Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Eleventh lecture, January 6, 2025

Tests and confidence intervals for 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 presented 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 constants and the expected value of the residuals ε_i is $\mathbb{E}[\varepsilon_i] = 0$.

Linear model, assumptions for parametric tests and confidence intervals

We now consider testing the parameters of a linear regression model and calculating confidence intervals for the estimated parameters under classical assumptions.

- 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 $\mathbb{E}[\varepsilon_i] = 0$.
- The residuals have the same variance $\mathbb{E}[\varepsilon_i^2] = \sigma^2$.
- The residuals are uncorrelated, i.e., $\rho(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$.
- The residuals are normally distributed.

Slope of the regression line

Testing the slope of the regression line

The null hypothesis:

$$H_0 \colon b_1 = b_1^0$$

(typically null hypothesis $b_1 = 0$ is tested).

Possible alternative hypotheses:

$$H_1 \colon b_1 > b_1^0 \text{ (one tailed)},$$
$$H_1 \colon b_1 < b_1^0 \text{ (one tailed)},$$
$$H_1 \colon b_1 \neq b_1^0 \text{ (two tailed)}.$$

Testing the slope of the regression line

- t-test statistic
$$t = \frac{\hat{b}_1 - b_1^0}{s/(s_x\sqrt{n-1})},$$
where $s^2 = \mathrm{Var}(\hat{\varepsilon}) = \frac{1}{n-2}\sum_{i=1}^{n}(\hat{\varepsilon}_i)^2$ (see previous lecture) and s_x^2 is the sample variance of the variable x.
- Under the null hypothesis H_0, the test statistic follows Student's t-distribution with $n-2$ degrees of freedom.
- Under the null hypothesis H_0, the expected value of the test statistic is $\mathbb{E}[t] = 0$.
- 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.

Slope of the regression line, confidence interval

Under the normality assumption on the residuals, the $(1-\alpha)\cdot 100\%$ confidence interval for the slope of the regression line can be given as

$$\left(\hat{b}_1 - t_{n-2,\alpha/2}\frac{s}{s_x\sqrt{n-1}}, \hat{b}_1 + t_{n-2,\alpha/2}\frac{s}{s_x\sqrt{n-1}}\right),$$

where $s^2 = \mathrm{Var}(\hat{\varepsilon})$, s_x^2 is the sample variance of the variable x, t_{n-2} is Student's t distribution with $n-2$ degrees of freedom, and $t_{n-2,\alpha/2}$ is the $(1-\alpha/2)\cdot 100$ percentile of the $t(n-2)$ distribution.

Intercept/constant term

Testing the constant term of the regression line

The null hypothesis:

$$H_0\colon b_0 = b_0^0.$$

Possible alternative hypotheses:

$$H_1\colon b_0 > b_0^0 \text{ (one tailed)},$$
$$H_1\colon b_0 < b_0^0 \text{ (one tailed)},$$
$$H_1\colon b_0 \neq b_0^0 \text{ (two tailed)}.$$

Testing the constant term of the regression line

- t-test statistic
$$t = \frac{\hat{b}_0 - b_0^0}{\frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{s_x\sqrt{n(n-1)}}},$$
where $s^2 = \mathrm{Var}(\hat{\varepsilon}) = \frac{1}{n-2}\sum_{i=1}^{n}(\hat{\varepsilon}_i)^2$ and s_x^2 is the sample variance of the variable x.
- Under the null hypothesis H_0, the test statistic follows Student's t-distribution with $n-2$ degrees of freedom.
- Under the null hypothesis H_0, the expected value of the test statistic is $\mathbb{E}[t] = 0$.
- 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.

Intercept, confidence interval

Under normality assumption, $(1-\alpha)\cdot 100\%$ confidence interval for the constant term of the regression line can be given as

$$\left(\hat{b}_0 - t_{n-2,\alpha/2}\frac{s\sqrt{\sum_{i=1}^n x_i^2}}{s_x\sqrt{n(n-1)}}, \hat{b}_0 + t_{n-2,\alpha/2}\frac{s\sqrt{\sum_{i=1}^n x_i^2}}{s_x\sqrt{n(n-1)}}\right),$$

where $s^2 = \mathrm{Var}(\hat{\varepsilon})$, s_x^2 is the sample variance of the variable x, t_{n-2} is Student's t-distribution with $n-2$ degrees of freedom, and $t_{n-2,\alpha/2}$ is the $(1-\alpha/2)\cdot 100$ percentile of the $t(n-2)$ distribution.

