Bioinformatics and Biostatistics BB2440: Biostatistics
Lecture 6: Correlation, Regression, ANOVA

Timo Koski & Jan Enger

TK

16.09.2013

KTH Matematik

Outline of Lecture 6.

- Correlation

Outline of Lecture 6.

- Correlation
- Correlation does not imply causation

KTH VETENSKAP OCH KONST
KTH Matematik

Outline of Lecture 6.

- Correlation
- Correlation does not imply causation
- Linear Regression

KTH Matematik

Outline of Lecture 6.

- Correlation
- Correlation does not imply causation
- Linear Regression
- Analysis of Variance (= ANOVA)

KTH VETENSKAP OCH KONST

KTH Matematik

Basic Concepts of Correlation

We are looking at paired data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We can look at them as a table, too.

	Sample			
	1	2	...	n
Variable x	x_1	x_2	...	x_n
Variable y	y_1	y_2	...	y_n

KTH VETENSKAP OCH KONST

KTH Matematik

Basic Concepts of Correlation

A **correlation** *exists between two variables when one of them is related to the other in some way.*

	Sample			
	1	2	...	n
Variable x	x_1	x_2	...	x_n
Variable y	y_1	y_2	...	y_n

KTH VETENSKAP OCH KONST

KTH Matematik

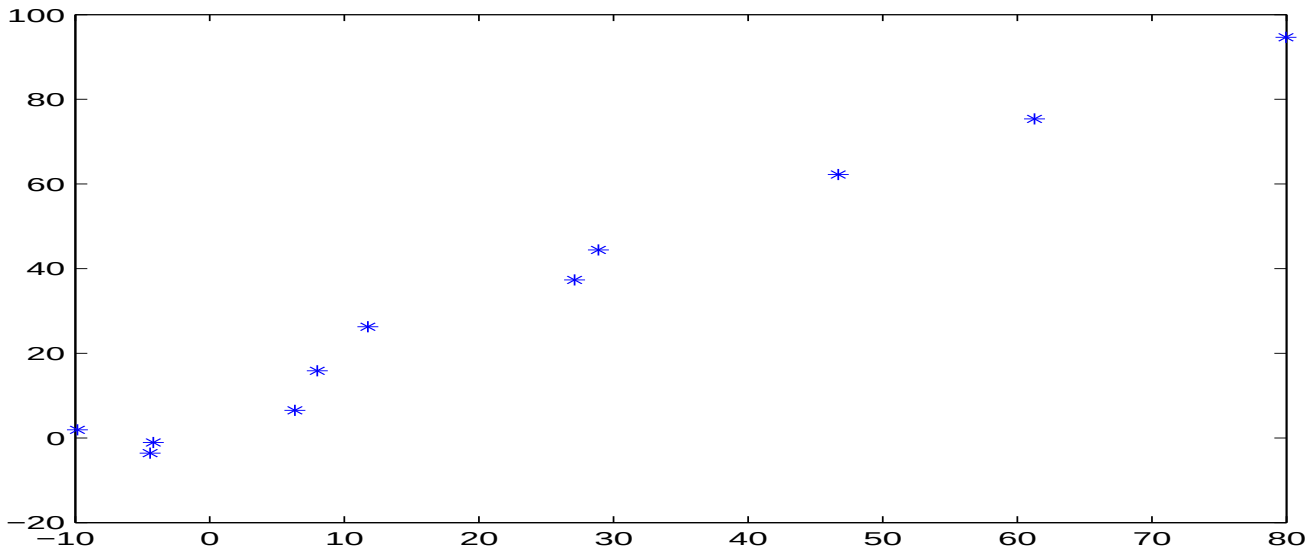
Scatterplot

A **scatterplot** *is a graph in which the paired samples* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ *are plotted with a horizontal* x *-axis and a vertical* y*-axis. Each individual* (x, y) *pair is plotted as a single point.*

KTH Matematik

Scatterplot

A **scatterplot** *is a graph in which the paired samples* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ *are plotted with a horizontal* x *-axis and a vertical* y*-axis. Each individual* (x, y) *pair is plotted as a single point (*$*$ *in the figure).*



The Correlation Coefficient

The **correlation coefficient** *r measures the strength of linear association between paired x and y sample values.*

KTH VETENSKAP OCH KONST

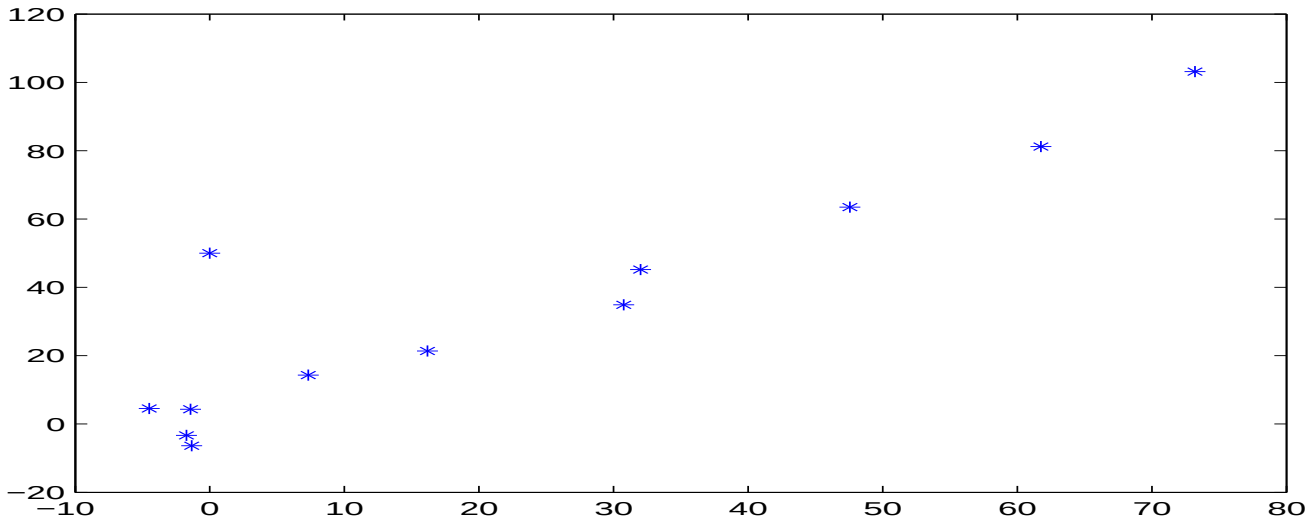
KTH Matematik

The Correlation Coefficient: requirements for validity

- random sample
- visual examination of the scatterplot must confirm that the points approximate a straight line pattern.
- Any outliers (=points lying far away from the other data points) must be removed if they are errors. The effects of any other outliers should be considered

KTH VETENSKAP OCH KONST
KTH Matematik

Outlier



KTH Matematik

The Correlation Coefficient

The **correlation coefficient** *r measures the strength of linear association between paired x and y sample values.*

KTH VETENSKAP OCH KONST

KTH Matematik

Covariance

The *covariance* between x- and y-values in $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$ is

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

and where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.
The intuitive interpretation is: if large sample values of x tend to go together with large sample values of y, then $(x_i - \bar{x})(y_i - \bar{y})$ is most often positive, and c_{xy} will tend to be positive.
The other way, if if small sample values of x tend to go together with small sample values of y, then $(x_i - \bar{x})(y_i - \bar{y})$ is most often negative, and c_{xy} will tend to be negative.

KTH VETENSKAP OCH KONST
KTH Matematik

Covariance and Correlation Coefficient

We standardize (to get back to the original units of measurement)

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

with s_x and s_y,

$$s_x = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2}, s_y = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

and get the *correlation coefficient* as

Definition

$$r \stackrel{\text{def}}{=} \frac{c_{xy}}{s_x s_y},$$

KTH Matematik

Correlation Coefficient

By some algebra one can get that

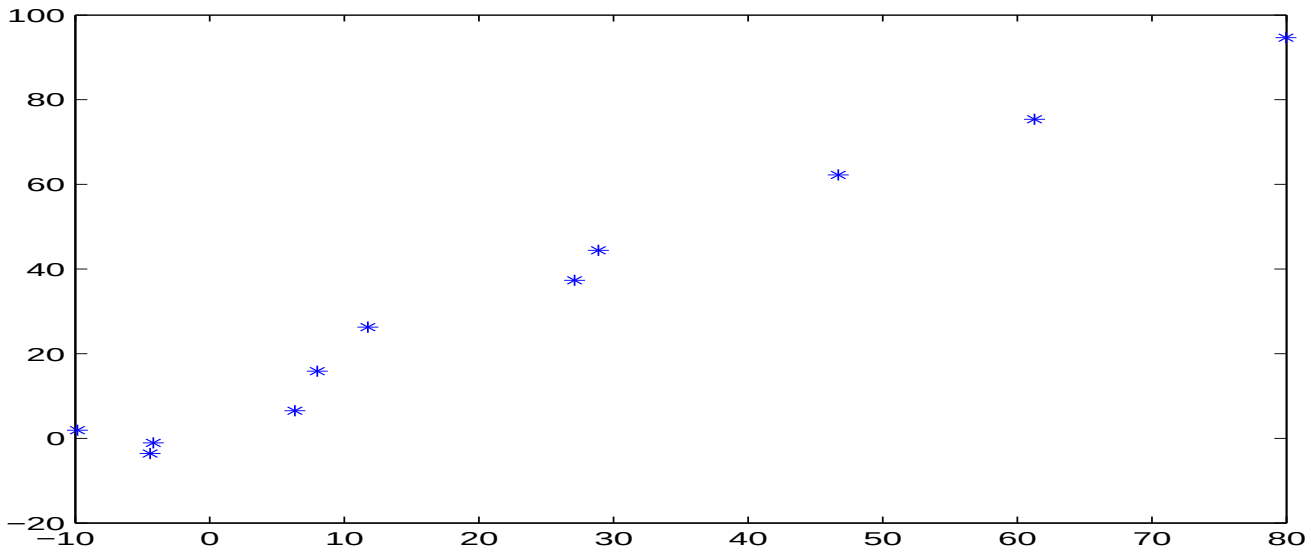
$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{j=1}^{n} y_j^2 - (\sum_{i=1}^{n} y_i)^2}}$$

which may perhaps be a more friendly or clever form for simple electronic calclulators.

