Simple linear regression

Statistics II (LIX002X05)

University of Groningen, Faculty of Arts, Information Science

Wilbert Heeringa

Variables

- Experiment Van Bezooijen & Heeringa (2006): measure intuitions of non-linguists about dialects in the Netherlands and Flanders.
- Task: rate the dialect distance compared to standard Dutch per province in a map: 0=no distance, 100=maximal distance.
- 140 Dutch subjects were involved in the experiment.

Average intuitive linguistic distances per province compared to standard Dutch.

Variables

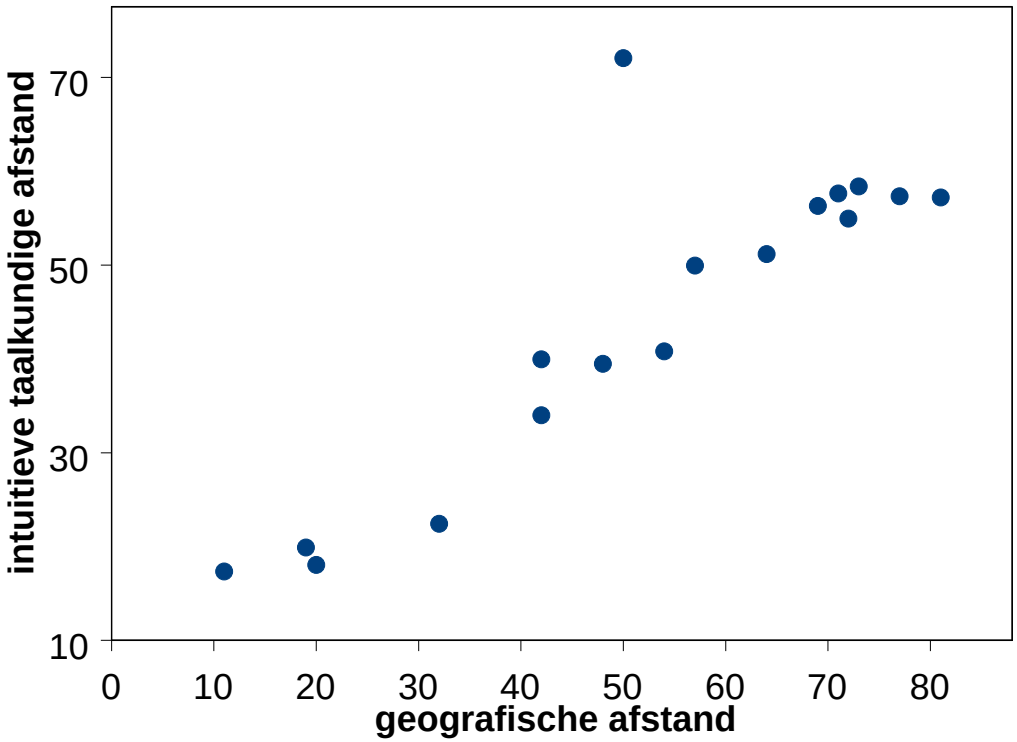
- Geographic distances between provinces and standard Dutch can also be measured.
- We locate standard Dutch at the position of Haarlem (situated west of Amsterdam).
- We measure as-the-crow-flies distances between the geographic centers of the provinces and Haarlem in millimeters scaled between 0 and 100.

Geographic distances between the geographic centers of the provinces and standard Dutch (located at the position of Haarlem).

Variables

- **Dependent variable** or **response variable**:
  measures the result of an investigation (intuitive linguistic distance).
- **Independent variable** or **explanatory variable** or **predictor variable**:
  tries to explain the observed results (geographic distance).
- The relation between the response variable and the explanatory variable are visualized in a scatter plot.

Scatter plot



Draw the explanatory variable on the x-axis, and the response variable on the y-axis. The points are the provinces.

Scatter plot

- Shape:
  linear relations, clusters.
- Direction:
  - Positive relation:
    high values of one variable tend to correspond with high values of another variable.
  - Negative relation:
    high values of one variable tend to correspond with low values
- Strength:
  Relationship is strong when points are found close to the straight line.

Correlation

- The correlation measures the direction and strength of a linear relationship between two quantitative variables.
- Correlation is rendered by r.
- Assume we have data about variables x en y for n individuals. Variable x has mean $\overline{x}$ and standard deviation s_x. Variable y has mean $\overline{y}$ and standard deviation s_y. The correlation r between x and y is:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

or:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i}$$

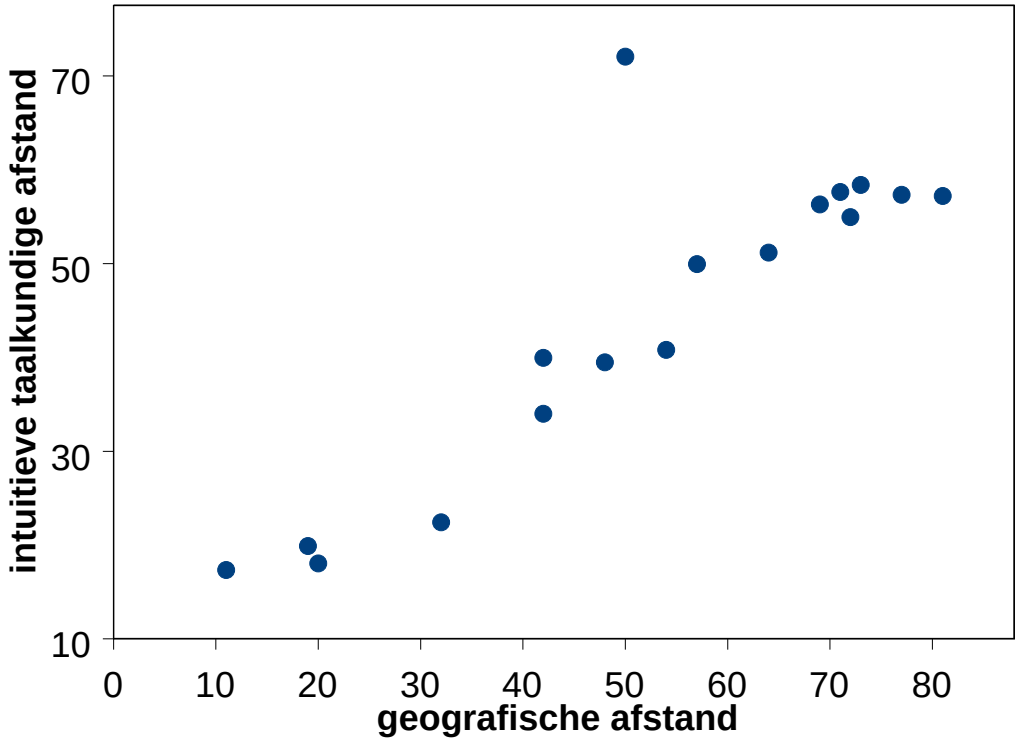
Correlation

- Both variables should be quantitative. There is no distinction between the explanatory variable and the response variable.
- The correlation r is a number between -1 and 1. The closer r is to -1 or 1, the stronger the linear relationship:
  - $r = 0$:
    no correlation;
  - $r = +1$:
    perfect positive correlation, all data points are on a straight line with a positive slope;
  - $r = -1$:
    perfect negative correlation, all data points are on a straight line with a negative slope.

Correlation

- The correlation measures the strength of a **linear** relationship between two variables only. Therefore, draw a scatter plot in advance.
- The value r is not influenced by changes in the unit of the two variables.
- Correlation may be misleading in case of outliers, influential observations or a non-linear relationship.
- Therefore, create always a scatter plot.