Predicting

Predicting the values of variable y

A prediction $\tilde{y}$ for the value of the variable y, when x has value $\tilde{x}$, can be given as

$$\tilde{y}|\tilde{x} = \hat{b}_0 + \hat{b}_1\tilde{x}.$$

The more there are observations, the smaller the variance σ^2 is, and the closer $\tilde{x}$ is to the sample mean of x, then the better (more accurate) the prediction is. Note that $\tilde{x}$ should be on the range of the observed values of the variable x.

Predicting the values of variable y

Under normality assumption, a $(1-\alpha)\cdot 100\%$ confidence interval for the value of y, when x has value $\tilde{x}$, can be given as

$$\hat{b}_0 + \hat{b}_1\tilde{x} \pm t_{n-2,\alpha/2}\, s\sqrt{1 + \frac{1}{n} + \frac{(\tilde{x}-\overline{x})^2}{(n-1)s_x^2}},$$

where $s^2 = \mathrm{Var}(\hat{\varepsilon})$, s_x^2 is the sample variance of the variable x, t_{n-2} is Student's t-distribution with $n-2$ degrees of freedom, and $t_{n-2,\alpha/2}$ is the $(1-\alpha/2)\cdot 100$ percentile of the $t(n-2)$ distribution.

Predicting the expected value of variable y

A prediction $\hat{\mu}_y$ for the expected value $\mathbb{E}[y]$, when x has value $\tilde{x}$, can be given as

$$\hat{\mu}_y|\tilde{x} = \hat{b}_0 + \hat{b}_1\tilde{x}.$$

Remarks:

- Note that $\tilde{y}|\tilde{x}$ estimates the value of a random variable while $\hat{\mu}_y|\tilde{x}$ estimates the expected value (constant). The estimate $\tilde{y}|\tilde{x}$ estimates the values of the variable on an individual level when x has value $\tilde{x}$, while the estimate $\hat{\mu}_y|\tilde{x}$ estimates the mean value of the variable y when x has value $\tilde{x}$.
- Even though the estimates are the same, the corresponding confidence intervals are not! The confidence interval for the value of y is wider. It is easier to predict average behavior than to predict individual values.

Predicting the expected value of variable y

Under normality assumption, a $(1-\alpha)\cdot 100\%$ confidence interval for $\mathbb{E}[y]$, when x has value $\tilde{x}$, can be given as

$$\hat{b}_0 + \hat{b}_1\tilde{x} \pm t_{n-2,\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{(\tilde{x}-\overline{x})^2}{(n-1)s_x^2}},$$

where $s^2 = \mathrm{Var}(\hat{\varepsilon})$, s_x^2 is the sample variance of the variable x, t_{n-2} is Student's t-distribution with $n-2$ degrees of freedom, and $t_{n-2,\alpha/2}$ is the $(1-\alpha/2)\cdot 100$ percentile of the $t(n-2)$ distribution.

Numerical example

Last lecture, we obtained the regression model

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x, \quad \hat{b}_0 = 10723.87 \text{ and } \hat{b}_1 = -0.9386$$

for the cookie sales of Brand Y (dependent variable) with respect to the cookie sales of Brand X (explanatory variable). We wish to derive the 95% confidence interval for the sales when 5500 units of Brand X cookies are sold.

On the condition that $\hat{c} = 5500$ units of Brand X cookies are sold, the prediction of the sales of Brand Y cookies is

$$\tilde{j}|\tilde{c} = \hat{b}_0 + \hat{b}_1 \tilde{c} = 10723.87 - 0.9386 \cdot 5500 = 5561.57.$$

The corresponding confidence interval can be given as

$$\hat{b}_0 + \hat{b}_1 \tilde{c} \pm t_{n-2,\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(\tilde{c}-\overline{c})^2}{(n-1)s_c^2}} = 5561.57 \pm 257.974,$$

where we plugged in the values $t_{n-2,\alpha/2} = t_{10,0.025} = 2.228$, $\overline{c} = 5567.833$, $s_c = 302.95$, and $s^2 = 11948.42$.

$\therefore$ If 5500 units of Brand X cookies are sold, then the prediction for the sales of Brand Y cookies is 5562 units. A 95% confidence interval for the prediction is (5308,5816).