KTH Matematik

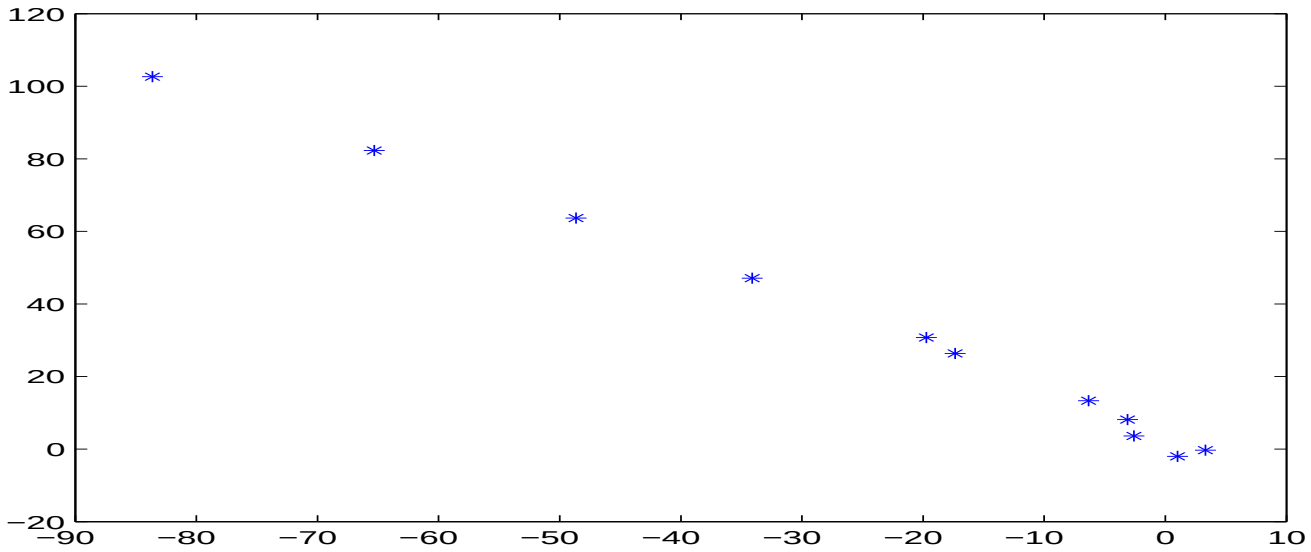
Coefficent of correlation

Here $r = 0.9989$



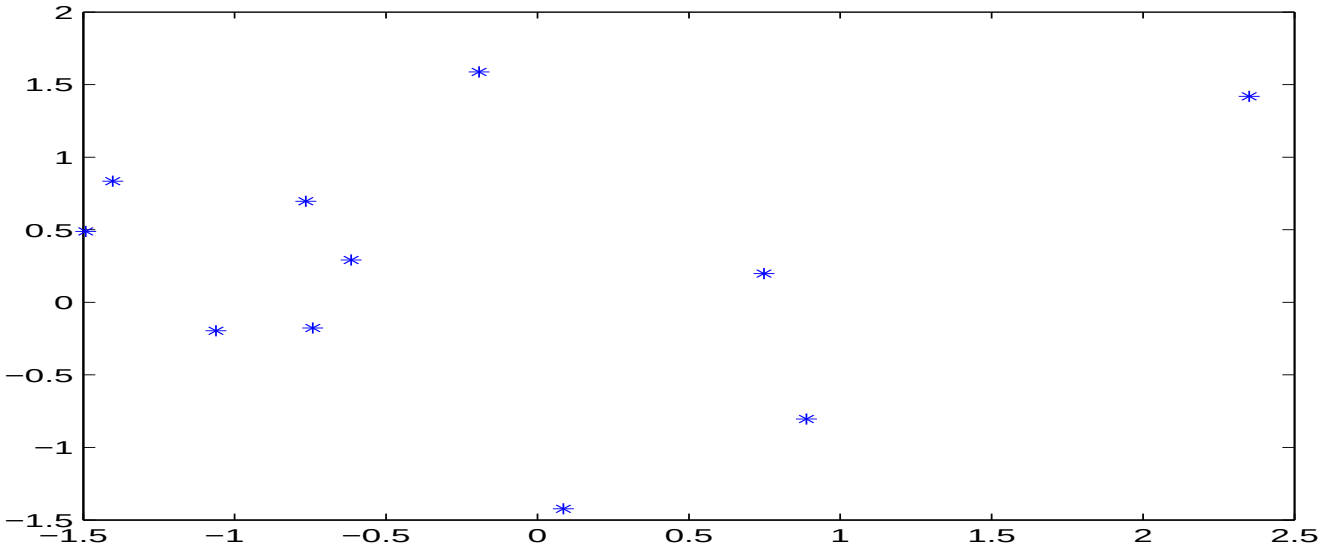
Coefficent of correlation

Here $r = -0.9977$



Coefficent of correlation

Here $r = 0.063$



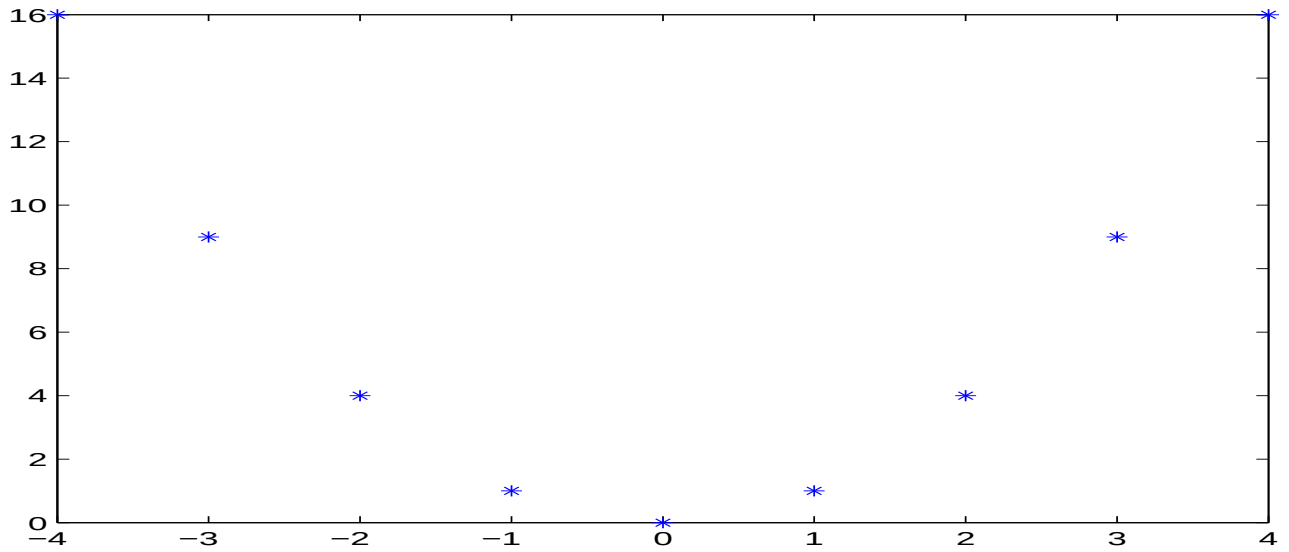
Coefficent of correlation: general properties

- *r is always between -1 and 1, i.e., $-1 \leq r \leq 1$. $r = 1$: perfect positive correlation, $r = -1$: perfect negative correlation, $r = 0$: x and y are non-correlated. (Recall the discussion of the sign of c_{xy})*
- *the value of r does not change if all values of the other variable are converted to a different scale*
- *r is not changed if x values are interchanged with y values.*
- ***r measures the strength of a linear association, it is not designed measure the strength of an association that is not linear.***

KTH Matematik

A non-linear association and the coefficent of correlation

Here $y = x^2$ is plotted for $x = -4, -3, \ldots 3, 4$. $r = 0.000$!



KTH Matematik

Hypothesis test for correlation

$\rho =$ correlation coefficient in the population

$$H_o : \rho = 0, \quad H_1 : \rho \neq 0,$$

Test statistic

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Critical values are found from the t-distribution with $n-2$ degrees of freedom. Then we follow the standard procedure.

KTH Matematik

Correlation and Causation

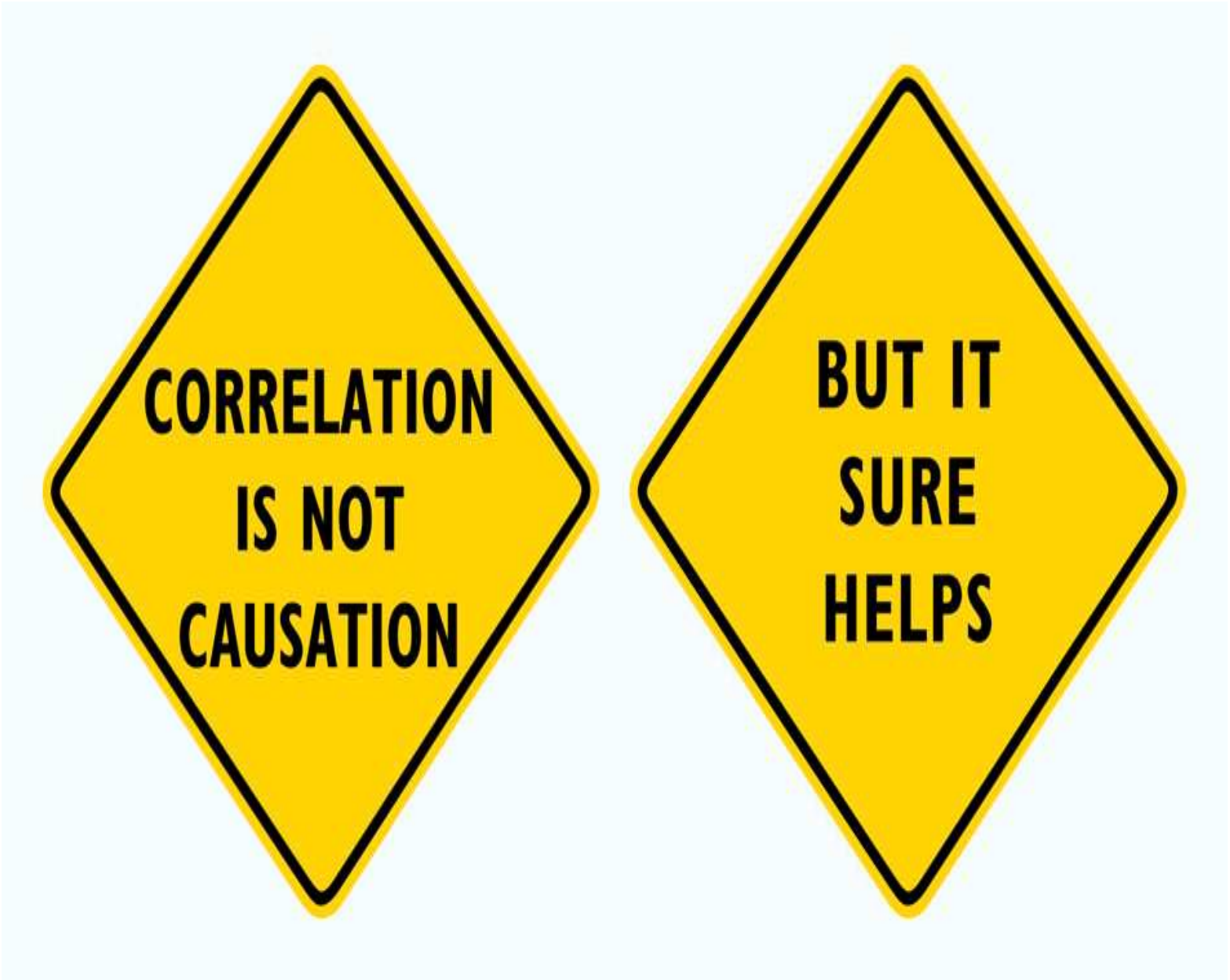
Correlation does not imply causation *is a phrase used to emphasize that a correlation between two variables does not necessarily imply that one causes the other. The counter assumption, that correlation proves causation, is considered a questionable cause logical fallacy in that two events occurring together are taken to have a cause-and-effect relationship.*

For a more thoroughgoing analysis, see

Bill Shipley: Cause and Correlation in Biology, *Cambridge University Press 2000*

