B34.UC2

Numerical Computation and Statistics in Engineering

Unit 6: Regression Analysis

Scatterplots

Scatterplots show the relationship between two quantitative variables. Examples are

- fuel consumption per speed;
- fuel consumption per weight;
- spending for leisure per income;
- etc.

Scatterplots

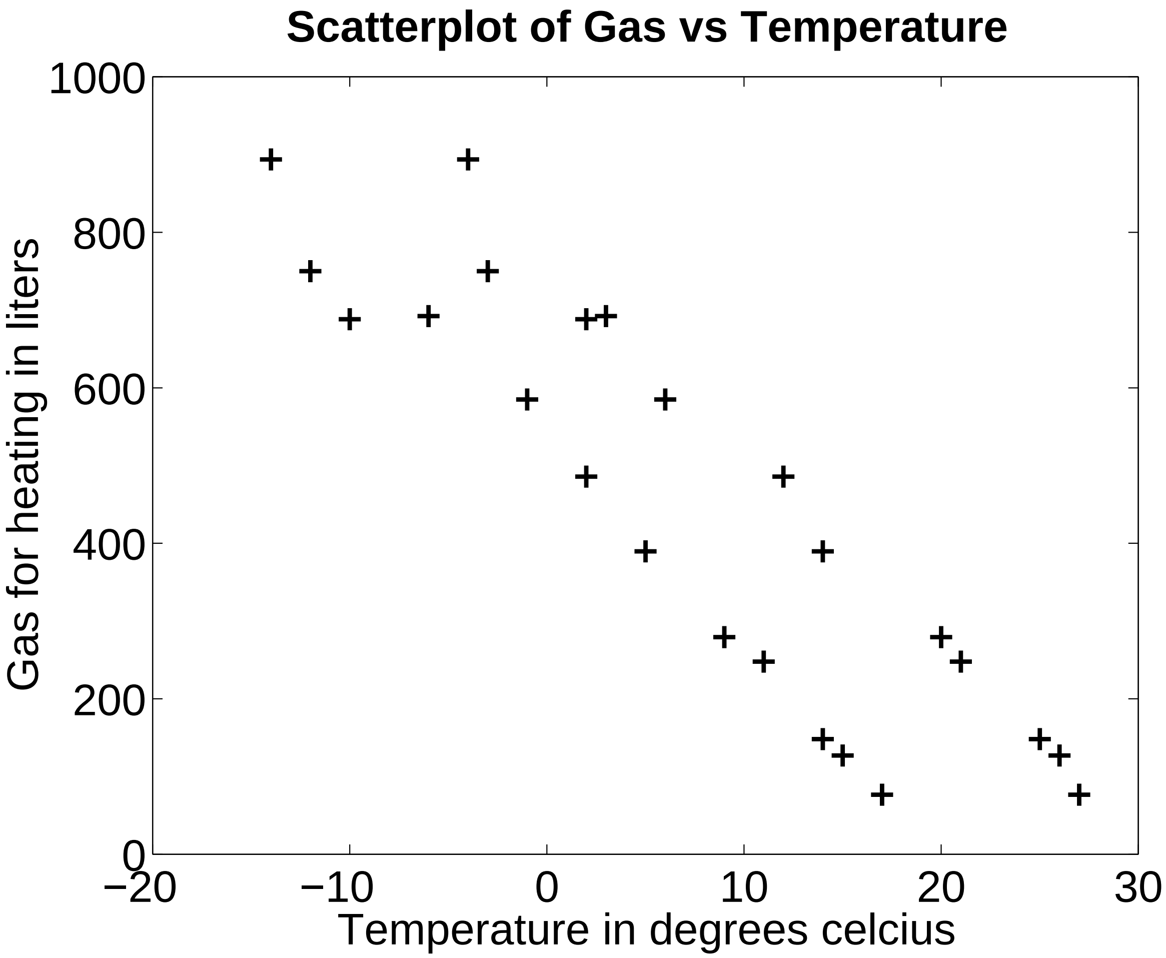
Example. We consider the following average high and low temperatures in Montreal (in degrees Celsius), and the measured gas consumption for heating (in liters) for a house:

	Jan.	Feb.	March	April	Mai	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
high	-4	-3	3	12	20	25	27	26	21	14	6	2
low	-14	-12	-6	2	9	14	17	15	11	5	-1	-10
gas	894	750	692	486	279	148	77	127	248	390	584	688

The following is a data plot of the gas consumption against the low and

Scatterplots

hight temperatures. Both suggest a strong linear relationship.



□

Scatterplots

In general, interpreting a scatterplot we first look for an overall pattern, and describe form, direction, and strength.

Two variables are *positively associated* if above average values of one variable tend to result in above average values of the other. Two variables are *negatively associated* if above average values of one variable tend to result in below average values of the other.

Correlation

Correlation measures strength and direction of the *linear* relationship between two data sets. If the random variables are x and y, each consisting of n individual data, then the correlation r between x and y is defined as