Bootstrap confidence intervals

Bootstrap confidence intervals for the regression coefficients

Consider the estimated residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ and the fitted values $\hat{y}_1, \ldots, \hat{y}_n$ of the regression model. Collect a new sample $\check{\varepsilon}_1, \ldots, \check{\varepsilon}_n$ by picking n data points randomly with replacement from $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$. Form a bootstrap sample

$$(x_1, \check{y}_1), \ldots, (x_n, \check{y}_n),$$

where

$$\check{y}_i = \hat{y}_i + \check{\varepsilon}_i.$$

Calculate estimates for the regression coefficients b_0 and b_1 from the bootstrap sample. Repeat this several times, for example 999 times. Order now all the estimates (the original ones and the 999 bootstrap estimates) from the smallest to the largest. Now an estimate for the 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.

Prediction, bootstrap confidence intervals

A prediction $\hat{\mu}_y$ for the expected value $\mathbb{E}[y]$, when x has value $\tilde{x}$, was given as

$$\hat{\mu}_y|\tilde{x} = \hat{b}_0 + \hat{b}_1\tilde{x}.$$

Consider bootstrap estimates for the regression coefficients b_0 and b_1. One can calculate bootstrap confidence intervals for $\hat{\mu}_y|\tilde{x}$ by replacing $\hat{b}_0$ and $\hat{b}_1$ by bootstrap estimates in the formula above. That is then repeated, for example, 999 times. After that, all the 1000 predictions are ordered and bootstrap confidence intervals are obtained.

Coefficient of determination, bootstrap confidence intervals

Bootstrap samples

$$(x_1, \check{y}_1), \ldots, (x_n, \check{y}_n)$$

can be used also for calculating bootstrap confidence intervals for the coefficient of determination of the model. Coefficient of determination is estimated (separately) from every bootstrap sample. One can use, for example, 999 bootstrap samples. After that, all the 1000 estimates are ordered and bootstrap confidence intervals are obtained.

Bootstrap confidence intervals, alternative approach

Instead of bootstrapping from the estimated residuals, one may take bootstrap samples directly from the original observations $(x_1, y_1), \ldots, (x_n, y_n)$. Parameter estimates are then calculated from the bootstrap samples, the estimates are ordered and bootstrap confidence intervals are obtained.

Multivariate linear regression

Multiple linear model

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 a p-dimensional x. (Here, x_i is a p-vector.) Assume that $p < n$ and that the values of the variable y depend linearly on the values of the variable x.

A multiple linear model can be presented in the following way

$$y_i = b_0 + b_1(x_i)_1 + b_2(x_i)_2 + \cdots + b_p(x_i)_p + \varepsilon_i, \quad i \in \{1, \ldots, n\}, \tag{1}$$

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

The model (1) can also be expressed in vectorized form as

$$y_i = b_0 + b^{\mathrm{T}} x_i + \varepsilon_i, \quad i \in \{1, \ldots, n\},$$

where $b = [b_1, \ldots, b_p]^{\mathrm{T}}$ and $x_i = [(x_i)_1, \ldots, (x_i)_p]^{\mathrm{T}}$.

Linear model, general assumptions

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

- The measurement of the values x_i is error-free.
- The values $(x_i)_s$, $(x_i)_k$, $s \neq k$, are mutually independent.
- 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 $\mathbb{E}[\varepsilon_i] = 0$.
- The residuals have the same variance $\mathbb{E}[\varepsilon_i^2] = \sigma^2$, $i = 1, \dots, n$.
- The residuals are uncorrelated, i.e., $\rho(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$.

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

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

Multiple linear regression

Multiple linear regression

The multiple linear model

$$y_i = b_0 + b^{\mathrm{T}} x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

has the following parameters: regression coefficients b_0 and $b = (b_1, \ldots, b_p)^{\mathrm{T}}$ 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 $i = 1, \ldots, n$, the linear model can be given as

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

where $\mathbb{E}[y_i] = b_0 + b^{\mathrm{T}} x_i$ is the systematic part and ε_i is the random part of the model.

Regression plane

The systematic part of the linear model

$$\mathbb{E}[y_i] = b_0 + b^{\mathrm{T}} x_i$$

defines the regression plane

$$y = b_0 + b^{\mathrm{T}} x.$$

The variance of the residuals $\mathbb{E}[\varepsilon_i^2] = \sigma^2$ describes the deviation of the observed points from the regression plane.

The aim in multiple linear regression analysis is to find estimates for the regression coefficients b_0 and $b = (b_1, \ldots, b_p)^{\mathrm{T}}$. The estimates should be such that the estimated regression plane would explain the variation of the values of the dependent variable with great accuracy.

