Statistics for Data Science

Wintersemester 2023/24

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Ninth lecture, December 18, 2023

Stochastic independence

Independence

Two random variables are independent if the result of one does not in any way help us predict the result of the other.

Formally, if $\mathbb{P}(x \in A, y \in B) = \mathbb{P}(x \in A)\mathbb{P}(y \in B)$ for all events A and B, then the random variables x and y are independent.

If the above does not hold, then the random variables x and y are dependent.

In statistics, the dependence of random variables is of great interest:

- The dependence between the unemployment rate and the GDP growth rate.
- The dependence between alcohol consumption and the price of alcohol.
- The dependence between lung cancer incidences and smoking.

Linear dependence

Let x and y be random variables. Let

$$y = ax + b, \quad a, b \in \mathbb{R}, \ a \neq 0.$$

Then the random variable y is a linear combination of the variable x and thus the variables x and y are (completely) linearly dependent. Linear dependence between two variables can be measured, for example, using the Pearson correlation coefficient.

Linear dependence

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y). Then the sample covariance

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

estimates the population covariance

$$\sigma_{xy} = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])]$$

and

$$\hat{\rho}(x, y) = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

estimates the Pearson correlation coefficient

$$\rho(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Linear dependence

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y).

- If the variables x and y are independent, then

$$\mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \mathbb{E}[x - \mathbb{E}[x]]\mathbb{E}[y - \mathbb{E}[y]] = 0$$

  and the Pearson correlation coefficient $\rho(x, y) = 0$.
- If $y = ax + b$, $a > 0$ and $b \in \mathbb{R}$, then $\rho(x, y) = 1$.
- If $y = ax + b$, $a < 0$ and $b \in \mathbb{R}$, then $\rho(x, y) = -1$.

In general, the Pearson correlation coefficient is a measure of the strength of linear dependence between two random variables. The coefficient $\rho \in [-1, 1]$.

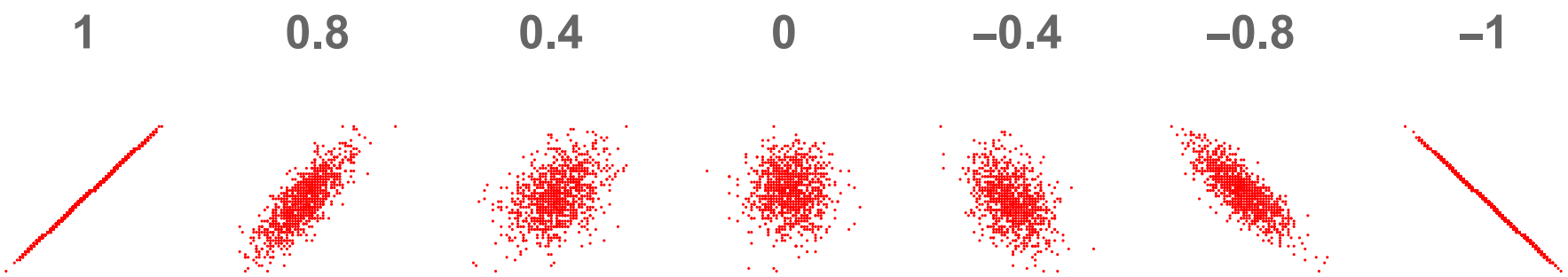
Pearson correlation coefficient

Note that linear independence does not guarantee independence.

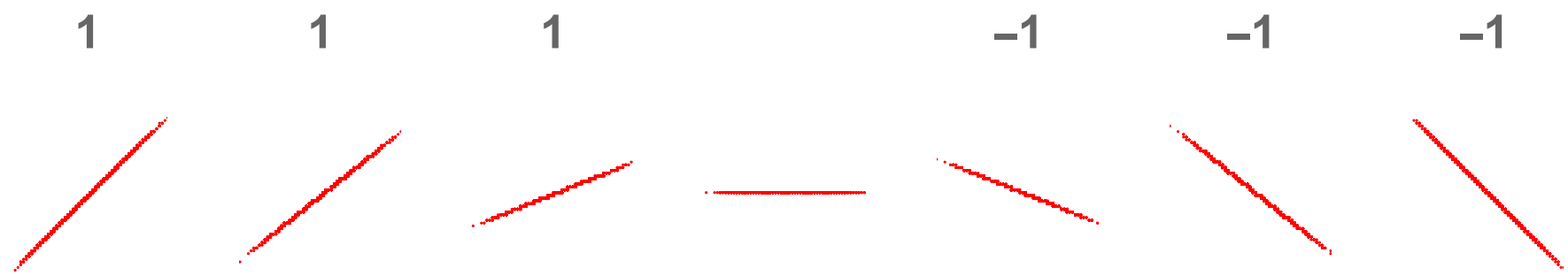
For example, if $x \sim \mathcal{U}([-1, 1])$ and $y = x^2$, then the (linear) correlation between the variables x and y is 0, even though they do depend on each other.

Recall that normally distributed random variables are uncorrelated if and only if they are independent.

Example 1

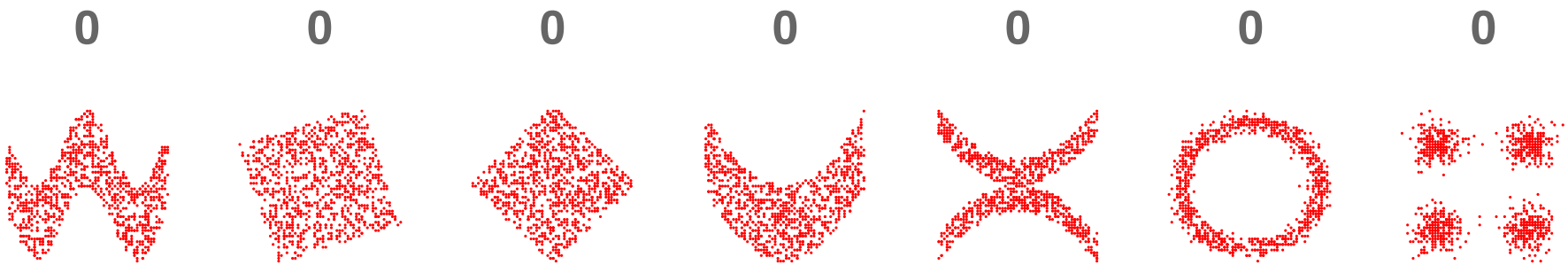


Example 2



Example 3

Correlation coefficients



Probability density function of a bivariate normal distribution:

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2(x,y)}\sigma_x\sigma_y}$$
$$\times \exp\left(-\frac{1}{2(1-\rho^2(x,y))}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho(x,y)\frac{(x-\mu_x)}{\sigma_x}\frac{(y-\mu_y)}{\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right)\right).$$

Parametric confidence interval

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y). Assume that (x, y) follows a bivariate normal distribution. Let

$$I = \frac{(1 + \hat{\rho}(x, y)) - (1 - \hat{\rho}(x, y)) \exp(2z_{\alpha/2}/\sqrt{n-3})}{(1 + \hat{\rho}(x, y)) + (1 - \hat{\rho}(x, y)) \exp(2z_{\alpha/2}/\sqrt{n-3})}$$

and let

$$u = \frac{(1 + \hat{\rho}(x, y)) - (1 - \hat{\rho}(x, y)) \exp(-2z_{\alpha/2}/\sqrt{n-3})}{(1 + \hat{\rho}(x, y)) + (1 - \hat{\rho}(x, y)) \exp(-2z_{\alpha/2}/\sqrt{n-3})},$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \frac{\alpha}{2})$ is the $(1 - \alpha/2) \cdot 100$ percentile of the standard normal distribution.

