25	55.00	65.50	76.00
y	12	0	67.00	4.44	15.37	45.00	51.00	67.50	81.00	89.00

MTB > covariance c1 c2

Covariances: x; y

	x	y
x	221.273	
y	188.818	236.182

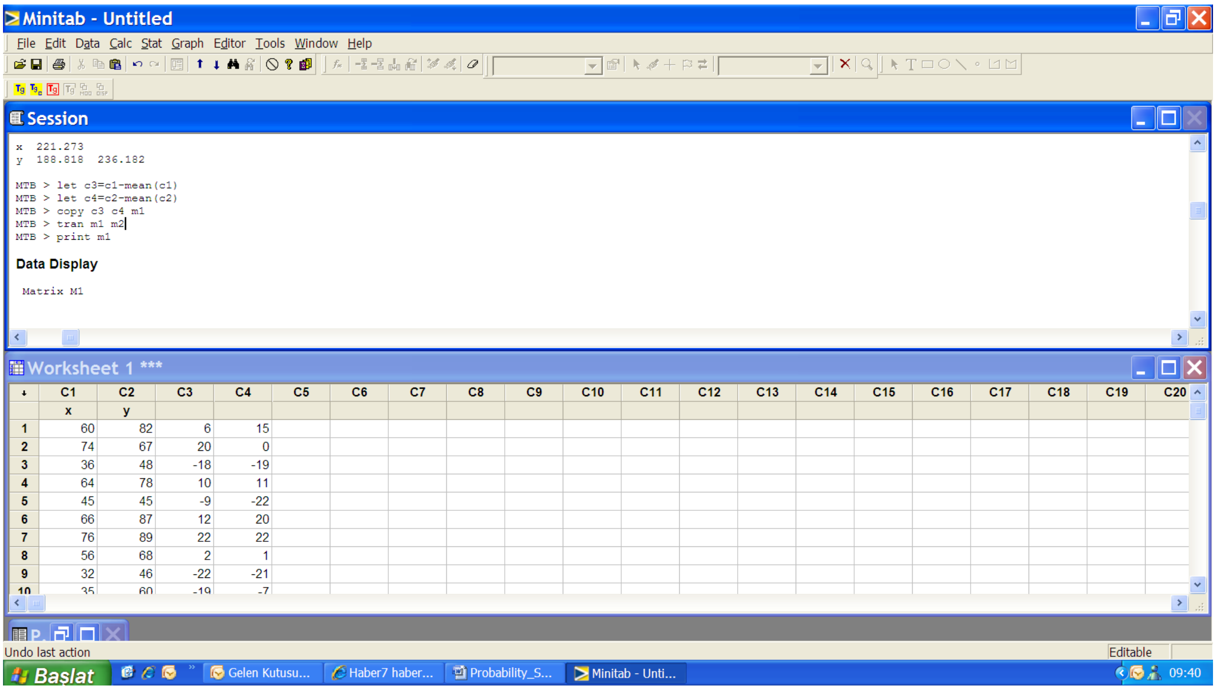
$$r = \frac{s_{xy}}{s_x s_y} = \frac{188.818}{(14.875)(15.368)} = 0.826$$

MTB > corre c1 c2

Correlations: x; y

Pearson correlation of x and y = 0.826
P-Value = 0.001

Matrix computation of Covariance




```
MTB > let c3=c1-mean(c1)
MTB > let c4=c2-mean(c2)
MTB > copy c3 c4 m1
MTB > tran m1 m2
MTB > print m1
```

Data Display

Matrix M1

6	15
20	0
-18	-19
10	11
-9	-22
12	20
22	22
2	1
-22	-21
-19	-7
-4	-4
0	4

```
MTB > print m2
```

Data Display

```
 Matrix M2

 6  20  -18  10   -9  12  22  2  -22  -19  -4  0
15   0  -19  11  -22  20  22  1  -21   -7  -4  4
```

```
MTB > mult m2 m1 m3
MTB > print m3
```

Data Display **Sum of squares and product matrix**

```
 Matrix M3

2434  2077
2077  2598
```

```
MTB > let k1=1/(count(c1)-1)
MTB > mult m3 k1 m4
MTB > print m4
```

Data Display

```
 Matrix M4

221.273  188.818
188.818  236.182
```

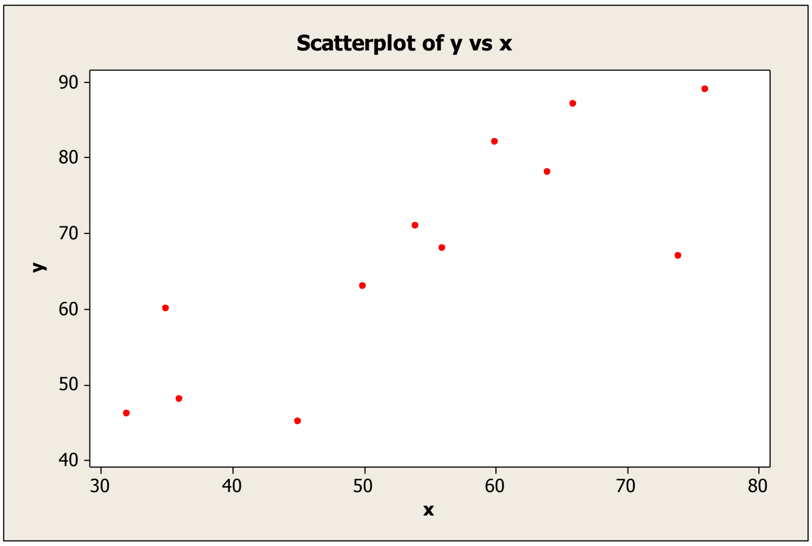
```
MTB > cova c1 c2
```

Covariances: x; y

```
             x        y
x  221.273
y  188.818  236.182
```

The Regression Model

Regression analysis is helpful in ascertaining the probable form of the relationship between variables, and the <u>ultimate objective when this method of analysis is employed usually</u> is to *<u>predict</u>* or *<u>estimate</u>* the value of one variable corresponding to a given value of another variable



The pattern made by the points plotted on the scatter diagram usually suggests the basic nature of the relationship between two variables. As we look at the above figure for example, the points seem to be scattered around an invisible straight line.

