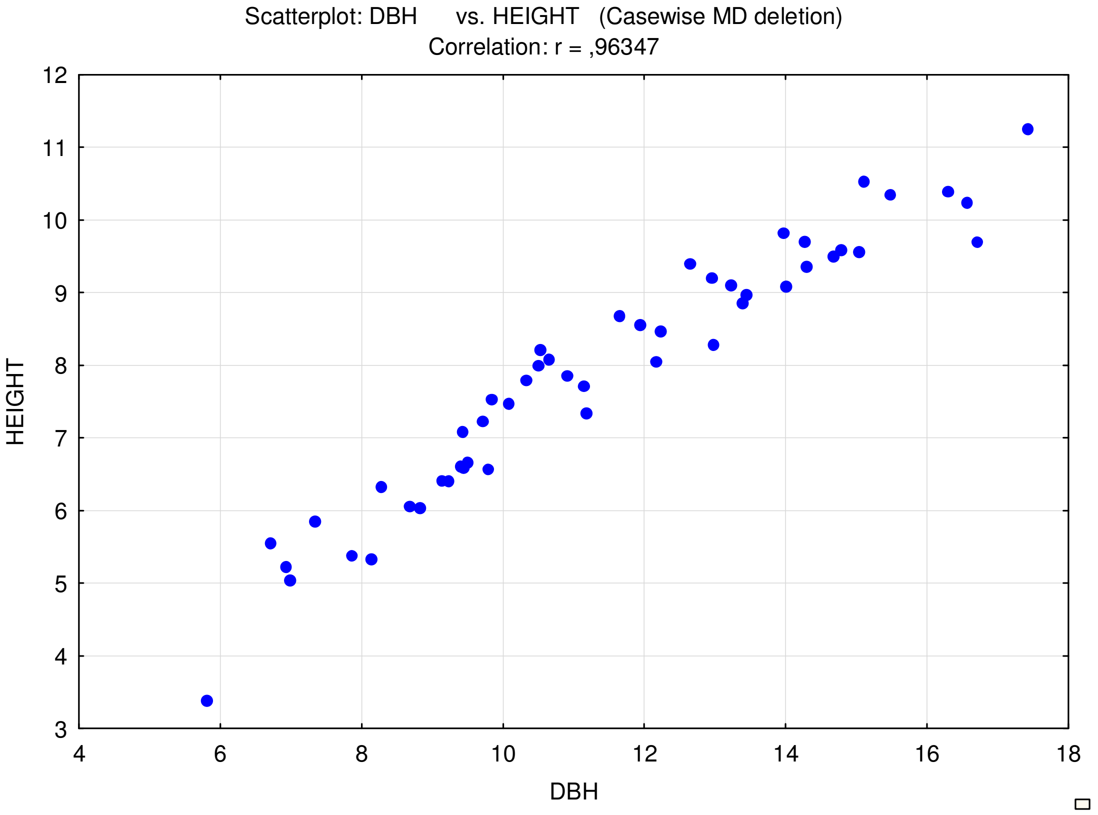
Two Dimensional Distributions and Linear Correlation

In most applications we observe more than one variable at each individual or sampling unit.

The first step of analysis: Scatterplot

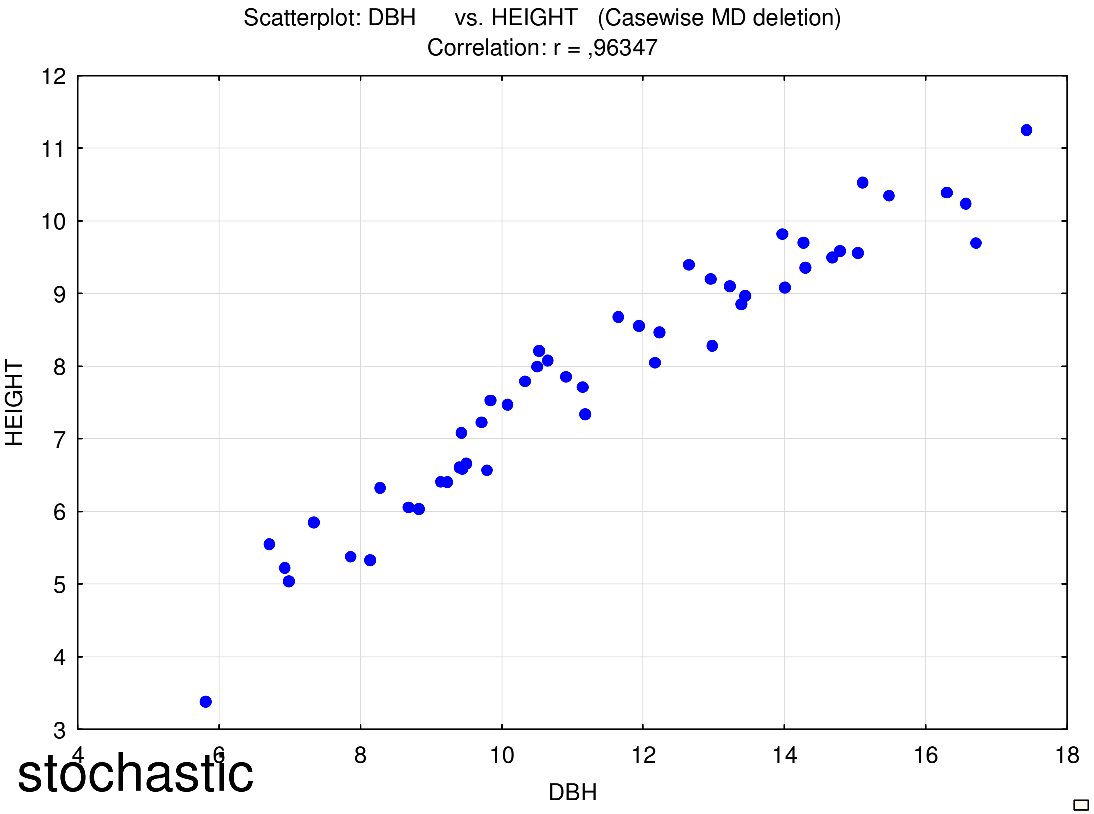


Correlation

Is a measure of relationship between two variables or measured data.

A measure of the strength and direction of **linear** dependence between two variables

Height (y; m) vs. Diameter at Breast Height - DBH (x; cm):

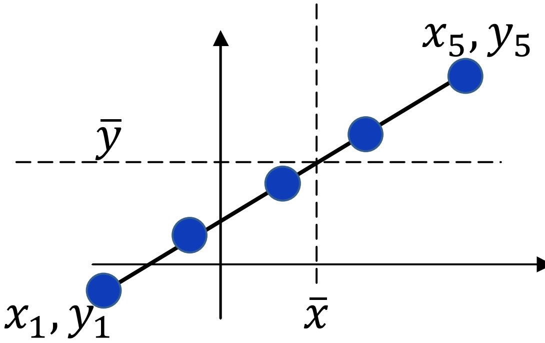


The relationship is stochastic

- Imperfect
- Appears as a scatterplott .

A functional linear relationship

- The equation : $y_i = a + bx_i; \quad a - Intercept, b - slope$
- Graph: All points (x_i, y_i) lie on the line
- $\frac{(y_i - \bar{y})}{(x_i - \bar{x})} = b$



- If $b > 0$

$$\left.\begin{array}{l}(x_i - \bar{x}) > 0 \Longrightarrow (y_i - \bar{y}) > 0 \\ (x_i - \bar{x}) < 0 \Longrightarrow (y_i - \bar{y}) < 0\end{array}\right\} allways!$$

Deviations of values of the variables X and Y from the correspondent means have always **the same** sign

- If $b < 0$

$$\left.\begin{array}{l}(x_i - \bar{x}) > 0 \Longrightarrow (y_i - \bar{y}) < 0 \\ (x_i - \bar{x}) < 0 \Longrightarrow (y_i - \bar{y}) > 0\end{array}\right\} allways!$$

Deviations of values of the variables X and Y from the correspondent means have always different signs