If the sample size n is large, then (l, u) estimates a level $(1 - \alpha)$ confidence interval for the Pearson correlation coefficient. Note that this confidence interval can only be used under the assumption of bivariate normal distribution. Note also that the confidence intervals for the Pearson correlation coefficient provided by different statistical softwares are almost always based on normality assumption.

Nonparametric confidence interval

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y). One can use bootstrapping to obtain nonparametric confidence intervals for the Pearson correlation coefficient:

1. Pick a new random sample of size n from the observed values $(x_1, y_1), \ldots, (x_n, y_n)$ with replacement, such that the new values are selected one-by-one and the selected observation is returned back to the original sample. (Note that this means that the same observation can be selected multiple times.)
2. Calculate the Pearson correlation coefficient for the new sample formed in the previous step.

Continued on the next slide!

3. Repeat the previous steps several times and order the obtained estimates from the smallest to the largest. Include also the original estimate of the Pearson correlation coefficient.
4. Calculate an estimate for a $(1-\alpha)\%$ confidence interval by selecting a lower bound l that is smaller than (or equal to) $1-\frac{\alpha}{2}$ of the ordered estimates and an upper bound u that is larger than (or equal to) $1-\frac{\alpha}{2}$ if the estimates. (Assume, for example, that there are 999 bootstrap estimates. Then, in total, there are 1000 estimates – the original one and the 999 new ones. Now, an estimated 90% confidence interval (l, u) is obtained by choosing the 50^{th} ordered estimate as l and the 951^{st} estimate as u. An estimate for the 95% confidence interval (l, u) is obtained by choosing the 25^{th} estimate as l and the 976^{th} estimate as u.)

One sample test for the Pearson correlation coefficient

One sample test for the Pearson correlation coefficient

The one sample test for the Pearson correlation coefficient compares the Pearson correlation coefficient to a given constant.

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y). Assume that (x, y) follows a bivariate normal distribution.

The null hypothesis: $H_0 : \rho(x, y) = \rho_0$.

The possible alternative hypotheses:

$$H_1 : \rho(x, y) > \rho_0 \text{ (one tailed)},$$
$$H_1 : \rho(x, y) < \rho_0 \text{ (one tailed)},$$
$$H_1 : \rho(x, y) \neq \rho_0 \text{ (two tailed)}.$$

One sample test for the Pearson correlation coefficient

- The test statistic

$$z = \frac{\operatorname{ar\,tanh}(\hat{\rho}(x,y)) - \operatorname{ar\,tanh}(\rho_0)}{\sqrt{\frac{1}{n-3}}} = \frac{\frac{1}{2}\log\left(\frac{1+\hat{\rho}(x,y)}{1-\hat{\rho}(x,y)}\right) - \frac{1}{2}\log\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\sqrt{\frac{1}{n-3}}}.$$

- If the sample size n is large, then under the null hypothesis, the test statistic z approximately follows the standard normal distribution.
- The expected value of the test statistic is 0.
- Large absolute values of the test statistic suggest that the null hypothesis H_0 is false.
- The null hypothesis H_0 is rejected if the p-value is small enough.

Two sample test for Pearson correlation coefficients

Two sample test for Pearson correlation coefficients

The two sample test (correlation comparison test) compares the Pearson correlation coefficients of two independent samples.

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y) and let $(z_1, w_1), \ldots, (z_m, w_m)$ be i.i.d. observations of a bivariate random variable (z, w). Assume that (x, y) follows a bivariate normal distribution with Pearson correlation coefficient $\rho(x, y)$ and that (z, w) follows a bivariate normal distribution with Pearson correlation coefficient $\rho(z, w)$. Assume that (x_i, y_i) and (z_j, w_j) are independent for all i, j.

The null hypothesis $H_0 : \rho(x, y) = \rho(z, w)$.

The possible alternative hypotheses:

$$H_1 : \rho(x, y) > \rho(z, w) \text{ (one tailed)},$$
$$H_1 : \rho(x, y) < \rho(z, w) \text{ (one tailed)},$$
$$H_1 : \rho(x, y) \neq \rho(z, w) \text{ (two tailed)}.$$

Two sample test for Pearson correlation coefficients

- The test statistic

$$z = \frac{\frac{1}{2}\log\left(\frac{1+\hat{\rho}(x,y)}{1-\hat{\rho}(x,y)}\right) - \frac{1}{2}\log\left(\frac{1+\hat{\rho}(z,w)}{1-\hat{\rho}(z,w)}\right)}{\sqrt{\frac{1}{n-3}+\frac{1}{m-3}}}.$$

- If n and m are large, then under the null hypothesis, the test statistic z approximately follows the standard normal distribution.
- The expected value of the test statistic is 0.
- Large absolute values of the test statistic suggest that the null hypothesis H_0 is false.
- The null hypothesis H_0 is rejected if the p-value is small enough.

Parametric significance test

Parametric significance test

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y). Assume that (x, y) follows a bivariate normal distribution.

The null hypothesis $H_0 : \rho(x, y) = 0$.

The possible alternative hypotheses:

$$H_1 : \rho(x, y) > 0 \text{ (one tailed)},$$
$$H_1 : \rho(x, y) < 0 \text{ (one tailed)},$$
$$H_1 : \rho(x, y) \neq 0 \text{ (two tailed)}.$$

Parametric significance test

- The test statistic

$$t = \hat{\rho}(x, y)\sqrt{\frac{n-2}{1-\hat{\rho}(x, y)^2}}.$$

- Under the null hypothesis, the test statistic follows Student's t-distribution with $n - 2$ degrees of freedom.
- The expected value of the test statistic is 0.
- Large absolute values of the test statistic suggest that the null hypothesis H_0 does not hold.
- The null hypothesis H_0 is rejected if the p-value is small enough.

Nonparametric significance test

Nonparametric significance test

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y).

The null hypothesis $H_0 : \rho(x, y) = 0$.

The possible alternative hypotheses:

$$H_1 : \rho(x, y) > 0 \text{ (one tailed)},$$
$$H_1 : \rho(x, y) < 0 \text{ (one tailed)},$$
$$H_1 : \rho(x, y) \neq 0 \text{ (two tailed)}.$$

Nonparametric significance test

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y). The significance of the observed Pearson sample correlation coefficient under the null hypothesis can be assessed using a Monte Carlo permutation test:

1. Form n new pairs $(x_1, y_1^*), \ldots, (x_n, y_n^*)$ from the original observed values $(x_1, y_1), \ldots, (x_n, y_n)$ such that each original y_j is used exactly once in the new sample.
2. Calculate the Pearson correlation coefficient $\hat{\rho}(x, y^*)$ for the sample $(x_1, y_1^*), \ldots, (x_n, y_n^*)$.
3. Repeat steps 1 and 2 several times and estimate the probability of the estimate $\hat{\rho}(x, y)$ under the null hypothesis using the values from step 2. That is, calculate the percentage of the estimates in step 2 that
   - satisfy $\hat{\rho}(x, y^*) \geq \hat{\rho}(x, y)$ (one tailed $H_1 : \rho(x, y) > 0$);
   - satisfy $\hat{\rho}(x, y^*) \leq \hat{\rho}(x, y)$ (one tailed $H_1 : \rho(x, y) < 0$);
   - satisfy $|\hat{\rho}(x, y^*)| \geq |\hat{\rho}(x, y)|$ (two tailed $H_1 : \rho(x, y) \neq 0$).

