Linear regression

Lecturer:
Francesco Dolce

Department of Applied Mathematics
Faculty of Information Technology
Czech Technical University in Prague

Probability and Statistics

BIE-PST, WS 2023/24, Lecture 12

Content

- **Probability theory:**
  - Events, probability, conditional probability, Bayes' Theorem, independence of events.
  - Random variables, distribution function, functions of random variables, characteristics of random variables: expected value, variance, moments, generating function, quantiles, critical values, important discrete and continuous distributions.
  - Random vectors, joint and marginal distributions, functions of random vectors, independence of random variables, conditional distribution, conditional expected value, covariance and correlation.
  - Markov's and Chebyshev's inequality, weak law of large numbers, strong law of large numbers, Central limit theorem.
- **Mathematical statistics:**
  - Point estimators, sample mean, sample variance, properties of point estimators, Maximum likelihood method.
  - Interval estimators, hypothesis testing, one-sided vs. two-sided alternatives, linear regression, estimators of regression parameters, testing of linear model.

Recap

Based on a **random sample** of i.i.d. random variables $X_1, \dots, X_n$ from a parametric distribution F_θ we can:

- Estimate the parameters using **point estimates** – sample mean, sample variance, etc.
- Find **confidence intervals** – regions, where the parameter lies with a large probability:

$$\mathrm{P}(L < \theta < U) = 1 - \alpha.$$

- Test **hypotheses** – verify whether statements about parameters may or may not be true, with a given maximal probability of wrongful rejection.

Covariance and correlation

Suppose we want to examine the connection between two variables.

Sometimes we expect that there is a relation, sometimes we can assume there is not.

Examples

- Heights of sons and heights of fathers.
- Bodily weight and height.
- Mean temperature and latitude from city to city.
- Income and the number of years spent studying.
- Number of storks and number of newborns in a city (see literature).

Covariance and correlation

Suppose we want to examine the connection between two variables.

Sometimes we expect that there is a relation, sometimes we can assume there is not.

Examples

- Heights of sons and heights of fathers.
- Bodily weight and height.
- Mean temperature and latitude from city to city.
- Income and the number of years spent studying.
- Number of storks and number of newborns in a city (see literature).

First we model this connection using **correlation**.

Covariance and correlation

The **covariance** of two random variables X and Y is defined as

$$\operatorname{cov}(X,Y) = \operatorname{E}\left((X - \operatorname{E}X)(Y - \operatorname{E}Y)\right)$$

and can be computed as

$$\operatorname{cov}(X,Y) = \operatorname{E}(XY) - \operatorname{E}X\,\operatorname{E}Y.$$

Covariance and correlation

The **covariance** of two random variables X and Y is defined as

$$\operatorname{cov}(X,Y) = \mathrm{E}\left((X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)\right)$$

and can be computed as

$$\operatorname{cov}(X,Y) = \mathrm{E}\,(XY) - \mathrm{E}\,X\,\mathrm{E}\,Y.$$

The **correlation coefficient** is defined as

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var} X}\sqrt{\operatorname{var} Y}}$$

and gives a measure of the **linear dependence** between X and Y.

Covariance and correlation

Theorem

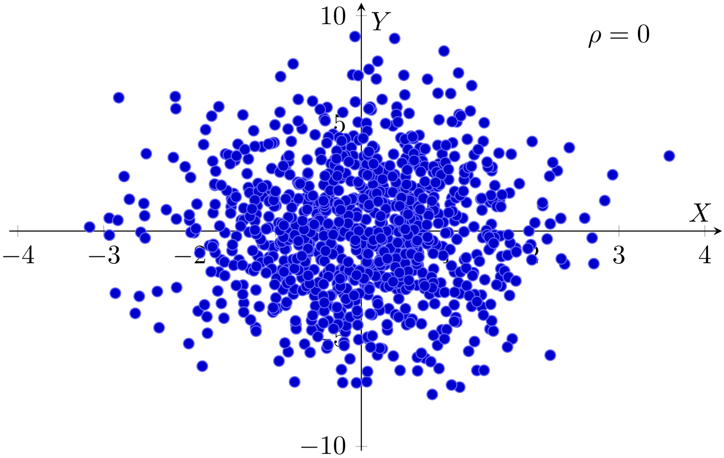
For the correlation coefficient $\rho_{X,Y}$ it holds that

1. $\rho_{X,Y} \in [-1, 1]$.
2. *If X and Y are independent, then $\rho_{X,Y} = 0$.*
3. *If $Y = a + bX$ for $b > 0$, then $\rho_{X,Y} = 1$.*
4. *If $Y = a + bX$ for $b < 0$, then $\rho_{X,Y} = -1$.*

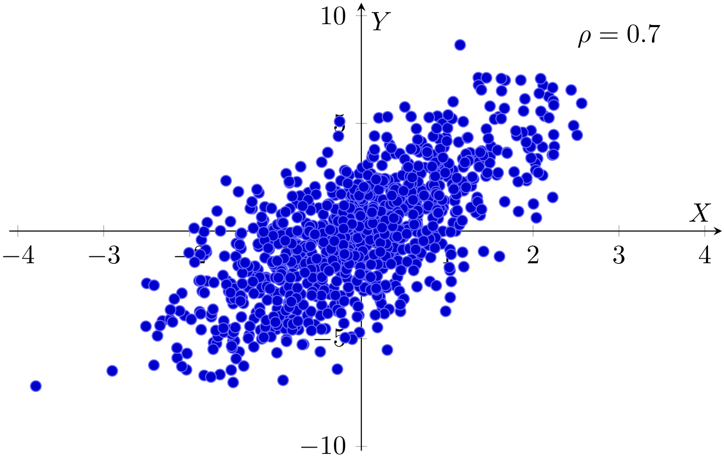
Proof

See lecture 6. □

Correlation – sample of 1000 values



Correlation – sample of 1000 values



Covariance and correlation – estimation

Based on a random sample of pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, the covariance can be estimated using the **sample covariance**:

$$s_{X,Y} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n).$$

Covariance and correlation – estimation

Based on a random sample of pairs $(X_1, Y_1), \dots, (X_n, Y_n)$, the covariance can be estimated using the **sample covariance**:

$$s_{X,Y} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n).$$

The correlation coefficient can be estimated using the **sample correlation coefficient** as

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y},$$

where $s_X = \sqrt{s_X^2}$ and $s_Y = \sqrt{s_Y^2}$ are the sample standard deviations of X and Y, respectively.

