Statistics, Data Analysis, and Simulation SS 2019

08.128.730 Statistik, Datenanalyse und Simulation

Dr. Michael O. Distler
<distler@uni-mainz.de>

Mainz, Mai 13, 2019

Least squares method

History: Introduced by *Legendre*, *Gauß*, and *Laplace* at the beginning of the 19th century.
Therefore the least squares method is older than the more general maximum likelihood method.
From now on the measured values which have the property of random variables (data) will be labeled y_i.
n measurements of x will give $y_1, y_2, \ldots, y_n$:

$$y_i = x + \epsilon_i$$

ϵ_i are the deviations $y_i \leftrightarrow x$ (statistical error).

Least squares method

- The measured values deviate from the *true values*. The size of this difference is parameterized by the standard deviation σ.
- The y_i are a statistical sample which can be described by a probability distribution function.
- We also need a functional description (model) for the *true values*.
- This model may depend on additional variables a_j called parameters.
- These parameters cannot be measured directly.

The model is given as one or more equations

$$f(a_1, a_2, \ldots, a_p, y_1, y_2, \ldots, y_n) = 0$$

Least squares method

The model can be used to find corrections Δy_i to the measured values y_i.
The method of least squares requires the sum of squares of the residuals Δy_i to be minimal.
For the simplest case of uncorrelated data with equal standard deviation this means:

$$S = \sum_{i=1}^{n} \Delta y_i^2 = \text{Minimum}$$

$\longrightarrow$ indirect measurement of the parameters.

Least squares method

The **least squares method** has a number of **optimal statistical properties** and often leads to easy solutions. Other rules are possible, but generally lead to complicated solutions.

$$\sum_{i=1}^{n} |\Delta y_i| = \text{minimum} \quad \text{or} \quad \max |\Delta y_i| = \text{minimum}$$

Least squares method

General case:

- The Data is written as a n-vector $\mathbf{y}$.
- Different standard deviations and correlations are taken care of by use of a covariance matrix $\mathbf{V}$.

Least squares method using matrices:

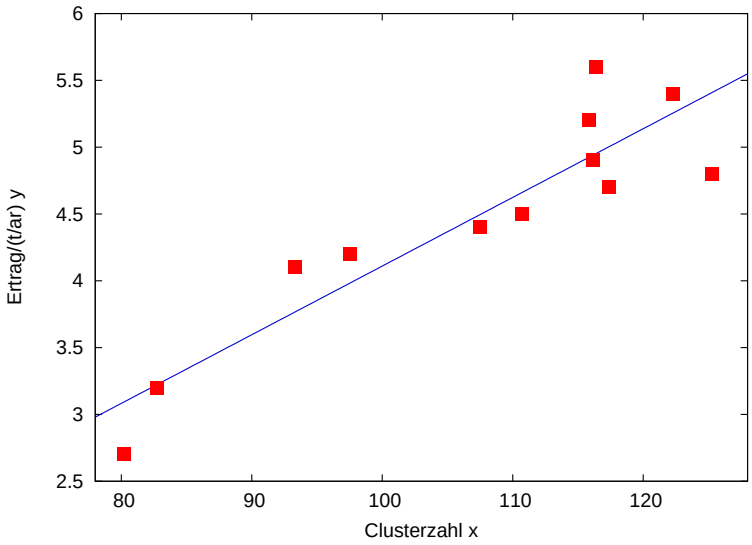
$$S = \Delta\mathbf{y}^T \mathbf{V}^{-1} \Delta\mathbf{y}$$

Here $\Delta\mathbf{y}$ is the vector of residuals.

Least squares method

Example: In wine-growing the amount of wine harvested in autumn is measured in tons per 100 m^2 (t/ar). It is known that the annual yield can be predicted fairly well in July, by determining the average number of berries which have been formed per bunch.

year	yield (y_i)	cluster (x_i)
1971	5.6	116.37
1973	3.2	82.77
1974	4.5	110.68
1975	4.2	97.50
1976	5.2	115.88
1977	2.7	80.19
1978	4.8	125.24
1979	4.9	116.15
1980	4.7	117.36
1981	4.1	93.31
1982	4.4	107.46
1983	5.4	122.30