Remark. A more accurate procedure can be obtained by using the permutation test without simulations: instead of simulating new pairs, all the $n!$ possible combinations are used. The probability of $\hat{\rho}(x, y)$ under the null hypothesis is estimated using all $n!$ correlation coefficients.

Spearman (rank) correlation coefficient

Monotonic dependence

Let x and y be random variables. Let $y = g(x)$, where g is a monotonic (increasing or decreasing) function. Then the variable y is a monotonic function of the variable x and the variables x and y are (completely) monotonically dependent.

The monotonic dependence between two random variables can be measured using the Spearman rank correlation coefficient.

Spearman correlation coefficient

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y). Let $R(x_i)$, $i \in \{1, \ldots, n\}$, be the rank of the observation x_i in the sample $x_1, \ldots, x_n$ and let $R(y_i)$, $i \in \{1, \ldots, n\}$, be the rank of the observation y_i in the sample $y_1, \ldots, y_n$.

The Spearman rank correlation coefficient $\rho_S(x, y)$ is the Pearson correlation coefficient calculated for the rank sample

$$(R(x_1), R(y_1)), \ldots, (R(x_n), R(y_n)).$$

The Spearman correlation coefficient is a measure of the strength of monotonic dependence between the two random variables. The coefficient $\rho_S \in [-1, 1]$.

Confidence intervals for the Spearman correlation coefficient can be estimated using bootstrap.

Significance test

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y).

The null hypothesis $H_0 : \rho_S(x, y) = 0$.

The possible alternative hypotheses:

$$H_1 : \rho_S(x, y) > 0 \text{ (one tailed)},$$
$$H_1 : \rho_S(x, y) < 0 \text{ (one tailed)},$$
$$H_1 : \rho_S(x, y) \neq 0 \text{ (two tailed)}.$$

Significance test

- The test statistic

$$t = \hat{\rho}_S(x, y)\sqrt{\frac{n-2}{1-\hat{\rho}_S(x, y)^2}}.$$

- If n is large, then under the null hypothesis the test statistic t approximately follows Student's t-distribution with $n-2$ degrees of freedom. If the sample size is small, statistical software can be used to calculate exact p-values for the test statistic.
- The expected value of the test statistic is 0.
- Large absolute values of the test statistic suggest that the null hypothesis H_0 is not true.
- The null hypothesis H_0 is rejected if the p-value is small enough.

The significance of the Spearman rank correlation coefficient can alternatively be tested using the permutation test.

Spearman rank correlation coefficient

It is possible that some of the sample points have the same rank. In that case, all those points are assigned to have the median of the corresponding ranks. For example, if two observations have the same rank, corresponding to ranks 7 and 8, then both are assigned to have rank 7.5. If three observations have the same rank, corresponding to ranks 3, 4, and 5, then each is assigned to have rank 4.

Numerical example

Twin sisters were asked to rank different cookie brands according to the taste. The goal was to test, on significance level 5%, whether the cookie preferences were monotonically dependent. The null hypothesis is $\rho(x, y) = 0$.

rank	10	9	8	7	6	5	4	3	2	1
X (twin 1)	J	G	D	H	A	C	B	I	E	F
Y (twin 2)	G	H	D	C	A	B	J	E	I	F

Table: Cookie preferences of the twins.

The tabulated values can be converted to rank pairs:
$(6, 6)$, $(4, 5)$, $(5, 7)$, $(8, 8)$, $(2, 3)$, $(1, 1)$, $(9, 10)$, $(7, 9)$, $(3, 2)$, $(10, 4)$.

The sample standard deviations are $s_X = 3.02765$ and $s_Y = 3.02765$ and the sample covariance is $s_{XY} = 6.5$. The Spearman rank correlation coefficient is $\hat{\rho}_S(X, Y) = 0.7090909$. The test statistic has the value

$$t = \hat{\rho}_S(X, Y)\sqrt{\frac{n-2}{1-\hat{\rho}_S(X, Y)^2}} = \frac{0.7090909 \cdot \sqrt{8}}{1-(0.7090909)^2} = 2.844367.$$

Under the null hypothesis, the test statistic approximately follows Student's t-distribution with $10 - 2 = 8$ degrees of freedom. The critical values on significance level 5% are -2.306 and 2.306. Since $2.844 > 2.306$, the null hypothesis is rejected and the alternative hypothesis is accepted. The cookie preferences of the twins are monotonically dependent.

Q: What went wrong with the previous example?
A: The sample size in this example is quite small, so using asymptotic p-values is questionable. It would be better to use the exact p-value computed using statistical software or to use the permutation test.

Words of warning

Dependence $\neq$ linear dependence!

Dependence does not imply causation! See *Spurious Correlations:* https://www.tylervigen.com/spurious-correlations

Regression analysis

Regression analysis

The aim in regression analysis is to study how a dependent variable changes when one or more explanatory variables are varied. It can be used to study, e.g., if the number of violent crimes depends on alcohol consumption and if it does, how strong is this dependence.

- Does salary depend on the education level and if it does, how strong is this dependence?
- Does a parent's smoking have an effect on the height of a child and if it does, how strong is this dependence?
- Do crime rates depend on the income inequality level and if yes, how strong is this dependence?
- ⋮

Possible goals in regression analysis:

- Description of the dependence between the explanatory and dependent variables. What is the type of the relationship? How strong is the dependence?
- Predicting the values of the dependent variable.
- Controlling the values of the dependent variable.

Linear model

There are several different models that can be used in regression analysis. Today, we focus on linear regression.

Consider n observations (pairs) $(x_1, y_1), \ldots, (x_n, y_n)$ of (x, y). Assume that the values y_i are observed values of a random variable y and assume that the values x_i are observed non-random values of x. Assume that the values y_i depend linearly on the value x_i. A simple (one explanatory variable) linear model can be represented in the following way:

$$y_i = b_0 + b_1 x_i + \varepsilon_i, \quad i \in \{1, \ldots, n\},$$

where the regression coefficients b_0 and b_1 are unknown constant and the expected value of the residuals ε_i is $\mathbb{E}[\varepsilon_i] = 0$.