A stochastic linear relationship

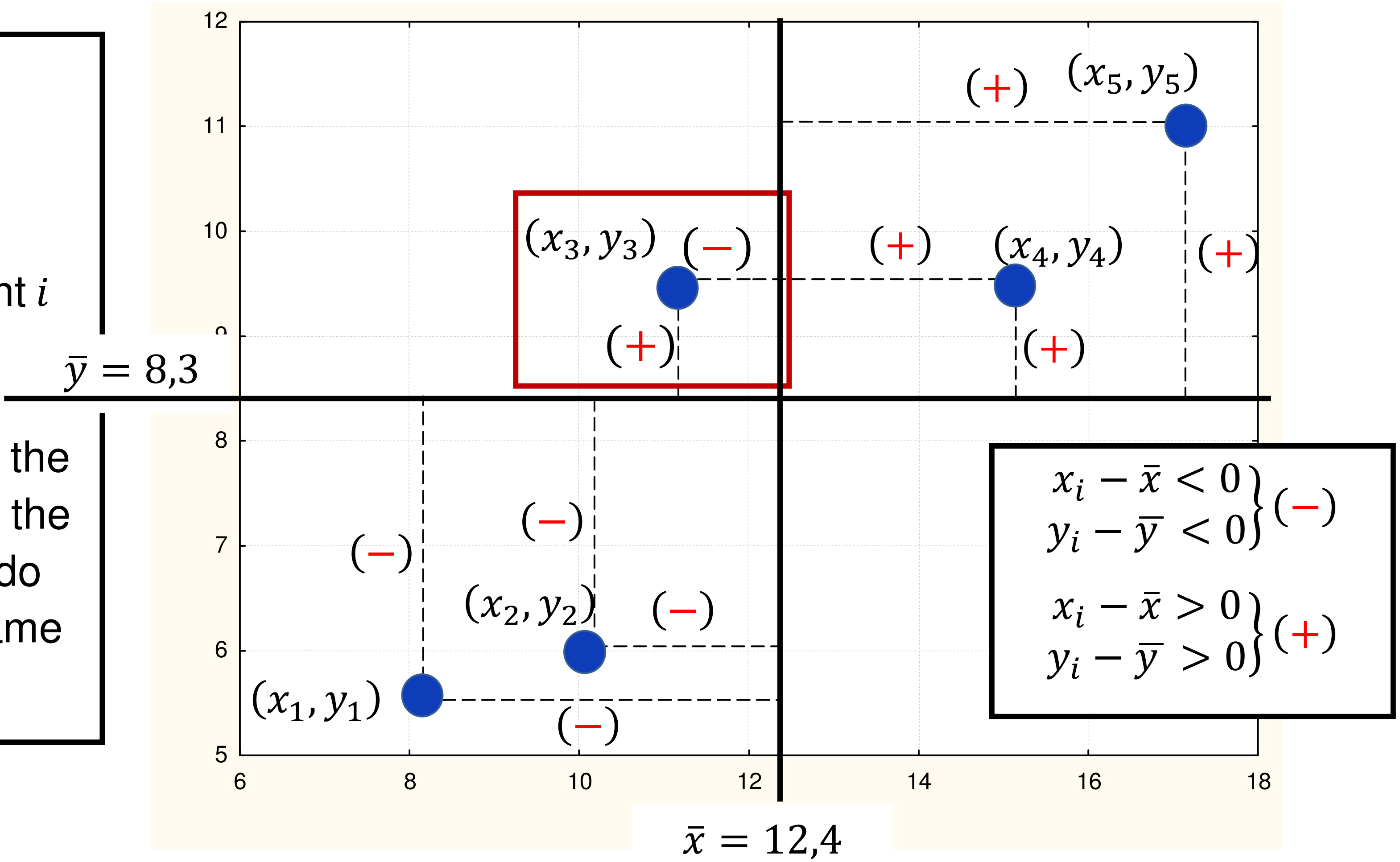
1.

$$\frac{(y_i - \bar{y})}{(x_i - \bar{x})}$$

is different for different i

2.

Deviations of values of the variables X and Y from the correspondent means do not have always the same signs



$$\left.\begin{array}{l} x_i - \bar{x} < 0 \\ y_i - \bar{y} < 0 \end{array}\right\} (-)$$

$$\left.\begin{array}{l} x_i - \bar{x} > 0 \\ y_i - \bar{y} > 0 \end{array}\right\} (+)$$

Deviations of values of the variables X and Y from the correspondent means do not have always the same signs

Covariance: A non-standardized measure of the linear relationship between two variables

If the random variable $\boldsymbol{X}$ has the expectation $E(X)$

and the random variable $\boldsymbol{Y}$ has the expectation $E(Y)$

then

$$Cov[X,Y] = E[(X - E[(X)])(Y - E[(Y)])]$$

is a covariance of X and Y.

Note:

$$Cov[X,X] = E[(X - E[(X)])(X - E[(X)])] = Var[(X)]$$

$$|Cov[X,X]| \leq \sqrt{Var[(X)]} \cdot \sqrt{Var[(Y)]}$$

Estimate of the **covariance**

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

where SP_{xy} is the Sum of the cross-products.

- A covariance is positive, if the higher values of X are combined with higher values of Y und lower with lower.

- A covariance is negative, if the higher values of X are combined with lower values of Y.

- If a covariance equals 0, there is no linear relationship between variables X and Y (**nonlinear relationships are possible!**).

- A covariance is **maximal**, if y_i is proportional to x_i

$$y_i = a + b \cdot x_i$$

Covariance: not meaningful enough, because the absolute value of the covariance depends on the scaling of the variables.

Correlation: Standardized Covariance: A dimensionless measure between -1 (negative relationship) and 1 (positive relationship)

Correlation coefficient ρ

(more precisely: the Pearson's Product moment correlation coefficient)

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}$$

where

σ_{XY} = covariance of X and Y

σ_X^2= variance of X

σ_Y^2= variance of Y

The Correlation encompasses only linear relationships!

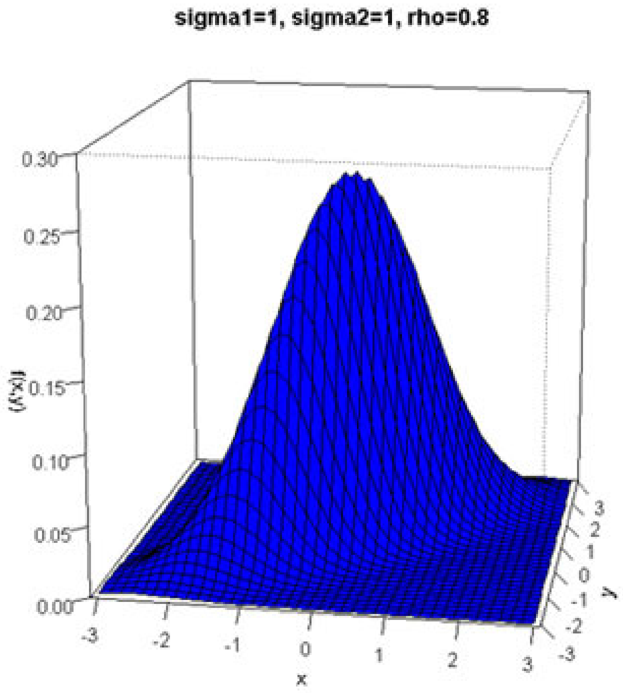
With help of a correlation we can conclude:

- Whether a relationship between two metric variables exists.
- How strong is this relationship.
- What direction has this relationship.

Assumptions of correlation analysis:

The data are **bivariate normally distributed.**

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot exp\left(-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right\}\right)$$