Least squares method

Straight line fit $f(x) = a + b \cdot x$ using *gnuplot*:

```
degrees of freedom (FIT_NDF) : 10
rms of residuals (FIT_STDFIT) = sqrt(WSSR/ndf) :
0.364062
variance of residuals (reduced chisquare) =
WSSR/ndf :  0.132541

Final set of parameters Asymptotic Standard Error
======================= ==========================
a = -1.0279 +/- 0.7836 (76.23%)
b = 0.0513806 +/- 0.00725 (14.11%)

correlation matrix of the fit parameters:
    a       b
 a  1.000
 b  -0.991  1.000
```

Parameter estimation

Estimation of the parameters **a** from measured data using a *linear* model.
The parameter vector **a** consists of p elements $a_1, a_2, \ldots, a_p$.
The measured values form the vector **y** (n random variables, $y_1, y_2, \ldots, y_n$).
The estimation value of y is a function of the variable x:

$$y(x) = f(x, \mathbf{a}) = a_1 f_1(x) + a_2 f_2(x) + \ldots + a_p f_p(x).$$

The estimation value of each single measurement y_i is

$$E[y_i] = f(x_i, \bar{\mathbf{a}}) = \bar{y}_i$$

Here the elements of $\bar{\mathbf{a}}$ are the true values of the parameters **a**.

Parameter estimation

The *residuals*

$$r_i = y_i - f(x_i, \mathbf{a})$$

have nice properties if $\mathbf{a} = \bar{\mathbf{a}}$:

$$E[r_i] = 0 \qquad E[r_i^2] = V[r_i] = \sigma_i^2.$$

The only requirements are that the p.d.f. of the residuals is unbiased and has finite variance.
So it is not required, that the residuals are *Gaussian distributed*.

Least squares: Normal equations

For now, all data has the same variance and is uncorrelated. Following the principle of least squares we minimize the sum of squares of the residuals varying the parameters $a_1, a_2, \ldots, a_p$:

$$S = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - a_1 f_1(x_i) - a_2 f_2(x_i) - \ldots - a_p f_p(x_i))^2$$

Conditions for the minimum:

$$\begin{aligned}
\frac{\partial S}{\partial a_1} &= 2 \sum_{i=1}^{n} f_1(x_i) \, (a_1 f_1(x_i) + a_2 f_2(x_i) + \ldots + a_p f_p(x_i) - y_i) = 0 \\
\ldots & \quad \ldots \\
\frac{\partial S}{\partial a_p} &= 2 \sum_{i=1}^{n} f_p(x_i) \, (a_1 f_1(x_i) + a_2 f_2(x_i) + \ldots + a_p f_p(x_i) - y_i) = 0
\end{aligned}$$

Least squares: Normal equations

We then write down the conditions as *normal equations*

$$
\begin{array}{lcccl}
a_1 \sum f_1(x_i)^2 & + & \ldots & + & a_p \sum f_1(x_i) f_p(x_i) = \sum y_i f_1(x_i) \\
a_1 \sum f_2(x_i) f_1(x_i) & + & \ldots & + & a_p \sum f_2(x_i) f_p(x_i) = \sum y_i f_2(x_i) \\
\ldots & & & & \\
a_1 \sum f_p(x_i) f_1(x_i) & + & \ldots & + & a_p \sum f_p(x_i)^2 = \sum y_i f_p(x_i)
\end{array}
$$

The solution of these normal equations is the least square estimate of the parameters $a_1, a_2, \ldots, a_p$.

Least squares: Matrix formalism

Matrix formalism and matrix algebra simplify the computation. The $n \times p$ values $f_j(x_i)$ form a $n \times p$ matrix. The p parameters a_j and the n measured values y_i form column vectors.

$$\mathbf{A} = \begin{pmatrix} f_1(x_1) & f_2(x_1) & \dots & f_p(x_1) \\ f_1(x_2) & f_2(x_2) & \dots & f_p(x_2) \\ \dots & & & \\ \dots & & & \\ f_1(x_n) & f_2(x_n) & \dots & f_p(x_n) \end{pmatrix} \quad \mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$$

Least squares: Matrix formalism

The n-vector of the residuals is

$$\mathbf{r} = \mathbf{y} - \mathbf{A}\mathbf{a}.$$

Considering the sum S we get

$$\begin{aligned} S = \mathbf{r}^T\mathbf{r} &= (\mathbf{y} - \mathbf{A}\mathbf{a})^T(\mathbf{y} - \mathbf{A}\mathbf{a}) \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{a}^T\mathbf{A}^T\mathbf{y} + \mathbf{a}^T\mathbf{A}^T\mathbf{A}\mathbf{a} \end{aligned}$$