Simple linear model

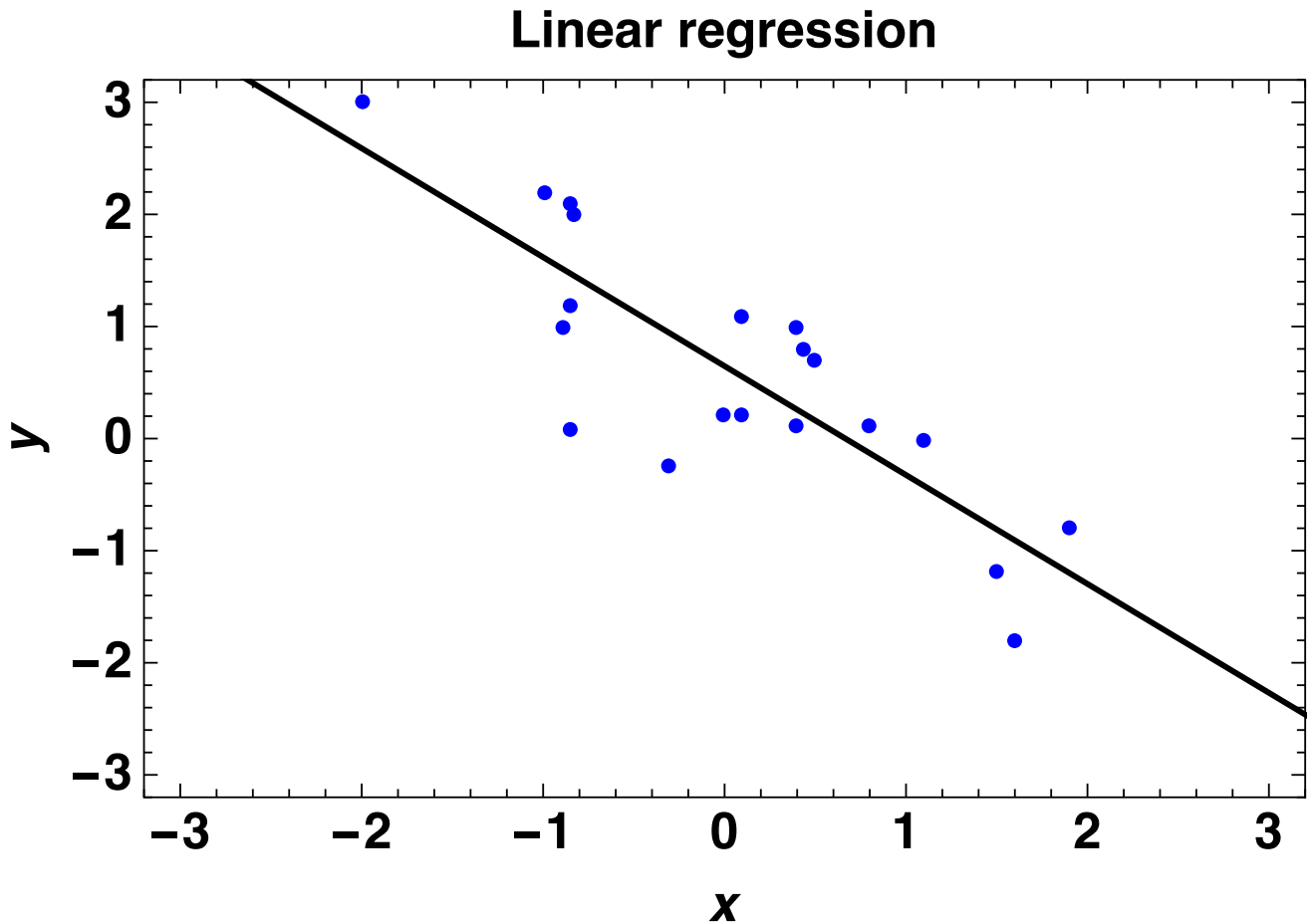


Figure: As the values of the variable x increase, the values of the variable y decrease.

Linear model, assumptions

The following assumptions are usually made when simple linear models are considered.

- The measurement of the values x_i is error-free.
- The residuals are independent of the values x_i.
- The residuals are independently and identically distributed (i.i.d.).
- The expected value of the residuals is $\mathbb{E}[\varepsilon_i] = 0$, $i \in \{1, \ldots, n\}$.
- The residuals have the same variance $\mathbb{E}[\varepsilon_i^2] = \sigma^2$, $i \in \{1, \ldots, n\}$.
- The residuals are uncorrelated, i.e., $\rho(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$.

Under these assumptions, the variable y has the following properties:

- The expected value $\mathbb{E}[y_i] = b_0 + b_1 x_i$, $i \in \{1, \ldots, n\}$.
- The variance $\mathrm{Var}(y_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$, $i \in \{1, \ldots, n\}$.
- The correlation coefficient $\rho(y_i, y_j) = 0$, $i \neq j$.

Linear regression

Linear regression

The linear model

$$y_i = b_0 + b_1 x_i + \varepsilon_i, \quad i \in \{1, \ldots, n\},$$

has the following parameters: the regression coefficients b_0 and b_1, and the variance of the residuals $\mathbb{E}[\varepsilon_i^2] = \sigma^2$. These parameters are usually unknown and must be estimated from the observations.

Under the assumption $\mathbb{E}[\varepsilon_i] = 0$ for all $i \in \{1, \ldots, n\}$, the linear model can be given as

$$y_i = \mathbb{E}[y_i] + \varepsilon_i, \quad i \in \{1, \ldots, n\},$$

where $\mathbb{E}[y_i] = b_0 + b_1 x_i$ is the so-called systematic part and ε_i is the random part of the model.

Regression line

The systematic part

$$\mathbb{E}[y_i] = b_0 + b_1 x_i$$

of the linear model defines the regression line

$$y = b_0 + b_1 x,$$

where

- b_0 is the intersection of the regression line and the y-axis;
- the slope b_1 tells us how much the independent variable y changes when the explanatory variable x grows by one unit;
- the variance of the residuals $\mathbb{E}[\varepsilon_i^2] = \sigma^2$ describes the deviation of the observed values form the regression line.

The aim in linear regression analysis is to find estimates for the regression coefficients b_0 and b_1. The estimates should be such that the estimated regression line would explain the variation of the values of the dependent variable with great accuracy.

Least squares method

In the so-called l_2 regression (least squares method), the least squares estimates are

$$\hat{b}_1 = \frac{s_{xy}}{s_x^2} = \hat{\rho}(x, y)\frac{s_y}{s_x}$$

and

$$\hat{b}_0 = \overline{y} - \hat{b}_1\overline{x}.$$

These estimates minimize the sum of squared differences

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

The least squares estimates now give an estimated regression line

$$\begin{aligned}\hat{y} = \hat{b}_0 + \hat{b}_1 x &= \overline{y} - \hat{b}_1\overline{x} + \hat{\rho}(x, y)\frac{s_y}{s_x}x \\ &= \overline{y} + \hat{\rho}(x, y)\frac{s_y}{s_x}(x - \overline{x}).\end{aligned}$$

Properties of the estimated regression line:

- If $\hat{\rho}(x, y) > 0$, then the line is increasing.
- If $\hat{\rho}(x, y) < 0$, then the line is decreasing.
- If $\hat{\rho}(x, y) = 0$, then the line is horizontal.