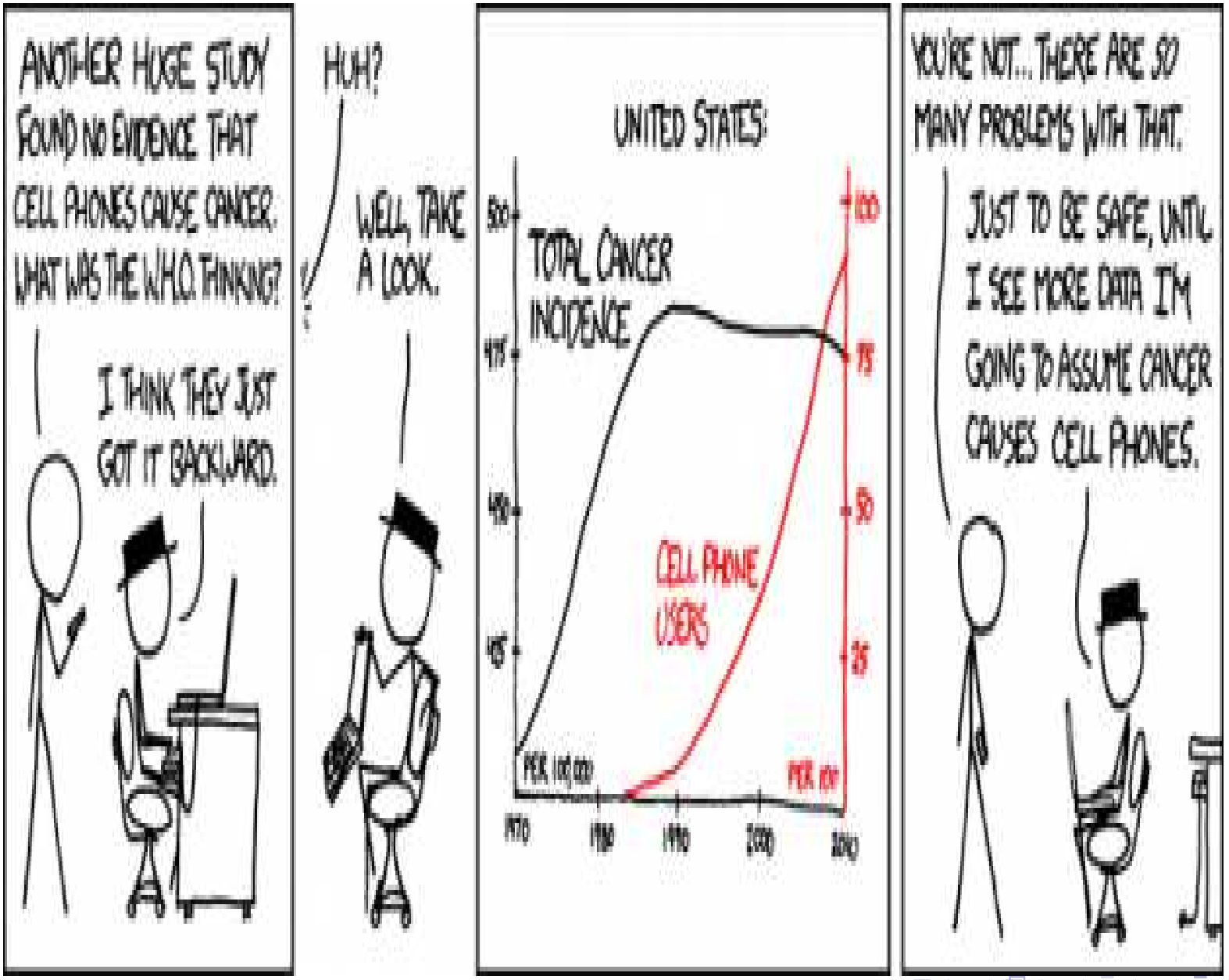
KTH VETENSKAP OCH KONST
KTH Matematik

Correlation and Causation



KTH Matematik

Correlation and Causation



Regression

We have now discussed paired data with the goal of determining whether there is a significant (= we can reject $H_o : \rho = 0$) linear correlation between two variables.
Now the goal is to describe the association between two variables by finding the graph and the straight line that represents the association.
This straight line is called the **regression line**.

KTH VETENSKAP OCH KONST
KTH Matematik

Regression: what is in the name ?

Sir Francis Galton, 1822-1911, bought family records that contained the heights of 205 sets of parents and their adult children. If the parents were short their children were slightly taller, on the other hand, if the parents were tall then the children were slightly shorter. This lead Galton to invent the word regression. *Regression was defined as the process of returning to the mean*. In the experiment the smallest parents had offspring who were bigger and closer to the mean. The largest parents had offspring who were smaller and once again closer to the mean.
During Galton's studies and experiments he invented words such as eugenics and regression. Galton first thought that breeding two smart people would produce an even smarter person. He also thought that breeding two tall people would produce an even taller person.

KTH Matematik

Regression line: terminology

$$\widehat{y} = a + bx \tag{1}$$

is called the **regression line**. Here x is called the **independent variable** or **predictor variable** or **explanatory variable** and $\widehat{y}$ is called the **dependent variable** or **response variable**. a is called the **intercept** and b is the **slope**. Later we will find a and b as sample estimates of *population intercept* α and *population slope* β.

KTH VETENSKAP OCH KONST

KTH Matematik

Galton and his students: biometrics

$$\widehat{y} = a + bx \tag{2}$$

The predictor/explanatory variable

$$x = \frac{\text{mother's height} + \text{father's height}}{2}.$$

The dependent variable or response variable y is the height of the child. The slope b measures heritability and the intercept a is like an average of enviromental effects.

Examples

- x = systolic reading, y = diastolic blood pressure
- x = cholesterol level, y = weight
- x = tree circumference, y = tree height
- e.t.c.

KTH VETENSKAP OCH KONST
KTH Matematik

Regression: marginal change

$$\widehat{y} = a + bx \tag{3}$$

When x changes by one unit we get the $\widehat{y}^{\dagger} = a + b(x+1)$. We have the **marginal change** $\triangle y$

$$\triangle y = \widehat{y}^{\dagger} - \widehat{y} = b$$

Hence the slope b represents the change in response, when the explanatory variable is changed by one unit.

KTH VETENSKAP OCH KONST
KTH Matematik

Method of Least Squares

We have paired data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We find a and b by the **method of least squares**, i.e., we minimize

$$Q = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{4}$$

as a function of a and b.

KTH VETENSKAP OCH KONST

KTH Matematik

Method of Least Squares

We minimize

$$Q = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{5}$$

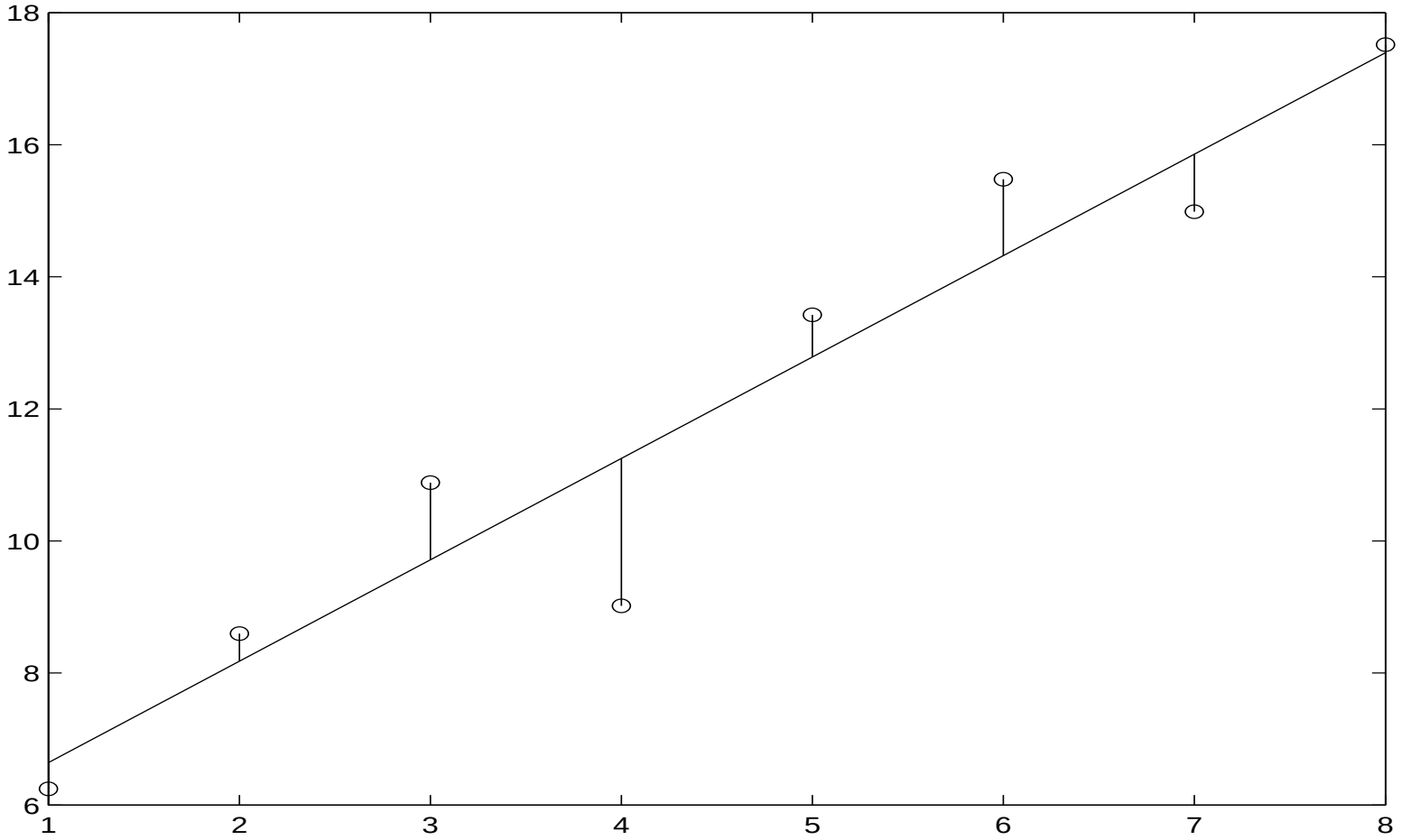
Geometrically this means that we minimize the sum of the squared vertical distances between observations y_i and regression line.

KTH VETENSKAP OCH KONST

KTH Matematik

Method of Least Squares

Minimize the sum of the squared vertical distances between observations y_i and regression line.



Method of Least Squares

$$Q = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{6}$$

By differentiating Q w.r.t. a and b and by setting the derivatives equal to 0, we find the slope and the intercept as