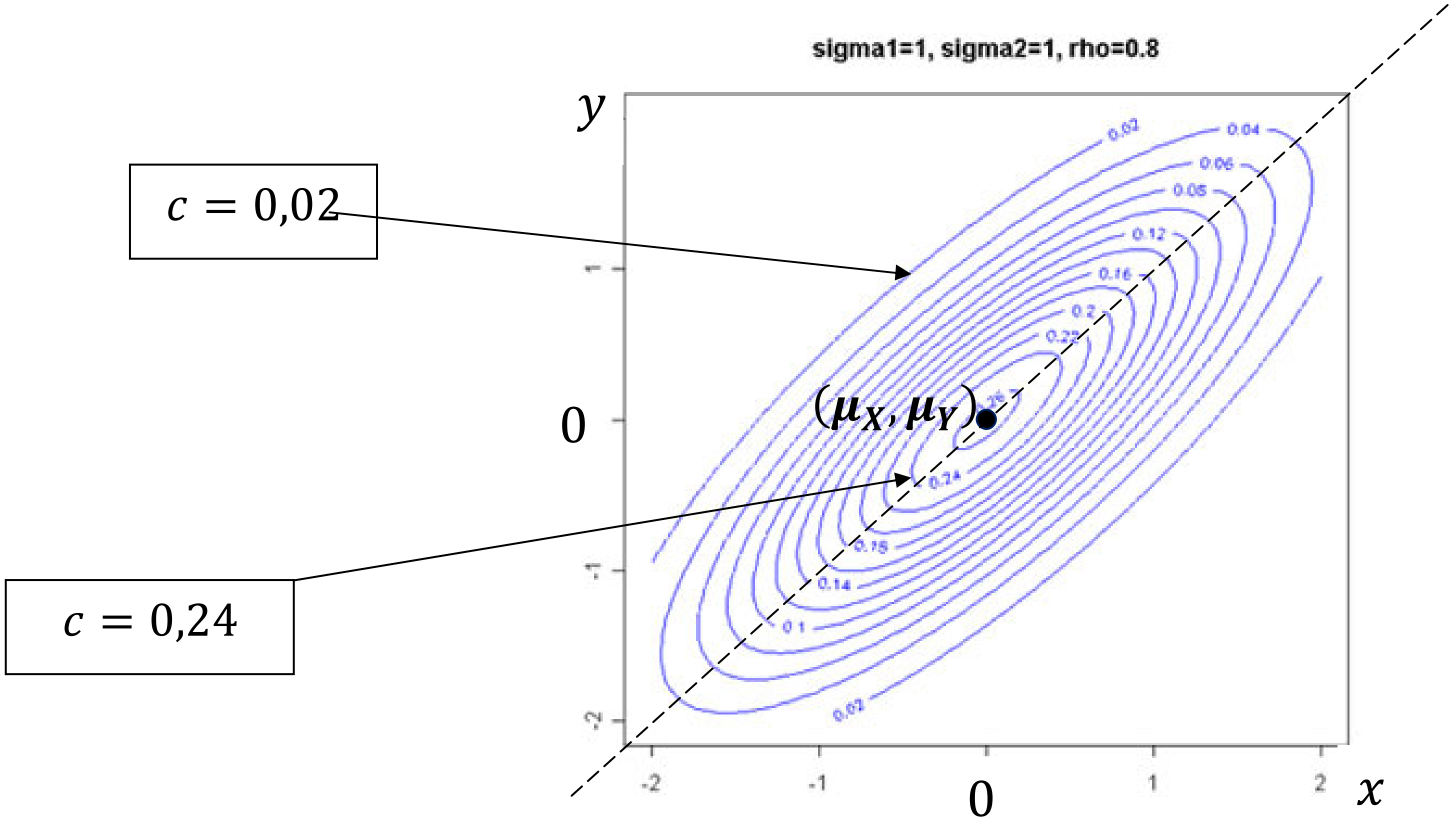


http://web.neuestatistik.de/demo/Demo_DE/MOD_100238/html/comp_100637.ht

The greater the correlation coefficient ρ, the narrower the figure

The graph shows the bivariate density function with the centre at (0,0).

So $\mu_X = 0; \mu_Y = 0$



$c = 0{,}02$

$c = 0{,}24$

If the correlation between X and Y is positive,
then the slope of the main axis of ellipse is positive

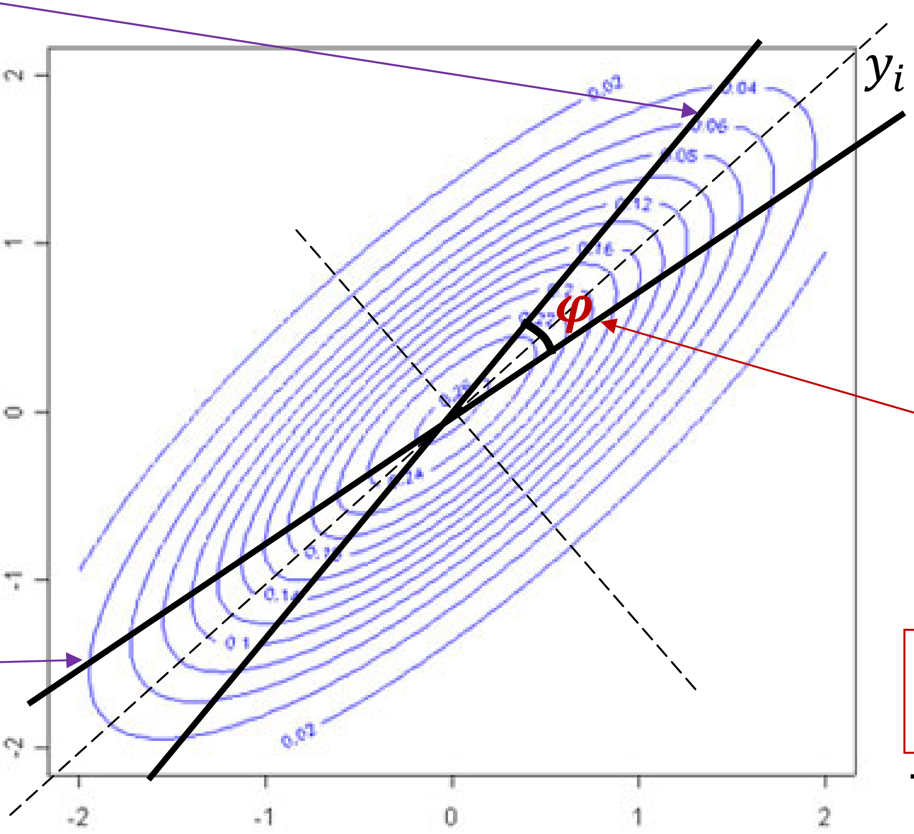
http://web.neuestatistik.de/demo/Demo_DE/MOD_100238/html/comp_100637.ht

Each single variable has a univariate normal distribution.
The contour plot of the **joint probability** density gives curves with constant density $f(x, y) = c$ for different c values.

The **conditional expectation** of X given Y

$$E[X|Y] = E[X] + \rho\frac{\sigma_X}{\sigma_Y}(Y - E[Y])$$

is a linear function of Y.



$y_i = a + bx_i$

$$tg\varphi = \frac{1-\rho^2}{\rho}\cdot\frac{\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}$$

$$\boxed{\rho = \pm 1 \;\Rightarrow\; \varphi = 0!}$$

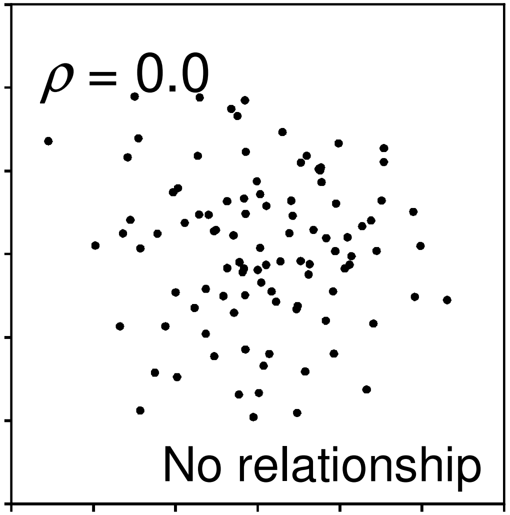
The ellipse becomes more and more elongated as the correlation approaches one.

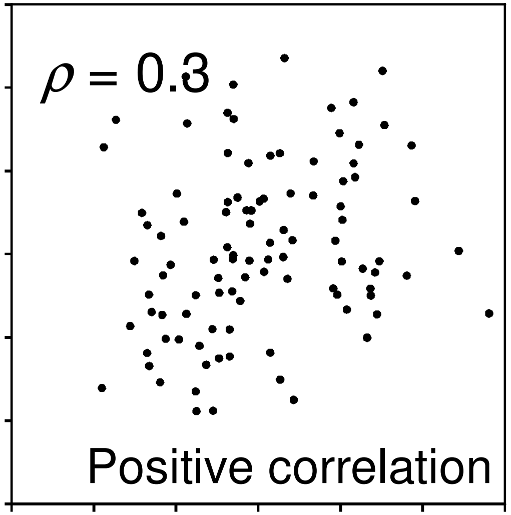
The **conditional expectation** of Y given X

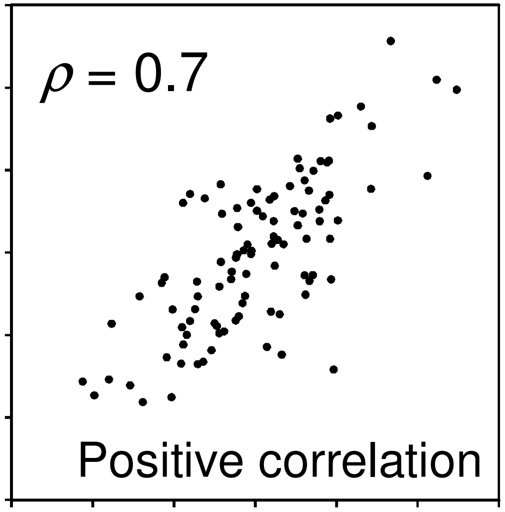
$$E[Y|X] = E[Y] + \rho\frac{\sigma_Y}{\sigma_X}(X - E[X])$$