Fitted values and residuals

The fitted value of the variable y_i, i.e., the value given to the variable y by the regression line at points x_i, is

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i, \quad i \in \{1, \ldots, n\}.$$

The residual $\hat{\varepsilon}_i$ of the estimated model is the difference

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i \in \{1, \ldots, n\},$$

between the observed value y_i (of the variable y) and the fitted value $\hat{y}_i$.

Note that $y_i = \hat{y}_i + \hat{\varepsilon}_i$, $i \in \{1, \ldots, n\}$.

The regression model explains the observed values of the dependent variables the better, the closer the fitted values are to the observed ones. In other words, the regression model explains the observed values of the dependent variable the better, the smaller the residuals of the estimated model are.

Example

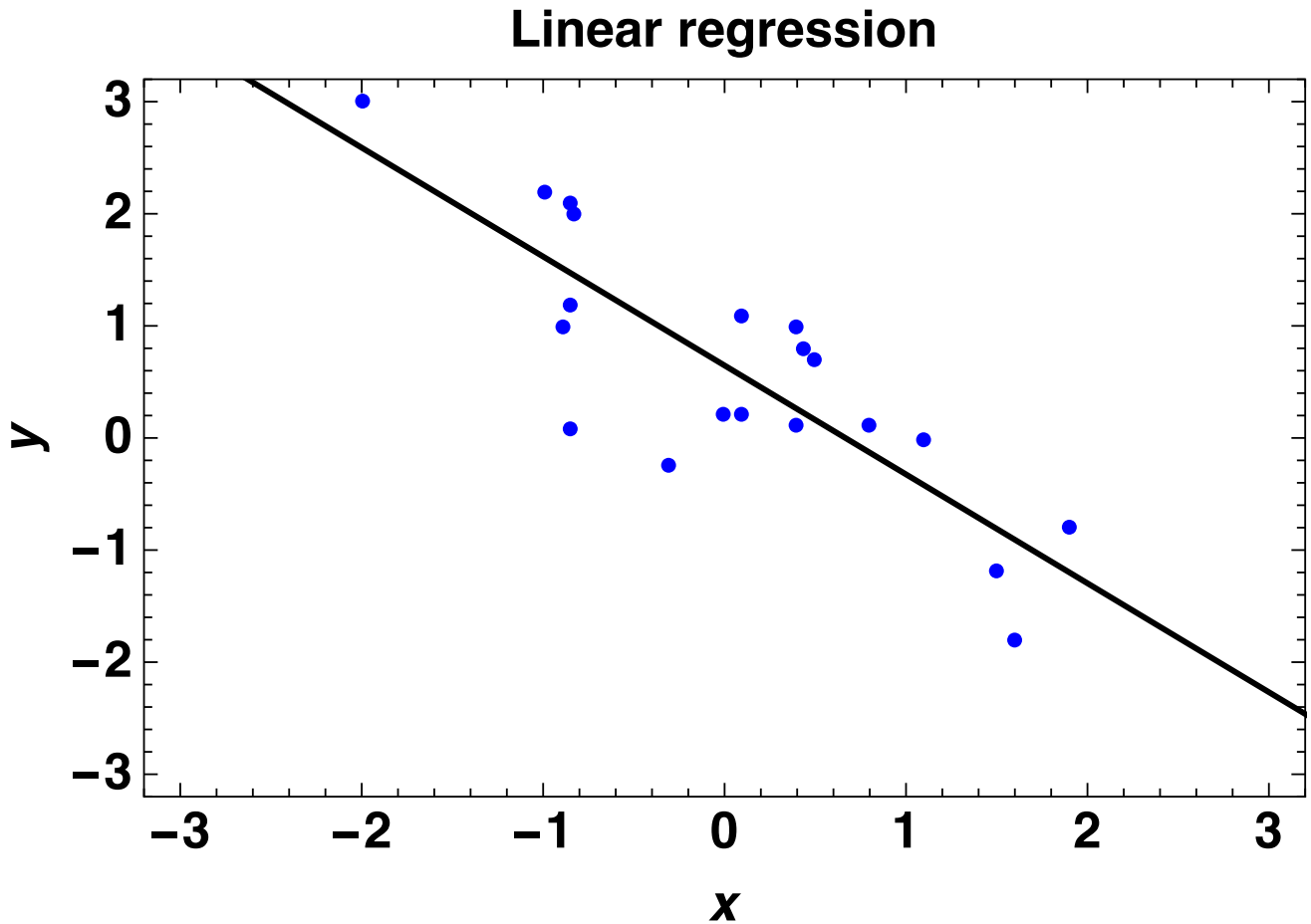


Figure: The estimated regression line minimizes the squared sum of the residuals.

Numerical example

We wish to model the dependence of the sales of Brand X cookies and Brand Y cookies. We assume that the sales are linearly dependent, and try to apply linear regression.

Brand X	Brand Y
5673	5489
4892	5987
5735	5362
5382	5738
5982	4988
5487	5576
5764	5481
5933	4999
5298	5832
5561	5591
5721	5298
5386	5632

Table: Monthly sales of Brand X and Brand Y cookies.

The sample standard errors $s_X = 302.95$ and $s_Y = 302.85$, the sample covariance $s_{XY} = -86145.95$, and the sample means $\overline{X} = 5567.833$ and $\overline{Y} = 5497.75$. The estimated regression parameters

$$\hat{b}_1 = \frac{s_{XY}}{s_X^2} = \frac{-86145.95}{302.95^2} = -0.938\ldots$$

and

$$\hat{b}_0 = \overline{Y} - \hat{b}_1\overline{X} = 5497.75 - (-0.938\ldots) \cdot 5567.833 = 10723.87.$$

An estimated regression model can now be given as

$$\hat{Y}_i = 10723.87 - 0.938X_i.$$

Fit	Actual	Residual
5399.040	5489	89.96
6132.108	5987	-145.11
5340.845	5362	21.15
5672.181	5738	65.82
5109.004	4988	-121.00
5573.625	5576	2.38
5313.625	5481	167.37
5154.997	4999	-156.00
5751.025	5832	80.97
5504.166	5591	86.83
5353.986	5298	-55.99
5668.426	5632	-36.43

Table: Fitted values and actual sales of Brand X cookies. The residuals $\hat{\epsilon}_i = y_i - \hat{y}_i$ have been tabulated as well.