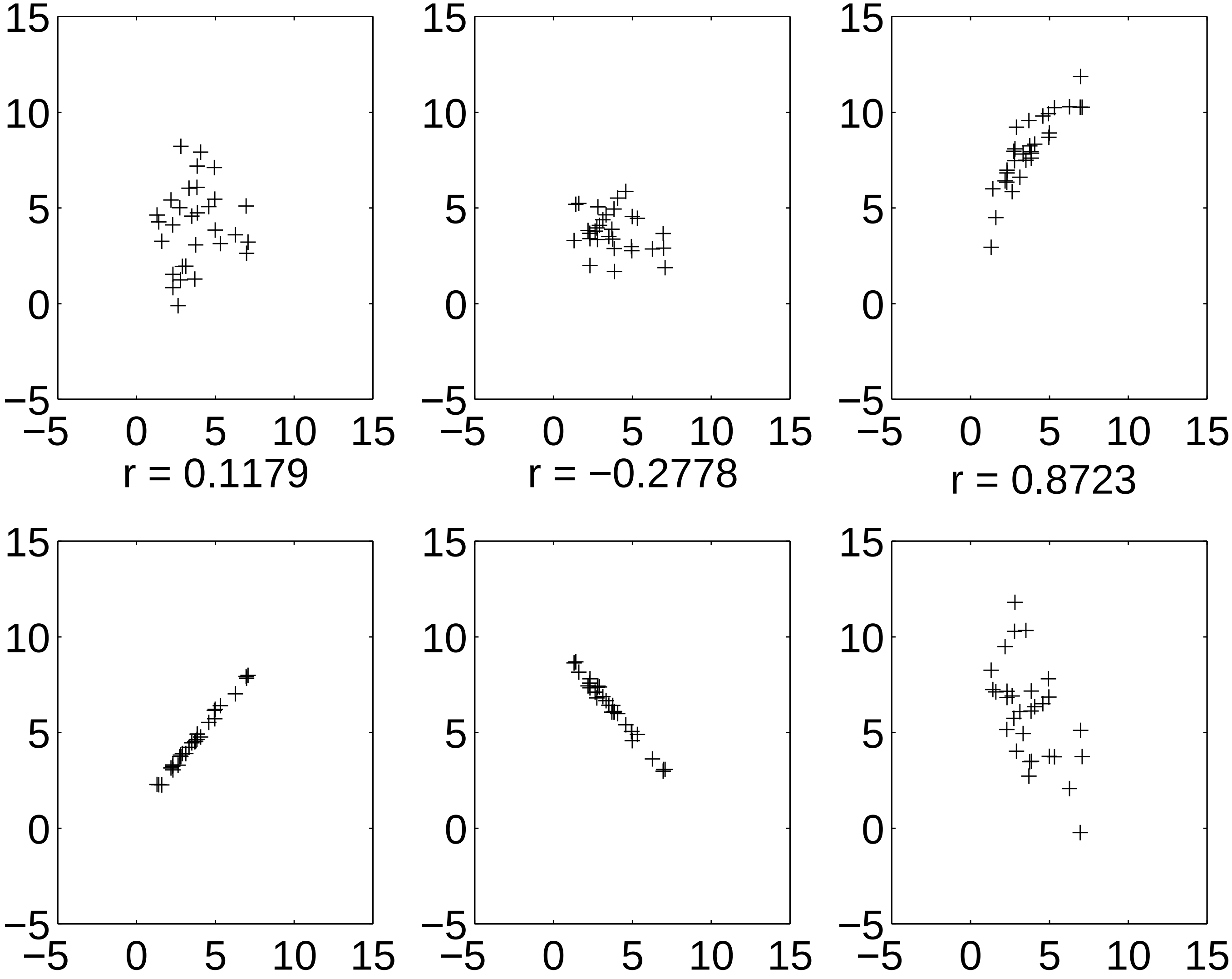
$$r = \frac{1}{n-1} \sum_i \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_x and s_y the sample errors. Correlation has the following properties:

- r does not have a dimension;
- the correlation is invariant under change of units of measurements;
- r is also independent from interchanging the role of x and y;
- r is *always* a number between -1 and 1.

Values close to -1 or 1 indicate strong linear relationships.

Correlation



r = 0.1179

r = −0.2778

r = 0.8723

r = 0.9961

r = −0.9928

r = −0.5705

Least Squares Regression Line

One of the variables is usually considered to be an *explanatory variable* (often denoted x), and the other a *response variable* (often denoted y).

A *regression line* is a line that describes how the response variable y changes when the explanatory variable x changes. Regression lines are used, among others, for prediction.

We use the notation

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

to denote a regression line for the data set x and y. Here $\hat{y}$ stands for the *predicted* value of the response variable (as opposed to the *observed* value). The quantity

$$\epsilon_i = \hat{y}_i - y_i = (\hat{\beta}_1 x_i + \hat{\beta}_0) - y_i$$

is called the error (or residual) of the observed data y_i, and is the vertical (!) distance between y_i and the regression line.

Least Squares Regression Line

The *least squares regression line* is the regression line that minimizes the sum

$$\sum\nolimits_i \epsilon_i^2 = \sum\nolimits_i (\hat{y}_i - y_i)^2 \,.$$

The least squares regression line $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$ has slope

$$\hat{\beta}_1 = r\frac{s_y}{s_x} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})(x_i - \bar{x})}$$

and intercept

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

- In least squares regression analysis the role of x and y are distinct. Changing the role of x and y gives *different* regression lines.
- A change of one standard deviation in x corresponds to a change of r standard deviation in y.
- Least square regression lines are sensitive to outliers.

Least Squares Regression Line

How do we find the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for the least squares regression line?
We want to minimize $E(\hat{\beta}_0, \hat{\beta}_1) = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$.
Then

$$\begin{aligned} \frac{\partial E}{\partial \hat{\beta}_0} &= \sum_i 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) \qquad \text{and} \\ \frac{\partial E}{\partial \hat{\beta}_1} &= \sum_i 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i)\,. \end{aligned}$$

Setting both lines equal to 0 and simplifying gives

$$\begin{aligned} 0 &= \sum_i (y_i - \hat{\beta}_1 x_i) - n\hat{\beta}_0 \\ &= -n\hat{\beta}_0 - \hat{\beta}_1 \sum_i x_i + \sum_i y_i \qquad \text{and} \\ 0 &= -\sum_i x_i y_i + \hat{\beta}_0 \sum_i x_i + \hat{\beta}_1 \sum_i x_i^2\,. \end{aligned}$$

Solving this system of linear equations in $\hat{\beta}_0$ and $\hat{\beta}_1$ gives the formulae above.