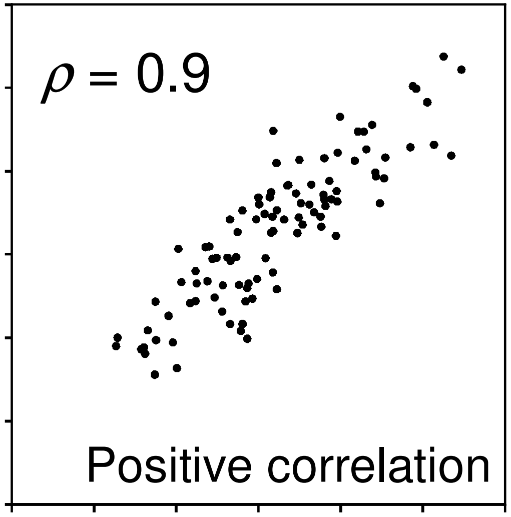
is a linear function of X

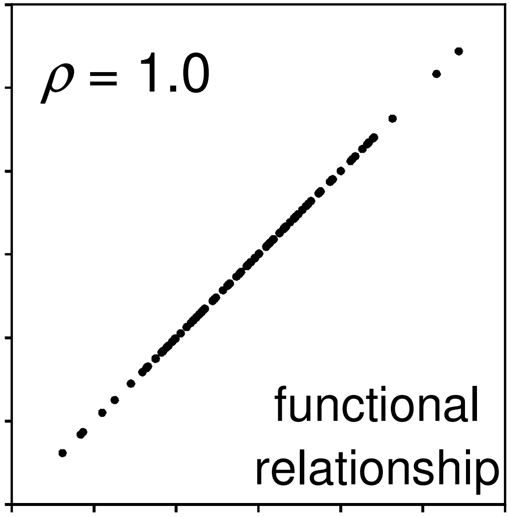
Positive Relationships



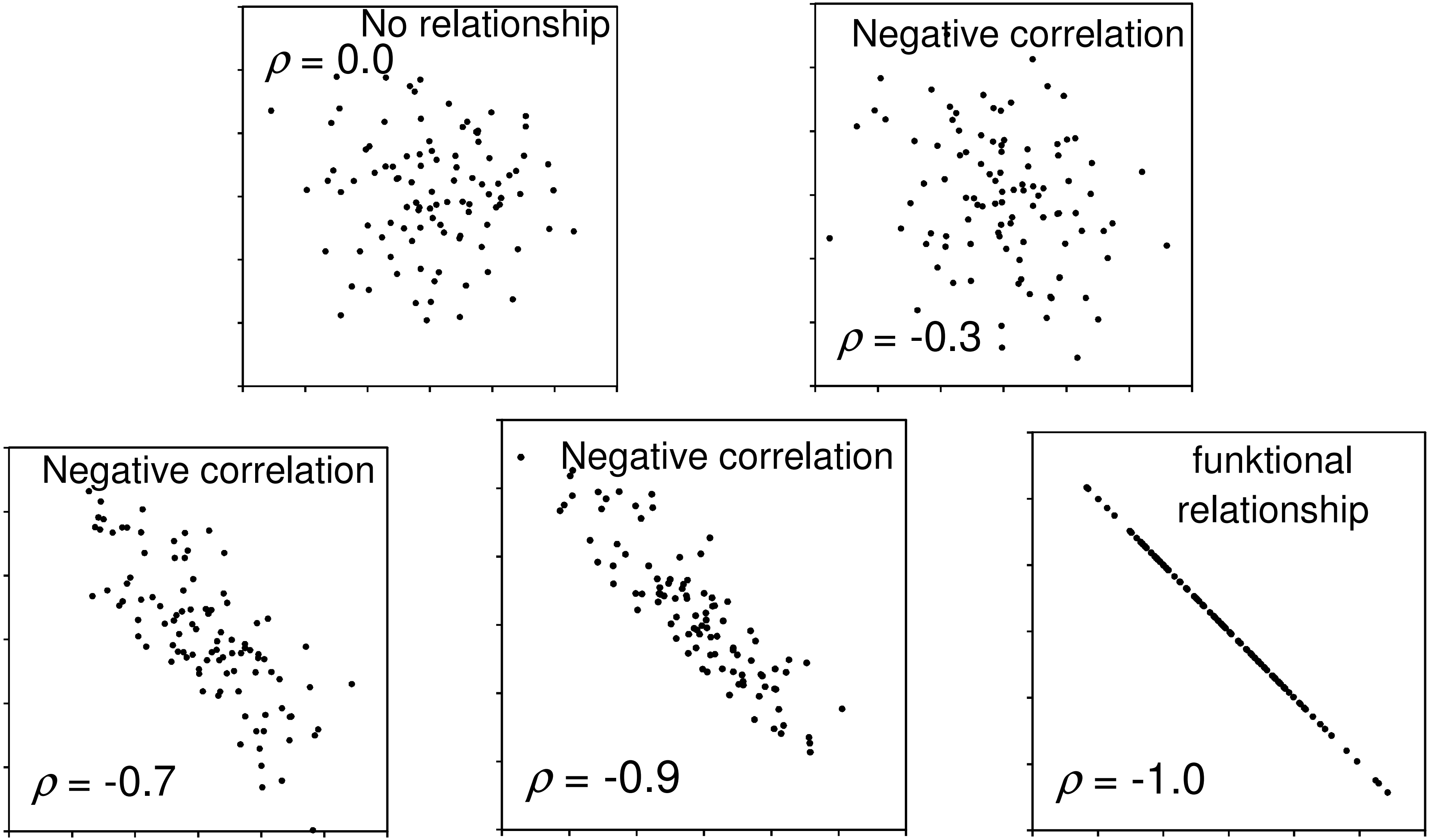








Negative Relationships



Estimating the correlation coefficient

(x_i, y_i) is the i-th sampling pair

$$r = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$

s_{xy}: Estimate of the **covariance**

$s_{xy} = \dfrac{SP_{xy}}{n-1}$ where $SP_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})$ (Sum of the cross-products)

s_x^2 and s_y^2: Estimates of the **variances**:

$s_x^2 = \dfrac{SS_x}{n-1}$ where $SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$ (Sum of squared deviations of X)

$s_y^2 = \dfrac{SS_y}{n-1}$ and $SS_y = \sum_{i=1}^{n}(y_i - \bar{y})^2$(Sum of squared deviations of Y)

$$\boldsymbol{r} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{SP_{xy}}{(n-1)\sqrt{\frac{SS_x}{n-1}\cdot\frac{SS_y}{n-1}}} = \boxed{\frac{\boldsymbol{SP_{xy}}}{\sqrt{\boldsymbol{SS_x \cdot SS_y}}}}$$

Example.

X - Variable: DBH $\bar{x} = 11.434$

Y- Variable: Height $\bar{y} = 7.884$

$\boldsymbol{i}$	$\boldsymbol{x_i}$	$\boldsymbol{y_i}$	$\boldsymbol{x_i - \overline{x}}$	$\boldsymbol{(x_i - \overline{x})^2}$	$\boldsymbol{y_i - \overline{y}}$	$\boldsymbol{(y_i - y)^2}$	$\boldsymbol{(x_i - \overline{x}) \cdot (y_i - y)}$
1	7.34	5.850	-4.094	16.759	-2.034	4.137	8.327
2	10.08	7.470	-1.354	1.833	-0.415	0.172	0.561
3	9.79	6.568	-1.644	2.702	-1.317	1.734	2.164
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
48	6.93	5.222	-4.504	20.284	-2.663	7.090	11.992
49	10.91	7.854	-0.524	0.274	-0.031	0.001	0.016
50	11.18	7.336	-0.254	0.064	-0.548	0.301	0.139
				$\sum = 425.305$		$\sum = 147.379$	$\sum = 241.217$

$$r = \frac{SP_{xy}}{\sqrt{SS_x \cdot SS_y}} = \frac{241.217}{\sqrt{425.305 \cdot 147.379}} = \boxed{0.96347}$$

Important: Correlation does not imply causation!

The symbolic diagram of the relationship of two variables X and Y

$X \longrightarrow Y$ (X causes Y)

Example: Fertilization causes yield

$X \longleftrightarrow Y$ (X causes Y and Y causes X: bidirectional causation)

Example: Increased pressure is associated with increased temperature.