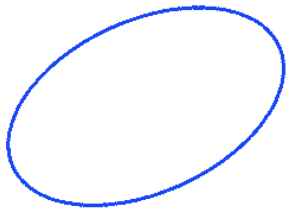
Correlation

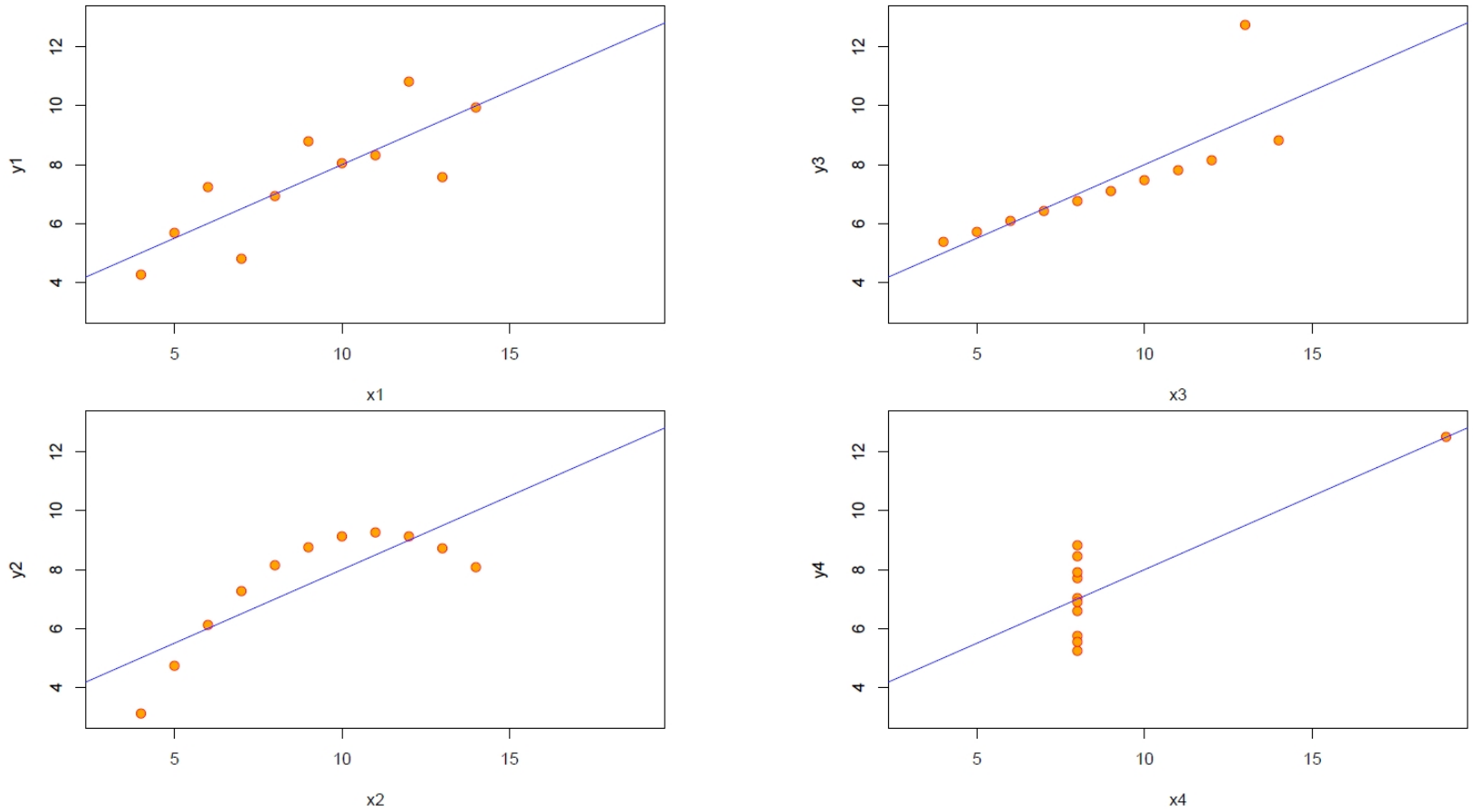


Correlation is not resistent, outliers may strongly change r: $r = 0.87$ when Friesland is included, and $r = 0.98$ when Friesland is excluded.

Inference for correlation

- Correlation coefficient:
  measure of strength and direction of the linear relationship between two variables.
- Assumption:
  the significance test assumes that the data has a bivariate (i.e. two-dimensional) normal distribution.
- y is normally distributed for each value of x, and x is normally distributed for each value of y.
- When the points in a scatter plot show an ellips, we found a bivariate normal distribution:





In each of the graphs the variable y has the same mean, standard deviation, correlation and regression line. (graphs Hartmut Fitz)

Inference for correlation

- Hypotheses of a significance test for ρ:

  $H_0 : \rho = 0$ (no linear relationship)

  $H_a : \rho > 0$; the p-value is $P(T \geq t)$
  $H_a : \rho < 0$; the p-value is $P(T \leq t)$
  $H_a : \rho \neq 0$; the p-value is $2P(T \geq |t|)$

  where T is a stochastic variable with the $t(n-2)$ distribution.

- The test statistic t is:

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

  with sample size n, sample correlation r and $n-2$ degrees of freedom.

Inference for correlation

- Hypotheses of significance test for ρ in our example:

  $H_0 : \rho = 0$
  (no linear relationship between geographic distance and intuitive linguistic distance)
  $H_a : \rho > 0$
  (positive linear relationship between geographic distance and intuitive linguistic distance)

- The test statistic t is:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.873\sqrt{17-2}}{\sqrt{1-(0.873^2)}} = 6.932$$

Inference for correlation

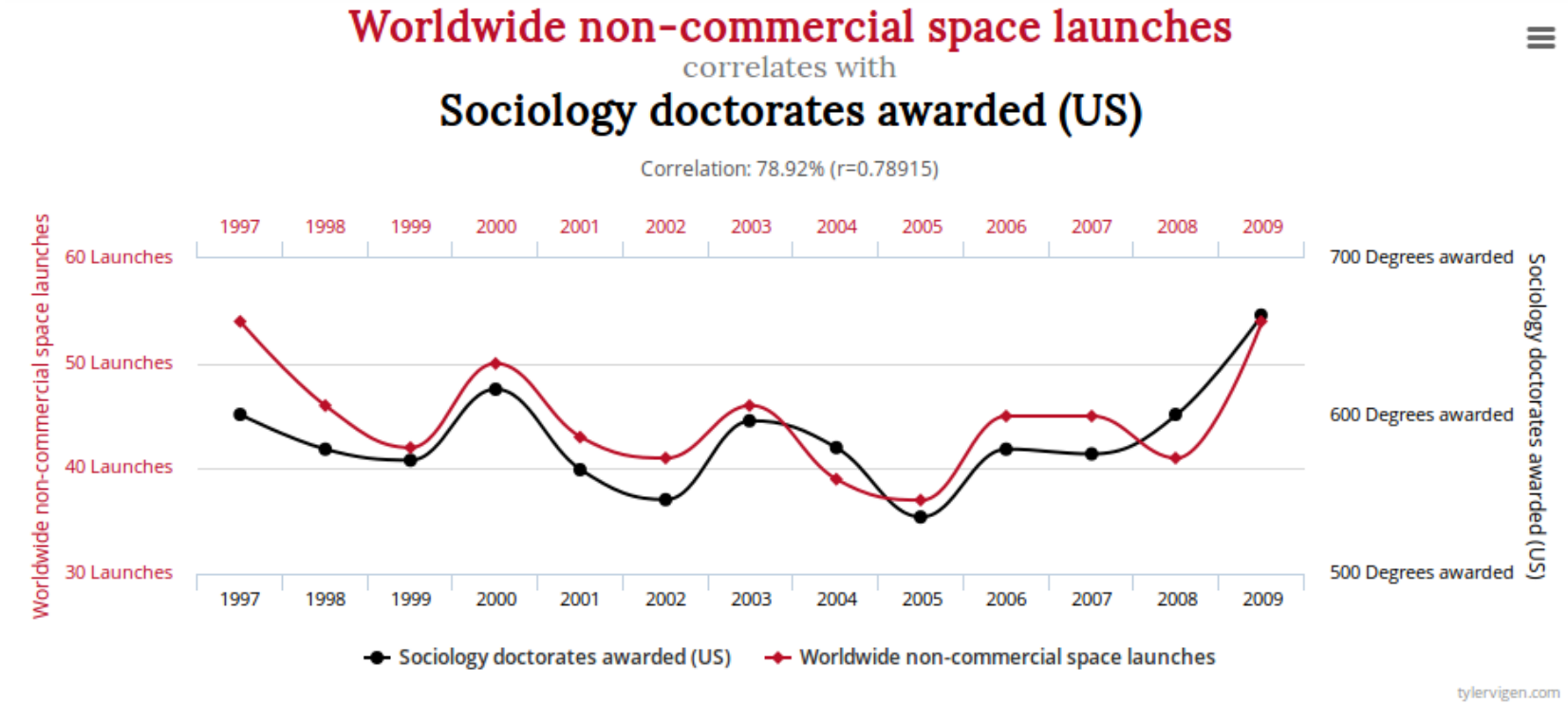
- The number of the degrees of freedom is $n - 2 = 17 - 2 = 15$.
- Go to to Vassarstats website at `http://www.vassarstats.net/tabs.html` and choose t *to* P.
- Enter t=6.932 and df=15, and choose *right tail*. We find $p < 0.0001$.
- We reject H_0 and accept H_a.