Least squares method

Let $\beta = (b_0, b_1, \ldots, b_p)^{\mathrm{T}}$. Let X be an $n \times (p+1)$ data matrix, where the elements of the first column are all equal to 1 and where the columns $2, \ldots, p+1$ are the observations x_i. Let $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$ be an $n \times 1$ data vector.

The least squares estimates for b_0 and $b = (b_1, \ldots, b_p)^{\mathrm{T}}$ are given by

$$\hat{\beta} = (\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_p)^{\mathrm{T}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y.$$

These estimates minimize the sum of the squared differences

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

Remark. We assumed above that the matrix $X^{\mathrm{T}}X$ is non-singular. If $X^{\mathrm{T}}X$ is singular, then some of the explanatory variables must be fully linearly dependent. In that case, some of the variables can be excluded from the analysis without losing any information.

Fits and residuals

The least squares estimates now give an estimated regression plane

$$\hat{y} = \hat{b}_0 + \hat{b}^{\mathrm{T}} x.$$

The fitted values of the variable y_i, i.e., the values given to the variable y by the regression plane at point x_i, are

$$\hat{y}_i = \hat{b}_0 + \hat{b}^{\mathrm{T}} x_i, \quad i = 1, \ldots, n.$$

The residuals $\hat{\varepsilon}_i$ of the estimated model are the differences

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

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

The regression model explains the observed values of the dependent variable 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.

Residual mean square estimation

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

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

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

Sums of squares

The total sum of squares (SST)

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

measures total variation of the observed values y_i. The error sum of squares (SSE)

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

measures the variation of the residuals $\hat{\varepsilon}_i$. The model sum of squares (SSM)

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

measures the part of the variation of the dependent variable y that is explained by the regression model.

Coefficient of determination

The coefficient of determination

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

measures the proportion of SST explained by the model.

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

Numerical example

The effect of nonpareils and chocolate chops on the mass of cookies is examined in a lab.

Nonpareil	Chocolate chip	Mass
15	5	24
13	7	28
12	9	26
11	7	27
10	10	29
9	12	31
17	2	19
16	4	21
12	8	25
3	15	36

Table: The number of nonpareils and chocolate chips, as well as the measured masses of a sample of cookies.

The least squares estimates for the regression coefficients $(b_0, b_1, b_2)^{\mathrm{T}}$ can be calculated using

$$X = \begin{bmatrix} 1 & 15 & 5 \\ 1 & 13 & 7 \\ 1 & 12 & 9 \\ 1 & 11 & 7 \\ 1 & 10 & 10 \\ 1 & 9 & 12 \\ 1 & 17 & 2 \\ 1 & 16 & 4 \\ 1 & 12 & 8 \\ 1 & 3 & 15 \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} 24 \\ 28 \\ 26 \\ 27 \\ 29 \\ 31 \\ 19 \\ 21 \\ 25 \\ 36 \end{bmatrix}.$$

The estimates are

$$(\hat{b}_0, \hat{b}_1, \hat{b}_2)^{\mathrm{T}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y = (29.9718, -0.6562, 0.5533)^{\mathrm{T}}.$$

Now one obtains the fits $\hat{y}_i = \hat{b}_0 + \hat{b}^{\mathrm{T}}x_i$ for the mass and can calculate the residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$.

Nonpareil	Chocolate chip	Mass	Fit	Residual
15	5	24	22.8953	1.1047
13	7	28	25.3143	2.6857
12	9	26	27.0771	-1.0771
11	7	27	26.6267	0.3733
10	10	29	28.9428	0.0572
9	12	31	30.7056	0.2944
17	2	19	19.9230	-0.9230
16	4	21	21.6858	-0.6858
12	8	25	26.5238	-1.5238
3	15	36	36.3027	-0.3027

Table: The effect of nonpareils and chocolate chips on the mass. Also the fitted values and residuals are tabulated.

The sample mean of the mass $\overline{y} = 26.6$ and the total sum of squares

$$SST = \sum_{i=1}^{n}(y_i - \overline{y}) = \sum_{i=1}^{10}(y_i - 26.6)^2 = 214.4.$$

The error sum of squares

$$SSE = \sum_{i=1}^{10}(\hat{\varepsilon}_i)^2 = 13.5586$$

and the model sum of squares

$$SSM = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = \sum_{i=1}^{10}(\hat{y}_i - 26.6)^2 = 200.8307.$$

Thus, the coefficient of determination is

$$R^2 = \frac{SSM}{SST} = \frac{200.8307}{214.4} = 0.9367 = 93.67\%.$$

Multivariate linear regression

Multivariate linear model

Consider n observations (pairs) $(x_1, y_1), \ldots, (x_n, y_n)$ of (x, y). Assume that the values y_i are the observed values of a q-variate random vector y and assume that the values x_i are observed non-random values of a p-variate x. Assume that $p < n$ and that the values of the variable y depend linearly on the variable x.