Therefore pressure causes temperature; temperature causes temperature.

$$\begin{tikzcd}
 & X \\
Z \arrow[ur] \arrow[dr] & \\
 & Y
\end{tikzcd}$$

(Third factor Z (the common-causal variable) causes both X and Y)

There is some lurking variable, which is simply a hidden third variable that affects both causes of the correlation

Example: A city's ice cream sales. These sales are highest when the rate of drownings in city swimming pools is highest. To allege that ice cream sales cause drowning, or vice-versa, would be to imply a spurious relationship between the two.

In reality, a heat wave may have caused both. The heat wave is an example of a hidden or unseen variable, also known as a confounding variable.

Summary: A correlation can be taken as evidence for a **possible** causal relationship, but cannot indicate **what** the causal relationship, if any, might be.

Correlations don‘t indicate, which variable causes the other – both variables are equal (symmetry of variables).

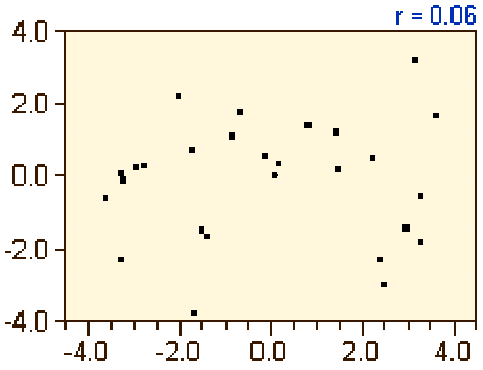
Correlation can only detect the strength (correlation coefficient value) and the direction (sign of the correlation coefficient) of the linear relationship between two variables.

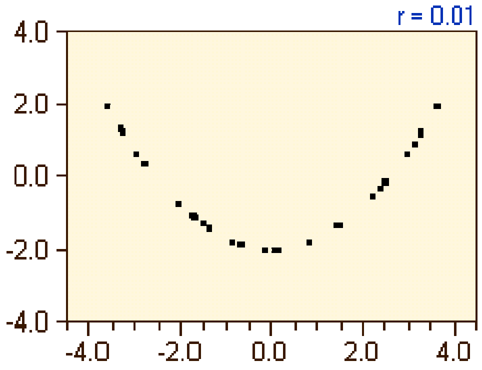
In General:

Sampling is not appropriate for detecting the causation of a relationship. It is only an indicator of their existence and can be used to build hypotheses about causal relationships.

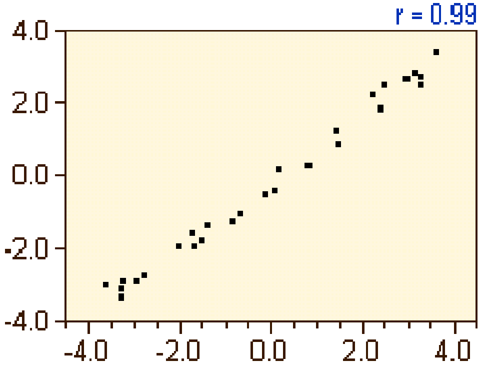
These hypotheses must be tested in experiments, where the causal factor is systematically varied and all other factors are held constant.

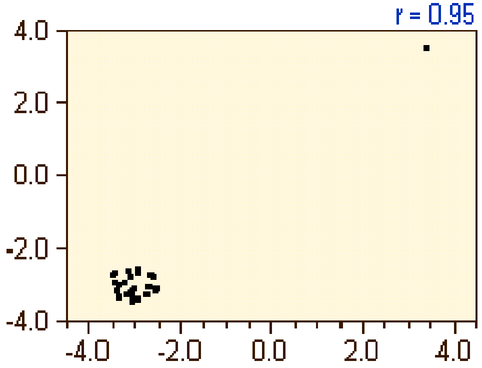
Important: The correlation indicates the strength of a linear relationship between two variables, but its value generally does not completely characterize their relationship.





A high correlation coefficient can be caused by outliers





http://www.statistics4u.com/fundstat_germ/cc_correlation.html

Correlation: Characterizes the **strength** of a stochastic relationship.

Regression: Characterizes the **kind** of a stochastic relationship.

Regression analysis is also used to understand which amount of the dispersion of the independent variables is related to the dependent variable, and to explore the forms of these relationships.

Tasks of regression analysis:

- Interpretation of Scatterplot
- Prediction and forecasting

The „Regression“ Notation – Regression toward the mean

(see Annex, P. 52)

Francis Galton: “Regression towards mediocrity in hereditary stature”

- Height of parents and children (as adults)

- Linear Relationship

- The offspring of parents who lie at the tails of the distribution will tend to lie closer to the centre, the mean, of the distribution.

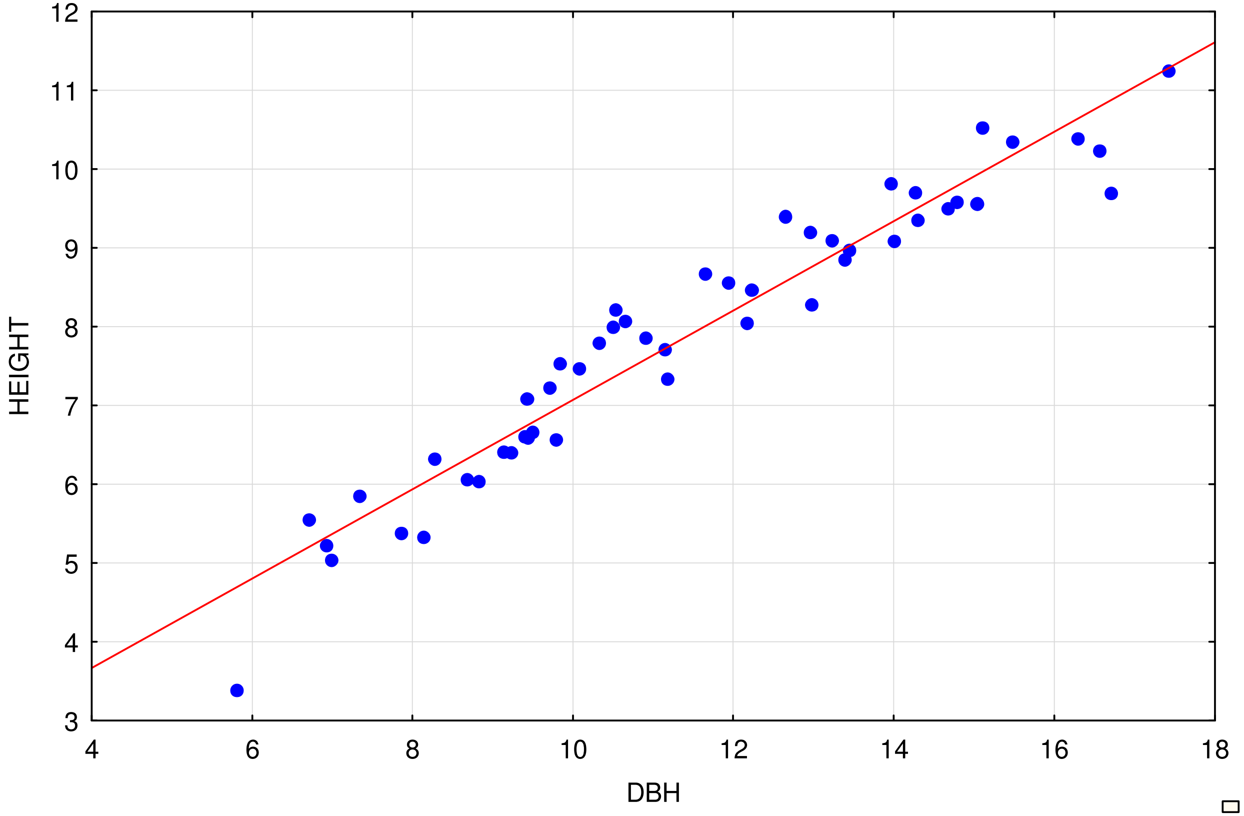
- "Regression" (regress) toward the mean

Linear Regression

Scatterplot: DBH vs. HEIGHT (Casewise MD deletion)