Sample covariance and correlation – properties

The sample covariance can be rewritten as

$$\begin{aligned} s_{X,Y} &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}_n\bar{Y}_n\right) \\ &= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}_n\bar{Y}_n\right). \end{aligned}$$

From the law of large numbers it follows that it is a consistent estimator of the covariance.

Sample covariance and correlation – properties

The sample covariance can be rewritten as

$$\begin{aligned} s_{X,Y} &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}_n\bar{Y}_n\right) \\ &= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}_n\bar{Y}_n\right). \end{aligned}$$

From the law of large numbers it follows that it is a consistent estimator of the covariance.

Because the sample variances are consistent estimators of the real variances, the sample correlation is therefore a consistent estimator of the correlation coefficient itself.

Estimating the correlation – example

Example – comparing heights of fathers and sons

Suppose we want to estimate the correlation between the heights of fathers and their sons. We have observed five pairs of fathers and their sons, now adults. Their heights were measured as follows:

height of father [cm]	X_i	172	176	180	184	186
height of son [cm]	Y_i	178	183	180	188	190

Estimating the correlation – example

Example – comparing heights of fathers and sons

Suppose we want to estimate the correlation between the heights of fathers and their sons. We have observed five pairs of fathers and their sons, now adults. Their heights were measured as follows:

height of father [cm]	X_i	172	176	180	184	186
height of son [cm]	Y_i	178	183	180	188	190

We have computed the following characteristics from the data:

$$\sum_{i=1}^{n} X_i = 898, \qquad \sum_{i=1}^{n} Y_i = 919,$$

$$\sum_{i=1}^{n} X_i^2 = 161412, \qquad \sum_{i=1}^{n} Y_i^2 = 169017,$$

$$\sum_{i=1}^{n} X_i Y_i = 165156.$$

Estimating the correlation – example

Example – comparing heights of fathers and sons, continued

From the observed characteristics we compute the sample means, variances and the covariance:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{898}{5} = 179.6, \qquad \bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{919}{5} = 183.8,$$

$$s_X^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2\right) = \frac{1}{4}\left(161412 - 5 \cdot 179.6^2\right) = 32.8,$$

$$s_Y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}_n^2\right) = \frac{1}{4}\left(169017 - 5 \cdot 183.8^2\right) = 26.2,$$

$$s_{X,Y} = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}_n\bar{Y}_n\right) = \frac{1}{4}(165156 - 5 \cdot 179.6 \cdot 183.8) = 25.9.$$

Estimating the correlation – example

Example – comparing heights of fathers and sons, continued

From the observed characteristics we compute the sample means, variances and the covariance:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{898}{5} = 179.6, \qquad \bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{919}{5} = 183.8,$$

$$s_X^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2\right) = \frac{1}{4}\left(161412 - 5\cdot 179.6^2\right) = 32.8,$$

$$s_Y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}_n^2\right) = \frac{1}{4}\left(169017 - 5\cdot 183.8^2\right) = 26.2,$$

$$s_{X,Y} = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}_n\bar{Y}_n\right) = \frac{1}{4}\left(165156 - 5\cdot 179.6\cdot 183.8\right) = 25.9.$$

The sample correlation coefficient is obtained as

$$r_{X,Y} = \frac{s_{X,Y}}{\sqrt{s_X^2 s_Y^2}} = \frac{25.9}{\sqrt{32.8\cdot 26.2}} \doteq 0.883.$$

Estimating the correlation – example

Example – comparing heights of fathers and sons, continued

From the observed characteristics we compute the sample means, variances and the covariance:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{898}{5} = 179.6, \qquad \bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{919}{5} = 183.8,$$

$$s_X^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2\right) = \frac{1}{4}\left(161412 - 5\cdot 179.6^2\right) = 32.8,$$

$$s_Y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}_n^2\right) = \frac{1}{4}\left(169017 - 5\cdot 183.8^2\right) = 26.2,$$

$$s_{X,Y} = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_iY_i - n\bar{X}_n\bar{Y}_n\right) = \frac{1}{4}\left(165156 - 5\cdot 179.6\cdot 183.8\right) = 25.9.$$

The sample correlation coefficient is obtained as

$$r_{X,Y} = \frac{s_{X,Y}}{\sqrt{s_X^2 s_Y^2}} = \frac{25.9}{\sqrt{32.8\cdot 26.2}} \doteq 0.883.$$

We can conclude that there is a positive correlation between the height of sons and their fathers.

Estimating the correlation – example

Example – comparing heights of fathers and sons, continued

From the observed characteristics we compute the sample means, variances and the covariance:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{898}{5} = 179.6, \qquad \bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{919}{5} = 183.8,$$

$$s_X^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2\right) = \frac{1}{4}\left(161412 - 5\cdot 179.6^2\right) = 32.8,$$

$$s_Y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}_n^2\right) = \frac{1}{4}\left(169017 - 5\cdot 183.8^2\right) = 26.2,$$

$$s_{X,Y} = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_iY_i - n\bar{X}_n\bar{Y}_n\right) = \frac{1}{4}\left(165156 - 5\cdot 179.6\cdot 183.8\right) = 25.9.$$

The sample correlation coefficient is obtained as

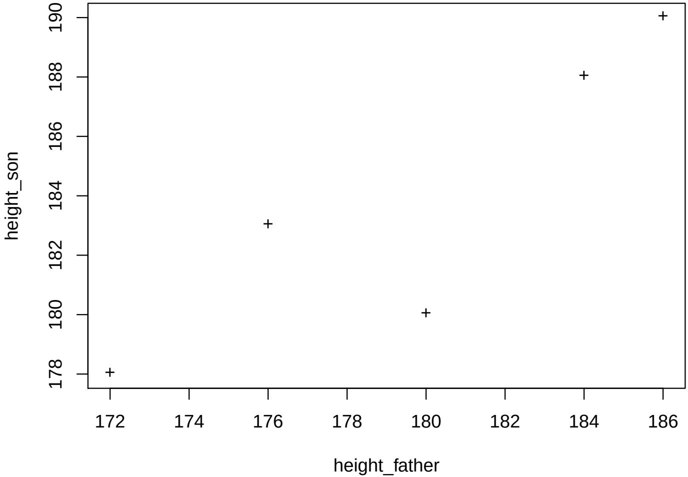
$$r_{X,Y} = \frac{s_{X,Y}}{\sqrt{s_X^2 s_Y^2}} = \frac{25.9}{\sqrt{32.8\cdot 26.2}} \doteq 0.883.$$

We can conclude that there is a positive correlation between the height of sons and their fathers.