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{7}$$

$$a = \bar{y} - b\bar{x} \tag{8}$$

KTH VETENSKAP OCH KONST

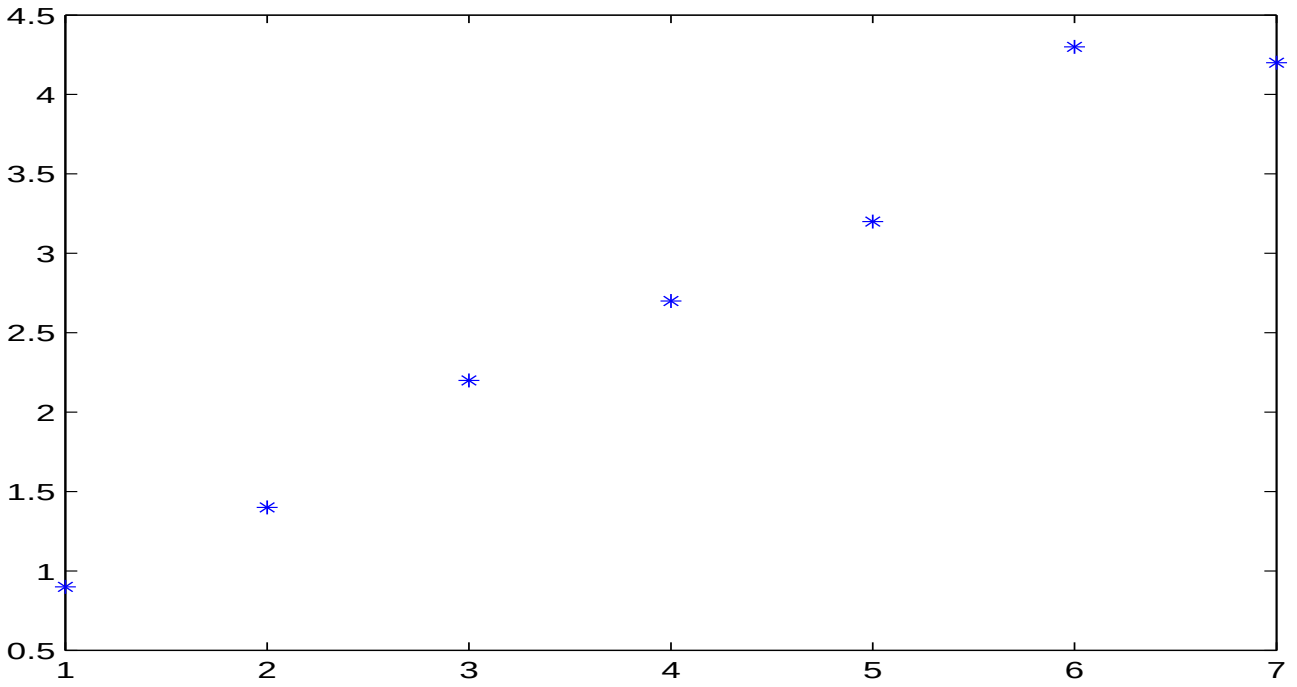
KTH Matematik

Scatterplot

An example

$$x = 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7$$

$$y = 0.9 \quad 1.4 \quad 2.2 \quad 2.7 \quad 3.2 \quad 4.3 \quad 4.2$$

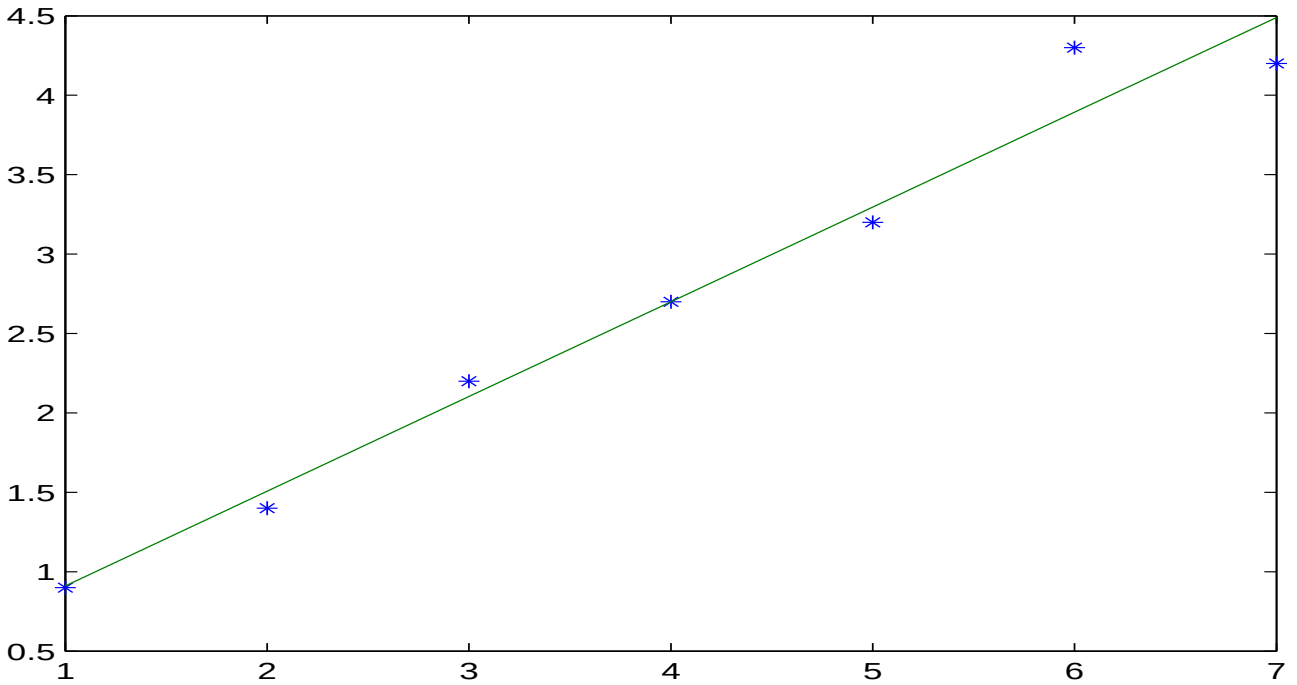


KTH Matematik

Linear regression

For the paired samples above the regression line is

$$\widehat{y} = 0.3143 + 0.5964x,$$



Residuals

The vertical distances e_i from y_i to regression line at x_i,

$$e_i \stackrel{\text{def}}{=} y_i - \widehat{y} = y_i - a - bx_i$$

when a and b are computed by the formulae above, are called **residuals**.

$$\underline{\text{Residual} = \text{observed y - predicted y}}$$

Q_0 is defined as

$$Q_0 = \sum_{i=1}^{n} e_i^2.$$

and is called the sum of the residual squares. For the example above

$$Q_0 = 0.2796$$

KTH VETENSKAP OCH KONST

KTH Matematik

Residuals

Since

$$Q = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

and its minimum value is

$$Q_0 = \sum_{i=1}^{n} e_i^2.$$

we can see the least squares method as minimizing the sum of residual squares.

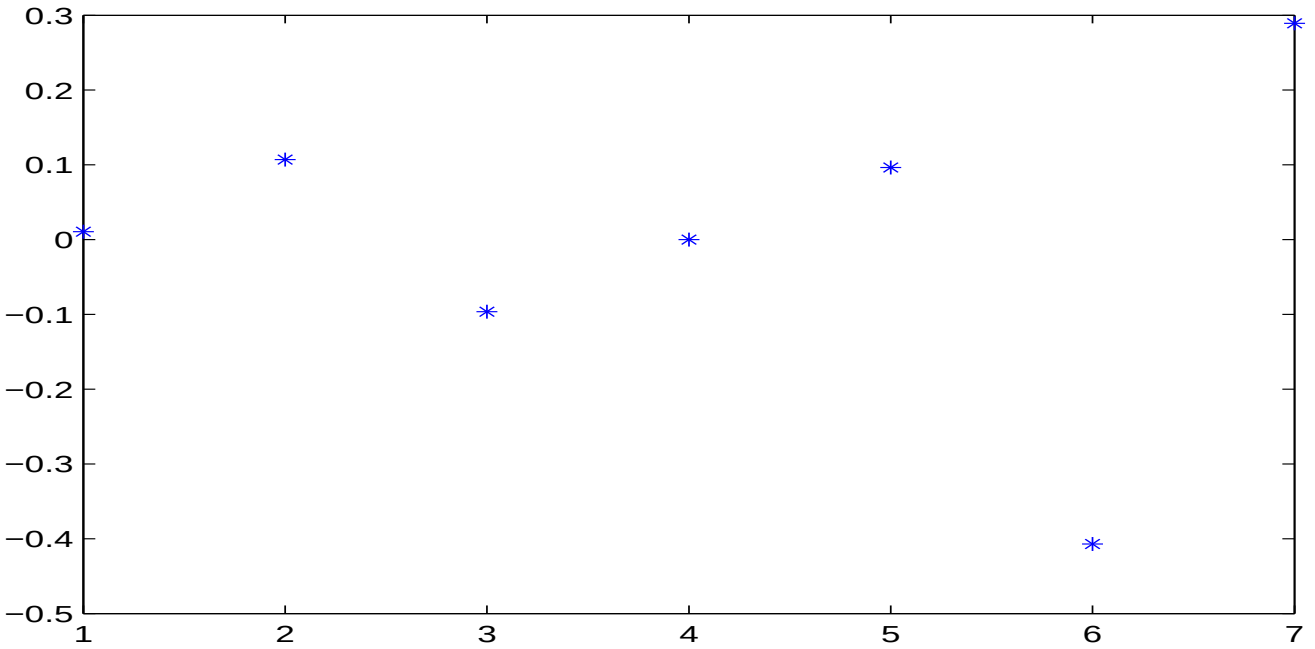
KTH VETENSKAP OCH KONST

KTH Matematik

Residual plot

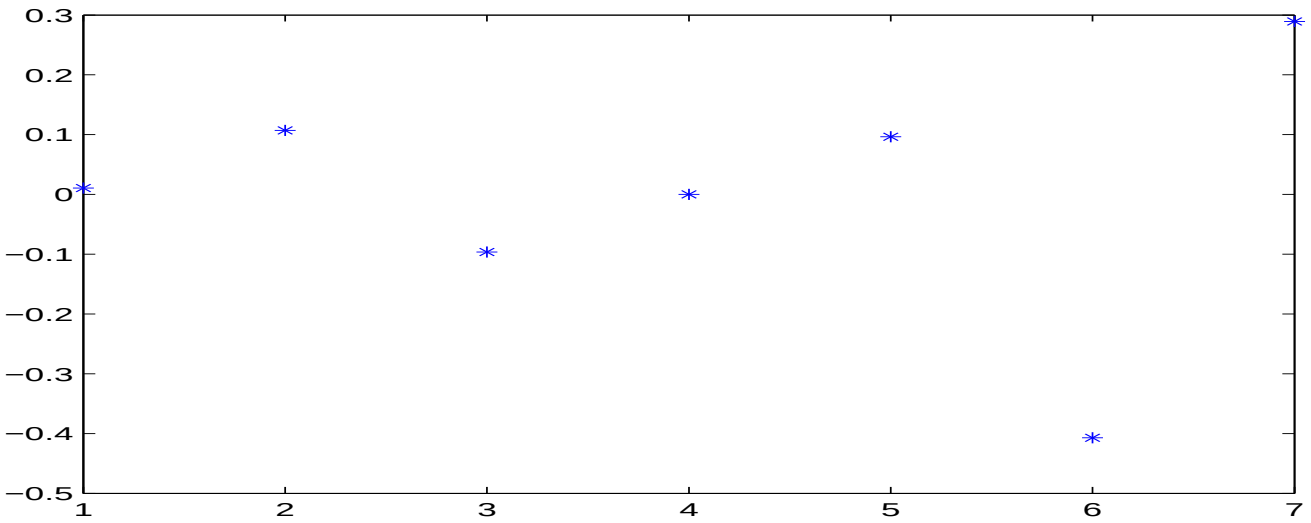
Definition

A **residual plot** *is a scatterplot of the pairs* (x_i, ε_i) $i = 1, 2, \ldots n$.