HEIGHT = 1,3996 + ,56716 * DBH

Correlation: r = ,96347



Regression Line Equation

Equation of a line

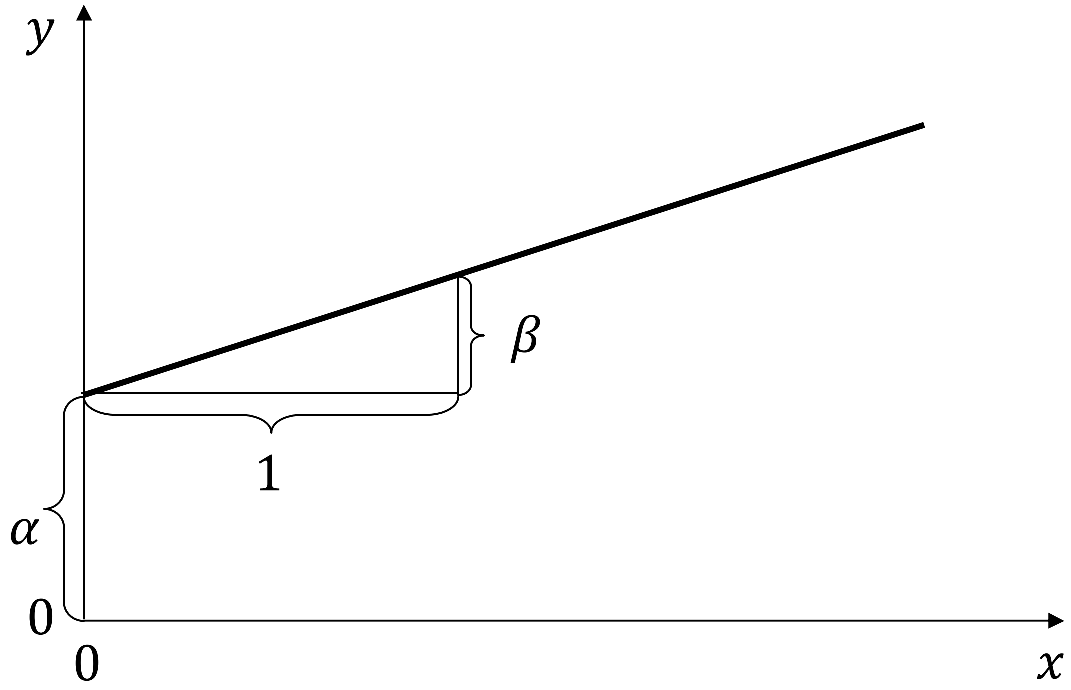
$$E[Y] = \alpha + \beta x$$

where

$E[Y]$ = Expectation of Y

α = Intercept

β = Slope

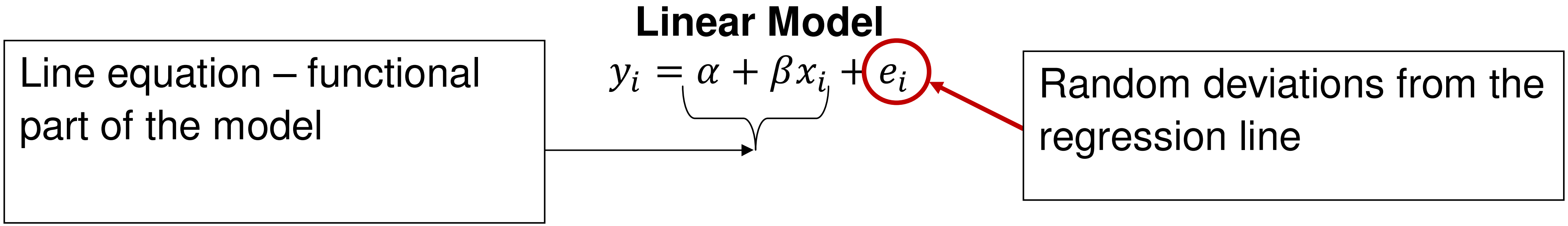


In contrast to the correlation the variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are not equal.

There is a clear difference between the dependent and independent variable (asymmetry).

X = **independent variable** (endogenous variable, explanatory variable, Regressor, Predictor):

Y = **dependent variable** (exogenous variable, response variable, Regressand)



Linear Model

$$y_i = \underbrace{\alpha + \beta x_i}_{} + e_i$$

Line equation – functional part of the model

Random deviations from the regression line

where

y_i = the i-th value of the dependent variable

x_i = the i-th value of the independent variable

e_i = Residual: Deviation of the i-th value of the dependent variable from the regression line

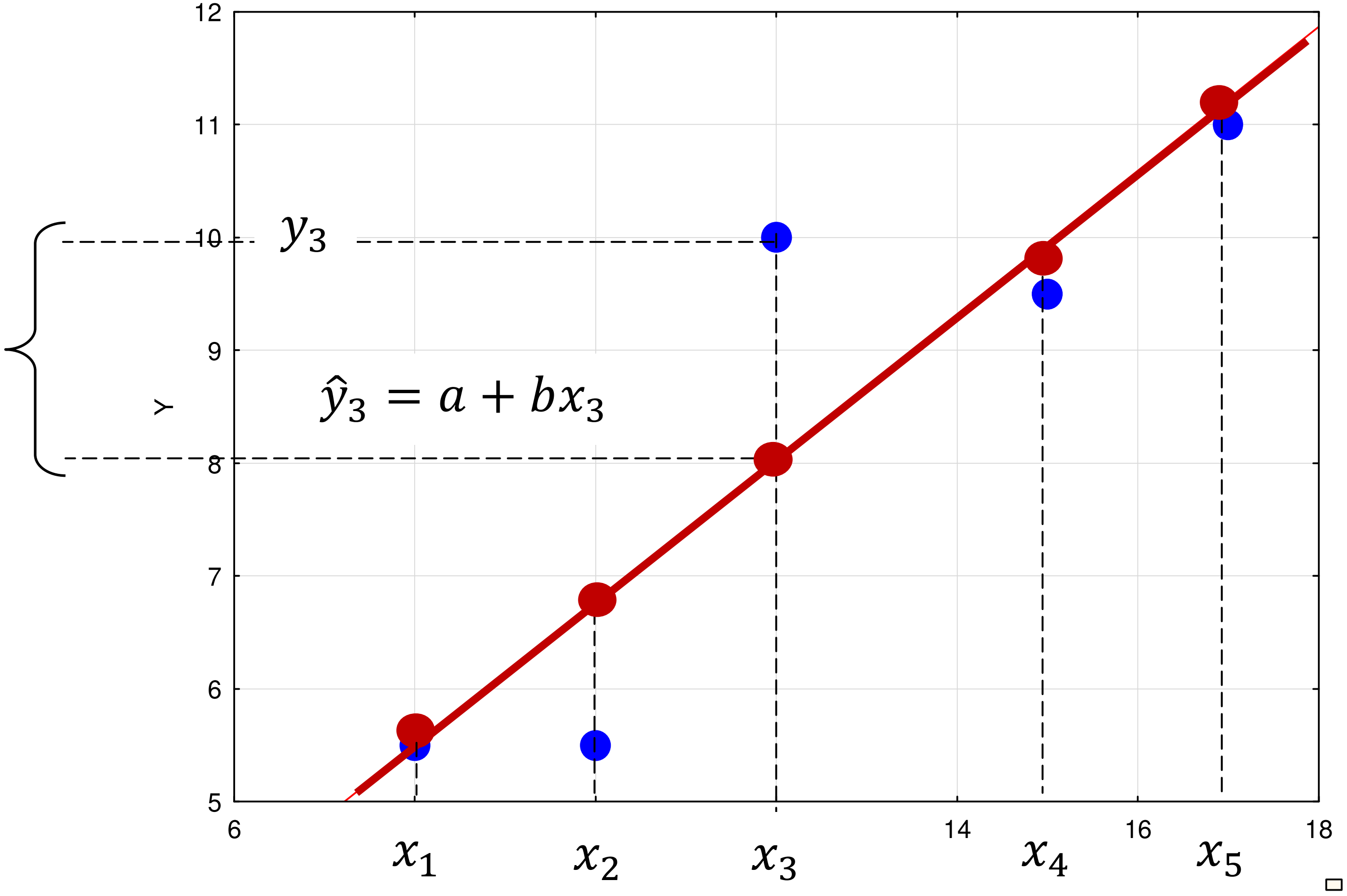
Residuals are considered as realizations of the stochastically independent random variables $E_i \sim N(0, \sigma)$

! In regression model the independent variables have no error.

Illustration of estimated deviations $\hat{e}_i$ of an individual value y_i from the estimated value $\hat{y}_i$ of the regression line

a – estimate of the parameter α ; b – estimate of the parameter β

$$\hat{e}_3 = y_3 - (a + bx_3)$$



● - measured value

● - estimated value

Least squares method

$y_i - (a + bx_i)$ – the deviation of an individual value from the regression line

a – estimated value of α

b – estimated value of β

Sum of squares

$$SS_{Error} = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

⇒ Criterion ⇒ minimize!