A multivariate linear model can be given as

$$y_i = b_0 + B^{\mathrm{T}} x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the elements of a $q \times 1$ vector b_0 and $p \times q$ regression matrix B are unknown constants and the expected value of the residuals ε_i is $\mathbb{E}[\varepsilon_i] = 0$.

Linear model, general assumptions

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

- The measurement of the values x_i is error-free.
- The values $(x_i)_s, (x_i)_k$, $s \neq k$, are mutually independent.
- 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 $\mathbb{E}[\varepsilon_i] = 0$, $i = 1, \ldots, n$.
- The residuals have the same covariance matrix $\mathbb{E}[\varepsilon_i \varepsilon_i^{\mathrm{T}}] = \Sigma$, $i = 1, \ldots, n$.
- The residuals are uncorrelated, i.e., $\rho((\varepsilon_i)_k, (\varepsilon_j)_k) = 0$ for all k and for all $i \neq j$.

Generalized least squares

Let $\beta = (b_0, b_1, \ldots, b_p)^{\mathrm{T}}$. Let X be an $n \times (p+1)$ data matrix, where the elements of the first column are all equal to 1 and where the columns $2, \ldots, p+1$ are the observations x_i. Let Y be an $n \times q$ data matrix, where the columns are the observations y_i.

Now the regression parameters b_0 and B can be estimated using

$$\hat{\beta} = [\hat{b}_0, \hat{B}^{\mathrm{T}}]^{\mathrm{T}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y.$$

Fits and residuals

The fitted values of the variable y_i, i.e., the values given to the variable y by the regression model at points x_i, are

$$\hat{y}_i = \hat{b}_0 + \hat{B}^{\mathrm{T}} x_i, \quad i = 1, \ldots, n.$$

The fits can also be expressed as a matrix

$$\hat{Y} = X\hat{\beta}.$$

The residuals $\hat{\varepsilon}_i$ of the estimated model are the differences

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

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

Trace correlation and determinant correlation

Assume that the matrix Y is centered so that the columns of Y have zero mean. (That is, the sample mean is subtracted from the original observations.) Let X be as above, and let $\hat{\beta}$ be calculated for the centered data. Let

$$\hat{Y} = X\hat{\beta},$$

$$\hat{E} = Y - X\hat{\beta}$$

and let

$$D = (Y^{\mathrm{T}}Y)^{-1}\hat{E}^{\mathrm{T}}\hat{E}.$$

It is straightforward to see that the matrix $\hat{E}^{\mathrm{T}}\hat{E}$ ranges between zero, when all the variation of Y is explained by the regression model, and $Y^{\mathrm{T}}Y$, when no part of the variation in Y is explained by X. Therefore $I - D$ varies between the identity matrix and the zero matrix. It can be shown that all the eigenvalues of $I - D$ lie between 1 and 0.

Trace correlation and determinant correlation

It would be desirable that a multivariate coefficient of determination would range between zero and one. This is obtained by either using trace correlation r_T or determinant correlation r_D:

$$r_T^2 = \frac{1}{p}\mathrm{tr}(I - D),$$

$$r_D^2 = \det(I - D).$$

Note that the coefficient r_D is zero if and only if at least one of the eigenvalues of $I - D$ is zero, while r_T is zero if and only if all the eigenvalues of $I - D$ are zero.

It is possible to construct parametric tests and confidence intervals for the parameters in multiple and multivariate regression analysis. Alternatively, one can consider bootstrapping.

Selecting variables

Selecting variables

In multiple and multivariate regression analysis, the explanatory variables are usually assumed to be independent. Perfect independence is rarely achieved if more than one explanatory variables are used. Still, the explanatory variables may not be highly correlated. Multicollinearity makes the model unstable and complicates assessing the effects of different explanatory variables separately.

Variance inflation factor

The variance inflation factor (VIF) can be used to measure the multicollinearity of the explanatory variables. The VIF for the explanatory variable $(x_i)_k$ is defined as

$$VIF_k = \frac{1}{1 - R_k^2},$$

where R_k^2 is the coefficient of determination for a model where $(x_i)_k$ is the dependent variable and the rest of $(x_i)_s$ are explanatory variables. VIF is calculated separately for each explanatory variable $(x_i)_k$. If the variable $(x_i)_k$ is independent from the other explanatory variables, then $VIF = 1$. On the other hand, $VIF \geq 10$ suggests that multicollinearity is present.