Least Squares Regression Line

For our data relating the use of gas for heating we find the following parameters for the least squares regression lines:

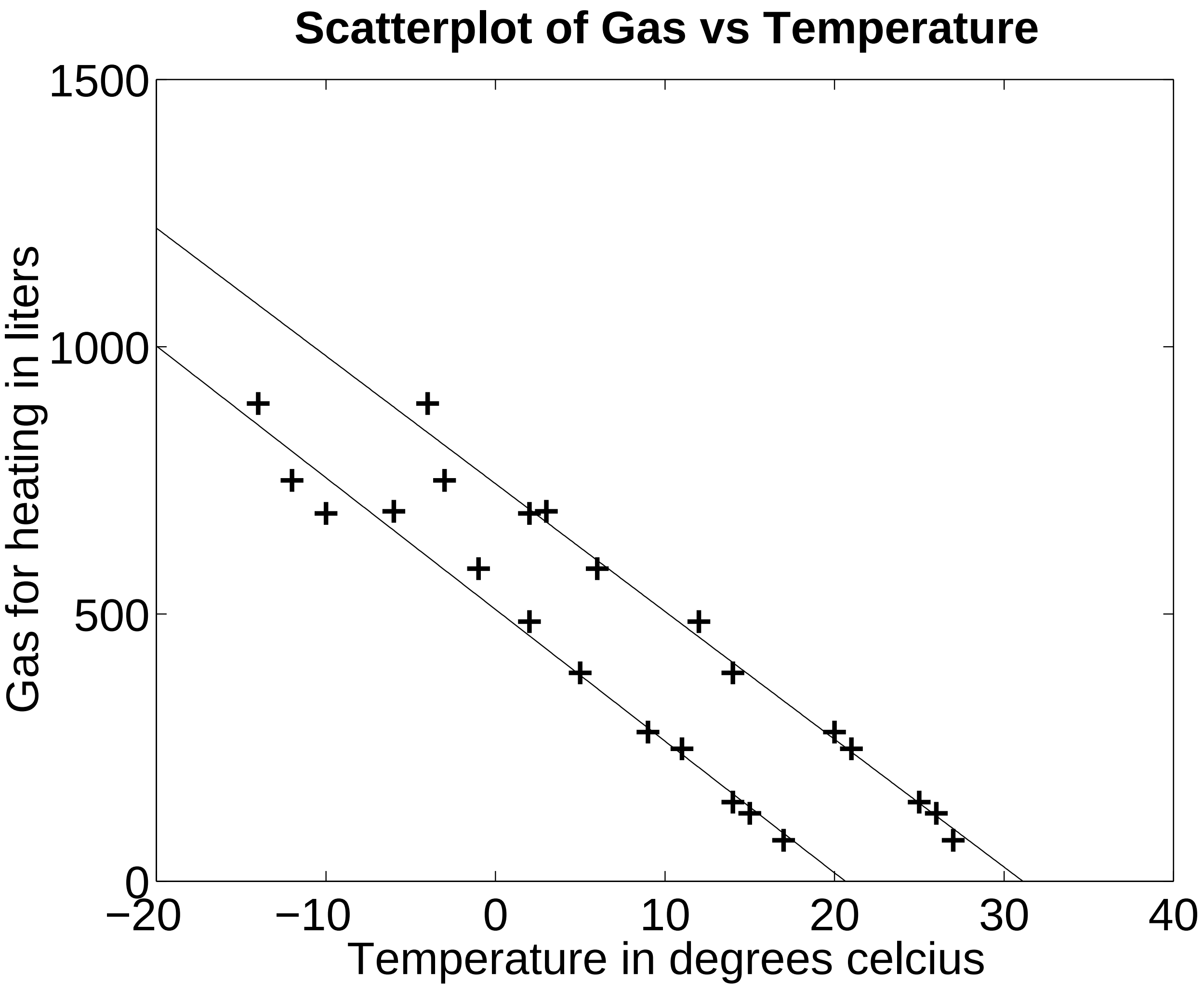
	r	s_x	s_y	$\hat{\beta}_1$	$\hat{\beta}_0$
high	-0.9940	11.4213	247.7005	-23.9072	743.8478
low	-0.9912	11.0495	247.7005	-24.6422	508.6054

The two equations for the regression lines are thus

$$\begin{aligned} \hat{y} &= -23.9072x + 743.8478 \quad \text{and} \\ \hat{y} &= -24.6422x + 508.6054\,. \end{aligned}$$

Least Squares Regression Line

The following diagram shows the data and the regression lines:



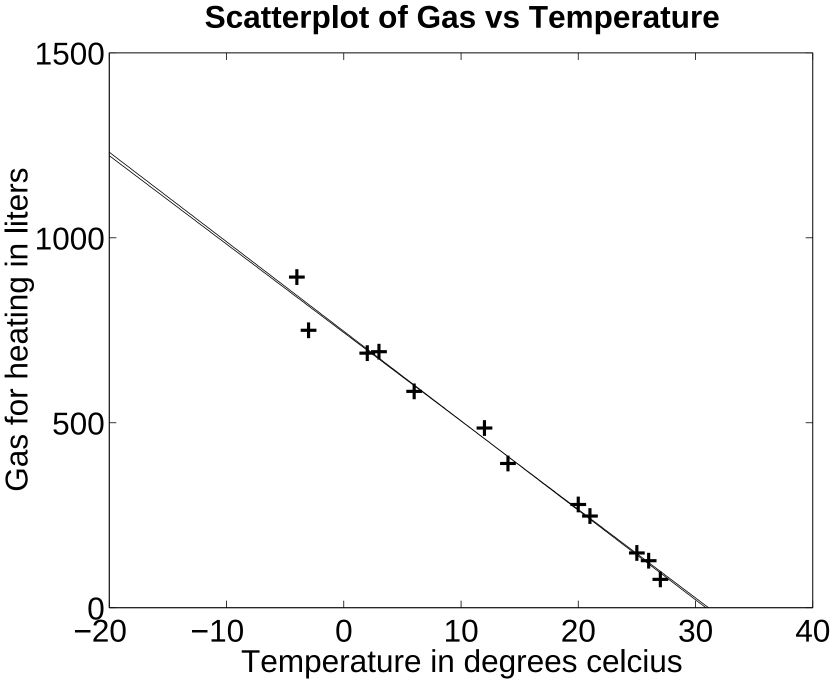
Least Squares Regression Line

If we interchange the role of x and y then we get the regression line

$$\hat{x} = \hat{\alpha}_1 y + \hat{\alpha}_0$$

with $\hat{\alpha}_1 = r\frac{s_x}{s_y}$ and $\hat{\alpha}_0 = \bar{x} - \hat{\alpha}_1\bar{y}$. This yields the equation

$$\hat{x} = -0.0412y + 30.8903 \qquad \text{or} \qquad y = -24.1967\hat{x} + 747.4423\,.$$