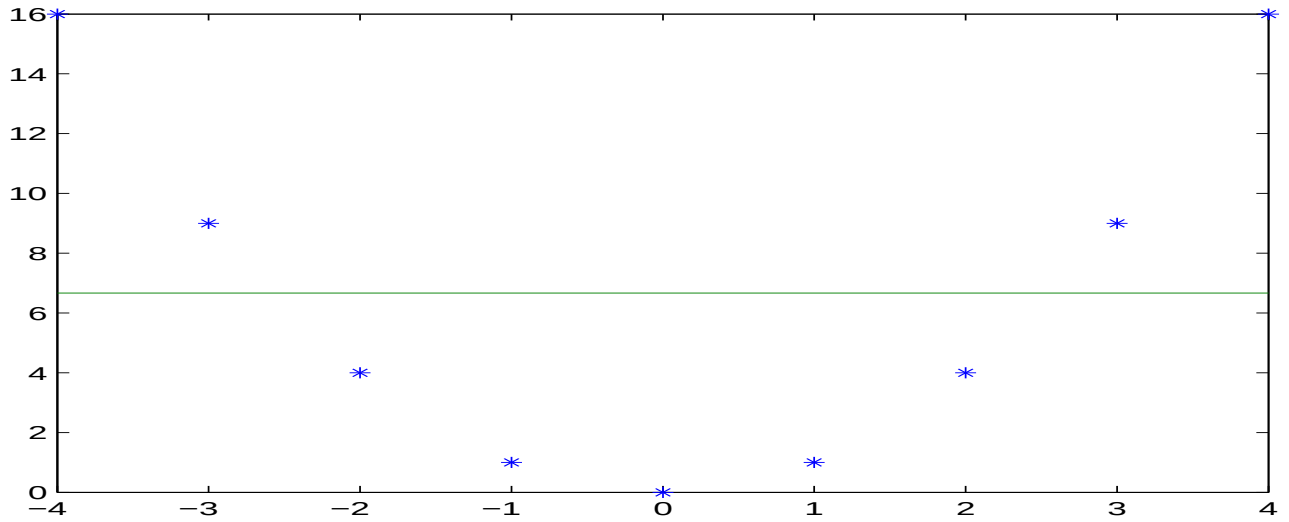


Residual plot

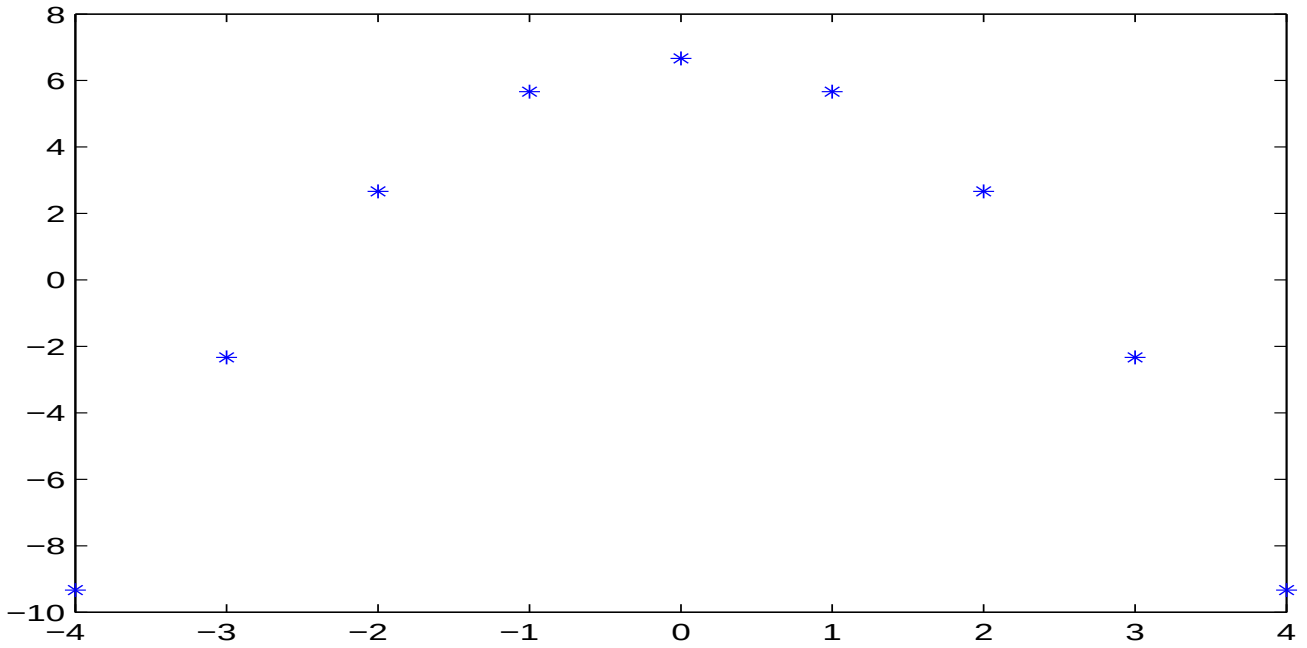
If the residual plot reveals no pattern, the regression equation is a good representation of the association between the two variables. If the residual plot reveals some systematic pattern, the regression equation is not a good representation of the association between the two variables. Here is the residual plot for the data in the example.



Regression for $y = x^2$



Regression for $y = x^2$ and the residual plot



A Question

Having found from samples a regression line like

$$\widehat{y} = 0.3143 + 0.5964x,$$

we might want to be assured, e.g., of that the slope $b = 0.5964$ is significantly different from 0. If we want a confidence interval or a statistical test, we need a statistical model.

KTH VETENSKAP OCH KONST

KTH Matematik

Statistical model

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n \tag{9}$$

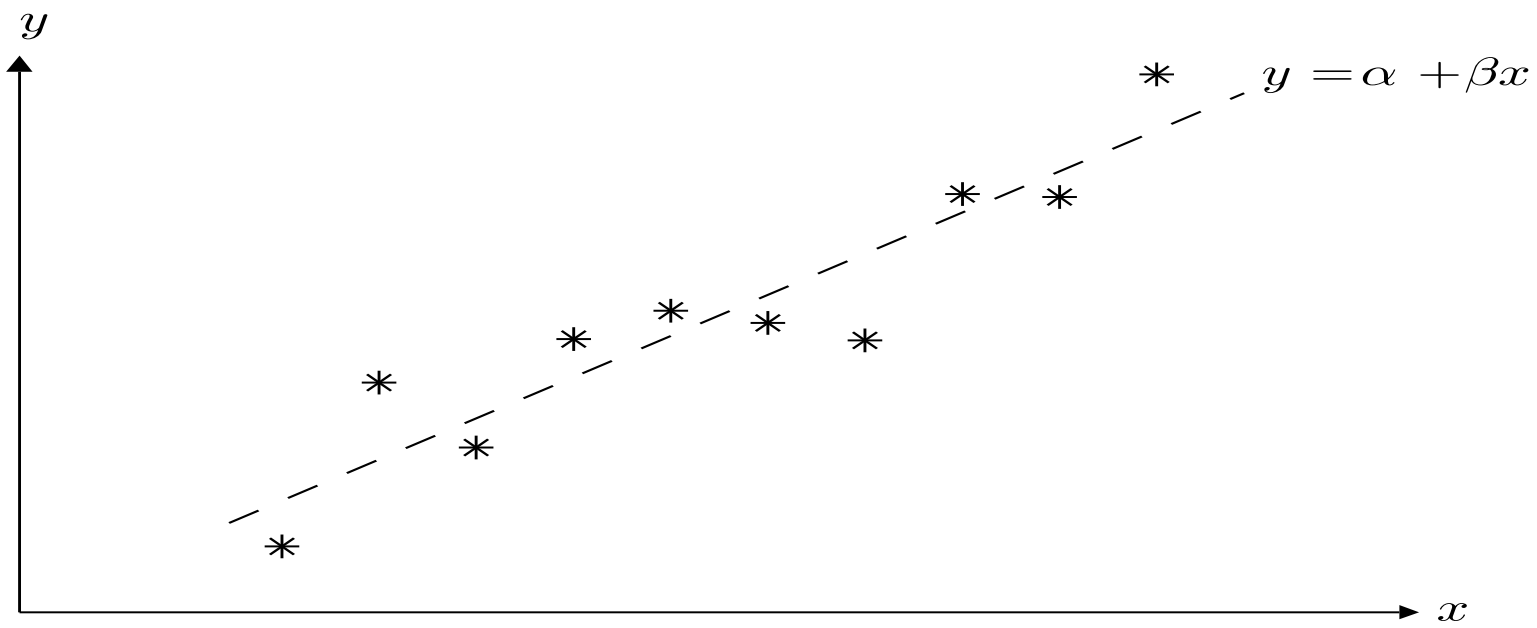
where $x_i, i = 1, 2, \ldots, n$ are known predictor values, $\varepsilon_i,\ i = 1, 2, \ldots, n$ are assumed to be independent and have normal distribution with mean 0 and variance σ^2, $\varepsilon_i \sim N(0, \sigma)$. Here α and β are unknown population parameters. a and b above are the respective estimates by means of the method of least squares.

KTH VETENSKAP OCH KONST

KTH Matematik

Theoretical regression line

$$\widehat{Y} = E(Y_i) = \alpha + \beta x_i, \quad V(Y_i) = \sigma^2$$



Statistical model & residuals

The residuals

$$e_i = y_i - a - bx_i$$

are computed from data using a and b (as given by the formulas above) and are thus observable.
The random variables

$$\varepsilon_i = Y_i - (\alpha + \beta x_i)$$

are not observable, they are statistical disturbances that push our outcomes y_i away from the theoretical regression line. Of course, we may regard e_is as estimates of ε_is.

KTH VETENSKAP OCH KONST
KTH Matematik

Theoretical regression line

$$V(Y_i) = \sigma^2$$

is a new parameter to be estimated from the paired samples.

$$Q_0 = \sum_{i=1}^{n} e_i^2.$$

is the sum of squared residuals and the estimate of σ^2 is

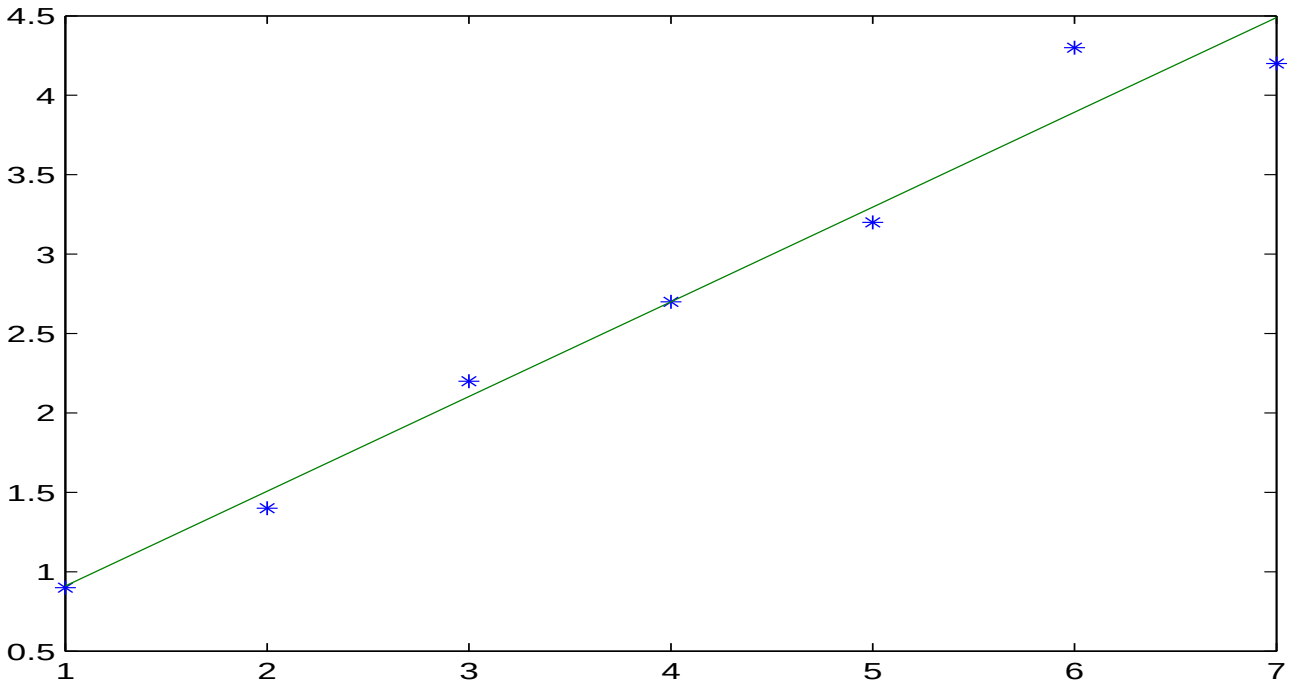
$$s^2 = \frac{Q_0}{n-2}.$$

KTH VETENSKAP OCH KONST

KTH Matematik

Example

$$y = 0.3143 + 0.5964x, \quad Q_0 = 0.2796, \quad s^2 = 0.0559$$



KTH Matematik

Confidence interval for α

Confidence interval for α with the degree of confidence $= p$ is given by

$$(a - E, a + E)$$

where

$$E = t_{(p/2)}(n-2)s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

and $t_{(p/2)}(n-2)$ is the critical value with significance level p from the t-distribution with $n-2$ degres of freedom.