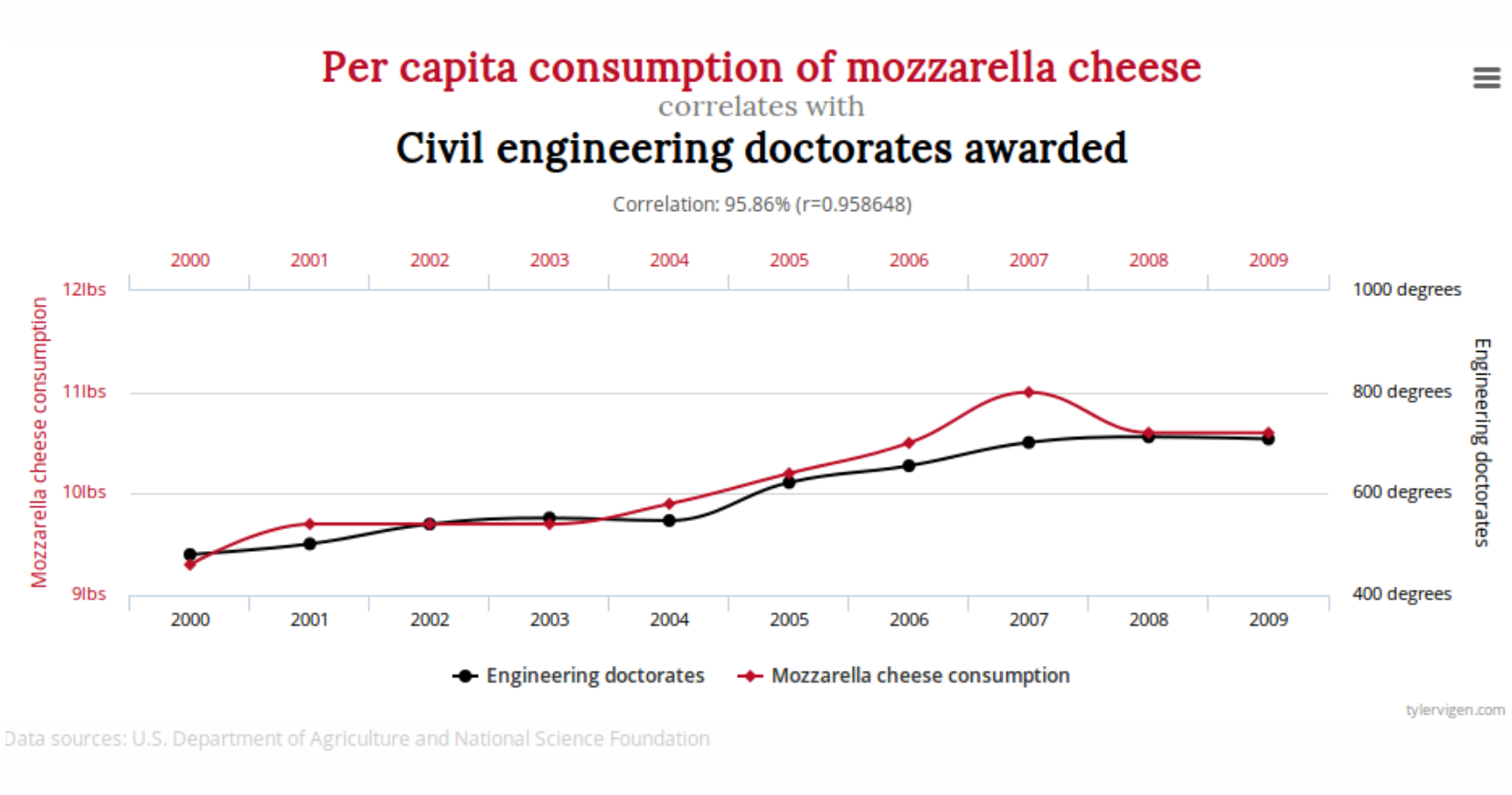
Spurious correlations

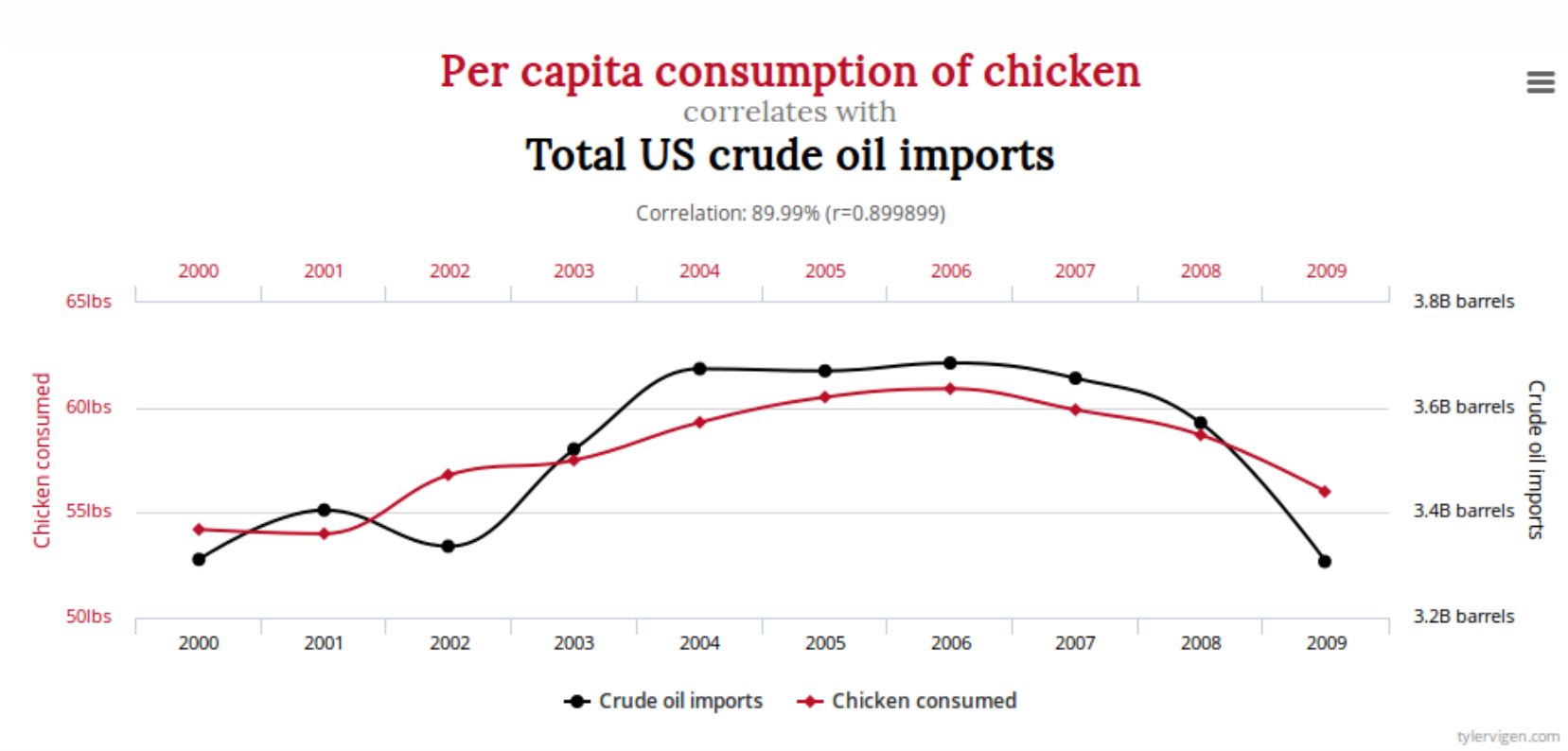
- A strong correlation does not necessarily imply a causal relationship, sometimes hidden variables play a role.
- The effect of hidden variables may occur as the result of joint dependency when changes in both the independent and dependent variable is caused by changes in hidden variables.
- Hidden variable:
  variable which has an important influence on the relationships between the variables in a study, but which is not included in the set of variables to be studied.
- Example:
  there is a correlation between shoe size and reading ability of children.
- Explanation:
  both are determined by age. Age is the hidden variable.