Example

The following data represent the number of members in the EC Council of Ministers of current EC members and of potential EC members, and the populations of the member states:

	number	population
		(in 1,000,000)
Germany	29	82.038
Great-Britain	29	59.247
France	29	58.966
Italy	29	57.610
Spain	27	39.394
Netherlands	13	15.760
Grece	12	10.533
Belgium	12	10.213

	number	population
		(in 1,000,000)
Portugal	12	9.980
Sweden	10	8.854
Austria	10	8.082
Denmark	7	5.313
Finland	7	5.160
Irland	7	3.744
Luxembourg	4	0.429

The list of candidates is as follows, with the agreed number of seats

Example

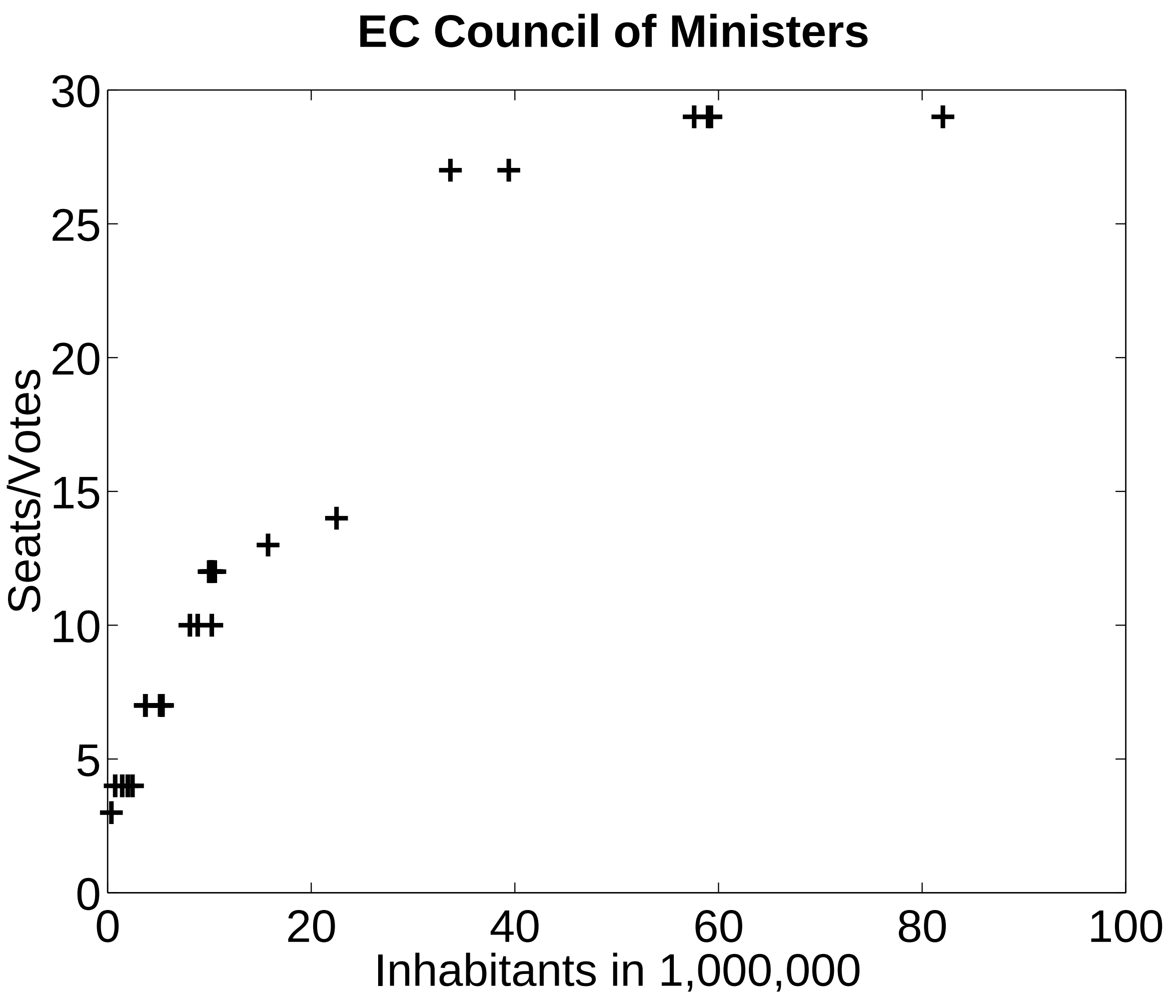
according to the EC meeting in Nice end of 2000:

	number	population
		(in 1,000,000)
Poland	27	33.667
Rumania	14	22.489
Czech Republic	12	10.290
Hungaria	12	10.092
Bulgaria	10	10.230
Slovakia	7	5.393

	number	population
		(in 1,000,000)
Lithuania	7	3.701
Letvia	4	2.439
Slovenia	4	1.978
Estonia	4	1.446
Cyprus	4	0.752
Malta	3	0.379

Example

The following diagram shows a scatterplot for these data:

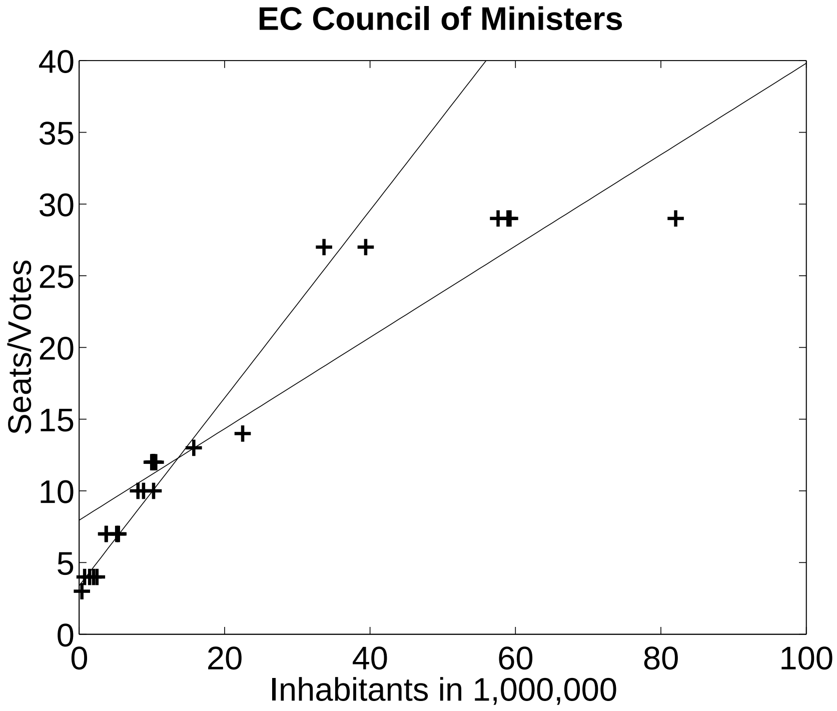


Example

For the regression lines we find

$$\hat{y} = 0.3186x + 7.9581 \qquad \text{and} \qquad \hat{y} = 0.6542x + 3.3924\,.$$

The correlation is 0.8930 for the first data set, and 0.9700 for the second.



The combined data set does suggest a logarithmic relationship.

Non-Linear Relationships

The following four data sets are from Frank J. Anscombe, Graphs in statistical analysis, *The American Statistician* 27: 17–21, 1973.

x_1	10	8	13	9	11	14	6	4	12	7	5
y_1	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
x_2	10	8	13	9	11	14	6	4	12	7	5
y_2	9.14	8.14	8.74	8.77	9.26	8.1	6.13	3.10	9.13	7.26	4.74
x_3	10	8	13	9	11	14	6	4	12	7	5
y_3	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
x_4	8	8	8	8	8	8	8	8	8	8	19
y_4	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

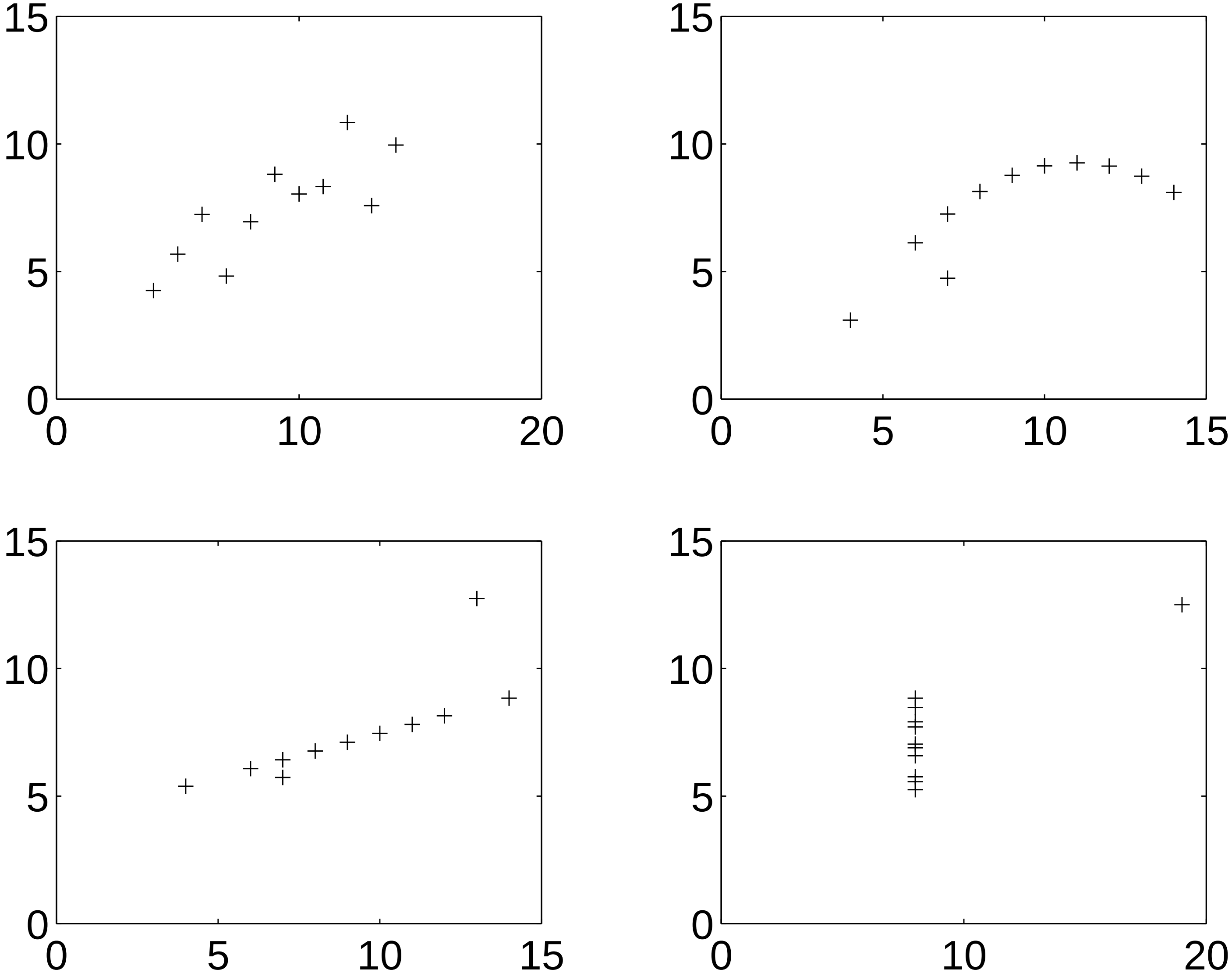
Non-Linear Relationships

The correlation and least squares lines are shown in the following list:

$$\begin{array}{ll} r_1 = 0.8164 & \hat{y}_1 = 0.5001x_1 + 3.0001 \\ r_2 = 0.8162 & \hat{y}_2 = 0.5000x_2 + 3.0009 \\ r_3 = 0.8163 & \hat{y}_3 = 0.4997x_3 + 3.0025 \\ r_4 = 0.8165 & \hat{y}_4 = 0.4999x_4 + 3.0017 \end{array}$$

However, when plotting the data we realize that only the third and fourth represent strong linear relationships (both with one influential outlier). The first data set represents a moderate linear relationship, while the second represents a curved relationship.

Non-Linear Relationships



Confidence Intervals

We want to gain confidence in the least squares regression line. For this we have to make a couple of assumptions about ϵ:

- For the random variable ϵ we make the assumption that $E(\epsilon) = 0$.
- The standard deviation of ϵ is a constant σ.
- The distribution of ϵ is normal.
- Errors associated with different observations are independent.

Under the assumptions above let s^2 be

$$\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} \, .$$

Then $\frac{(n-2)s^2}{\sigma^2}$ has a chi-square distribution with $n-2$ degrees of freedom, and s^2 is an unbiased estimator for σ^2.

The standard deviation of ϵ can be interpreted as follows: We expect most observations y to lie within $2s$ of their least squares predicted value $\hat{y}$.

Confidence Intervals

If we make the four assumptions above then $\hat{\beta}_1$ has normal distribution with mean the true slope of the line and standard deviation

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}}\,.$$

Here we use the notation

$$S_{uv} = \sum\nolimits_i (u_i - \bar{u})(v_i - \bar{v}) = \sum\nolimits_i u_i v_i - \frac{1}{n} \sum\nolimits_i u_i \sum\nolimits_i v_i$$

and note that

$$\sum (y_i - \hat{y}_i)^2 = S_{yy} - \hat{\beta}_1 S_{xy}\,.$$

Confidence Intervals

We want to find confidence intervals for the slope. The random variable

$$\frac{\hat{\beta}_1 - \beta}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta}{s/\sqrt{S_{xx}}}$$

has a t-distribution with $n-2$ degrees of freedom (we cannot use the normal distribution since $\sigma_{\hat{\beta}_1}$ is estimated by $s_{\hat{\beta}_1}$). Thus, the endpoints of a $(1-\alpha)100\%$ confidence interval for the slope β_1 are

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} s_{\hat{\beta}_1} \qquad \text{where } s_{\hat{\beta}_1} = s/\sqrt{S_{xx}}\,.$$

Note that the endpoints of the interval are again of the form

$$\text{point estimator} \pm t_{n-2,\alpha/2} \text{ estimated standard error of the estimator}$$

Confidence Intervals

Example. The following data are given:

x_i	1	2	3	4	5	6
y_i	1	2	2	4	4	6

Here $\sum x_i = 21$, $\sum y_i = 19$, $\sum x_i^2 = 91$, $\sum y_i^2 = 77$ and $\sum x_i y_i = 83$. It follows that

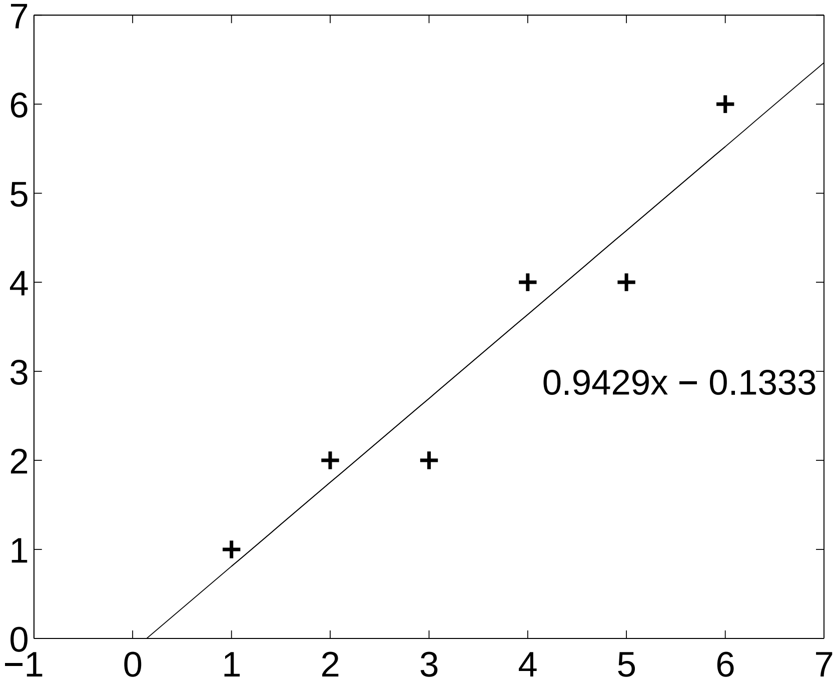
$$\begin{aligned} S_{xx} &= \sum x_i^2 - \frac{1}{6}(\sum x_i)^2 \;=\; 91 - \frac{21^2}{6} \\ &= 17.5\,, \quad \text{and} \\ S_{xy} &= \sum x_i y_i - \frac{1}{6}(\sum x_i)(\sum y_i) \;=\; 83 - \frac{21 \cdot 19}{6} \\ &= 16.5\,. \end{aligned}$$

Confidence Intervals

Thus we calculate for slope and intercept of the least squares regression line

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0.9429\,,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{6} \cdot 19 - 0.9429 \cdot \frac{21}{6} = -0.13\,.$$