In multiple and multivariate regression models the aim is to select variables such that the coefficient of determination is as high as possible and the explanatory variables are as independent as possible. VIF (or some other measure of dependence) can be used in selecting the variables. Variables can be added and removed one by one and the changes in VIF and coefficient of determination can be tracked.

Cookie example continues

In this example VIF is used to assess multicollinearity of nonpareils and chocolate chips.

Nonpareil	Chocolate chip
15	5
13	7
12	9
11	7
10	10
9	12
17	2
16	4
12	8
3	15

Table: Cookie data, number of nonpareils and chocolate chips.

The sample standard deviation for nonpareil $s_x = 4.022161$ and chocolate chips $s_y = 3.842742$, the sample means $\overline{x} = 11.8$ and $\overline{y} = 7.9$, and the sample correlation coefficient $\hat{\rho}(x, y) = -0.9647379$ are needed. Fit

$$\hat{y}_i = \overline{y} + \hat{\rho}(x, y)\frac{s_y}{s_x}(x_i - \overline{x}) = 7.8 + (-0.9647379)\frac{3.842742}{4.022161}(x_i - 11.8).$$

Total sum of squares $SST = 113$, error sum of squares $SSE = 9.307418$, and model sum of squares $SSM = 123.6926$. Coefficient of determination

$$R^2 = \frac{SSM}{SST} = \frac{123.6926}{133} = 0.9300195$$

and

$$VIF = \frac{1}{1 - R^2} = \frac{1}{1 - 0.930\ldots} = 14.28969.$$

Words of warning

- Regression models should not be used to predict any values outside 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 outlines (i.e., it is non-robust).

Parameter identification for non-linear models
and the maximum likelihood estimator

Linear regression is a prototypical example of a parameter identification problem. In addition to linear models, one may also be interested in parameter identification for other types of models.

Example

Given i.i.d. normally distributed data $y_1, \ldots, y_n \sim \mathcal{N}(\mu, \sigma^2)$, estimate μ and σ^2.

Example

Given data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$, find parameters $a, b, c \in \mathbb{R}$ such that

$$y_i = ax_i^2 + bx_i + c + \varepsilon_i, \quad i \in \{1, \ldots, n\},$$

where the residuals ε_i satisfy $\mathbb{E}[\varepsilon_i] = 0$.

Example

Given data $y \in \mathbb{R}^k$, find the unknown parameter $x \in \mathbb{R}^d$ such that $y = Ax + \varepsilon$, where the residual ε satisfies $\mathbb{E}[\varepsilon] = 0$.

Let $y_1, \ldots, y_n$ be the data, which are i.i.d. random variables. We assume that these follow a parameter-dependent probability distribution with PDF (resp. PMF) $f(x, y)$ for some realization of the parameter $x \in X$. (With a slight abuse of notation, one might write $y_1, \ldots, y_n \overset{\text{i.i.d.}}{\sim} f(x, \cdot)$ for some unknown $x \in X$.)

Thus we are interested in identifying the value of the parameter $x \in X$, which (in some sense) best approximates the data out of the set

$$\mathscr{F} = \{f(x, y) \mid x \in X\}$$

containing all the possible candidates for the PDFs $f(x, y)$ which could have generated the data.

A common method to estimate parameters in a parametric models is the maximum likelihood method.

Maximum likelihood

Let $y_1, \ldots, y_n$ be i.i.d. with the PDF $f(x, y)$.

Definition

The likelihood function is defined by

$$\mathcal{L}_n(x) = \prod_{i=1}^{n} f(x, y_i).$$

The log-likelihood function is defined by $\ell_n(x) = \log \mathcal{L}_n(x)$.

The likelihood function is simply the joint density of the data, except that we treat it as a function of the parameter x. The likelihood function is not a density function: in general, the function $\mathcal{L}_n(x)$ does not integrate to 1 with respect to x.