The sample correlation coefficient can be computed in R using

`cor(height_father,height_son)`.

Estimating the correlation – example



Testing for zero correlation

We want to be able to determine whether the correlation between the variables is **statistically significant**.

Testing for zero correlation

We want to be able to determine whether the correlation between the variables is **statistically significant**.

Theorem

*When observing **independent normally distributed pairs**, then when $\rho_{X,Y} = 0$, the statistic*

$$T = \frac{r_{X,Y}}{\sqrt{1 - r_{X,Y}^2}} \sqrt{n-2}$$

has the Student's t-distribution with $n-2$ degrees of freedom.

Proof

See literature. □

Testing for zero correlation

We want to be able to determine whether the correlation between the variables is **statistically significant**.

Theorem

When observing ***independent normally distributed pairs****, then when $\rho_{X,Y} = 0$, the statistic*

$$T = \frac{r_{X,Y}}{\sqrt{1 - r_{X,Y}^2}} \sqrt{n-2}$$

has the Student's t-distribution with $n-2$ degrees of freedom.

Proof

See literature. □

We can then test the hypothesis $H_0 : \rho_{X,Y} = 0$ and reject it in favor of $H_A : \rho_{X,Y} \neq 0$ on level of significance α if $|T| > t_{\alpha/2,n-2}$, i.e., if the standardised sample correlation coefficient differs significantly from zero.

Testing for zero correlation – example

Example – comparing heights of fathers and sons, continued

Is there a significant correlation between the heights of fathers and their sons? Test on $\alpha = 5\%$.

We obtain

$$T = \frac{r_{X,Y}}{\sqrt{1 - r_{X,Y}^2}} \sqrt{n-2} \doteq \frac{0.883}{\sqrt{1 - 0.883^2}} \sqrt{3} \doteq 3.267.$$

Testing for zero correlation – example

Example – comparing heights of fathers and sons, continued

Is there a significant correlation between the heights of fathers and their sons? Test on $\alpha = 5\%$.

We obtain

$$T = \frac{r_{X,Y}}{\sqrt{1 - r_{X,Y}^2}} \sqrt{n-2} \doteq \frac{0.883}{\sqrt{1 - 0.883^2}} \sqrt{3} \doteq 3.267.$$

The critical value $t_{\alpha/2,n-2} = t_{0.025,3} = 3.182$, thus

$$3.267 = |T| > t_{0.025,3} = 3.182.$$

We reject the null hypothesis that there is no correlation on level of significance 5%.

Testing for zero correlation – example

Example – comparing heights of fathers and sons, continued

Is there a significant correlation between the heights of fathers and their sons? Test on $\alpha = 5\%$.

We obtain

$$T = \frac{r_{X,Y}}{\sqrt{1 - r_{X,Y}^2}} \sqrt{n-2} \doteq \frac{0.883}{\sqrt{1 - 0.883^2}} \sqrt{3} \doteq 3.267.$$

The critical value $t_{\alpha/2,n-2} = t_{0.025,3} = 3.182$, thus

$$3.267 = |T| > t_{0.025,3} = 3.182.$$

We reject the null hypothesis that there is no correlation on level of significance 5%. We say

that there is a *statistically significant* positive correlation between the heights of fathers and the heights of their sons.

Testing for zero correlation – example

Example – comparing heights of fathers and sons, continued

We can test the non-correlation in R using `cor.test`:

```
> cor.test(height_father,height_son)


        Pearson's product-moment correlation

data:  height_father and height_son
t = 3.267, df = 3, p-value = 0.04688
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.00564631 0.99229297
sample estimates:
      cor
0.8835115
```

The p-value is smaller than $\alpha = 5\%$, thus we reject the hypothesis that there is no correlation on level of significance 5%. Alternatively we can decide based on the t-statistic $T = 3.267$.

Linear regression

We are often also interested in observing and evaluating the dependence of a random variable Y on an explanatory variable x, which is not random.

Linear regression

We are often also interested in observing and evaluating the dependence of a random variable Y on an explanatory variable x, which is not random.

Examples

- The number of cars passing a bridge during various times of the day.
- Body height depending on the age of a person.
- Body weight depending on the height of a person.
- The wind speed depending on the altitude.

Linear regression

We are often also interested in observing and evaluating the dependence of a random variable Y on an explanatory variable x, which is not random.

Examples

- The number of cars passing a bridge during various times of the day.
- Body height depending on the age of a person.
- Body weight depending on the height of a person.
- The wind speed depending on the altitude.

Suppose there is a linear dependence of a random variable $Y = Y(x)$ on an explanatory variable x. We measure n independent observations $Y_i = Y(x_i)$ at points $x_1, \ldots, x_n$ and thus we obtain pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$.

Linear regression

We are often also interested in observing and evaluating the dependence of a random variable Y on an explanatory variable x, which is not random.

Examples

- The number of cars passing a bridge during various times of the day.
- Body height depending on the age of a person.
- Body weight depending on the height of a person.
- The wind speed depending on the altitude.

Suppose there is a linear dependence of a random variable $Y = Y(x)$ on an explanatory variable x. We measure n independent observations $Y_i = Y(x_i)$ at points $x_1, \ldots, x_n$ and thus we obtain pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$.

Based on these pairs we want to analyze the linear dependence of $Y = Y(x)$ on x.

Regression model

For the description of the linear dependence we can use the **linear regression model**

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad i = 1, \ldots, n,$$

where:

Regression model

For the description of the linear dependence we can use the **linear regression model**

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad i = 1, \ldots, n,$$

where:

- x_i are given values – not all equal,
- ε_i are i.i.d. zero mean random variables (experimental errors, often $\mathrm{N}(0, \sigma^2)$),
- α and β are unknown parameters.

Regression model

For the description of the linear dependence we can use the **linear regression model**

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad i = 1, \dots, n,$$

where:

- x_i are given values – not all equal,
- ε_i are i.i.d. zero mean random variables (experimental errors, often N$(0, \sigma^2)$),
- α and β are unknown parameters.

It follows that:

$$\mathrm{E}\, Y_i = \alpha + \beta x_i, \qquad \operatorname{var} Y_i = \operatorname{var} \varepsilon_i = \sigma^2.$$

Regression model

For the description of the linear dependence we can use the **linear regression model**

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad i = 1, \dots, n,$$

where:

- x_i are given values – not all equal,
- ε_i are i.i.d. zero mean random variables (experimental errors, often N$(0, \sigma^2)$),
- α and β are unknown parameters.