Confidence interval for β

Confidence interval for β with the degree of confidence $= p$ is given by

$$(b - E, b + E)$$

where

$$E = t_{(p/2)}(n-2)s/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and $t_{(p/2)}(n-2)$ is the critical value with significance level p from the t-distribution with $n-2$ degres of freedom.

Test of Hypothesis β

$$H_0: \ \beta = 0$$

$$H_1: \ \beta \neq 0$$

Reject H_0 at level 0.05 if the interval

$$(b - E, b + E)$$

does not include 0.

$$E = t_{0.025}(n-2)\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

KTH VETENSKAP OCH KONST

KTH Matematik

Confidence interval for the regression line at x_0

Confidence interval for the regression line $\widehat{y}_0 = \alpha + \beta x_0$ at x_0 with the degree of confidence $= p$ is given by

$$(a + bx_0 - E, a + bx_0 + E)$$

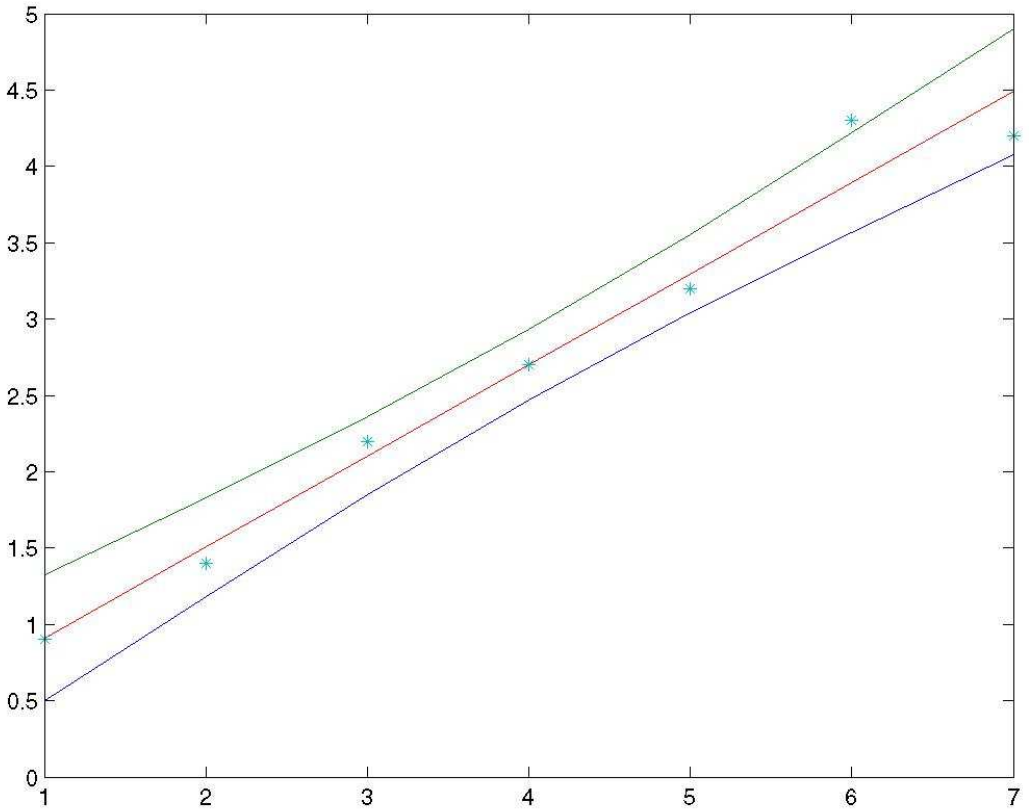
where

$$E = t_{(p/2)}(n-2)s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

and $t_{(p/2)}(n-2)$ is the critical value with confidence level p from the t-distribution with $n-2$ degres of freedom.

Confidence interval for the regression line of y given x_0

In the example above



Prediction interval for an individual y

The interval

$$(a + bx_0 - E, a + bx_0 + E)$$

with

$$E = t_{0.025}(n-2)s \cdot \sqrt{1 + \frac{1}{n} + (x_0 - \bar{x})^2 / S_{xx}}$$

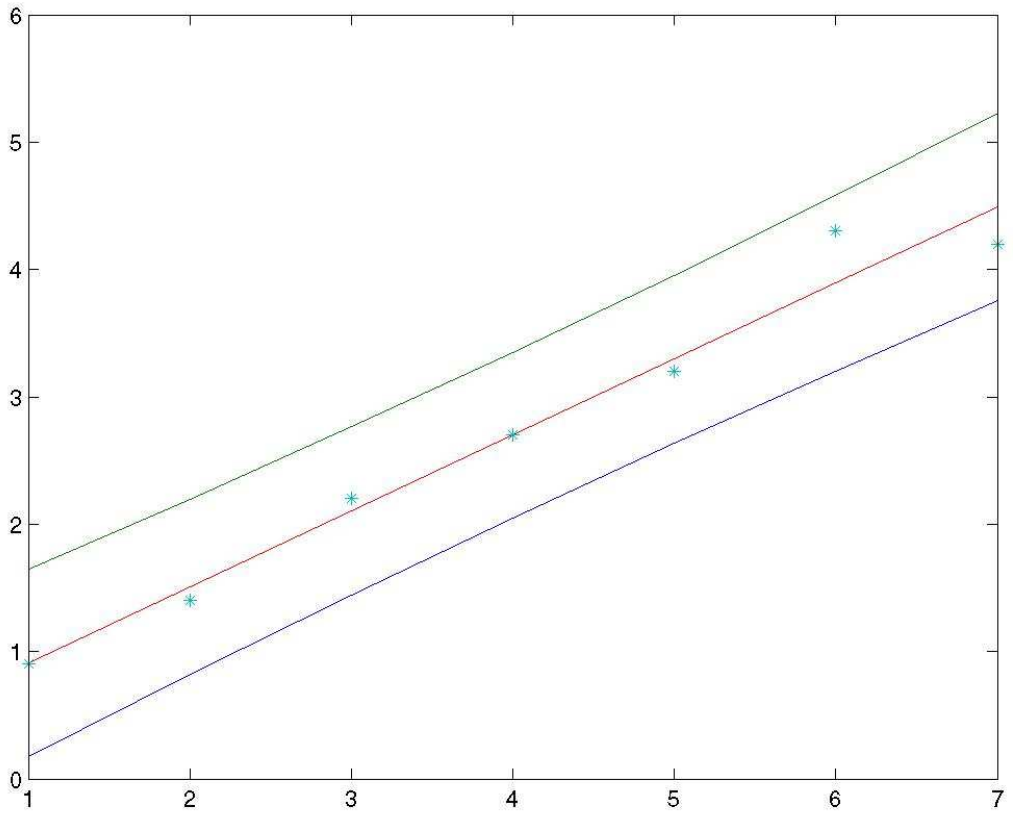
is called 95% prediction interval for y given the predictor value x_0.

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

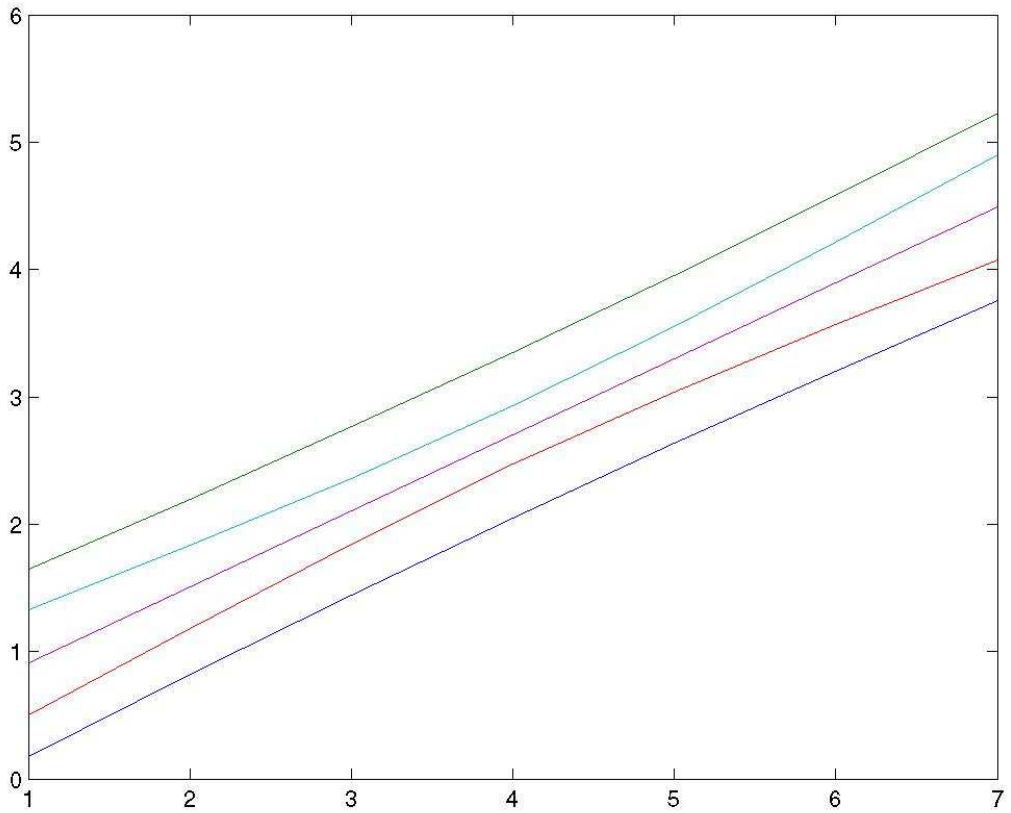
KTH VETENSKAP OCH KONST

KTH Matematik

Prediction interval for an individual y in the example



Prediction interval and the confidence interval for the regression line, the regression line



A Chemistry Example

Paint surfaces with $x=$ different dilutions of of lacquer paint by petroleum sprit and

Dilution	Drying Time	
10	8.3	8.0
20	8.0	8.3
30	7.3	7.5
40	6.9	6.5
50	6.2	5.9

KTH Matematik

Chemistry Example: Number crunching

The estimate of β is

$$b = \frac{\sum_1^{10}(x_i - \bar{x})y_i}{\sum_1^{10}(x_i - \bar{x})^2} = \frac{-113}{2000} = -0.0565$$