Definition

The maximum likelihood (ML) estimator is defined as a maximizer of the likelihood function

$$\hat{x}_{\mathrm{ML}} = \arg\max_{x \in X} \mathcal{L}_n(x).$$

Remarks:

- The ML estimator satisfies

$$\mathcal{L}_n(\hat{x}_{\mathrm{ML}}) \geq \mathcal{L}_n(x) \quad \text{for all } x \in X.$$

  It answers the question: *Which value of the unknown x is the most likely to produce the measured data?*
- The ML estimator may not be unique.
- The maximum of the log-likelihood $\ell_n(x)$ occurs at the same point as the maximum of $\mathcal{L}_n(x)$. It is often more convenient to work with

$$\hat{x}_{\mathrm{ML}} = \arg\max_{x \in X} \ell_n(x).$$

- Multiplying $\mathcal{L}_n(x)$ by any positive constant c (not depending on x) will not change the ML estimator, so the constants in the likelihood function are often dropped.

Example

Assume that $y_1, \dots, y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ for some unknown mean parameter $\mu \in \mathbb{R}$. The likelihood function for $\mu \in \mathbb{R}$ is given by

$$\mathcal{L}_n(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(y_i-\mu)^2} = \frac{1}{(2\pi)^{n/2}} \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

and the log-likelihood is given by

$$\ell_n(\mu) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

Differentiating this with respect to μ yields

$$\ell_n'(\mu) = \sum_{i=1}^{n}(y_i - \mu) = n(\overline{y}_n - \mu).$$

Setting this to 0 yields the ML estimator $\hat{\mu}_{\mathrm{ML}} = \overline{y}_n$. Thus the ML estimator coincides with the empirical mean of $y_1, \dots, y_n$.

Example

Consider n observations (pairs) $(x_1, y_1), \dots, (x_n, y_n)$ 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 values x_i through a simple linear model

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

where the residuals are assumed to be Gaussian $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, $\sigma > 0$. Writing $b = (b_0, b_1)$, the ML estimator $\hat{b} = \hat{b}_{\mathrm{ML}}$ is given by

$$\hat{b}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \hat{\rho}(x, y) \frac{s_y}{s_x},$$
$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}.$$

Example

Assume that $y_1, \ldots, y_n \in \mathbb{R}^k$ are i.i.d. realizations of random variable y, which come from some mathematical model

$$y_i = F(x) + \varepsilon_i,$$

where $x \in \mathbb{R}^d$ is the unknown parameter, $F\colon \mathbb{R}^d \to \mathbb{R}^k$ is a function, and $\varepsilon_1, \ldots, \varepsilon_k$ are i.i.d. realizations of measurement noise ε with PDF ρ_{n}.

Now

$$\begin{aligned}\mathbb{P}(y \in B) = \mathbb{P}(F(x) + \varepsilon \in B) = \mathbb{P}(\varepsilon \in B - F(x)) &= \int_{B-F(x)} \rho_{\mathrm{n}}(t)\,\mathrm{d}t \\ &= \int_B \rho_{\mathrm{n}}(t - F(x))\,\mathrm{d}t \quad \text{for all events } B.\end{aligned}$$

This means that $f(x, y_i) = \rho_{\mathrm{n}}(y_i - F(x))$, and the likelihood function is

$$\mathcal{L}_n(x) = \prod_{i=1}^{n} \rho_{\mathrm{n}}(y_i - F(x)).$$

Example

Assume that $y \in \mathbb{R}^k$ is an observation of the mathematical model

$$y = F(x) + \varepsilon,$$

where $x \in \mathbb{R}^d$ is the unknown parameter and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ is Gaussian measurement noise, with $\sigma > 0$ and I is the $k \times k$ identity matrix. In this case, the noise has the PDF

$$\rho_{\mathrm{n}}(\varepsilon) = \frac{1}{(2\pi\sigma^2)^{k/2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\|\varepsilon\|^2}$$

and the likelihood function is given by

$$\mathcal{L}_n(x) = \frac{1}{(2\pi\sigma^2)^{k/2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\|y-F(x)\|^2}.$$

The ML estimator can therefore be found as the *minimizer*(!)

$$\hat{x}_{\mathrm{ML}} = \underset{x \in \mathbb{R}^d}{\arg\min} \, \|y - F(x)\|^2.$$

Example

Assume that $y \in \mathbb{R}^k$ is an observation of the linear mathematical model

$$y = Ax + \varepsilon,$$

where $x \in \mathbb{R}^d$ is the unknown parameter, $A \in \mathbb{R}^{k \times d}$ is a matrix, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ is Gaussian measurement noise, with $\sigma > 0$ and I is the $k \times k$ identity matrix. This corresponds to $F(x) = Ax$ in the previous example, with the likelihood

$$\mathcal{L}_n(x) = \frac{1}{(2\pi\sigma^2)^{k/2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\|y - Ax\|^2}$$

and ML estimator

$$\hat{x}_{\mathrm{ML}} = \arg\min_{x \in \mathbb{R}^d} \|y - Ax\|^2.$$

If $A^{\mathrm{T}}A$ is invertible, then the ML estimator is precisely the least squares solution

$$A^{\mathrm{T}}A\hat{x}_{\mathrm{ML}} = A^{\mathrm{T}}y.$$

(If $A^{\mathrm{T}}A$ is not invertible, then the ML estimator is not unique.)