We search some line for the scatterplot, which runs possibly near to the measured values.

$$SS_{Error} = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 = min$$

Estimate of a and b to minimize the SS_{Error}

$$\frac{\partial SS_{Error}}{\partial b} = 2\sum_{i=1}^{n} [y_i - (a + bx_i)](-1)\, x_i = 0$$

$$\frac{\partial SS_{Error}}{\partial a} = 2\sum_{i=1}^{n} [y_i - (a + bx_i)](-1) = 0$$

⇒ normal equations

$$\frac{\partial SS_{Error}}{\partial a} = 2\sum_{i=1}^{n}[y_i - (a + bx_i)](-1) = 0$$

$$\sum_{i=1}^{n}[y_i - (a + bx_i)] =$$

$$= \sum_{i=1}^{n} y_i - \left(\sum_{i=1}^{n} a + b\sum_{i=1}^{n} x_i\right) =$$

$$n\bar{y} - n(a + b\bar{x}) = 0$$

$$\boxed{a = \bar{y} - b\bar{x}} \Rightarrow$$

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

$$\frac{\partial SS_{Error}}{\partial b} = 2\sum_{i=1}^{n}[y_i - (a + bx_i)](-1)\, x_i = 0$$

$$\sum_{i=1}^{n}[y_i - (a + bx_i)]\, x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - b\left[\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i\right] =$$

$$= \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} - b\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i}{n}\right.$$

$$= SP_{xy} - b \cdot SS_x = 0$$

$$\boxed{b = \frac{SP_{xy}}{SS_x}}$$

Summary: Estimate of the regression line

(method of least squares):

$$b = \frac{SP_{xy}}{SS_x}$$

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

where

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

and

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

Estimated regression line:

$$\hat{y} = a + bx$$

Exercise:

X - variable: DBH $\bar{x} = 11.434$

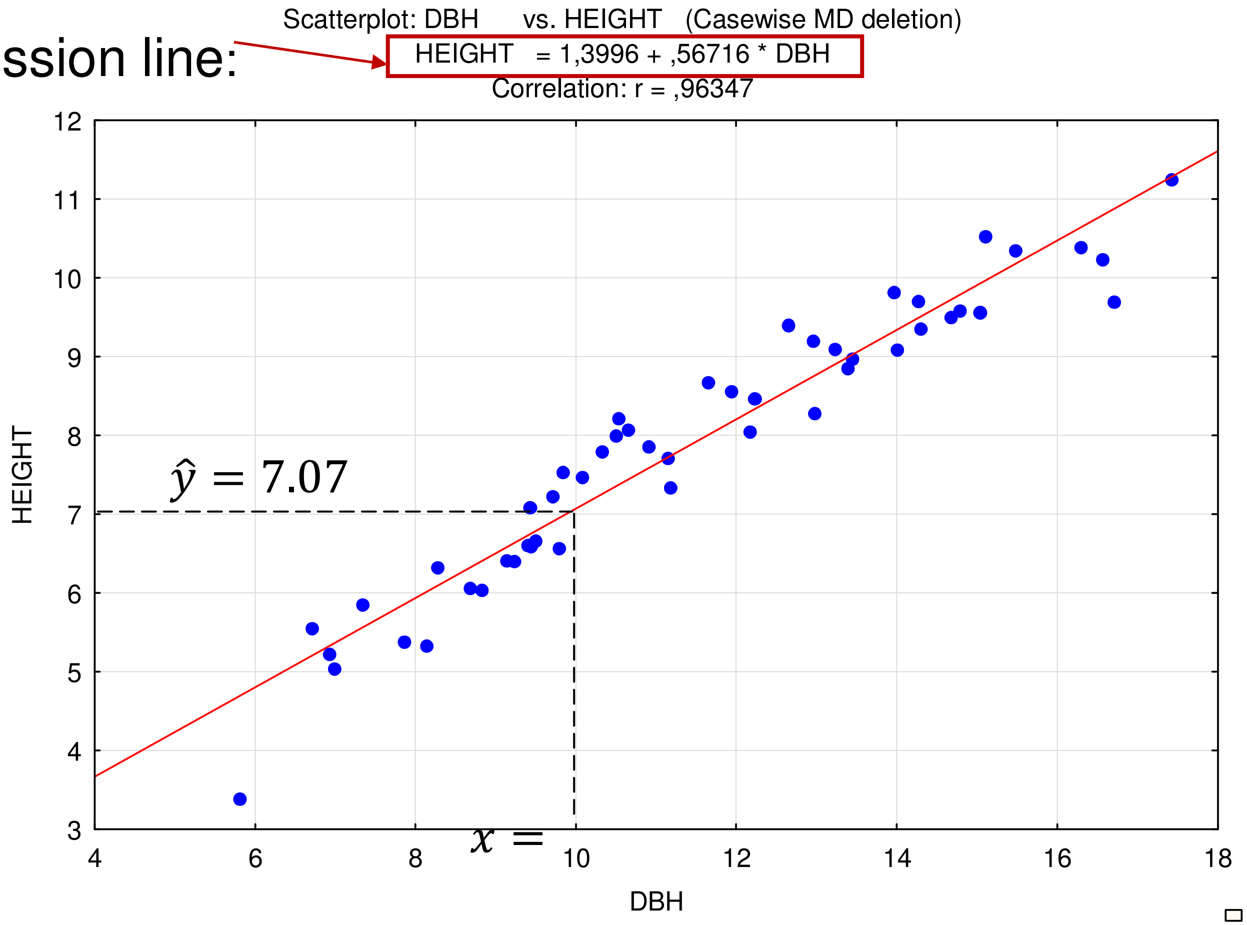
Y- variable: Height $\bar{y} = 7.884$

$\boldsymbol{i}$	$\boldsymbol{x_i}$	$\boldsymbol{y_i}$	$\boldsymbol{x_i - \bar{x}}$	$(\boldsymbol{x_i - \bar{x}})^2$	$\boldsymbol{y_i - \bar{y}}$	$(\boldsymbol{y_i - y})^2$	$(\boldsymbol{x_i - \bar{x}}) \cdot (\boldsymbol{y_i - \bar{y}})$
1	7.34	5.850	-4.094	16.759	-2.034	4.137	8.327
2	10.08	7.470	-1.354	1.833	-0.415	0.172	0.561
3	9.79	6.568	-1.644	2.702	-1.317	1.734	2.164
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
48	6.93	5.222	-4.504	20.284	-2.663	7.090	11.992
49	10.91	7.854	-0.524	0.274	-0.031	0.001	0.016
50	11.18	7.336	-0.254	0.064	-0.548	0.301	0.139
				$\sum = 425.305$		$\sum = 147.379$	$\sum = 241.217$

$$b = \frac{241.217}{425.305} = 0.567; \quad a = 7.884 - 0.567 \cdot 11.434 = 1.40$$

Estimated regression line: $\hat{y} = 1.40 + 0.567 \cdot x$

Estimated regression line:



Interpretation: If DBH increases by one cm then we expect an increase of HEIGHT by about 0,567 m

Prediction: If DBH=10 cm we expect the HEIGHT$= 1.40 + 0.567 \cdot 10 = 7.07\ m$

Relationship between the correlation coefficient r and the regression coefficient b:

$$r = b\sqrt{\frac{SS_x}{SS_y}}$$

where

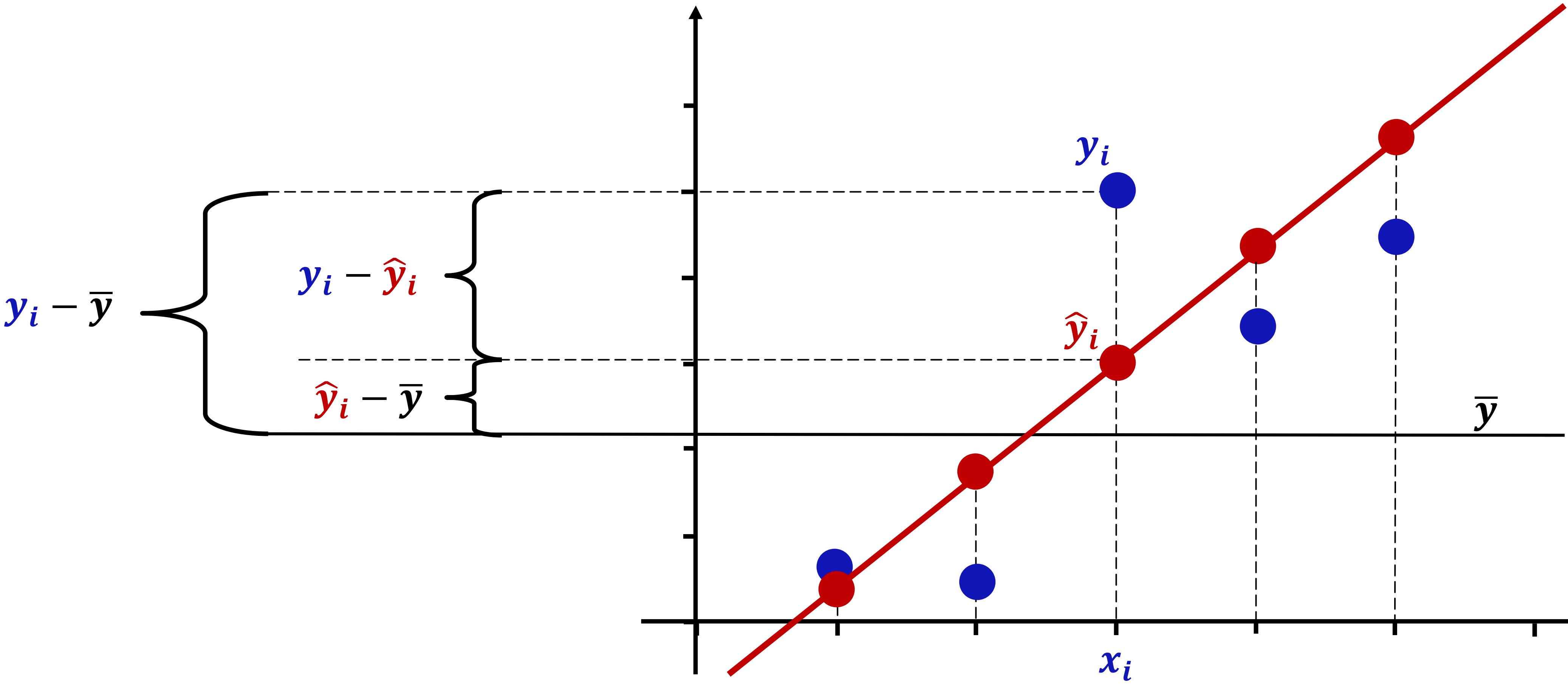
$$SS_x = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

and

$$SS_y = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

$$r = 0.567\sqrt{\frac{425.305}{147.379}} = 0.963$$

Decomposition of the variance of Y



Decomposition of the total deviation from the mean $(y_i - \bar{y})$

into the part, that can be explained by regression $(\hat{y}_i - \bar{y})$

and the deviations from the regression line (Error) $(y_i - \hat{y}_i)$

$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

where

$\hat{y}_i = a + bx_i$ – expected value at x_i

Deviations:

$y_i - \bar{y}$ = Total deviation from the mean

$y_i - \hat{y}_i$ = Deviation of observed value from the regression line (**Error**)

$\hat{y}_i - \bar{y}$ = Deviation of expected value from the mean (The part of deviation of the dependent variable y, that can be explained by the effect of the variable x).

Decomposition of the sum of squares

$$SS_y = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\Bigg(\underbrace{y_i - \hat{y}_i}_{Error} + \underbrace{\hat{y}_i - \bar{y}}_{Regression}\Bigg)^2 =$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

where

$$\hat{y} = a + bx$$

Sum of cross-products equals 0!

$$\hat{y}_i = \boxed{a} + bx_i = \underbrace{\bar{y} - b\bar{x}}_{a} + bx_i = \bar{y} + \boxed{b(x_i - \bar{x})}$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}\Big(y_i - \overbrace{\bar{y} - b(x_i - \bar{x})}\Big)\big(\overbrace{\cancel{\bar{y}} + b(x_i - \bar{x})} - \cancel{\bar{y}}\big) =$$

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

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

$$= bSP_{xy} - b^2SS_x = \frac{SP_{xy}}{SS_x}SP_{xy} - \frac{SP^2{}_{xy}}{SS^2{}_x}SS_x = 0$$

Decomposition of the sum of squares

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SS_y \quad = SS_{Regression} + \quad SS_{Error}$$

From $\hat{y}_i = \bar{y} + b(x_i - \bar{x})$

follows $\hat{y}_i - \bar{y} = b(x_i - \bar{x})$

So

$$\boxed{SS_{Regression}} = \sum^{n}(\hat{y}_i - \bar{y})^2 = b^2 \sum^{n}(x_i - \bar{x})^2 = \boxed{b^2 SS_x}$$

Goodness of fit of the least squares method:

Multiple correlation coefficient R

Multiple correlation coefficient is a measure fort h correlationbetween observed values y_i and theyr’s estimates $\hat{y}_i$ **:** $R = r_{y\hat{y}}$.

For a sample regression $r_{y\hat{y}} = r_{xy}$

$$\boxed{r_{y\hat{y}}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} =$$

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})(\cancel{a} + bx_i - \cancel{a} - b\bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\cancel{a} + bx_i - \cancel{a} - b\bar{x})^2}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})b(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 b \sum_{i=1}^{n}(x_i - \bar{x})^2}} =$$

$$\frac{\cancel{b}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\cancel{b}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{SP_{xy}}{SQ_x SQ_y} = \boxed{r_{xy}}$$

Coefficient of determination

The most important measure of the strength of the regression relation is the squared correlation coefficient or coefficient of determination R^2.

The coefficient of determination gives the amount of the total variability of the dependent variables (Y) that can be explained by the effect of the independent variable (X).

$$\boxed{R^2} = r_{y\hat{y}}^2 = r_{xy}^2 = \frac{SP_{xy}^2}{SS_x \cdot SS_y} = \frac{SP_{xy}^2}{SS_x^2} \cdot \frac{SS_x}{SS_y} = \frac{b^2 SS_x}{SS_y} =$$

$$\frac{SS_{Regression}}{SS_y} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \boxed{\frac{explaned\ variability}{total\ variability\ of\ Y}}$$

The coefficient of determination R^2 is a measure of how well the least squares equation

$$\hat{y} = a + bx$$

performs as a predictor of y.

- The higher the R^2, the more useful the model
- R^2 takes on values between 0 and 1.
- Essentially, R^2 tells us how much better we can do in predicting y by using the model and computing $\hat{y}$ than by just using the mean $\bar{y}$ as a predictor.
- The estimated value $\hat{y}$ depends on x because

  $$\hat{y} = a + bx$$

  Thus, we act as if x contains information about y.

- If we just use $\bar{y}$ to predict y, then we are saying that x does not contribute information about y and thus our predictions of y do not depend on x.

Example:

$$R^2 = \frac{SS_{Regression}}{SS_y} = \frac{136.81}{147.38} = 0.928 = 0.963^2$$

92.8% of the total variance of Height can be explained by the effect of the variable DBH.

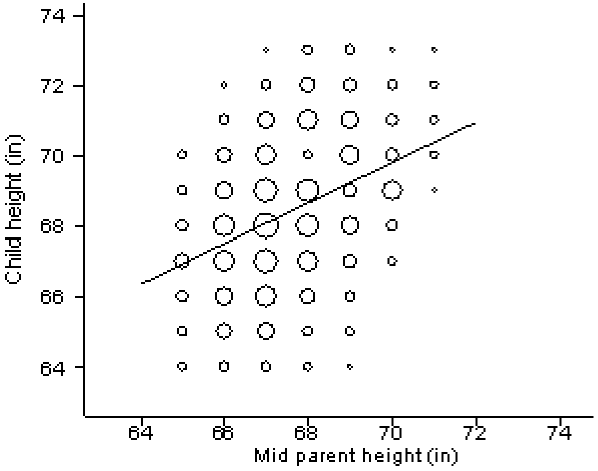
Annex

Regression towards the mean: Where does the name come from?

http://www-users.york.ac.uk/~mb55/talks/regmean.htm

The name “regression” comes from a paper by the Victorian geneticist and polymath Francis Galton (Galton 1886) entitled “Regression towards mediocrity in hereditary stature”. Galton set up a stand at the Great Exhibition, where he measured the heights of families attending.

He measured the heights of families attending. He adjusted the female heights by multiplying by 1.08. He then calculated the average height of the two parents, the “midheight”, and related it to the height of their adult children:



This plot is based on Galton’s original. The area of the circle represents the number of coincident points. The line is the regression of child height on midparent height. The means of both are the same, 68.2 inches.

Consider parents with midheight 70 inches. Their children had heights between 67 and 73 inches, and a mean height of 69.6 inches. The mean height of the subgroup of children was closer to the mean height of all children than the mean height of the subgroup of midparents was to the mean height of parents. Galton called this ‘regression towards mediocrity’.

The same thing happens if we start with the children. For example, for the children with height 70 inches, the mean height of their midparents is 67.9 inches. This is a statistical, not a genetic phenomenon.

Galton called this “regression towards mediocrity”. Because the word “mediocrity” has acquired adverse connotations since Galton’s time, we now call it “regression towards the mean”.