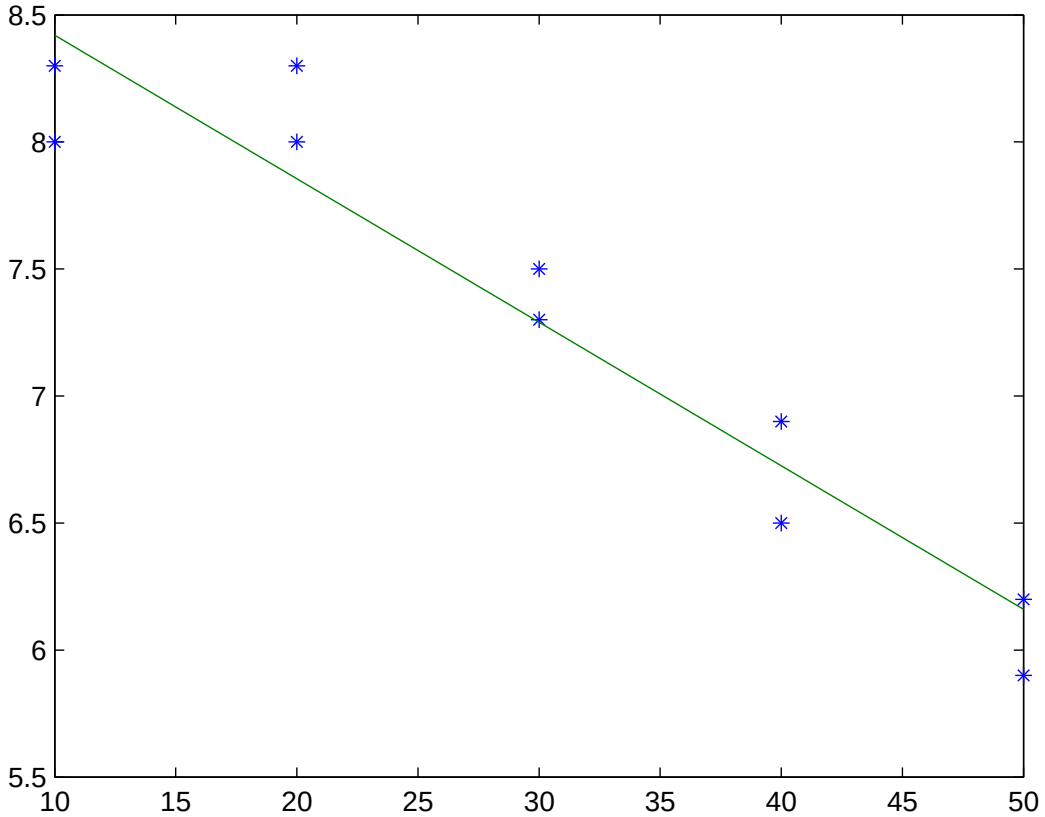
and

$$a = \bar{y} - \widehat{\beta}\bar{x} = 7.29 - (-0.0565) \cdot 30 = 8.985$$

and the regression line is

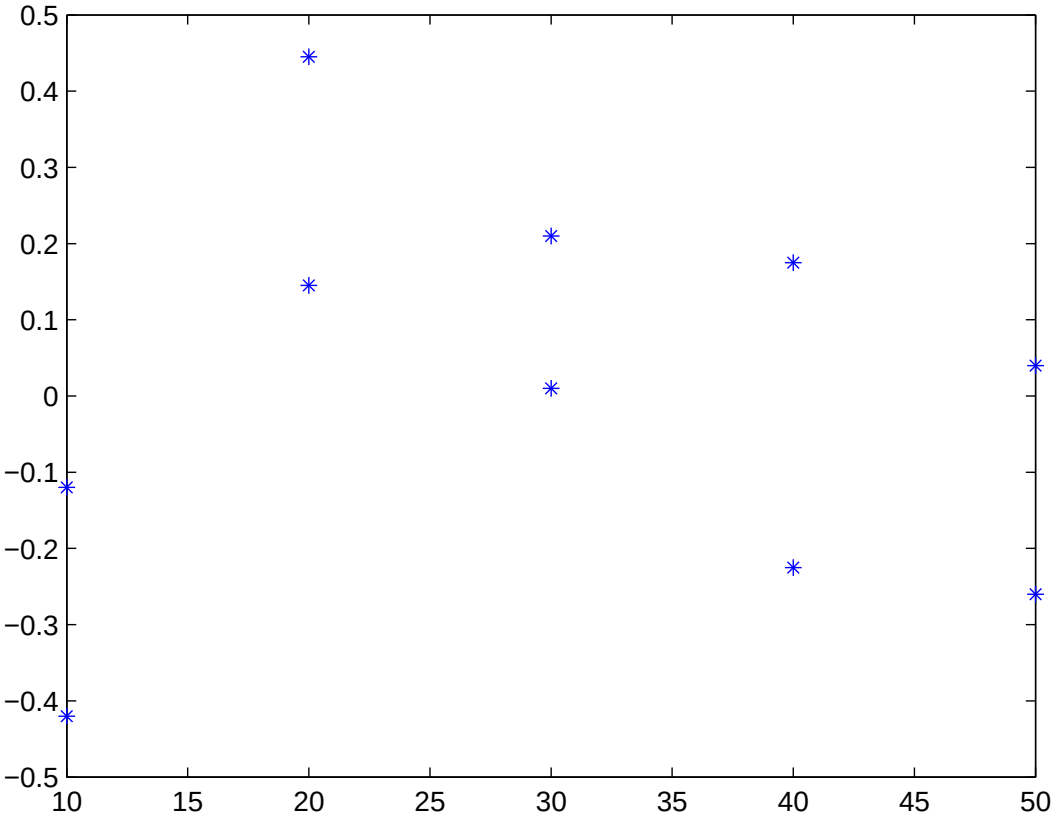
$$y = 8.985 - 0.0565 \cdot x$$

Chemistry Example: Scatterploot and the Regression Line



KTH VETENSKAP OCH KONST
KTH Matematik

Chemistry Example: Errorplot



Chemistry Example: CI for β

95%- confidence interval for β becomes

$$b + \pm t_{0.95,10-2} \frac{s}{\sqrt{\sum_1^{10}(x_i - \bar{x})^2}}.$$

$$s^2 = \frac{1}{10-2}\left(\sum_1^{10}(y_i - \bar{y})^2 - b^2 \sum_1^{10}(x_i - \bar{x})^2\right) =$$

$$\frac{1}{8}\left(538.43 - 10 \cdot 7.29^2 - (-0.0565)^2 \cdot 2000\right) = \frac{0.6045}{8} = 0.0755625.$$

KTH Matematik

We get

$$E = 2.31\frac{\sqrt{0.0755625}}{\sqrt{2000}} \approx 0.0142$$

and the interval

$$(-0.0565 - 0.0142, -0.0565 + 0.0142) = (-0.0707, -0.0423)$$

Hence $H_o : \beta = 0$ is rejected.

Multivariate regression

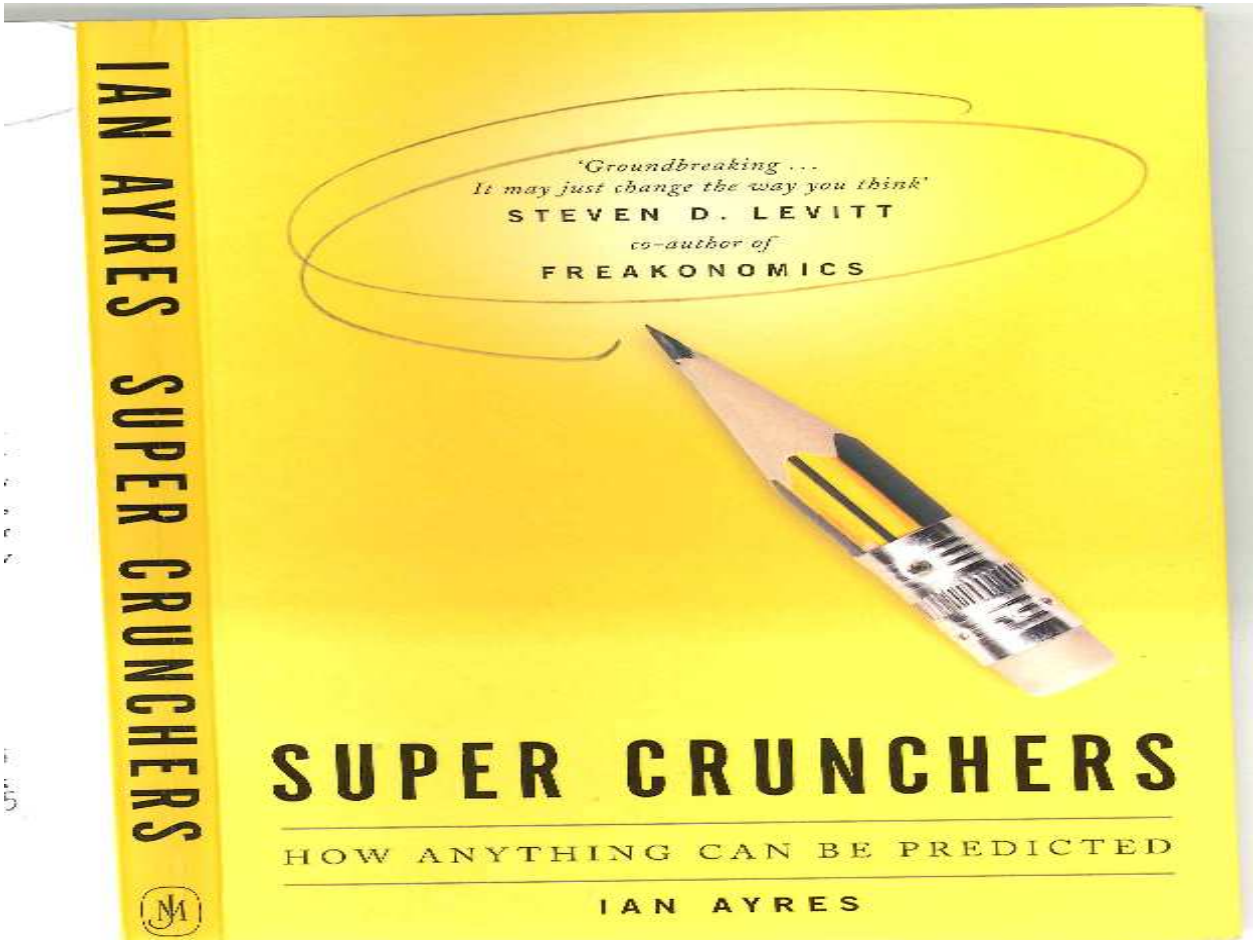
$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \tag{10}$$

where $x_j, j = 1, 2, \ldots, k$ are k predictors of the response y, variables whose values can be determined in advance of y values. Matrix calculus and computers are needed for treatment of multivariate regression.

KTH VETENSKAP OCH KONST

KTH Matematik

Multivariate regression



Analysis of Variance

- We have in the preceding treated testing of the equality between two means by the t-test.

Analysis of Variance

Analysis of variance (ANOVA) *is a method of testing the equality of three or more population means by analyzing sample variances.*

Typical applications:

- We treat one group with two aspirin tablets each day, and second group with one aspirin tablet a day, while a third group is given a placebo each day. We want to determine if there is sufficient evidence to support that the three groups have different mean blood pressure levels.

KTH VETENSKAP OCH KONST
KTH Matematik

Analysis of Variance

Analysis of variance (ANOVA) *is a method of testing the equality of three or more population means by analyzing sample variances.*

Typical applications:

- The analysis of microarray gene expression data typically tries to identify differential gene expression patterns in terms of differences of the population means between groups of arrays (e.g. treatments or biological conditions).

KTH Matematik

Analysis of Variance and e.g., Gene Expression Microarray Data

One question is how to make valid estimates of the relative expression for genes. Recognizing that there is inherent "noise" in microarray data, how does one estimate the error variation associated with an estimated change in expression, i.e., how does one construct the error bars? We demonstrate that ANOVA methods can be used to normalize microarray data and provide estimates of changes in gene expression that are corrected for potential confounding effects.