It follows that:

$$\mathrm{E}\, Y_i = \alpha + \beta x_i, \qquad \operatorname{var} Y_i = \operatorname{var} \varepsilon_i = \sigma^2.$$

We want to find estimators a and b of the parameters α and β such that the values

$$\hat{Y}_i = a + b x_i$$

are the best approximations of Y_i.

Least squares method

Parameters α and β are estimated using the **least squares method**.

Least squares method

Parameters α and β are estimated using the **least squares method**.

Good estimators a and b are such values which minimize the **residual sum of squares** S_e:

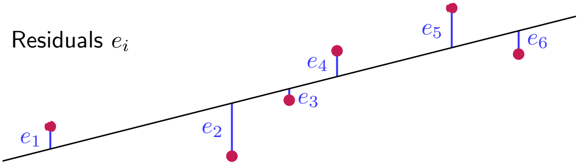
$$S_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (a + bx_i))^2.$$

Least squares method

Parameters α and β are estimated using the **least squares method**.

Good estimators a and b are such values which minimize the **residual sum of squares** S_e:

$$S_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (a + bx_i))^2.$$

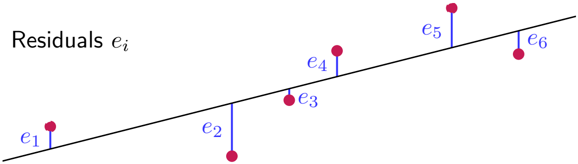


Least squares method

Parameters α and β are estimated using the **least squares method**.

Good estimators a and b are such values which minimize the **residual sum of squares** S_e:

$$S_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (a + bx_i))^2.$$



The estimated regression line $a + bx$ has the minimal sum of the second powers (squares) of the vertical distance from the measured values.

Estimating parameters of the regression line

Theorem

Point estimators of the regression parameters obtained by the least squares method are

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \quad \textit{and} \quad a = \bar{Y}_n - b\,\bar{x}_n,$$

where $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ *and* $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

An unbiased estimator of the variance $\operatorname{var} Y_i = \sigma^2$ *is*

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - a - bx_i)^2 = \frac{1}{n-2}\,S_e$$

and is called the ***residual variance***.

Estimating parameters of the regression line

Proof

We proceed for concrete observations $y_1, \ldots, y_n$:
By differentiating S_e with respect to a and b we find the minimum:

□

Estimating parameters of the regression line

Proof

We proceed for concrete observations $y_1, \ldots, y_n$:
By differentiating S_e with respect to a and b we find the minimum:

$$\frac{\partial S_e}{\partial a} = 0, \qquad \frac{\partial S_e}{\partial b} = 0.$$

□

Estimating parameters of the regression line

Proof

We proceed for concrete observations $y_1, \ldots, y_n$:
By differentiating S_e with respect to a and b we find the minimum:

$$\frac{\partial S_e}{\partial a} = 0, \qquad \frac{\partial S_e}{\partial b} = 0.$$

$$-2\sum_{i=1}^{n}(y_i - a - b \cdot x_i) = 0$$

$$-2\sum_{i=1}^{n}(y_i - a - bx_i)\, x_i = 0$$

□

Estimating parameters of the regression line

Proof

We proceed for concrete observations $y_1, \dots, y_n$:
By differentiating S_e with respect to a and b we find the minimum:

$$\frac{\partial S_e}{\partial a} = 0, \qquad \frac{\partial S_e}{\partial b} = 0.$$

$$-2\sum_{i=1}^{n}(y_i - a - b \cdot x_i) = 0 \quad \rightarrow \quad a = \bar{y}_n - b\,\bar{x}_n$$

$$-2\sum_{i=1}^{n}(y_i - a - bx_i)\,x_i = 0$$

□

Estimating parameters of the regression line

Proof

We proceed for concrete observations $y_1, \ldots, y_n$:
By differentiating S_e with respect to a and b we find the minimum:

$$\frac{\partial S_e}{\partial a} = 0, \qquad \frac{\partial S_e}{\partial b} = 0.$$

$$-2\sum_{i=1}^{n}(y_i - a - b \cdot x_i) = 0 \quad \rightarrow \quad a = \bar{y}_n - b\,\bar{x}_n$$

$$-2\sum_{i=1}^{n}(y_i - a - bx_i)\,x_i = 0 \quad \leftarrow$$

□

Estimating parameters of the regression line

Proof

We proceed for concrete observations $y_1, \ldots, y_n$:
By differentiating S_e with respect to a and b we find the minimum:

$$\frac{\partial S_e}{\partial a} = 0, \qquad \frac{\partial S_e}{\partial b} = 0.$$

$$-2\sum_{i=1}^{n}(y_i - a - b \cdot x_i) = 0 \quad \rightarrow \quad a = \bar{y}_n - b\,\bar{x}_n$$

$$-2\sum_{i=1}^{n}(y_i - a - bx_i)\,x_i = 0 \quad \swarrow$$

$$0 = \sum_{i=1}^{n} x_i y_i - \bar{y}_n \sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 + b\,\bar{x}_n \sum_{i=1}^{n} x_i$$

□

Estimating parameters of the regression line

Proof

We proceed for concrete observations $y_1, \ldots, y_n$:
By differentiating S_e with respect to a and b we find the minimum:

$$\frac{\partial S_e}{\partial a} = 0, \qquad \frac{\partial S_e}{\partial b} = 0.$$

$$-2\sum_{i=1}^{n}(y_i - a - b \cdot x_i) = 0 \quad \rightarrow \quad a = \bar{y}_n - b\,\bar{x}_n$$

$$-2\sum_{i=1}^{n}(y_i - a - bx_i)\,x_i = 0$$

$$0 = \sum_{i=1}^{n} x_i y_i - \bar{y}_n \sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 + b\,\bar{x}_n \sum_{i=1}^{n} x_i$$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{y}_n\bar{x}_n}{\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2}$$

□

Estimating parameters of the regression line

Proof

We proceed for concrete observations $y_1, \ldots, y_n$:
By differentiating S_e with respect to a and b we find the minimum:

$$\frac{\partial S_e}{\partial a} = 0, \qquad \frac{\partial S_e}{\partial b} = 0.$$

$$-2\sum_{i=1}^{n}(y_i - a - b\cdot x_i) = 0 \quad \rightarrow \quad a = \bar{y}_n - b\,\bar{x}_n$$

$$-2\sum_{i=1}^{n}(y_i - a - bx_i)\,x_i = 0$$

$$0 = \sum_{i=1}^{n} x_i y_i - \bar{y}_n \sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 + b\,\bar{x}_n \sum_{i=1}^{n} x_i$$