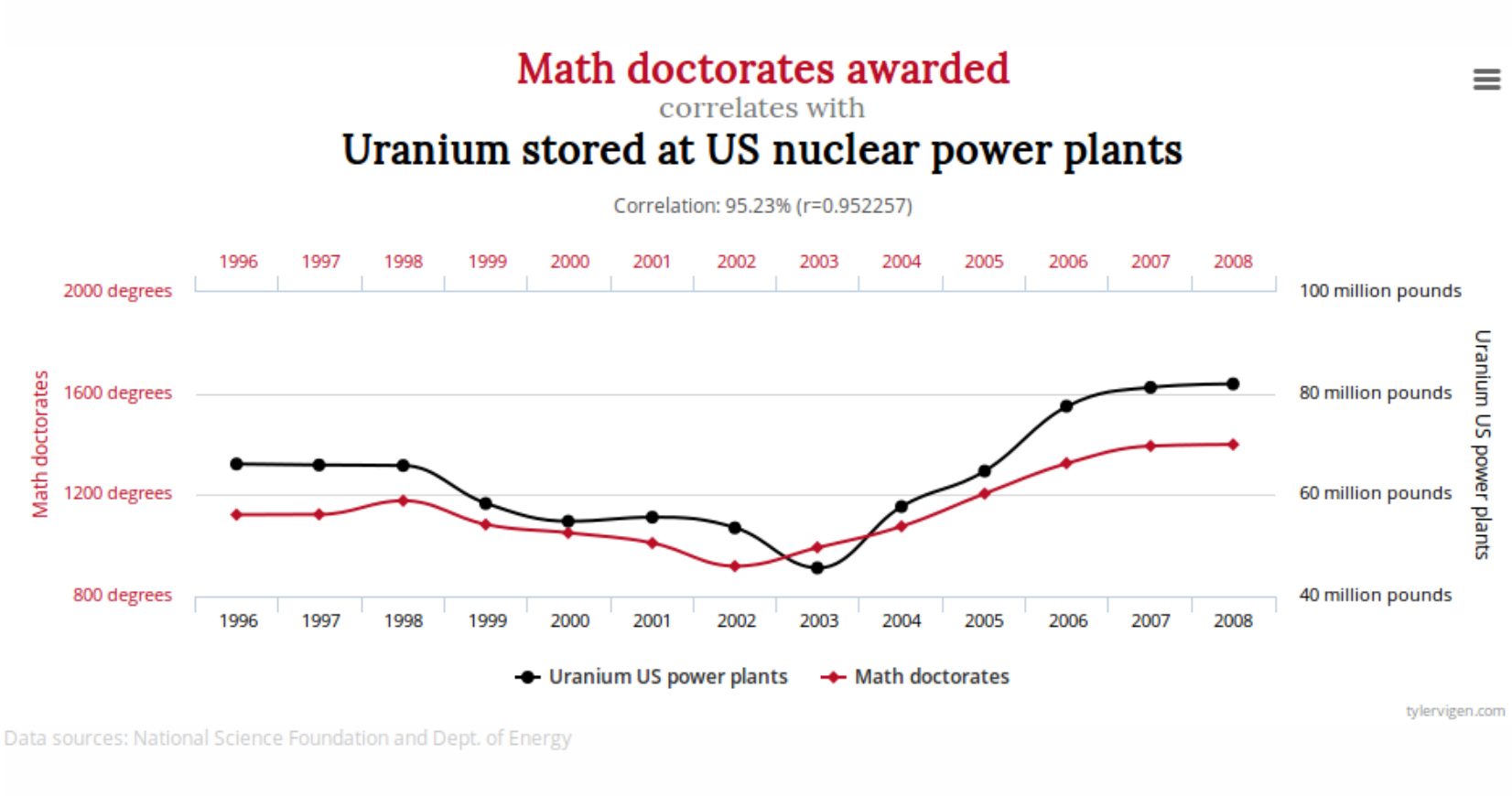
Spurious correlations

- Many examples of spurious correlations are given by Tyler Vigen (2015) in his book *Spurious Correlations* published in 2015.
- He uses **data dredging**: a computer program correlates one variable to hundreds of other variables.
- The correlations are **accidentally** and do not represent causal relationships.
- See: `http://tylervigen.com/spurious-correlations`.


Worldwide non-commercial space launches
correlates with
Sociology doctorates awarded (US)
Correlation: 78.92% (r=0.78915)
Worldwide non-commercial space launches
60 Launches
50 Launches
40 Launches
30 Launches
Sociology doctorates awarded (US)
700 Degrees awarded
600 Degrees awarded
500 Degrees awarded
1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
Sociology doctorates awarded (US)
Worldwide non-commercial space launches
tylervigen.com
Data sources: Federal Aviation Administration and National Science Foundation


Per capita consumption of mozzarella cheese
correlates with
Civil engineering doctorates awarded
Correlation: 95.86% (r=0.958648)
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
12lbs
11lbs
10lbs
9lbs
Mozzarella cheese consumption
1000 degrees
800 degrees
600 degrees
400 degrees
Engineering doctorates
Engineering doctorates
Mozzarella cheese consumption
tylervigen.com
Data sources: U.S. Department of Agriculture and National Science Foundation


Per capita consumption of chicken
correlates with
Total US crude oil imports
Correlation: 89.99% (r=0.899899)
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
65lbs
60lbs
55lbs
50lbs
Chicken consumed
3.8B barrels
3.6B barrels
3.4B barrels
3.2B barrels
Crude oil imports
Crude oil imports
Chicken consumed
tylervigen.com
Data sources: U.S. Department of Agriculture and Dept. of Energy


Math doctorates awarded
correlates with
Uranium stored at US nuclear power plants
Correlation: 95.23% (r=0.952257)
1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
2000 degrees
1600 degrees
1200 degrees
800 degrees
Math doctorates
100 million pounds
80 million pounds
60 million pounds
40 million pounds
Uranium US power plants
Uranium US power plants
Math doctorates
tylervigen.com
Data sources: National Science Foundation and Dept. of Energy

Robust alternative

- **Spearman's rank correlation coefficient** or Spearman's rho: Pearson's correlation coefficient between the $ranked$ variables.
- **Nonparametric**: does not require the population's distribution to be characterized by certain parameters.
- Sometimes also called **distribution-free tests**: they are based on fewer assumptions.
- Nonparametric tests are usually less powerful.

From correlation to regression

- Correlation **quantifies** the degree to which two variables x and y are related.
- If x and y are swapped, the r remains the same.
- Regression **describes** how one variable is related to another variable by means of a formula.
- Distinction between independent (or: predictor) variable x and dependent (or response) variable y; if x and y are swapped, the formula will change!
- The formula can be used to make predictions on y based on values of x.
- The correlation coefficient r is used in order to calculate the slope of the regression line.