Kerr, M. K., Martin, M. and Churchill, G.A. : Analysis of variance for gene expression microarray data, Journal of computational biology, *pp. 819-837, 2000.*

KTH VETENSKAP OCH KONST
KTH Matematik

Analysis of Variance

We deal with **one-way analysis of variance** *or* **single factor anlysis of variance**, *there is only one property describing the population.*

Definition

A **treatment** *or* **factor** *is a property that allows us to distinguish different populations from each other.*

One-Way Analysis of Variance

We have data that have been obtained so that a *factor* A is varied at k different levels $A_1,\ A_2, \ldots, A_k$. At level A_i we have n_i data values, $y_{i1}, y_{i2}, \ldots, y_{in_i}$.
Our statistical model is that these are outcomes of random variables $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$. We assume that all have $N(\mu_i, \sigma)$ distribution.
The quantities $\mu_1, \mu_2 \ldots, \mu_k$ are thus means at the different levels.
Our goal is to compare these means.

KTH VETENSKAP OCH KONST
KTH Matematik

One-Way Analysis of Variance: data table

Level	Observations				Mean	Sample variance
A_1	y_{11}	y_{12}	$\cdots$	y_{1n_1}	$\bar{y}_{1\cdot}$	s_1^2
A_2	y_{21}	y_{22}	$\cdots$	y_{2n_2}	$\bar{y}_{2\cdot}$	s_2^2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
A_k	y_{k1}	y_{k2}	$\cdots$	y_{kn_k}	$\bar{y}_{k\cdot}$	s_k^2

One-Way Analysis of Variance: notations

Here the notation is admittedly a complex one:

Level	Observations				Mean	S Sample variance	$y_{1.}$
A_1	y_{11}	y_{12}	$\cdots$	y_{1n_1}	$\bar{y}_{1.}$	s_1^2	

means that we have summed over the second index:

$$y_{1.} = \frac{1}{n_1}\left(y_{11} + y_{12} + \ldots + y_{1n_1}\right)$$

Thus

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}\left(y_{1j} - y_{1.}\right)^2$$

One-Way Analysis of Variance: notations

$$\bar{y}_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$$

is the grand mean, $N = n_1 + n_2 + \cdots + n_k$ is total number of samples.

$$\bar{y}_{..} = \frac{1}{N} (y_{11} + y_{12} + \ldots + y_{1n_1}$$

$$+ \ldots +$$

$$\ldots + y_{k1} + y_{k2} + \ldots + y_{kn_k})$$

KTH Matematik

One-Way Analysis of Variance

ANOVA estimates threee sample variances: a **total variance** based on all the **observation deviations from the grand mean**, an **error variance** based on all the **observation deviations from their appropriate treatment means** $(y_{i.})$ and a treatment variance. The **treatment variance** is based on the **deviations of treatment means from the grand mean**, the result being multiplied by the number of observations in each treatment to account for the difference between the variance of observations and the variance of means.

KTH Matematik

One-Way Analysis of Variance: The Hypothesis

$\mu_1, \mu_2, \ldots, \mu_k$ and σ^2 are unknown parameters. The statistical problem is to test whether all means are equal.

$$H_o: \ \mu_1 = \mu_2 = \cdots = \mu_k$$

This is the claim that all treatments, instruments e.t.c. are of equal quality.

KTH VETENSKAP OCH KONST

KTH Matematik

One-Way Analysis of Variance: ANOVA Table

Source	df	SS	MSS
Between samples	$k-1$	$\sum_{i=1}^{k} n_i(\bar{y}_{i.}-\bar{y}_{..})^2$	SS/df
Within samples	$N-k$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2$	$\widehat{\sigma}^2$=SS/df
Total	$N-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2$	

Source = source of variation, df= degrees of freedom, SS= sum of squares, MSS= mean sum of squares. and

$N = n_1 + n_2 + \cdots + n_k$, total number of samples.

KTH Matematik

ANOVA Table: treatment variance

$$\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

measures the dispersion of the means. If this sum is large, then we may suspect that the factor levels are systematically different.

KTH VETENSKAP OCH KONST

KTH Matematik

ANOVA Table: error variance

The second sum of squares can be written as

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2=\sum_{i=1}^{k}s_i^2(n_i-1)$$

where s_i^2 is the sample variance for level i. These are measurements of random variation, i.e., of σ^2.

KTH Matematik

ANOVA Table: the sums of squares

$$\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 = \sum_{i=1}^{k} s_i^2(n_i - 1)$$

By comparing these two sums of squares with each other we can decide if the levels are of equal value.

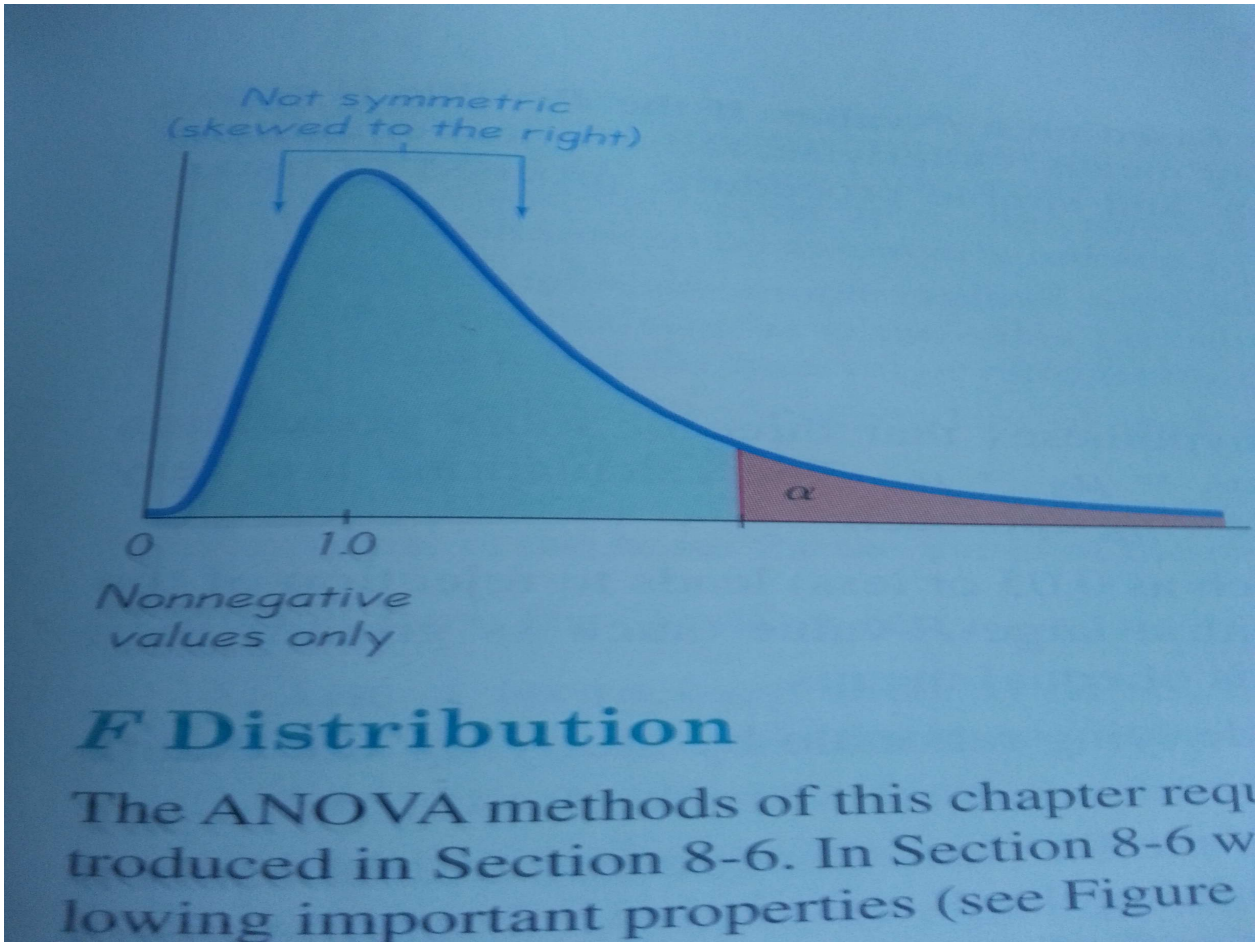
ANOVA Test Statistic

We should compare the two sums of squares by aid of their ratio so that the **test statistic** is taken as

$$F_A \overset{\text{def}}{=} \frac{\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2/(k-1)}{\widehat{\sigma}^2} = \frac{\text{MSS between}}{\text{MSS within}}$$

It can be shown by an extensive exercise in mathematical statistcs that F_A has an F-**distribution**, if H_o is true. (F for Sir R.A. Fisher).

ANOVA F-distribution



KTH Matematik

ANOVA Test Statistic

Test statistic:

$$F_A = \frac{\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2/(k-1)}{\widehat{\sigma}^2} = \frac{\text{MSS between}}{\text{MSS within}}$$

The hypothesis H_A is rejected if $F_A > F_p(k-1, N-k)$. The critical value $F_p(k-1, N-k)$ gives the level of significance p, if H_o is true, and is found in a table for percentiles of the F-distribution.

KTH Matematik

F- distribution quantiles $F_{0.05}(f_2, f_1)$ %

f_2/f_1	1	2	3	4	5	6	7	8	9	10
1	161	200	216	225	230	234	237	239	241	242

These are found in Norman and Streiner pp. 366−367

KTH VETENSKAP OCH KONST

KTH Matematik

ANOVA computations

These computations are simple and can in principle be done by hand. This is very time consuming and software is preferably used. There is statistical software, and even Excel can handle ANOVA tables.

KTH VETENSKAP OCH KONST

KTH Matematik

Rothamstead Experimental Station

Frank Yates, in his office in 1974, using the Millionaire calculating machine developed by and built for R.A. Fisher in the 1920's.

KTH VETENSKAP OCH KONST
KTH Matematik

ANOVA Table

The lowermost row in the ANOVA table has not been given any role so far. It is there for computational reasons: this row is often easier to calculate and and the other rows are obtainable by subtraction. When relying on computers and calculators this makes little difference.

KTH VETENSKAP OCH KONST

KTH Matematik

ANOVA: Example

Four instruments of measurement of length are compared. One operator measured one and the same length with each of the instruments. In the table we see the results.

Instrument	Observations			
A_1	1236	1238	1239	
A_2	1235	1234		
A_3	1236	1237	1238	
A_4	1233	1235	1234	1236

KTH Matematik

ANOVA: Example

Source	df	SS	MSS
Between instruments	3	$24\frac{3}{4}$	$33/4$
Within instruments	8	$12\frac{1}{6}$	$\widehat{\sigma}^2 = 73/48$
Total	11	$36\frac{11}{12}$	

Test statistic $F = \dfrac{33/4}{73/48} = 5.42 > 4.07 = F_{0.05}(3, 8)$. Hypothesis that the instruments are of equal value is thus rejected.

F-distribution, critical values

f_2 / f_1	1	2	3	4	5	6
1	161	200	216	225	230	234
2	18.5	19	19.2	19.2	19.3	19.3
3	10.1	9.55	9.28	9.12	9.01	8.94
4	7.71	6.94	6.59	6.39	6.26	6.16
5	6.61	5.79	5.41	5.19	5.05	4.95
6	5.99	5.14	4.76	4.53	4.39	4.28
7	5.59	4.74	4.35	4.12	3.97	3.87
8	5.32	4.46	4.07	3.84	3.69	3.58
9	5.12	4.26	3.86	3.63	3.48	3.37
10	4.96	4.1	3.71	3.48	3.33	3.22
11	4.84	3.98	3.59	3.36	3.2	3.09

KTH VETENSKAP OCH KONST

KTH Matematik

Sir Ronald A. Fisher the founder of biostatistics (computing critical values of the F-distribution)

KTH Matematik

End

KTH Matematik