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

□

Estimating parameters of the regression line

Proof

We proceed for concrete observations $y_1, \dots, y_n$:
By differentiating S_e with respect to a and b we find the minimum:

$$\frac{\partial S_e}{\partial a} = 0, \qquad \frac{\partial S_e}{\partial b} = 0.$$

$$-2\sum_{i=1}^{n}(y_i - a - b \cdot x_i) = 0 \quad \rightarrow \quad a = \bar{y}_n - b\,\bar{x}_n$$

$$-2\sum_{i=1}^{n}(y_i - a - bx_i)\,x_i = 0$$

$$0 = \sum_{i=1}^{n} x_i y_i - \bar{y}_n \sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 + b\,\bar{x}_n \sum_{i=1}^{n} x_i$$

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

By computing the matrix of second derivatives and showing that it is positive definite it can be proven that this point is indeed the minimum. □

Estimating parameters of the regression line

Proof

We proceed for concrete observations $y_1, \dots, y_n$:
By differentiating S_e with respect to a and b we find the minimum:

$$\frac{\partial S_e}{\partial a} = 0, \qquad \frac{\partial S_e}{\partial b} = 0.$$

$$-2\sum_{i=1}^{n}(y_i - a - b\cdot x_i) = 0 \quad \rightarrow \quad a = \bar{y}_n - b\,\bar{x}_n$$

$$-2\sum_{i=1}^{n}(y_i - a - bx_i)\,x_i = 0 \quad \leftarrow$$

$$0 = \sum_{i=1}^{n} x_i y_i - \bar{y}_n \sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 + b\,\bar{x}_n \sum_{i=1}^{n} x_i$$

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

By computing the matrix of second derivatives and showing that it is positive definite it can be proven that this point is indeed the minimum. For the proof of the unbiasedness of the estimator of the variance see literature. □

Estimating parameters of the regression line

✓ It can be shown that the above mentioned estimators are the best unbiased estimators of the regression parameters.

Estimating parameters of the regression line

✓ It can be shown that the above mentioned estimators are the best unbiased estimators of the regression parameters.

If we treated the explanatory variables as random, $X_1, \ldots, X_n$, the estimator of the regression parameter β can be given by means of estimators of variances and the covariance:

$$b = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2} = \frac{s_{X,Y}}{s_X^2} = r_{X,Y}\frac{s_Y}{s_X}\,,$$

Estimating parameters of the regression line

✓ It can be shown that the above mentioned estimators are the best unbiased estimators of the regression parameters.

If we treated the explanatory variables as random, $X_1, \dots, X_n$, the estimator of the regression parameter β can be given by means of estimators of variances and the covariance:

$$b = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2} = \frac{s_{X,Y}}{s_X^2} = r_{X,Y}\frac{s_Y}{s_X},$$

where $s_{X,Y}$ is the sample covariance and $r_{X,Y}$ is the sample correlation coefficient

$$s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n), \qquad r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y}$$

and s_X and s_Y are the sample standard deviations – square roots of sample variances

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \qquad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2.$$

Linear regression – example

Example – dependence of the heights of sons on the heights of their fathers

Suppose we want to model the linear dependence of the heights of sons on the heights of their fathers from the previous example. Their height was measured as follows:

height of father [cm]	x_i	172	176	180	184	186
height of son [cm]	Y_i	178	183	180	188	190

Linear regression – example

Example – dependence of the heights of sons on the heights of their fathers

Suppose we want to model the linear dependence of the heights of sons on the heights of their fathers from the previous example. Their height was measured as follows:

height of father [cm]	x_i	172	176	180	184	186
height of son [cm]	Y_i	178	183	180	188	190

We find the sample variance and covariance as follows:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = 32.8, \qquad s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = 25.9.$$

Linear regression – example

Example – dependence of the heights of sons on the heights of their fathers

Suppose we want to model the linear dependence of the heights of sons on the heights of their fathers from the previous example. Their height was measured as follows:

height of father [cm]	x_i	172	176	180	184	186
height of son [cm]	Y_i	178	183	180	188	190

We find the sample variance and covariance as follows:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = 32.8, \qquad s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = 25.9.$$

The parameters of the regression line are then estimated as

$$b = \frac{s_{X,Y}}{s_X^2} = \frac{25.9}{32.8} \doteq 0.79$$

$$a = \bar{Y}_n - b \cdot \bar{X}_n \doteq 183.8 - \frac{25.9}{32.8} \cdot 179.6 \doteq 41.98.$$

Linear regression – example

Example – dependence of the heights of sons on the heights of their fathers

Suppose we want to model the linear dependence of the heights of sons on the heights of their fathers from the previous example. Their height was measured as follows:

height of father [cm]	x_i	172	176	180	184	186
height of son [cm]	Y_i	178	183	180	188	190

We find the sample variance and covariance as follows:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = 32.8, \qquad s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = 25.9.$$

The parameters of the regression line are then estimated as

$$b = \frac{s_{X,Y}}{s_X^2} = \frac{25.9}{32.8} \doteq 0.79$$

$$a = \bar{Y}_n - b \cdot \bar{X}_n \doteq 183.8 - \frac{25.9}{32.8} \cdot 179.6 \doteq 41.98.$$

For every centimeter of difference between the fathers' height, we expect an average difference of 0.79 centimeters between their sons.

Linear regression – example

Example – dependence of the heights of sons on the heights of their fathers

Suppose we want to model the linear dependence of the heights of sons on the heights of their fathers from the previous example. Their height was measured as follows:

height of father [cm]	x_i	172	176	180	184	186
height of son [cm]	Y_i	178	183	180	188	190

We find the sample variance and covariance as follows:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = 32.8, \qquad s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = 25.9.$$

The parameters of the regression line are then estimated as

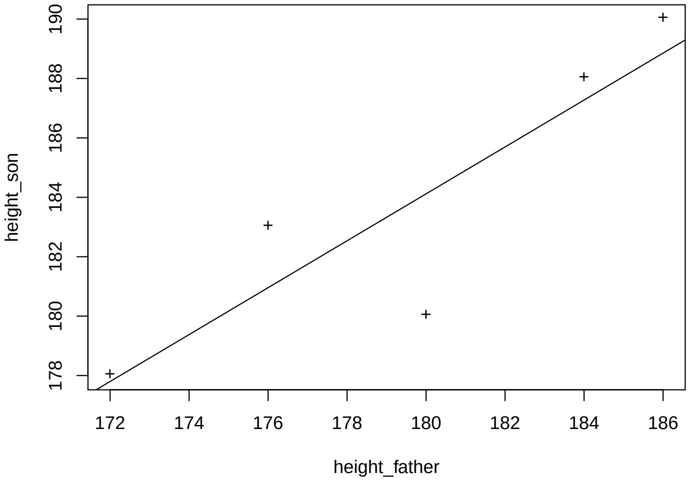
$$b = \frac{s_{X,Y}}{s_X^2} = \frac{25.9}{32.8} \doteq 0.79$$

$$a = \bar{Y}_n - b \cdot \bar{X}_n \doteq 183.8 - \frac{25.9}{32.8} \cdot 179.6 \doteq 41.98.$$

For every centimeter of difference between the fathers' height, we expect an average difference of 0.79 centimeters between their sons.