Simple Linear Regression Model

Sometimes the two variables, x and y are related in a particular way. It may be that the value of y depends on the value of x; that is, the value of x in some way explains the value of y. In these situations, we call y the *dependent variable*, while x is called the *independent variable.*

A linear model that relates two variables, *x* and y. It can be written as:

$$y = \beta_0 + \beta_1 x + e$$

- ***y* and *x* may be referred in one of the following ways:**

x	**y**
Independent Variable **Explanatory Variable** **Control Variable** **Predictor Variable** **Regressor**	**Dependent Variable** **Explained Variable** **Response Variable** **Predicted Variable** **Regressand**

- ***e* is referred to as the *error term* or *disturbance*. (Represents factors other than x that affect y. It is treated as unobservable.)**
- **β_0 is the *intercept.***
  (Gives the value of *y* when *x* = 0 and *u* = 0.)
- **β_1 is the *slope.***
  (Relates a change *in y* to a change in *x*)

Definition (Fitted Value):

The value of y when $x=x_i$ using estimates of the regression coefficients

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Definition (Residual):

The difference between the actual and fitted values of y_i

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Given a sample of data $\{(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, we want to find to estimates of the intercept and slope such that minimize the total amount of residuals. ***Negative and positive residuals will cancel each other out, so look at the square of the residuals.***

$$Sum = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

The least square criterion requires that *Sum* be a <u>minimum.</u>
The first order conditions are:

$$\frac{\partial Sum}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0$$

$$\frac{\partial Sum}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0$$

Normal Equations

Dividing each of these equations by -2 and expanding the summation, we get the so called normal equations

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_0 n = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

Solving these equations simultaneously gives the estimations of intercept and slope.

From the first normal equation we obtain

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

(Estimator of β_0)

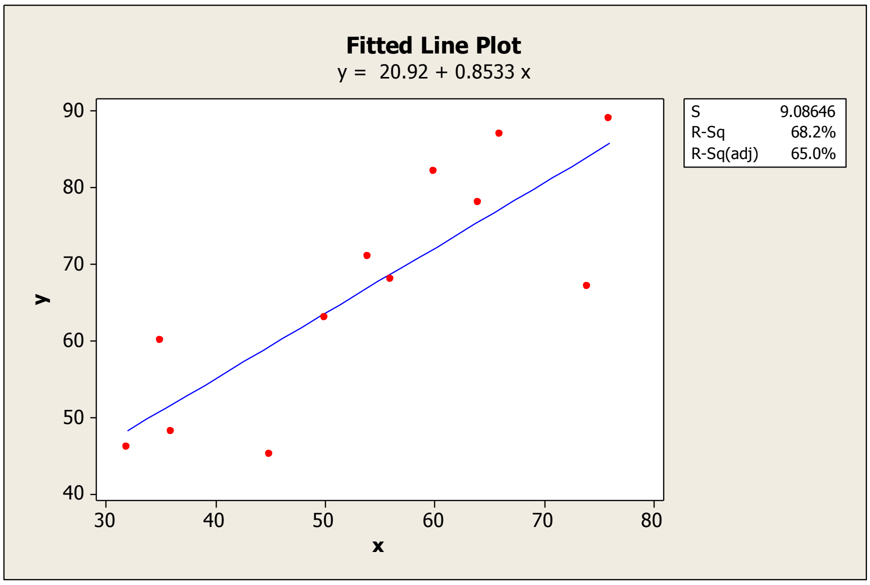
From the second normal equation we obtain

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

(Estimator of β_1)

Example:

x	y	x^2	y^2	xy
60	82	3600	6724	4920
74	67	5476	4489	4958
36	48	1296	2304	1728
64	78	4096	6084	4992
45	45	2025	2025	2025
66	87	4356	7569	5742
76	89	5776	7921	6764
56	68	3136	4624	3808
32	46	1024	2116	1472
35	60	1225	3600	2100
50	63	2500	3969	3150
54	71	2916	5041	3834
648	**804**	**37426**	**56466**	**45493**



$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\hat{\beta}_1 = \frac{12(45493) - (648)(804)}{12(37426) - (648)^2} = 0.8533$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 = 804/12 - 0.8533(648/12) = 20.92$$

Normal Equations

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_0 n = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

$$648\hat{\beta}_1 \; + 4\hat{\beta}_0 = 804$$

$$37426\hat{\beta}_1 + 648\hat{\beta}_0 = 45493$$

$\hat{\beta}_1$ = **0.8533**, $\hat{\beta}_0$ = **20.92**

The least squares equation is

$$\hat{y} = 20.92 + 0.8533x$$

The coefficient of determination

The coefficient of determination, which is equal to the explained sum of squares divided by the total sum of squares, is

$$r^2 = \frac{\beta_1^2\left(\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n\right)}{\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 / n} =$$

$$= \frac{(0.8533)^2\{(37426) - (648)^2/12\}}{56466 - (804)^2/12} = \frac{1772.246}{2598} = 0.68215$$

Pearson correlation of x and y = 0.826 ***