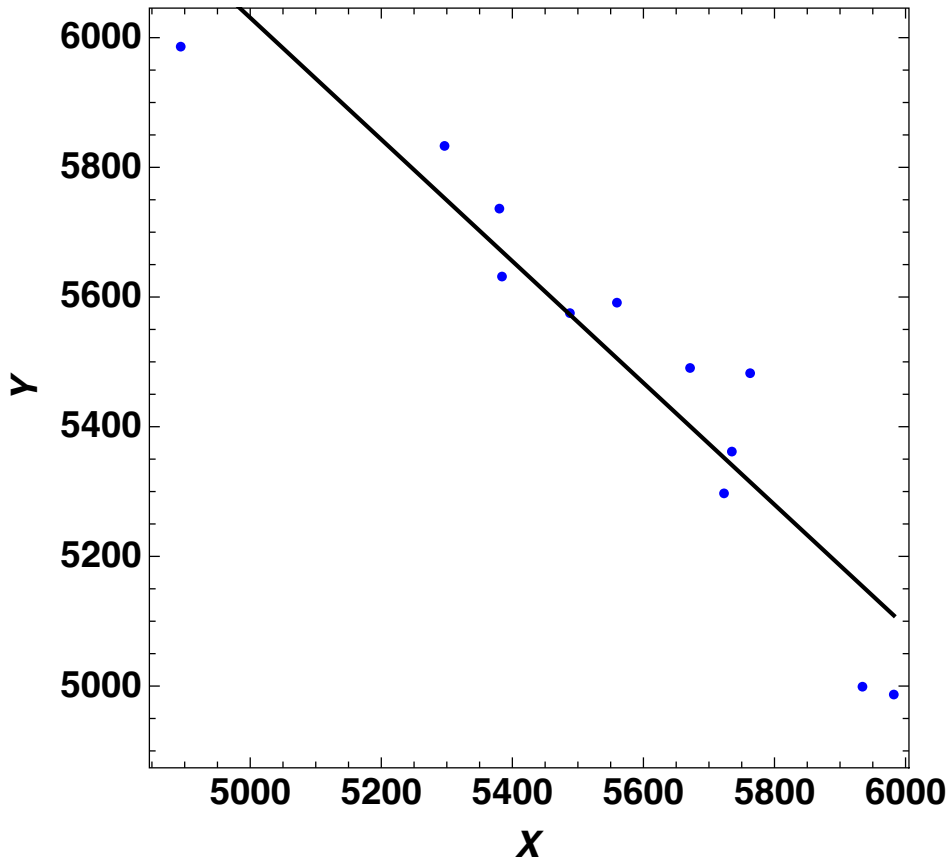


Figure: Brand Y cookies, sales and fit.

Residual mean square estimation

If the assumptions of the linear model hold, then an unbiased estimate of $\mathrm{Var}(\varepsilon_i) = \sigma^2$ is

$$\mathrm{Var}(\hat{\varepsilon}) = \frac{1}{n-2}\sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2 = \frac{1}{n-2}\sum_{i=1}^{n}(\hat{\varepsilon}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

In the formula above, the number of the estimated parameters (b_0 and b_1) is subtracted from the sample size n.

Error sum of squares

Consider the total sum of squares (SST)

$$\sum_{i=1}^{n}(y_i - \overline{y})^2,$$

and the error sum of squares (SSE)

$$\sum_{i=1}^{n}(\hat{\varepsilon}_i)^2.$$

It can be shown that

$$SSE = \sum_{i=1}^{n}(\hat{\varepsilon}_i)^2 = (1 - \hat{\rho}(x,y)^2)\sum_{i=1}^{n}(y_i - \overline{y})^2 = (1 - \hat{\rho}(x,y)^2)SST.$$

Since $\hat{\rho}(x,y) \in [-1,1]$, we have that SSE$\leq$SST.

Error sum of squares

The error sum of squares SSE is 0 if and only if all the observed values lie on the same line. In this case, the linear regression model explains the values of the dependent variable perfectly.

The error sum of squares SSE equals the total sum of squares if and only if the sample correlation coefficient $\hat{\rho}(x, y) = 0$. In this case, the linear regression model fails to explain any part of the values of y.

Model sum of squares

The model sum of squares SSM is defined as

$$SSM = SST - SSE.$$

The model sum of squares SSM describes the part of variation of the observed values of y that is explained by the regression model.

There holds

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2,$$

and since $\overline{y} = \overline{\hat{y}}$, the equation can be given as

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \overline{\hat{y}})^2.$$

Coefficient of determination

The coefficient of determination is defined as

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSM}{SST}.$$

The coefficient of determination R^2 measures the proportion of SST explained by the model.

There holds $0 \leq R^2 \leq 1$, and the coefficient of determination is usually given as a percentage $100R^2\%$.

The coefficient of determination $R^2 = (\hat{\rho}(y, \hat{y}))^2$, where $\hat{\rho}(y, \hat{y})$ is the sample correlation coefficient of the observed values of the dependent variable and the corresponding fitted values. In a simple linear regression model with one explaining variable, $R^2 = (\hat{\rho}(y, x))^2$.

Properties of the coefficient of determination

The following conditions are equivalent:

- The coefficient of determination $R^2 = 1$.
- All the residuals vanish: $\hat{\varepsilon}_i = 0$, $i \in \{1, \ldots, n\}$.
- All the observations (x_i, y_i) lie on the same line.
- The sample correlation coefficient $\hat{\rho}(x, y) = \pm 1$.
- The regression model completely explains the variation of the observed values of the dependent variable y.

Properties of the coefficient of determination

The following conditions are equivalent:

- The coefficient of determination $R^2 = 0$.
- The regression coefficient $\hat{b}_1 = 0$.
- The sample correlation coefficient $\hat{\rho}(x, y) = 0$.
- The regression model fails completely in explaining the variation of the observed values of the dependent variable y.

Numerical example

The numerical example above continues. . .

Calculate the total sum of squares

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{1} 2(y_i - 5497.75)^2 = 889447.8,$$

the error sum of squares

$$SSE = \sum_{i=1}^{n} (\hat{\varepsilon}_i)^2 = 119484.2$$

and the model sum of squares

$$SSM = SST - SSE = 889447.8 - 119484.2 = 769963.6.$$

Now, the coefficient is determination

$$R^2 = \frac{SSM}{SST} = \frac{769963.6}{889447.8} = 0.865\ldots$$

Is this a good model?

About the assumptions: We assumed above that the values x_i of the explanatory variable x are non-random. In linear regression, the values x_i can very well also be assumed to be random.

Words of warning:

- The regression model should not be used to predict any values of the range of x. Tail behavior can differ from majority of the data.
- If there is nonlinear dependence between x and y, then linear regression is not a suitable approach.
- The least squares method (l_2 regression) is very sensitive to outliers (i.e., it is non-robust).