Computing ML estimates

In special cases the ML estimator $\hat{x}_{\mathrm{ML}}$ can be solved analytically, but more often, the optimization problem needs to be solved numerically. If the log-likelihood ℓ_n is twice continuously differentiable, one can use, e.g., the Newton–Raphson algorithm. Suppose that the parameter $x \in \mathbb{R}$ is one-dimensional. If x is a good guess for $\hat{x}_{\mathrm{ML}}$ (in the sense that $x \approx \hat{x}_{\mathrm{ML}}$, then Taylor's theorem implies that

$$0 = \ell_n'(\hat{x}_{\mathrm{ML}}) \approx \ell_n'(x) + (\hat{x}_{\mathrm{ML}} - x)\ell_n''(x).$$

Solving for $\hat{x}_{\mathrm{ML}}$ yields $\hat{x}_{\mathrm{ML}} = x - \frac{\ell_n'(x)}{\ell_n''(x)}$.

Repeating this process iteratively yields the following algorithm.

> Let $x_0 \in \mathbb{R}$ be an initial guess for $\hat{x}_{\mathrm{ML}}$.
> **for** $j = 1, 2, \ldots,$ **do**
>   Set $x_j = x_{j-1} - \frac{\ell_n'(x_{j-1})}{\ell_n''(x_{j-1})}$
> **until** $|\ell_n'(x_j)| < TOL$

In the multiparameter case $x \in \mathbb{R}^d$, the ML estimator $\hat{x}_{\mathrm{ML}}$ is a vector and the method is the following:

> Let $x_0 \in \mathbb{R}^d$ be an initial guess for $\hat{x}_{\mathrm{ML}}$.
> **for** $j = 1, 2, \ldots,$ **do**
> > Set $x_j = x_{j-1} - H(x_{j-1})^{-1} \nabla \ell(x_{j-1})$
>
> **until** $\|\nabla \ell_n(x_j)\| < TOL$

Here, $H(x)$ is the $d \times d$ Hessian matrix defined by

$$H(x) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} \ell_n(x) & \frac{\partial^2}{\partial x_1 \partial x_2} \ell_n(x) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_d} \ell_n(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} \ell_n(x) & \frac{\partial^2}{\partial x_2^2} \ell_n(x) & \cdots & \frac{\partial^2}{\partial x_2 \partial x_d} \ell_n(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_d \partial x_1} \ell_n(x) & \frac{\partial^2}{\partial x_d \partial x_2} \ell_n(x) & \cdots & \frac{\partial^2}{\partial x_d^2} \ell_n(x) \end{bmatrix}.$$

Remark. Depending on the application, any other reasonable or natural optimization procedure might also work: e.g., Gauss–Newton method, Levenberg–Marquardt method, conjugate gradient or Krylov subspace methods, (stochastic) gradient descent. . .

Properties of the ML estimator

The ML estimator has many desirable qualities under somewhat relaxed assumptions:

- The ML estimator is *consistent*: $\hat{x}_{\mathrm{ML}} \xrightarrow{P} x_\star$ as $n \to \infty$, where $x_\star$ denotes the true value of the parameter x.
- The ML estimator is *asymptotically normal*: $\frac{\hat{x}_{\mathrm{ML}} - x_\star}{\widehat{\mathtt{se}}} \xrightarrow{d} \mathcal{N}(0, 1)$.
- The ML estimator is *asymptotically optimal*: roughly, this means that among all well-behaved estimators, the ML estimator has the smallest variance, at least for large samples.

  $\vdots$

As the sample size $n \to \infty$, the ML estimator is an ideal estimator from a *frequentist* point of view.

However, in some applications one might have a limited amount of data and/or the data generation is not repeatable, so the asymptotic properties of the ML estimator may not be of much use. Next week, we will start discussing the *Bayesian paradigm*, where the fundamental conceit is that only a finite amount of data is available: probability is not defined as the limit of relative frequencies, but as a (subjective) degree of belief.