Confidence Intervals

An estimator for the variance s^2 of the error $\epsilon = y - \hat{y}$ is

$$\frac{\sum(y_i - \hat{y}_i)^2}{n-2} = S_{yy} - \hat{\beta}_1 S_{xy}\,.$$

With the data above we find that $S_{yy} = 77 - \frac{1}{6}(19)^2 = 16.8333$ so that

$$s^2 = 1.2755\,.$$

The endpoints of the 95% confidence interval for $\hat{\beta}_1$ are

$$\hat{\beta}_1 \pm t_{4,0.025} s_{\hat{\beta}_1} = 0.9429 \pm 2.776 \cdot \frac{\sqrt{1.2755}}{\sqrt{17.5}} = 0.9429 \pm 0.7494\,.$$

□

Multiple Linear Regression

By way of example we consider multiple linear regression. We consider the following table of apartment blocks sold in a big city:

#appartments	#floors	price (in 1,000,000)
60	10	78.2
40	5	45.4
80	10	100.0
30	6	35.7
60	3	80.5
40	6	42.8
90	12	120.4
80	7	90.5

We want to find the equation of a plane (!) allowing to predict the price z

Multiple Linear Regression

of an apartment block with x apartments and y floors using the method of least squares.

The equation of the plane is

$$\hat{z} = \hat{\alpha} x + \hat{\beta} y + \hat{\gamma}\,.$$

For observed data z_i we have the error

$$\epsilon_i = z_i - \hat{z}_i = z_i - \hat{\alpha} x_i - \hat{\beta} y_i - \hat{\gamma}\,,$$

and we try to minimize

$$E(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = \sum\nolimits_i \epsilon_i^2 = \sum\nolimits_i (z_i - \hat{\alpha} x_i - \hat{\beta} y_i - \hat{\gamma})^2\,.$$

Multiple Linear Regression

Then

$$\begin{aligned}
\frac{\partial E}{\partial \hat{\alpha}} &= \sum\nolimits_i 2(z_i - \hat{\alpha} x_i - \hat{\beta} y_i - \hat{\gamma})(-x_i) \\
\frac{\partial E}{\partial \hat{\beta}} &= \sum\nolimits_i 2(z_i - \hat{\alpha} x_i - \hat{\beta} y_i - \hat{\gamma})(-y_i) \\
\frac{\partial E}{\partial \hat{\gamma}} &= \sum\nolimits_i 2(z_i - \hat{\alpha} x_i - \hat{\beta} y_i - \hat{\gamma})(-1)
\end{aligned}$$

Thus we have to solve the system of linear equations

$$\begin{aligned}
0 &= -\sum z_i x_i + \hat{\alpha} \sum x_i^2 + \hat{\beta} \sum x_i y_i + \hat{\gamma} \sum x_i \\
0 &= -\sum z_i y_i + \hat{\alpha} \sum x_i y_i + \hat{\beta} \sum y_i^2 + \hat{\gamma} \sum y_i \\
0 &= -\sum z_i + \hat{\alpha} \sum x_i + \hat{\beta} \sum y_i + n\hat{\gamma}
\end{aligned}$$

Multiple Linear Regression

With the data above we find

$$\begin{array}{lcllcllcl} \sum x_i^2 & = & 33200 & \sum x_i y_i & = & 3840 & \sum x_i & = & 480 \\ \sum y_i^2 & = & 499 & \sum x_i z_i & = & 40197 & \sum y_i & = & 59 \\ \sum z_i^2 & = & 50499.4 & \sum y_i z_i & = & 4799.8 & \sum z_i & = & 593.5 \end{array}$$

Thus we have to solve the system of linear equations

$$\begin{pmatrix} 33200 & 3840 & 480 \\ 3840 & 499 & 59 \\ 480 & 59 & 8 \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} 40197.0 \\ 4799.8 \\ 593.5 \end{pmatrix}.$$

Linear algebra (or MATLAB) tells us that the solution is

$$\hat{\alpha} = 1.3067, \quad \hat{\beta} = 0.4813, \quad \hat{\gamma} = -7.7611\,.$$

Multiple Linear Regression

We can use the equation of the plane

$$\hat{y} = 1.3068x + 0.4813y - 7.7611$$

to predict the price of an apartment block with 50 apartments and 5 floors, which will have a predicted price of

$$\hat{z} = 1.3068 \cdot 50 + 0.4813 \cdot 5 - 7.7611 \approx 59.98$$

million.

The Previous Example in MATLAB

```
>> x = [60 40 80 30 60 40 90 80];
>> y = [10 5 10 6 3 6 12 7];
>> z = [78.2 45.4 100.0 35.7 80.5 42.8 120.4 90.5];
>> Sxx = sum(x.*x);
>> Syy = sum(y.*y);
>> Szz = sum(z.*z);
ans =
     1.0e+004 *
            3.2200      0.0499      5.0449
>> Sxy = sum(x.*y);
>> Sxz = sum(x.*z);
>> Syz = sum(z.*y);
>> [Sxx Syy Szz]
```

The Previous Example in MATLAB

```
>> [Sxy Sxz Syz]
ans =
     1.0e+004 *
            0.3840     4.0197     0.4800
>> [sum(x) sum(y) sum(z)]
ans =
     480.0000     59.0000     593.5000
>> A=[Sxx Sxy sum(x); Sxy Syy sum(y); sum(x) sum(y) 8]
A =
       32200          3840          480
        3840           499           59
         480            59            8
>> b=[Sxz; Syz; sum(z)]
```

The Previous Example in MATLAB

```
b =

  1.0e+004 *


    4.0197

    0.4800

    0.0594

>> inv(A)

ans =

    0.0005   -0.0024   -0.0127

   -0.0024    0.0267   -0.0556

   -0.0127   -0.0556    1.2996

>> (inv(A)*b)'

ans =

    1.3067     0.4813     -7.7611
```