The estimates can be called in R using `lm(height_son height_father)`.

Linear regression – example



Precision of the regression model

For evaluating the precision of a linear model we can use the **coefficient of determination R^2:**

$$R^2 = 1 - \frac{S_e}{S_T},$$

where S_e is the residual sum of squares and $S_T = (n-1)s_Y^2$:

$$S_T = \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2.$$

Precision of the regression model

For evaluating the precision of a linear model we can use the **coefficient of determination** R^2**:**

$$R^2 = 1 - \frac{S_e}{S_T},$$

where S_e is the residual sum of squares and $S_T = (n-1)s_Y^2$:

$$S_T = \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2.$$

The closer R^2 is to 1 the better the linear model fits the data. More precisely, it can be compared with the critical values of its proper distribution – see literature.

Precision of the regression model

For evaluating the precision of a linear model we can use the **coefficient of determination** R^2:

$$R^2 = 1 - \frac{S_e}{S_T},$$

where S_e is the residual sum of squares and $S_T = (n-1)s_Y^2$:

$$S_T = \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2.$$

The closer R^2 is to 1 the better the linear model fits the data. More precisely, it can be compared with the critical values of its proper distribution – see literature.

R^2 can be interpreted as the proportion of variability in the data which is explained by the regression model.

Testing linear independence

Often we want to test the hypothesis

$$H_0 : \beta = 0 \qquad \text{versus} \qquad H_A : \beta \neq 0.$$

Which equivalently means testing

$$H_0 : Y_i = \alpha + \varepsilon_i \qquad \text{versus} \qquad H_A : Y_i = \alpha + \beta x_i + \varepsilon_i.$$

Testing linear independence

Often we want to test the hypothesis

$$H_0 : \beta = 0 \qquad \text{versus} \qquad H_A : \beta \neq 0.$$

Which equivalently means testing

$$H_0 : Y_i = \alpha + \varepsilon_i \qquad \text{versus} \qquad H_A : Y_i = \alpha + \beta x_i + \varepsilon_i.$$

In fact we test whether Y actually does linearly depend on x or not.

Testing linear independence

Often we want to test the hypothesis

$$H_0 : \beta = 0 \qquad \text{versus} \qquad H_A : \beta \neq 0.$$

Which equivalently means testing

$$H_0 : Y_i = \alpha + \varepsilon_i \qquad \text{versus} \qquad H_A : Y_i = \alpha + \beta x_i + \varepsilon_i.$$

In fact we test whether Y actually does linearly depend on x or not. Testing can be based on the **two-sided confidence interval** for β. When the random errors ε_i are normally distributed, then the corresponding confidence interval can be found as:

$$\left(b - t_{\alpha/2,n-2} \frac{\sqrt{s^2}}{\sqrt{(n-1)s_X^2}} \, , \, b + t_{\alpha/2,n-2} \frac{\sqrt{s^2}}{\sqrt{(n-1)s_X^2}} \right),$$

where s^2 is the **residual variance** from the last theorem and $t_{\alpha/2,n-2}$ is the critical value of the Student's t-distribution with $n-2$ degrees of freedom.

We can then check whether 0 lies in the interval or not. Alternatively we can decide based on the p-value of the test.

Testing linear independence – example

Example – heights of fathers and sons, continued

We want to test whether the heights of sons depend significantly on the heights of their fathers. In R we can call the properties of a fitted linear model using `summary(lm())`:

```
> summary(lm(height_son~height_father))

Call:
lm(formula = height_son ~ height_father)

Residuals:
      1       2       3       4       5
 0.2012  2.0427 -4.1159  0.7256  1.1463

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    41.9817    43.4272   0.967   0.4050
height_father   0.7896     0.2417   3.267   0.0469 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.769 on 3 degrees of freedom
Multiple R-squared:  0.7806,    Adjusted R-squared:  0.7075
F-statistic: 10.67 on 1 and 3 DF,  p-value: 0.04688
```

The p-value corresponding to $H_0 : \beta = 0$ is 0.0469 and is smaller than $\alpha = 5\%$. On level of significance 5% we can thus reject the hypothesis that there is no dependence.

Prediction intervals

Suppose that we have estimated the parameters of the regression model from obtained data. For a new value x for which we do not know the value Y we may be interested in a **prediction** of Y and the **confidence interval** for the prediction.

Prediction intervals

Suppose that we have estimated the parameters of the regression model from obtained data. For a new value x for which we do not know the value Y we may be interested in a **prediction** of Y and the **confidence interval** for the prediction.

Prediction $\hat{Y}$:

$$\hat{Y} = a + b \cdot x.$$

Prediction intervals

Suppose that we have estimated the parameters of the regression model from obtained data. For a new value x for which we do not know the value Y we may be interested in a **prediction** of Y and the **confidence interval** for the prediction.

Prediction $\hat{Y}$:

$$\hat{Y} = a + b \cdot x.$$

$(1-\alpha)100\%$ confidence interval for the prediction

$$a + b \cdot x \pm t_{\alpha/2,n-2}\sqrt{s^2}\sqrt{\frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}.$$

Prediction intervals

Suppose that we have estimated the parameters of the regression model from obtained data. For a new value x for which we do not know the value Y we may be interested in a **prediction** of Y and the **confidence interval** for the prediction.

Prediction $\hat{Y}$:

$$\hat{Y} = a + b \cdot x.$$

$(1-\alpha)100\%$ confidence interval for the prediction

$$a + b \cdot x \pm t_{\alpha/2,n-2}\sqrt{s^2}\sqrt{\frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}.$$

If we plot the regression line and the boundaries of the confidence interval of the prediction as a function of x, we obtain the **pointwise confidence intervals**.