Method of least-squares

- Describe relationship between variables x and y in terms of a formula.
- Regression line:
  a straight line which describes the change of a variable y with respect to the change of a variable x.
- Assume response variable y (vertical axis) and explanatory variable x (horizontal axis). The equation of the straight line has the form:

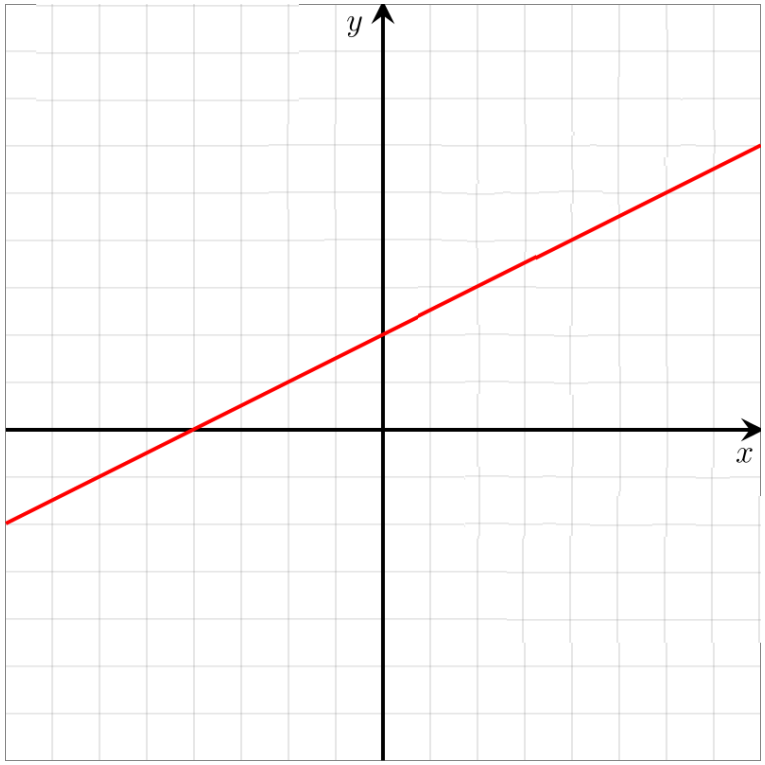
$$\hat{y} = a + bx$$

- Intercept a: the value of $\hat{y}$ for $x = 0$. Slope b: the extent to which $\hat{y}$ increases when x increases one unit.
- In our example:

$$\text{intuitive linguistic distance} = 9.082 + 0.672 \times \text{geographic distance}$$

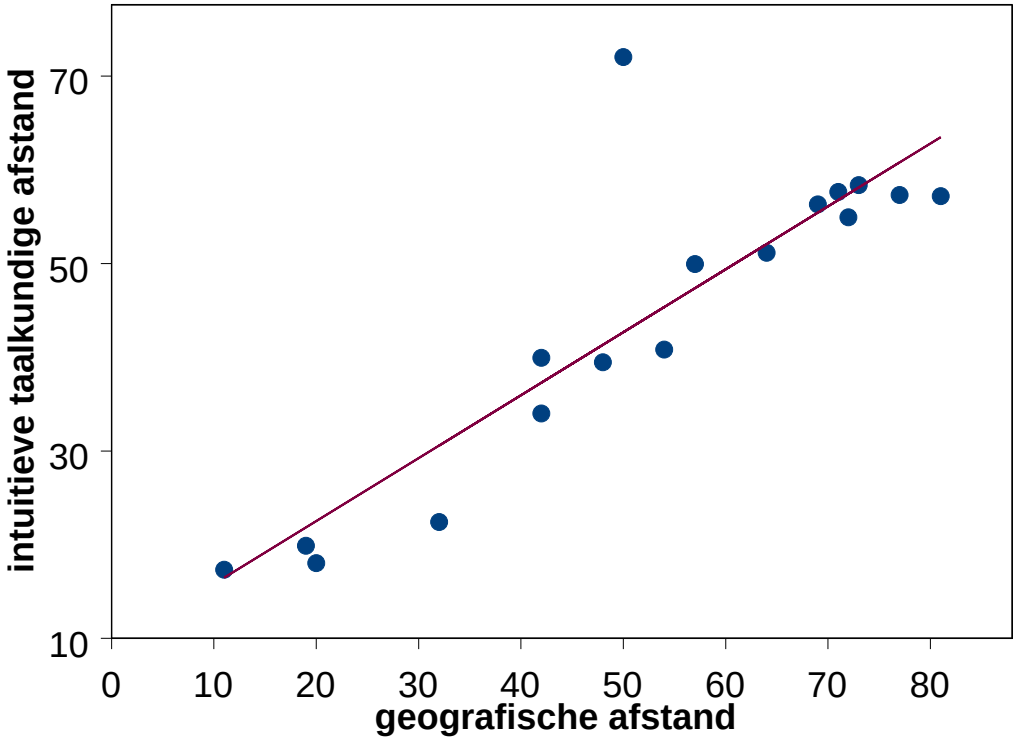
- How do we find the optimal a and b?

Method of least-squares

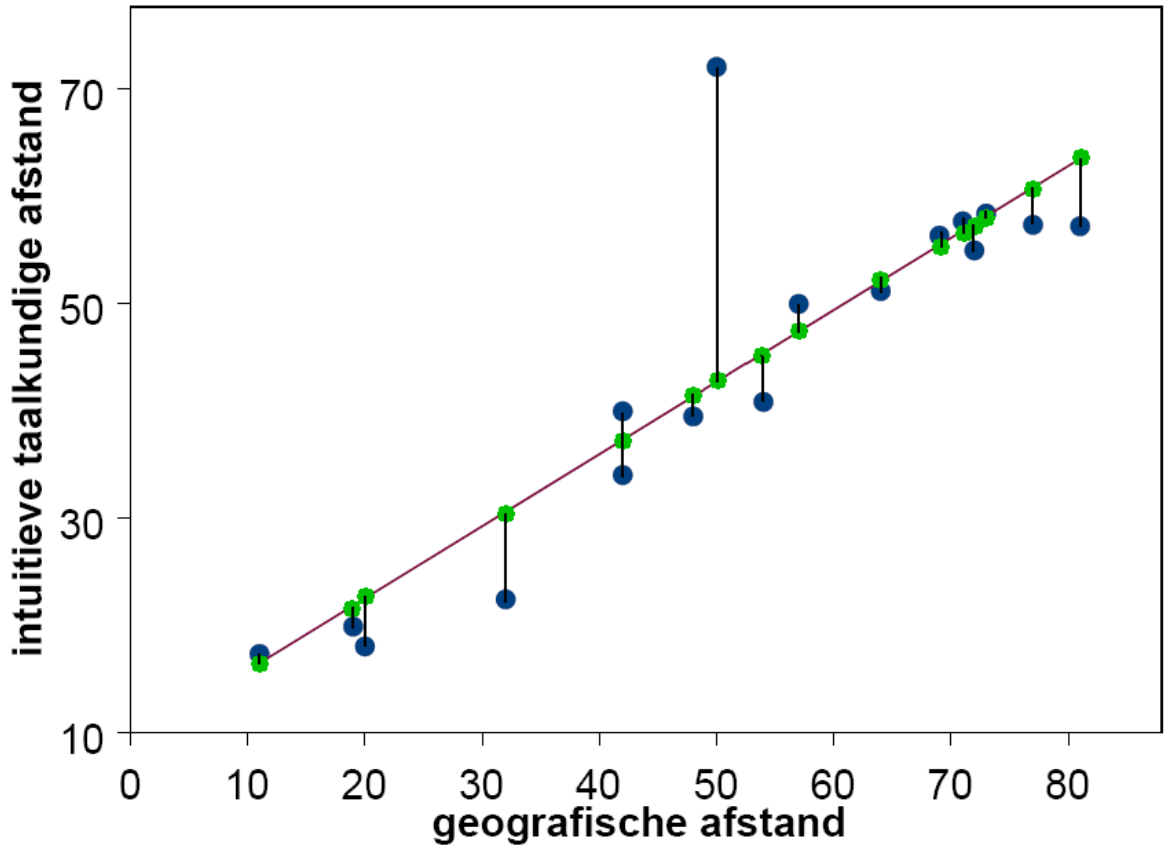


General formula: $y = a + bx$. Since the intercept $a = 2$ and the slope $b = 0.5$ the formula becomes: $y = 2 + 0.5x$ (picture adapted from Wikipedia, $Linear\ equation$).

Method of least-squares



Formula found by least-squares method:
intuitive linguistic distance $= 9.082 + 0.672 \times$ geographic distance



With the linear equation for each geographic distance the intuitive linguistic distance can be predicted. The differences between the predicted values and the observations are called **residues**.

Method of least-squares

- Least-squares regression line of y on x:
  the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.
- Assume we have data on an explanatory variable x and a response variable y for n individuals. Variable x has mean $\overline{x}$ and standard deviation s_x. Variable y has mean $\overline{y}$ and standard deviation s_y. The correlation between x and y is r. The equation of the least-squares regression line of y on x is:

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

- Slope:

$$b = r\frac{s_y}{s_x}$$

- Intercept:

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

Method of least-squares

- In our example:

$\overline{x}$=51.88235, $\overline{y}$=43.93353, s_x=21.71371, s_y=16.70869, r=0.8729689

- Slope:

$$b = r\frac{s_y}{s_x} = 0.8729689 \times \frac{16.70869}{21.71371} = 0.6717$$

- Intercept:

$$a = \overline{y} - b\overline{x} = 43.93353 - 0.6717 \times 51.88235 = 9.0816$$

- Equation:

$$\hat{y} = 9.0816 + 0.6717x$$

Method of least-squares

- Properties of regression line:
  - A change of one standard deviation in x corresponds with a change of r standard deviations in y.
  - The line always passses through the point $(\overline{x}, \overline{y})$ of the graph of y against x.