Example, linear regression

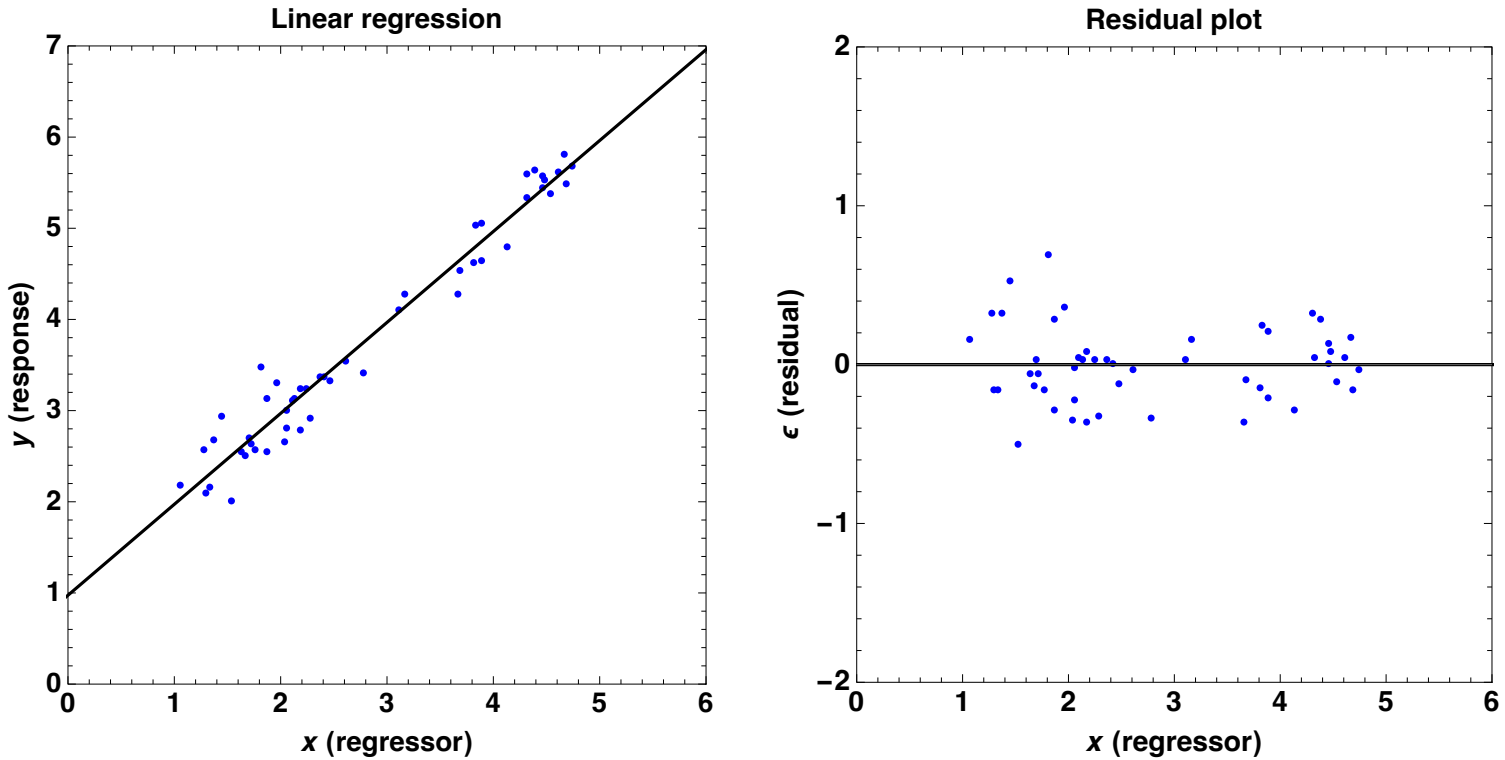


Figure: Estimated regression line and residuals.

Example, outlier

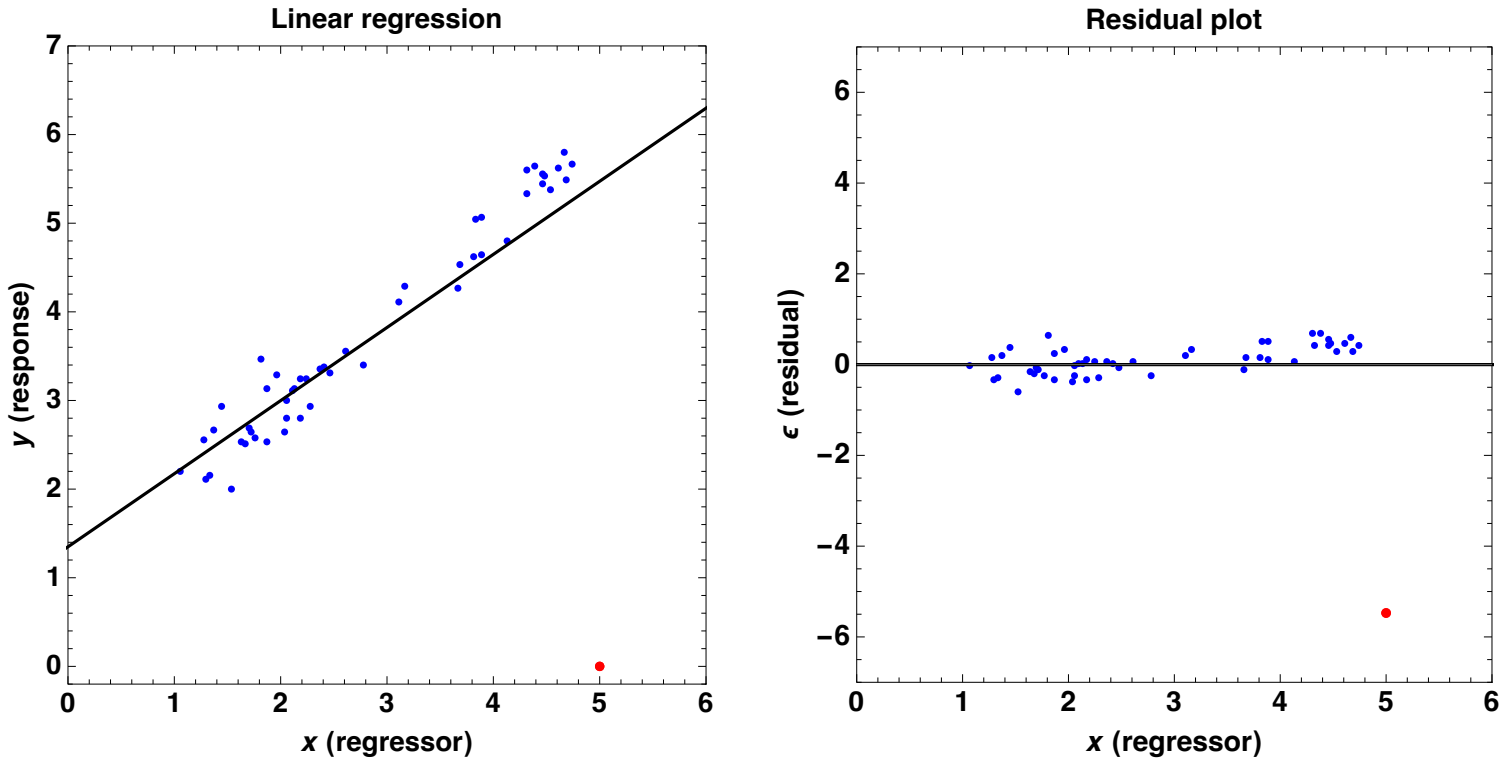


Figure: Estimated regression line and residuals. Note the effect of an outlier.