Example

Example. A study is made into the response time to 911 calls. It is measured the time (in minutes) it takes the ambulance to arrive at the scene against the distance (in kilometers) between the station and the scene. The following data are collected:

dist. x	3.4	1.8	4.6	2.3	3.1
time y	2.5	1.8	3.0	2.6	2.9
dist. x	5.2	0.6	2.9	2.7	4.0
time y	3.9	1.6	2.3	2.0	3.4
dist. x	2.3	1.0	6.3	4.5	3.5
time y	2.5	1.8	2.6	2.8	2.8

Example

To find the least squares regression line $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$ we first calculate

$$\begin{aligned} S_{xx} &= 33.5573 \\ S_{yy} &= 5.3933 \\ S_{xy} &= 10.0367 \end{aligned}$$

and thus

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{10.0367}{33.5573} = 0.2991\,, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 2.5667 - 0.2991 \cdot 3.2133 = 1.6056\,. \end{aligned}$$

and the least squares regression line is

$$\hat{y} = 0.2991x + 1.6056\,.$$

Next we look at the probability distribution of the random error compo-

Example

nent ϵ. For s^2 we find

$$s^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{2.3915}{13} = 0.1840\,.$$

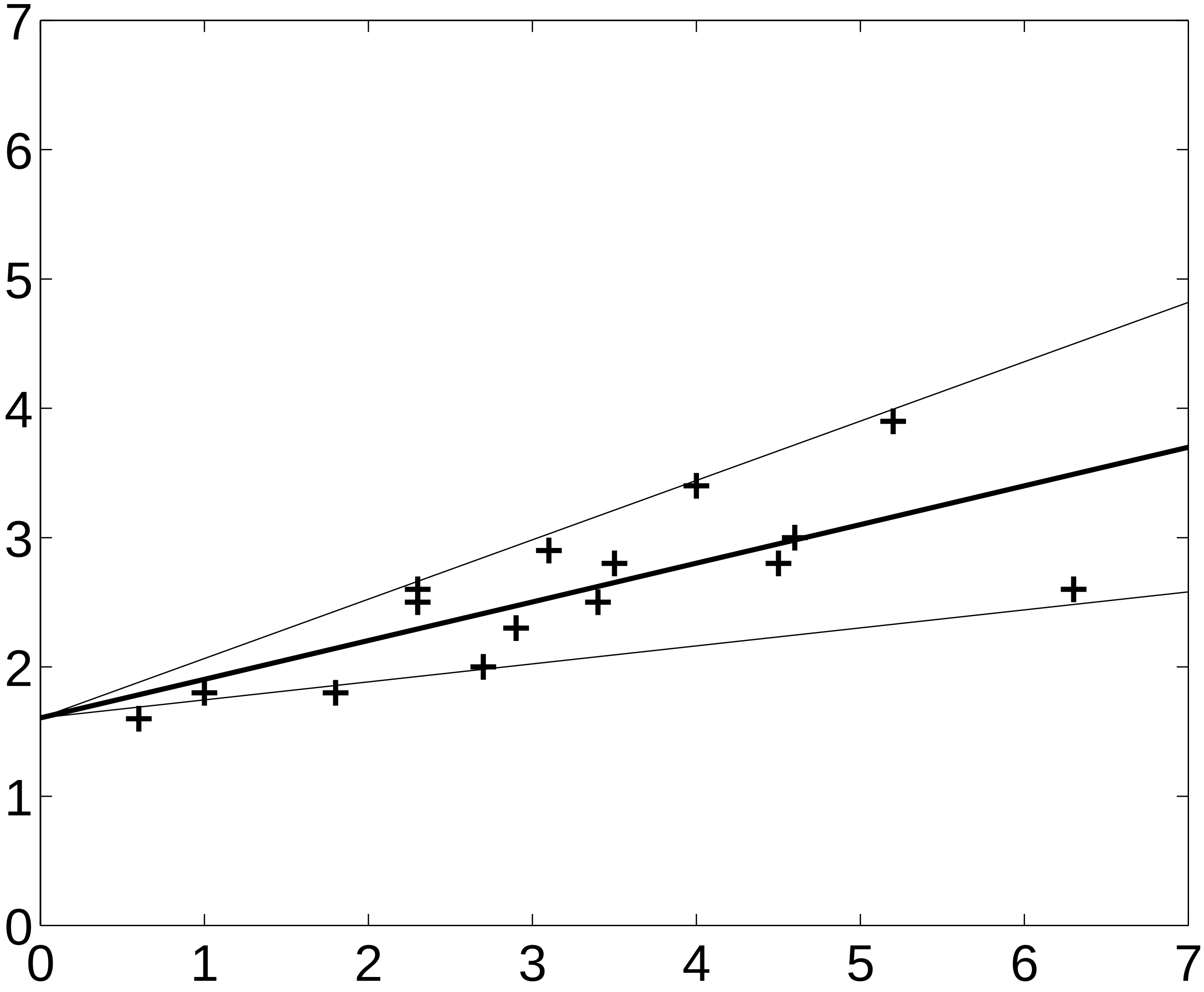
Thus $s = 0.4289$.

The endpoints of the 95% confidence interval for $\hat{\beta}_1$ are

$$\hat{\beta}_1 \pm t_{13,0.025}\frac{s}{\sqrt{S_{xx}}} = 0.2991 \pm 2.160\frac{0.4289}{\sqrt{33.5573}}$$

which gives the interval $[0.1392, 0.4590]$. The following diagram shows the data set, the least squares regression line, and the two boundary lines for the interval estimating $\hat{\beta}_1$:

Example



□

Summary

- Regression analysis aims to find relationships between variables where there is an estimated dependency.
- Correlation measures strength and direction of a linear relationship between two data sets. However, correlation is sensitive to outliers.
- The least squares regression line is the line that fits best best two data sets. This line minimizes the vertical distance between the observed values and the predicted values on the line.
- Multiple linear relationships and non-linear relationships can be tackled with similar methods.