Inference for regression

- We consider the least-squares line obtained on the basis of the sample as an estimate of the $actual$ regression line of the population.
- We found:

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

- Another notation:

$$\hat{y} = b_0 + b_1 x$$

- The intercept b_0 and the slope b_1 of the regression line of the sample are estimates of the intercept β_0 and the slope β_1 of the regression line of the population.
- Regression line of the population:

$$\mu_y = \beta_0 + \beta_1 x$$

Inference for regression

- We study the relationship between the geographic province distances (x) and the intuitive linguistic distances (y).
- The data is a random sample from a population that may exhibit a linear relationship between x and y.
- When we would repeat the experiment, we would get another sample that exhibits a linear relationship between x and y that is slightly different.
- For a particular value of x we find different response values for y across the samples.
- We describe the **population mean response** as a function of the explanatory variable x:

$$\mu_y = \beta_0 + \beta_1 x$$

- For any fixed x, the responses y follow a normal distribution with standard deviation σ.

Inference for regression

- DATA = FIT + RESIDUAL
- Sample data fits the model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

  where $\beta_0 + \beta_1 x_i$ is the mean response.
- The deviations ϵ_i are supposed to be independent and normally distributed with mean 0 and standard deviation σ.
- The parameters of the model are: β_0, β_1 and σ.
- We would like to use the least-squares line as a basis for inference regarding the population which the observations of our sample are taken from.
- The parameters are estimated by intercept b_0, slope b_1 and standard deviation s.

Inference for regression

- Slope (note the relationship with correlation):

$$b_1 = r\frac{s_y}{s_x}$$

- Intercept:

$$b_0 = \overline{y} - b_1\overline{x}$$

- Standard deviation: measures the variation of y around the population regression line:

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

$n-2$ is the number of the degrees of freedom: n is the number of individuals, and 2 is the number of β's which needs to be estimated; y_i's are the 'real' measurements, $\hat{y}_i$'s are the predicted measurements.

Inference for regression

- A level C confidence interval for the intercept β_0 is:

$$b_0 \pm t^* SE_{b_0}$$

- A level C confidence interval for the slope β_1 is:

$$b_1 \pm t^* SE_{b_1}$$

- t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ en t^*.

Inference for regression

- In our example a critical value t^* with the $t(n-2) = t(17-2) = t(15)$ distribution is required.
- Go to to Vassarstats website at `http://www.vassarstats.net` and choose Distributions, t-Distributions, and enter df=15.
- Look at the table at the right. For level $C = 0.95$ we look for $p = 1 - C = 1 - 0.95 = 0.05$. Look at the two-tailed column. We find $t^* = 2.1315$.

Inference for regression

- If b_0=9.082 and SE_{b_0}=5.427 the 95% confidence interval for intercept β_0 is:

$$((9.082 - (2.1315 \times 5.427)), (9.082 + (2.1315 \times 5.427)))$$

is

$$(-2.483, 20.647)$$

We are for 95% confident that β_0 is found between -2.483 and 20.647.

- If b_1=0.672 and SE_{b_1}=0.097 the 95% confidence interval for slope β_1 is:

$$((0.672 - (2.1315 \times 0.097)), (0.672 + (2.1315 \times 0.097)))$$

is

$$(0.465, 0.879)$$

We are for 95% confident that β_1 is found between 0.465 and 0.879.

Inference for regression

- Hypotheses of significance tests for β_1:

  $H_0 : \beta_1 = 0$ (expectation of y does not vary with x)

  $H_a : \beta_1 > 0$; the p-value is $P(T \geq t)$
  $H_a : \beta_1 < 0$; the p-value is $P(T \leq t)$
  $H_a : \beta_1 \neq 0$; the p-value is $2P(T \geq |t|)$

  where T a a stochastic variable with the $t(n-2)$ distribution.

- The test statistic t is:

$$t = \frac{b_1}{SE_{b_1}}$$

- The number of the degrees of freedom is $n - 2$.

Inference for regression

- Hypotheses of significance test of β_1 in our example:

  $H_0 : \beta_1 = 0$
  (no linear relationship between geographic distance and intuitive linguistic distance)
  $H_a : \beta_1 > 0$
  (positive linear relationship between geographic distance and intuitive linguistic distance)

- The test statistic t is:

$$t = \frac{b_1}{SE_{b_1}} = \frac{0.672}{0.097} = 6.928$$

Inference for regression

- The number of the degrees of freedom is $n - 2 = 17 - 2 = 15$.
- Go to to Vassarstats website at `http://www.vassarstats.net/tabs.html` and choose t *to* P.
- Enter t=6.928 and df=15, and choose *right tail*. We find $p < 0.0001$.
- We reject H_0 and accept H_a.

Assumptions

- 1. Linearity
  The residual plot should not show any obvious pattern. If you find a curve or another pattern there is no linearity.
- Residual plot:
  a scatter plot of the regression residuals against the explanatory variable. Using a residual plot enables us to judge the fit of the regression line.
- When performing a regression analysis in SPSS make a plot: *ZPRED (X) versus *ZRESID (Y).
- 2. No perfect multicollinearity
  There should be no perfect correlation between the predictors. In simple-regression we do not need to check for this assumption, since there is just one predictor.

Assumptions

- 3. Homoskedasticity
  The variability of the data should be approximately equal across the range of the predicted values. At each level of the predictors the variance of the residuals should be constant.
- The residuals need to roughly have a similar amount of deviation from the predicted values. A good residual plot essentially looks blob-like.
- At a deeper level homogeneity of variance and homoscedasticity are the same assumption.
- Both mean that the variance of the residuals is the same everywhere. I.e. that the residual variance around predicted scores is the same for all predicted values.
- In the case of t-tests or ANOVAs, the prediction is generally the group mean. In the case of multiple regression the prediction is that which is yielded by applying the regression equation to any given set of predictor values.
- See: `http://jeromyanglim.tumblr.com/post/45895285880/`.

Assumptions

- 4. Normality of residuals
  This assumption is the least important and sometimes even not mentioned.
- Perform a Shapiro-Wilk test on the residuals and make a normal quantile plot of the residuals.
- 5. Absence of influential datapoints
  Calculate Cook's distance. A rule-of-thumb is to be concerned about Cook's distance greater than one.
- The Cook's values are generated when doing the regression analysis. Simply make a bar plot of the values.
- Do not automatically remove outliers and influential points! Check the data, look for errors or explanations for the outliers. You should always have a substantive reason to remove outliers.

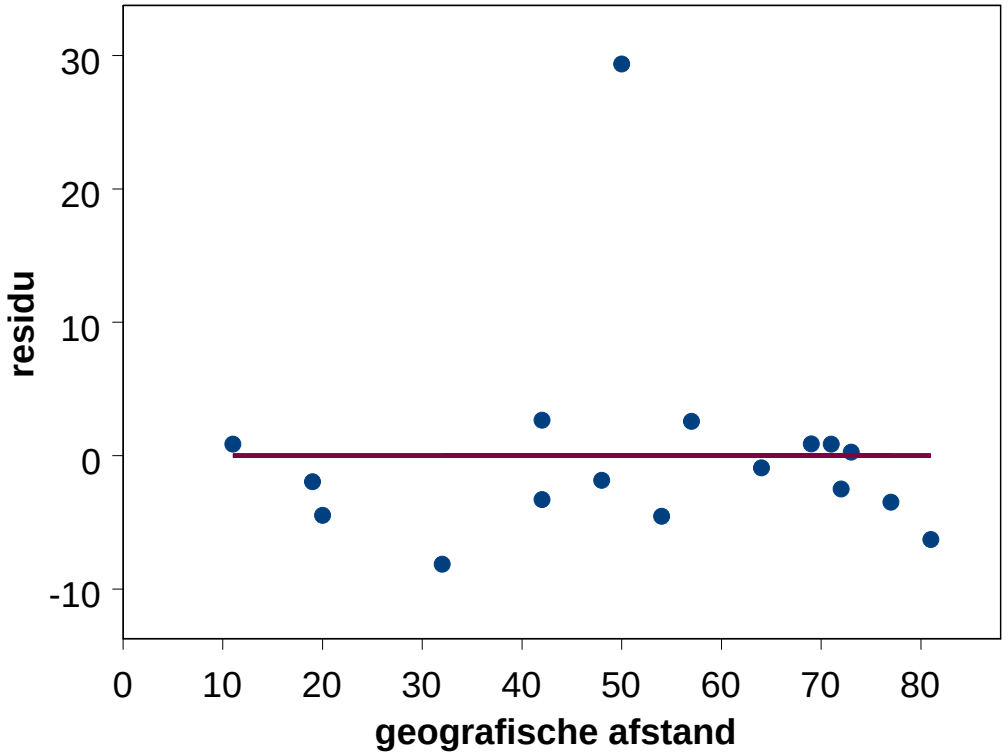
Assumptions

- 6. Independence
  All the values of the dependent variable (outcome variable) are independent of each other.