Prediction intervals

Suppose that we have estimated the parameters of the regression model from obtained data. For a new value x for which we do not know the value Y we may be interested in a **prediction** of Y and the **confidence interval** for the prediction.

Prediction $\hat{Y}$:

$$\hat{Y} = a + b \cdot x.$$

$(1-\alpha)100\%$ confidence interval for the prediction

$$a + b \cdot x \pm t_{\alpha/2,n-2}\sqrt{s^2}\sqrt{\frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}.$$

If we plot the regression line and the boundaries of the confidence interval of the prediction as a function of x, we obtain the **pointwise confidence intervals**.

We can also construct a band in which the regression line lies with a probability $1-\alpha$. Such band is called the **confidence band for the whole regression line**.

Prediction intervals

Suppose that we have estimated the parameters of the regression model from obtained data. For a new value x for which we do not know the value Y we may be interested in a **prediction** of Y and the **confidence interval** for the prediction.

Prediction $\hat{Y}$:

$$\hat{Y} = a + b \cdot x.$$

$(1-\alpha)100\%$ confidence interval for the prediction

$$a + b \cdot x \pm t_{\alpha/2,n-2}\sqrt{s^2}\sqrt{\frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}}.$$

If we plot the regression line and the boundaries of the confidence interval of the prediction as a function of x, we obtain the **pointwise confidence intervals**.

We can also construct a band in which the regression line lies with a probability $1-\alpha$. Such band is called the **confidence band for the whole regression line**. The corresponding expression is based on the Fisher's F-distribution (see literature), with $t_{\alpha/2,n-2}$ replaced with $\sqrt{2F_{\alpha/2,2,n-2}}$.

Regression prediction – example

Example – dependence of the heights of sons on the heights of their fathers

Suppose we want to estimate the expected height of a son whose father is 175 centimeters tall.

Regression prediction – example

Example – dependence of the heights of sons on the heights of their fathers

Suppose we want to estimate the expected height of a son whose father is 175 centimeters tall.

For given $x = 175$ cm, we want to predict $\hat{Y}$:

$$\begin{aligned} \hat{Y} &= a + b \cdot x \\ &\doteq 41.98 + 0.79 \cdot 175 \\ &\doteq 180.2 \text{ cm.} \end{aligned}$$

Regression prediction – example

Example – dependence of the heights of sons on the heights of their fathers

Suppose we want to estimate the expected height of a son whose father is 175 centimeters tall.

For given $x = 175$ cm, we want to predict $\hat{Y}$:

$$\begin{aligned}\hat{Y} &= a + b \cdot x \\ &\doteq 41.98 + 0.79 \cdot 175 \\ &\doteq 180.2 \text{ cm.}\end{aligned}$$

The 95% confidence interval for the prediction is then

$$(174.9, 185.5).$$

Example – concentration of lactic acid

It was studied how much lactic acid there is in 100 ml of new mothers' blood (values x_i) and their newborn children (values Y_i) directly after birth.

x_i	40	64	34	15	57	45
Y_i	33	46	23	12	56	40

We consider a linear dependence between the concentration in mothers' and their children's blood.

Example – concentration of lactic acid

It was studied how much lactic acid there is in 100 ml of new mothers' blood (values x_i) and their newborn children (values Y_i) directly after birth.

x_i	40	64	34	15	57	45
Y_i	33	46	23	12	56	40

We consider a linear dependence between the concentration in mothers' and their children's blood. The estimates of the regression parameters are:

$$b = \frac{\sum_{i=1}^{6}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{6}(x_i - \bar{x}_n)^2} = 0.8543$$

$$a = \bar{Y}_n - b\bar{x}_n = -1.3082$$

Example – concentration of lactic acid

It was studied how much lactic acid there is in 100 ml of new mothers' blood (values x_i) and their newborn children (values Y_i) directly after birth.

x_i	40	64	34	15	57	45
Y_i	33	46	23	12	56	40

We consider a linear dependence between the concentration in mothers' and their children's blood. The estimates of the regression parameters are:

$$b = \frac{\sum_{i=1}^{6}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{6}(x_i - \bar{x}_n)^2} = 0.8543$$

$$a = \bar{Y}_n - b\bar{x}_n = -1.3082$$

Let us test the hypothesis that the concentration in mother's blood does not influence the concentration in their children's blood: $H_0 : \beta = 0$ versus $H_A : \beta \neq 0$

Example – concentration of lactic acid

It was studied how much lactic acid there is in 100 ml of new mothers' blood (values x_i) and their newborn children (values Y_i) directly after birth.

x_i	40	64	34	15	57	45
Y_i	33	46	23	12	56	40

We consider a linear dependence between the concentration in mothers' and their children's blood. The estimates of the regression parameters are:

$$b = \frac{\sum_{i=1}^{6}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{6}(x_i - \bar{x}_n)^2} = 0.8543$$

$$a = \bar{Y}_n - b\bar{x}_n = -1.3082$$

Let us test the hypothesis that the concentration in mother's blood does not influence the concentration in their children's blood: $H_0 : \beta = 0$ versus $H_A : \beta \neq 0$
The 95% confidence interval for β is

$$(0.404, 1.305).$$

Example – concentration of lactic acid

It was studied how much lactic acid there is in 100 ml of new mothers' blood (values x_i) and their newborn children (values Y_i) directly after birth.

x_i	40	64	34	15	57	45
Y_i	33	46	23	12	56	40

We consider a linear dependence between the concentration in mothers' and their children's blood. The estimates of the regression parameters are:

$$b = \frac{\sum_{i=1}^{6}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{6}(x_i - \bar{x}_n)^2} = 0.8543$$

$$a = \bar{Y}_n - b\bar{x}_n = -1.3082$$

Let us test the hypothesis that the concentration in mother's blood does not influence the concentration in their children's blood: $H_0 : \beta = 0$ versus $H_A : \beta \neq 0$

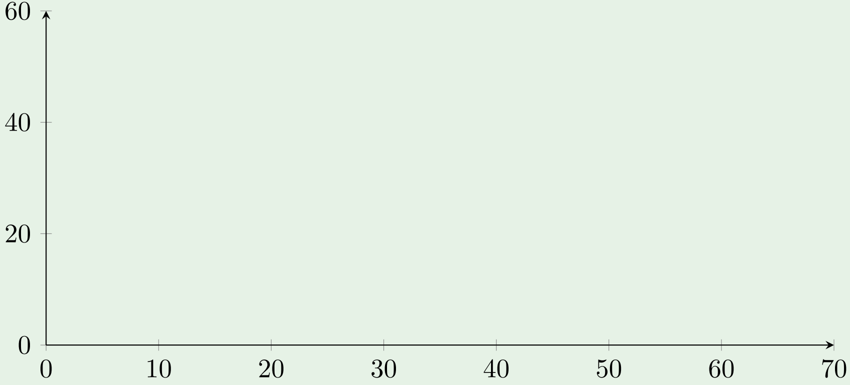
The 95% confidence interval for β is

$$0 \notin (0.404, 1.305).$$

This means that we reject the null hypothesis. The dependence is thus significant.