Example, heteroscedasticity

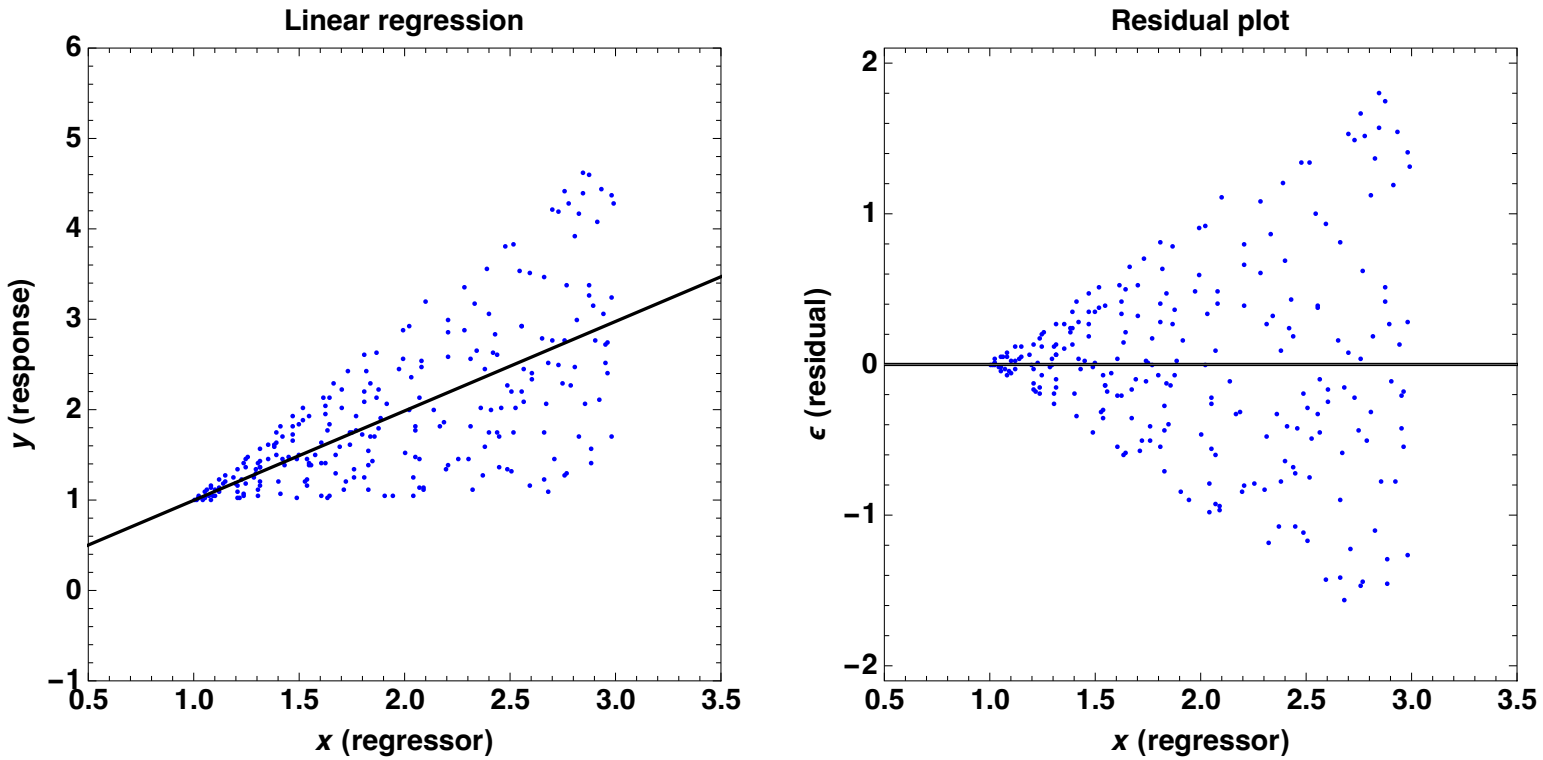


Figure: Estimated regression line and residuals. Note that the variance of the residuals increases.

Example, non-linear dependence

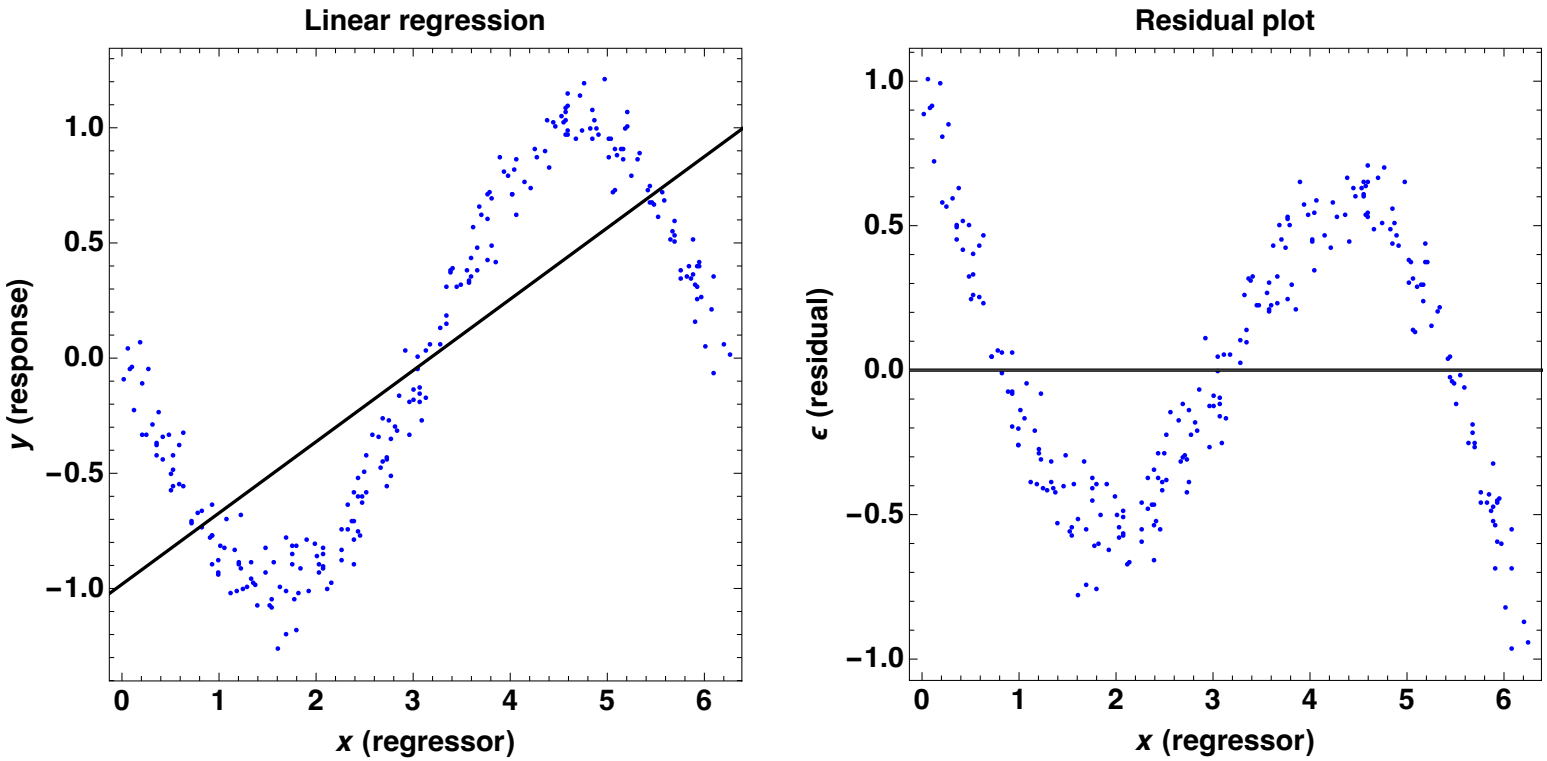


Figure: Estimated regression line and residuals. Note the clear non-linear dependence.