1. Linearity / 3. Homoskedasticity

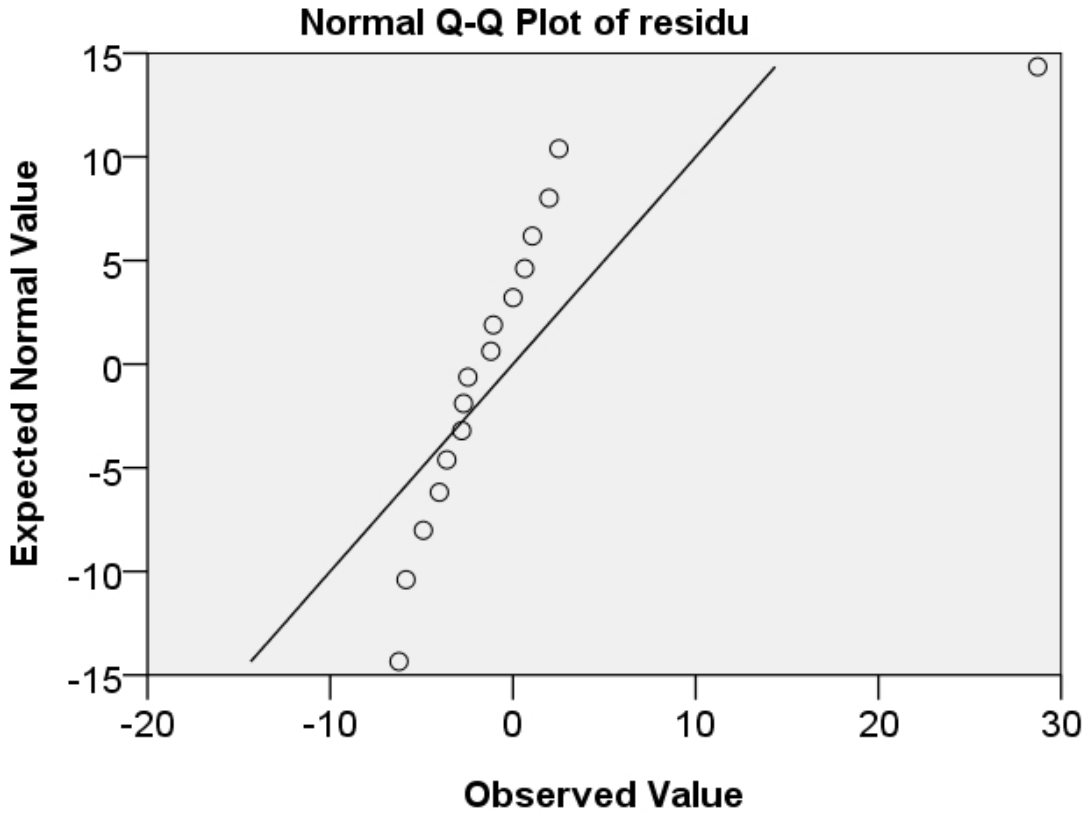
- If the regression line catches the overall pattern of the data, there should be $no\ pattern$ in the residuals: an unstructured horizontal band centered at zero (average of the residues) and symmetric at zero.

1. Linearity / 3. Homoskedasticity



Residual plot: the regression residues again the geographic distances (explanatory variable).

4. Normality of residuals

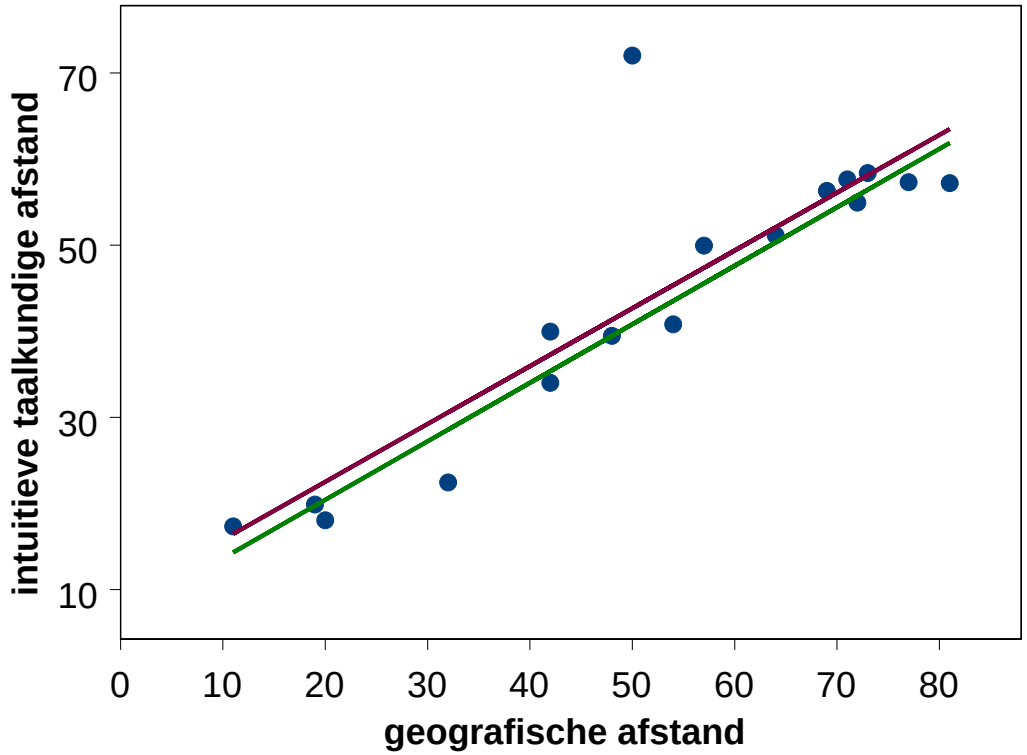


Normal quantile plot of the residues. The Shapiro-Wilk test gives a $p < 0.001$.