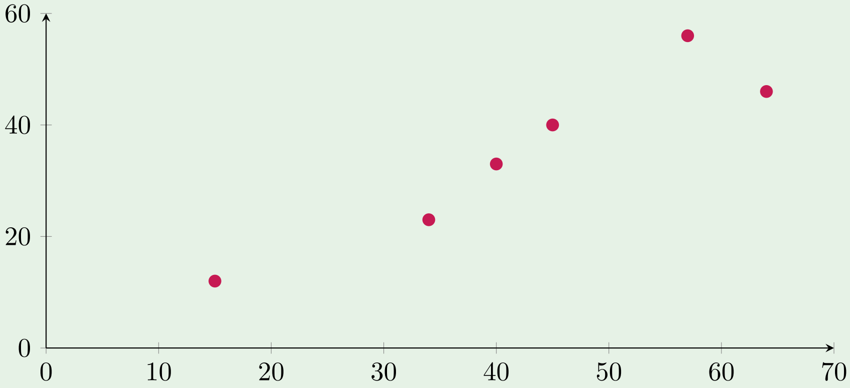
Example – concentration of lactic acid, continued

Let us plot the measured data, the estimated regression line and corresponding confidence bands:



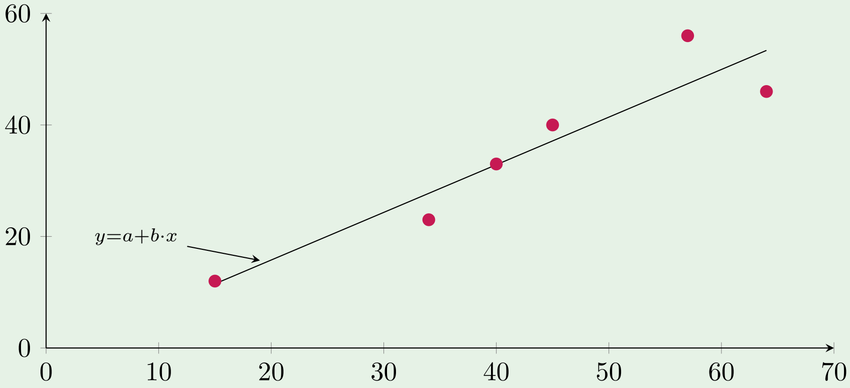
Example – concentration of lactic acid, continued

Let us plot the measured data, the estimated regression line and corresponding confidence bands:



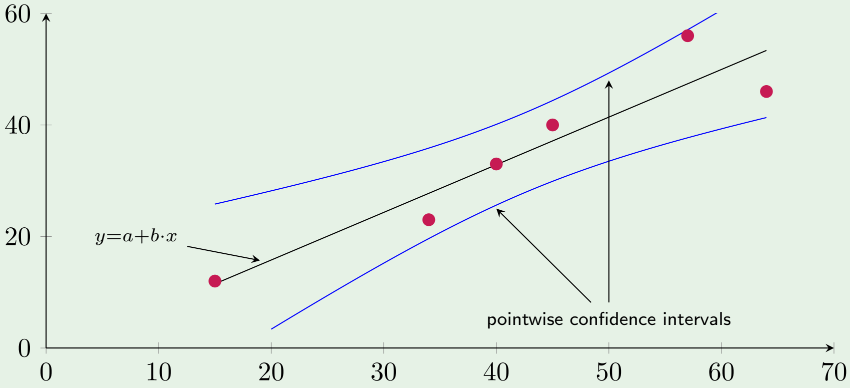
Example – concentration of lactic acid, continued

Let us plot the measured data, the estimated regression line and corresponding confidence bands:



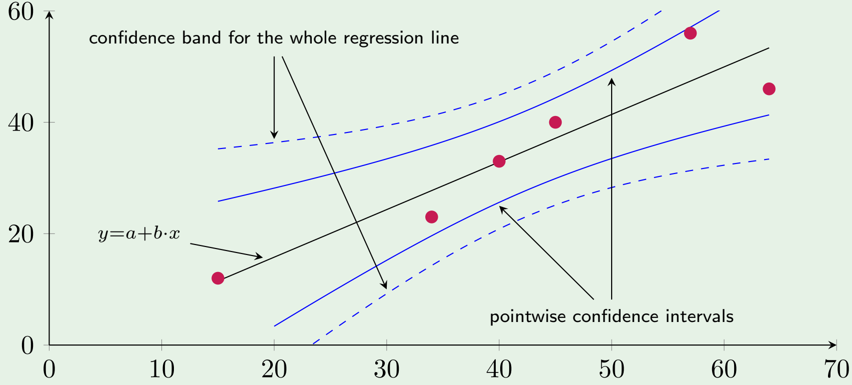
Example – concentration of lactic acid, continued

Let us plot the measured data, the estimated regression line and corresponding confidence bands:



Example – concentration of lactic acid, continued

Let us plot the measured data, the estimated regression line and corresponding confidence bands:



Recap

The **correlation coefficient** gives a measure of linear dependence between two random variables and is defined as

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var} X}\sqrt{\operatorname{var} Y}}.$$

It can be estimated using the **sample correlation coefficient** as

$$r_{X,Y} = \frac{s_{X,Y}}{s_X \cdot s_Y},$$

where $s_{X,Y}$ is the **sample covariance**.

If we want to model the dependence of Y on x taken as fixed, we can use **linear regression**. We assume that there is a linear dependence of the form

$$Y_i = \alpha + \beta x_i + \varepsilon_i,$$

where ε_i are independent zero-mean random errors and α and β are parameters which we want to estimate.

Given observed data, we obtain the estimators a and b of the parameters using the **least squares method** as:

$$b = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}_n\bar{y}_n}{\sum_{i=1}^n x_i^2 - n\bar{x}_n^2} = \frac{s_{X,Y}}{s_X^2} = r_{X,Y}\frac{s_Y}{s_X},$$

$$a = \bar{y}_n - b \cdot \bar{x}_n.$$