Conditions for the minimum

$$-2\mathbf{A}^T\mathbf{y} + 2\mathbf{A}^T\mathbf{A}\hat{\mathbf{a}} = 0$$

or using normal equations

$$(\mathbf{A}^T\mathbf{A})\hat{\mathbf{a}} = \mathbf{A}^T\mathbf{y}$$

The solution only requires standard methods of matrix algebra:

$$\boxed{\hat{\mathbf{a}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}}$$

Least squares: Matrix formalism

The covariance matrix is the quadratic $n \times n$ matrix

$$\mathbf{V}[\mathbf{y}] = \begin{pmatrix} \mathrm{var}(y_1) & \mathrm{cov}(y_1, y_2) & \dots & \mathrm{cov}(y_1, y_n) \\ \mathrm{cov}(y_2, y_1) & \mathrm{var}(y_2) & \dots & \mathrm{cov}(y_2, y_n) \\ \dots & \dots & & \dots \\ \mathrm{cov}(y_n, y_1) & \mathrm{cov}(y_n, y_2) & \dots & \mathrm{var}(y_n) \end{pmatrix}$$

Here the covariance matrix is just a diagonal matrix

$$\mathbf{V}[\mathbf{y}] = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & & \dots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

Least squares: Matrix formalism

Since the parameters are a linear function of the data $\hat{\mathbf{a}} = \mathbf{B}\mathbf{y}$ we can use the standard error propagation:

$$\mathbf{V}[\hat{\mathbf{a}}] = \mathbf{B}\mathbf{V}[\mathbf{y}]\mathbf{B}^T$$

with $\mathbf{B} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ we get

$$\mathbf{V}[\hat{\mathbf{a}}] = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{V}[\mathbf{y}]\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}$$

Here we have equal errors for all data points

$$\boxed{\mathbf{V}[\hat{\mathbf{a}}] = \sigma^2 (\mathbf{A}^T\mathbf{A})^{-1}}$$

Least squares: Matrix formalism

The sum $\hat{S}$ of squares of the residuals in the minimum is

$$\hat{S} = \mathbf{y}^T\mathbf{y} - 2\hat{\mathbf{a}}^T\mathbf{A}^T\mathbf{y} + \hat{\mathbf{a}}^T\mathbf{A}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} = \mathbf{y}^T\mathbf{y} - \hat{\mathbf{a}}^T\mathbf{A}^T\mathbf{y}.$$

The expectation value of $E[\hat{S}]$ is

$$\boxed{E[\hat{S}] = \sigma^2\,(n-p)}.$$

If the variance of the data is not known, we can use $\hat{S}$ to get an estimate

$$\hat{\sigma}^2 = \hat{S}/(n-p).$$

This is a good estimate for large values of $(n-p)$.

Least squares: Matrix formalism

After estimating the parameters using the linear least squares method we can calculate $f(x)$ for arbitrary x

$$\hat{y}(x) = f(x, \hat{\mathbf{a}}) = \sum_{j=1}^{p} \hat{a}_j f_j(x).$$

For the values x_i which belong to the measured values y_i we will get the predicted values using

$$\hat{\mathbf{y}} = \mathbf{A}\hat{\mathbf{a}}.$$

Using error propagation one gets the associated covariance matrix

$$\mathbf{V}[\hat{\mathbf{y}}] = \mathbf{A}\mathbf{V}[\mathbf{a}]\mathbf{A}^T = \sigma^2\, \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$$

Least squares: Matrix formalism

If the data points are independent the covariance matrix is

$$\mathbf{V}[\mathbf{y}] = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}$$

The sum of squares of the residuals is now

$$S = \sum_i \frac{r_i^2}{\sigma_i^2} = \text{Minimum}$$

We define a weight matrix $\mathbf{W}(\mathbf{y})$ which is the inverse of the covariance matrix

$$\mathbf{W}(\mathbf{y}) = \mathbf{V}[\mathbf{y}]^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & \dots & 0 \\ \dots & \dots & & \dots \\ 0 & 0 & \dots & 1/\sigma_n^2 \end{pmatrix}$$