5. Absence of influential datapoints

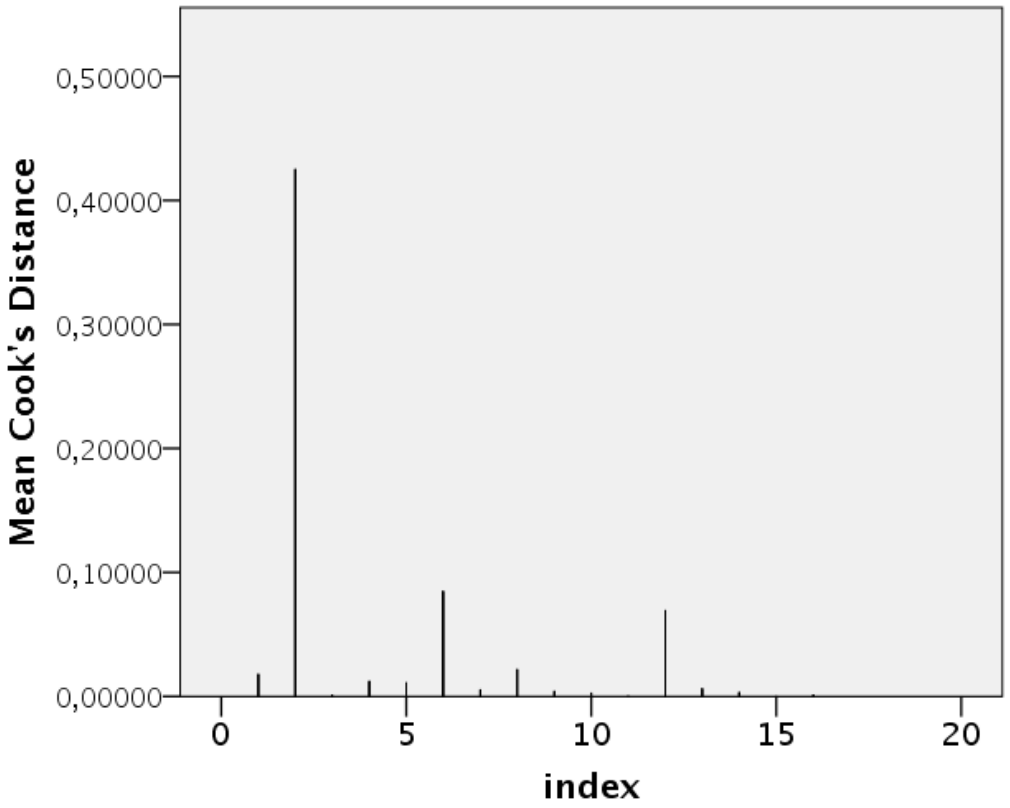
- Outlier:
  point vertically distant from the regression line and therefore, causing a large residue.
- In our example: Friesland.
- Like the correlation the least-squares regression is **not resistent** –i.e. sensitive– to influential observations.
- Influential observation:
  removing an influential observations causes a remarkable change in the position of the regression line; points distant from other points in x direction usually are influential.

5. Absence of influential datapoints



Least-squares regression line obtained on the basis of observations including Friesland (aubum) and excluding Friesland (green). The slope is about the same in these two cases.

5. Absence of influential datapoints



Plot with Cook's values. They vary between 0 and 0.42514 (all are < 1).

Results in SPSS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9,082	5,427		1,673	,115
	geografische afstand	,672	,097	,873	6,931	,000

a. Dependent Variable: intuitieve taalkundige afstand

Results in SPSS: b_0=9.082, b_1=0.672, s=8.41738. And: SE_{b_0}=5.427 and SE_{b_1}=0.097 (SE = standard error).

Determination coefficient

- The determination coefficient R^2 is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x.
- Formula:

$$R^2 = \frac{variance\ of\ predicted\ values\ \hat{y}}{variance\ of\ observed\ values\ y}$$

- In our example: $R^2 = (0.8729689)^2 = 0.7620747 = 76\%$.
- Outliers (check scatter plot) cause R^2 to be lower.

Results in SPSS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,873[a]	,762	,746	8,41738

a. Predictors: (Constant), geografische afstand

Inference for regression

- For each specific value of x, for example x^*, what will be the population **mean** response μ_y?

$$\mu_y = \beta_0 + \beta_1 x^*$$

- Since we estimate the expectation from the sample, this is:

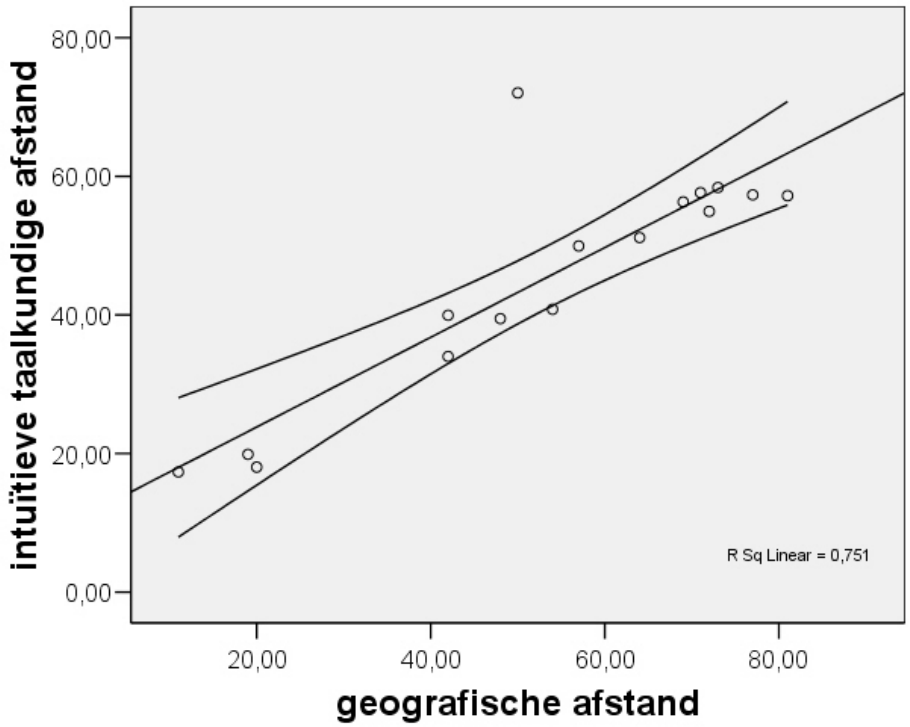
$$\mu_y = b_0 + b_1 x^*$$

- A level C confidence interval of the population mean response μ_y if x is x^* is:

$$\hat{\mu}_y \pm t^* SE_{\hat{\mu}}$$

where t^* is the value of the $t(n-2)$ distribution with area C between $-t^*$ en t^*.

Inference for regression



Confidence intervals of mean responses. We are for 95% confident that the u_y's are found between the given lower and upper boundary.

Inference for regression

- For each specific value of x, for example x^*, what response value $\hat{y}$ would one find when doing an experiment?

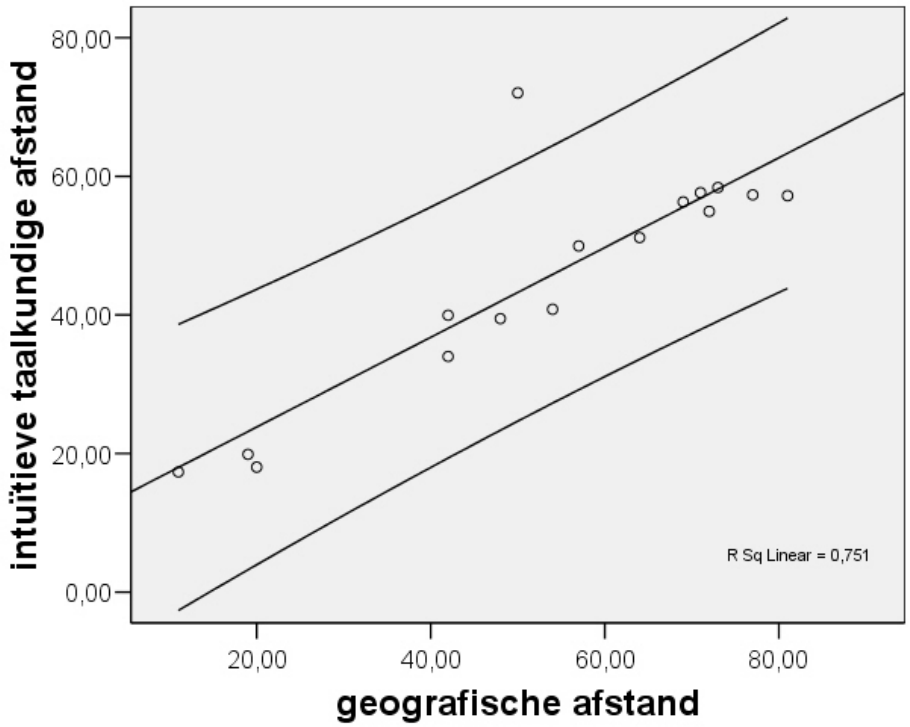
$$\hat{y} = b_0 + b_1 x^*$$

- A level C prediction interval for the predicted response value $\hat{y}$ if x is x^* is:

$$\hat{y} \pm t^* SE_{\hat{y}}$$

where t^* is the value of the $t(n-2)$ distribution with area C between $-t^*$ en t^*.

Inference for regression



Prediction intervals for individual observations. We are 95% that the $\hat{y}$'s are found between the given lower and upper boundary.