Least squares: Matrix formalism

The sum of squares of the weighted residuals is

$$S = \mathbf{r}^T \mathbf{W}(\mathbf{y})\mathbf{r} = (\mathbf{y} - \mathbf{A}\mathbf{a})^T \mathbf{W}(\mathbf{y})(\mathbf{y} - \mathbf{A}\mathbf{a})$$

which has to minimized. One gets

$$\boxed{\begin{aligned} \hat{\mathbf{a}} &= (\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{W}\mathbf{y} \\ \mathbf{V}[\hat{\mathbf{a}}] &= (\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1} \end{aligned}}$$

The sum of squares of the residuals for $\mathbf{a} = \hat{\mathbf{a}}$ is

$$\hat{S} = \mathbf{y}^T\mathbf{W}\mathbf{y} - \hat{\mathbf{a}}^T\mathbf{A}^T\mathbf{W}\mathbf{y}$$

with expectation value $\boxed{E[\hat{S}] = n - p}$.
The covariance matrix for the predicted values is

$$\mathbf{V}[\hat{\mathbf{y}}] = \mathbf{A}(\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^T$$

Least squares: Linear regression

For the linear regression we fit the function
$y = f(x, \mathbf{a}) = a_1 + a_2 x$.
The data y_i has been taken at certain values of x_i.

$$\mathbf{A} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \dots & \\ 1 & x_n \end{pmatrix} \qquad \mathbf{V} = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & & 0 \\ 0 & 0 & \sigma_3^2 & & 0 \\ \dots & & & & \dots \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}$$

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{pmatrix} \qquad \mathbf{W} = \mathbf{V}^{-1} \qquad w_{ii} = \frac{1}{\sigma_i^2}$$

Least squares: Linear regression

Solution:

$$\mathbf{A}^T\mathbf{W}\mathbf{A} = \begin{pmatrix} \sum w_i & \sum w_i x_i \\ \sum w_i x_i & \sum w_i x_i^2 \end{pmatrix} = \begin{pmatrix} S_1 & S_x \\ S_x & S_{xx} \end{pmatrix}$$

$$\mathbf{A}^T\mathbf{W}\mathbf{y} = \begin{pmatrix} \sum w_i y_i \\ \sum w_i x_i y_i \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$$

$$\begin{pmatrix} S_1 & S_x \\ S_x & S_{xx} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$$

$$\begin{aligned} \hat{\mathbf{a}} &= (\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{W}\mathbf{y} \\ \mathbf{V}[\hat{\mathbf{a}}] &= (\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1} \end{aligned}$$

$$\begin{pmatrix} S_1 & S_x \\ S_x & S_{xx} \end{pmatrix}^{-1} = \frac{1}{D}\begin{pmatrix} S_{xx} & -S_x \\ -S_x & S_1 \end{pmatrix} \quad \text{with } D = S_1 S_{xx} - S_x^2$$

Least squares: Linear regression

The estimate for the parameters is

$$\begin{aligned} \hat{a}_1 &= (S_{xx} S_y - S_x S_{xy})/D \\ \hat{a}_2 &= (-S_x S_y - S_1 S_{xy})/D \end{aligned}$$

and the covariance matrix is

$$\mathbf{V}[\hat{\mathbf{a}}] = \frac{1}{D} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S_1 \end{pmatrix} .$$

For the sum of squares of residuals one gets

$$\hat{S} = S_{yy} - \hat{a}_1 S_y - \hat{a}_2 S_{xy}$$

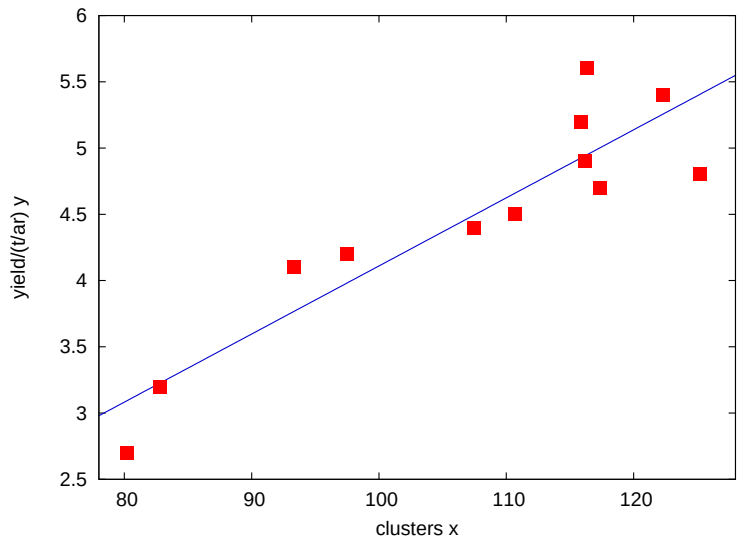
For the predicted value $\hat{y} = \hat{a}_1 + \hat{a}_2 x$ we get the variance by calculating:

$$V[\hat{y}] = V[\hat{a}_1] + x^2 V[\hat{a}_2] + 2xV[\hat{a}_1, \hat{a}_2] = (S_{xx} - 2xS_x + x^2 S_1)/D$$

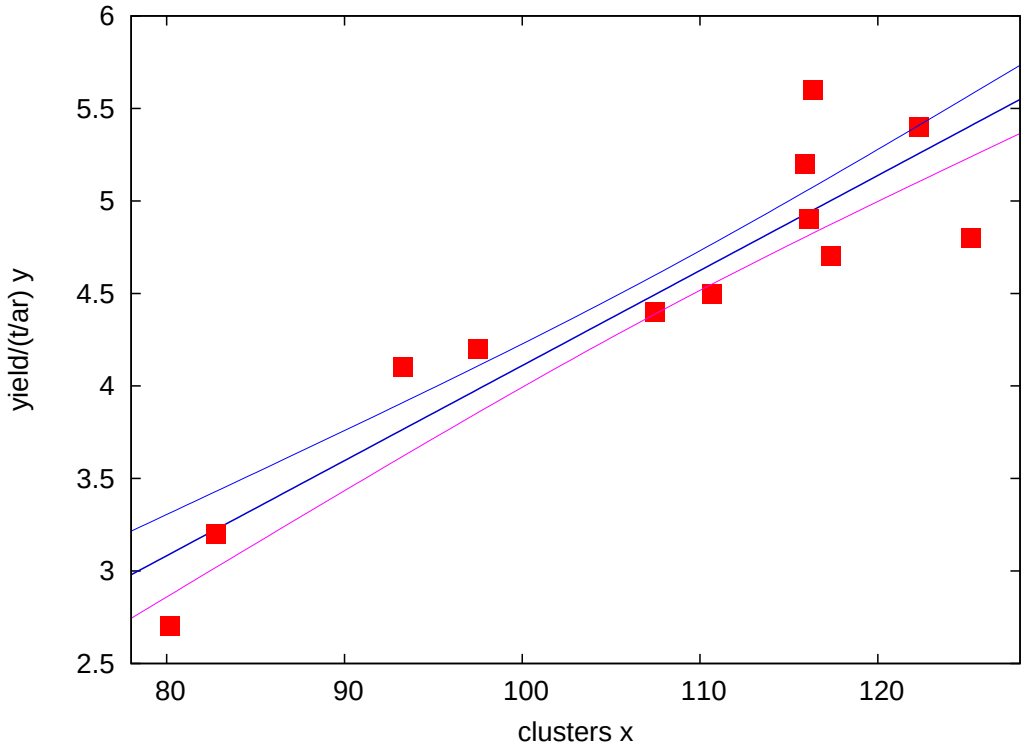
Least squares method

Example: In wine-growing the amount of wine harvested in autumn is measured in tons per 100 m^2 (t/ar). It is known that the annual yield can be predicted fairly well in July, by determining the average number of berries which have been formed per bunch.

year	yield (y_i)	cluster (x_i)
1971	5.6	116.37
1973	3.2	82.77
1974	4.5	110.68
1975	4.2	97.50
1976	5.2	115.88
1977	2.7	80.19
1978	4.8	125.24
1979	4.9	116.15
1980	4.7	117.36
1981	4.1	93.31
1982	4.4	107.46
1983	5.4	122.30



Least squares method: Wine-growing example



$$
\begin{aligned}
a_1 &= -1.0279 \pm 0.7836 \\
a_2 &= 0.0513806 \pm 0.00725 \\
\mathit{Errorband} : err(x) &= -1.02790 + 0.0513806x \\
&\quad \pm \sqrt{5.2561 \cdot 10^{-5} x^2 - 0.011259x + 0.61395}
\end{aligned}
$$