Linear least squares

1. The least squares principle

2. Linear least squares

3. Properties of least square estimates

4. Independent data
The least squares principle

A model with parameters is assumed to describe the data.

Principle of parameter estimation: minimize sum $S$ of squares of deviations $\Delta y_i$ between model and data!

Solution: derivatives of $S$ w.r.t. parameters = zero!

Different forms: sum of squared deviations, weighted sum of squared deviations, sum of squared deviations weighted with inverse covariance matrix:

\[ S = \sum_{i=1}^{n} \Delta y_i^2 \quad S = \sum_{i=1}^{n} \left( \frac{\Delta y_i}{\sigma_i} \right)^2 \quad S = \Delta y^T V^{-1} \Delta y \]

Example: mean value $\bar{y}$ of $n$ measured values $y_i$:

\[ S = \sum_{i=1}^{n} (y - y_i)^2 = \text{minimum} \quad \hat{y} = \sum_{i=1}^{n} y_i/n \quad \text{follows from } \text{grad } S = 0 \]
Systems of linear equations

Linear model: \[ A \cdot a = y \quad \text{and} \quad A \cdot a \approx y \]

with \( n \) elements of the measured vector \( y \) and \( p \) elements of the parameter vector \( a \).
The model of Linear Least Squares: \( y \cong A a \)

\[
\begin{align*}
  y &= \text{vector of measured data (} n \text{ elements)} \\
  A &= \text{matrix (fixed)} \\
  a &= \text{vector of parameters (} p \text{ elements)} \\
  r &= y - Aa = \text{vector of residuals} \\
  V[y] &= \text{covariance matrix of the data} \\
  W &= V[y]^{-1} \text{ weight matrix}
\end{align*}
\]

**Least Squares Principle**: minimize the expression

\[
S(a) = r^T W r = (y - Aa)^T W (y - Aa)
\]

with respect to \( a \).
Least Squares solution

Derivatives of expression $S(a)$:

$$\frac{1}{2} \text{ grad } S = \frac{1}{2} \frac{\partial S}{\partial a} = (-A^T W y + (A^T W A) a)$$

$$\frac{1}{2} \frac{\partial^2 S}{\partial a^2} = (A^T W A) = \text{ constant}$$

Solution (from $\partial S/\partial a = 0$)

$$-A^T W y + (A^T W A) a = 0$$

is linear transformation of the data vector $y$:

$$\hat{a} = (A^T W A)^{-1} A^T W y = B y$$

Covariance matrix of $a$ by ”error” propagation ($V[y] = W^{-1}$):

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

$$= (A^T W A)^{-1} = \text{ inverse of second derivate of } S$$

Solution vector $a$ and covariance matrix $V[y]$ are calculated by few matrix operations. No starting parameter values necessary, no iterations – a single step.
Properties of solution

Starting from **Principles**: methods of solution and properties of the solution are derived, which are valid under certain conditions.

**Conditions:**

- Data are unbiased: \( E[y] = A \) \( a_{\text{true}} \) (\( a_{\text{true}} = \) true parameter vector)
- Covariance matrix \( V[y] \) is known (correct) and finite

**Properties:**

- Estimated parameters are unbiased:

  \[
  E[\hat{a}] = (A^T W A)^{-1} A^T W E[y] = a_{\text{true}}
  \]

- In the class of unbiased estimates \( a^* \), which are linear in the data, the **Least Squares** estimates \( \hat{a} \) have the smallest variance (Gauß-Markoff theorem)

**Properties are not valid, if conditions violated.**
Simplification for independent (=uncorrelated) data

...assuming same variance \( \sigma^2 \) for all data.

Covariance matrix and weight matrix are diagonal:

\[
V(y) = \sigma^2 I_n \quad \quad \quad \quad W = \frac{1}{\sigma^2} I_n
\]

\((I_n \text{ is } n\text{-by-}n \text{ unit matrix}).\)

\[
\text{solution} \quad \hat{a} = C^{-1} A^T y \quad \text{with} \quad C = A^T A
\]

\[
\text{covariance matrix} \quad V(\hat{a}) = \sigma^2 C^{-1}
\]

Note: the solution \( \hat{a} \) does not depend on \( \sigma^2 \), but the covariance matrix is proportional to \( \sigma^2 \).
Properties of least square estimates

Basic assumptions on the properties of the data:

1. the data are unbiased: \( E[y] = A\alpha_{\text{true}} \) or \( E[y - A\alpha_{\text{true}}] = 0 \)

2. the variances are all the same: \( V[y - A\alpha_{\text{true}}] = \sigma^2 I_n \)

(i.e. special case of independent data of same precision is assumed).

No assumption is made on the distribution of the residuals (i.e. a Gaussian distribution is not required!)

Least squares estimates:

\[
\hat{a} = C^{-1} A^T y \quad \text{with} \quad C = A^T A \quad V[\hat{a}] = \sigma^2 C^{-1}
\]

First property: Least square estimates are unbiased.

Proof:

\[
E[\hat{a}] = C^{-1} A^T E[y] = C^{-1} A^T A \alpha_{\text{true}} = \alpha_{\text{true}}
\]
Gauß-Markoff Theorem

Consider class of linear estimates $\mathbf{a}^* = \mathbf{U} \mathbf{y}$, which are unbiased:

$$E[\mathbf{a}^*] = \mathbf{U} E[\mathbf{y}] = \mathbf{U} \mathbf{A} \mathbf{a}_{\text{true}} = \mathbf{a}_{\text{true}}, \quad V[\mathbf{a}^*] = \sigma^2 \mathbf{U} \mathbf{U}^T$$

Case of least squares: $\mathbf{U}_{LS} = \mathbf{C}^{-1} \mathbf{A}^T$ with $V[\hat{\mathbf{a}}] = \sigma^2 \mathbf{C}^{-1}$.

**Theorem:** The least square estimate $\hat{\mathbf{a}}$ has the property

$$V[\mathbf{a}^*]_{jj} \geq V[\hat{\mathbf{a}}]_{jj} \quad \text{for all } j,$$

i.e., the least squares estimate has the smallest possible error.

Proof: product $\mathbf{U} \mathbf{U}^T$ can be written in the form

$$\mathbf{U} \mathbf{U}^T = \mathbf{C}^{-1} + (\mathbf{U} - \mathbf{C}^{-1} \mathbf{A}^T)(\mathbf{U} - \mathbf{C}^{-1} \mathbf{A}^T)^T$$

$$= \mathbf{C}^{-1} + \mathbf{U} \mathbf{U}^T - \mathbf{U} \mathbf{A} \mathbf{C}^{-1} - \mathbf{C}^{-1} \mathbf{A}^T \mathbf{U}^T + \mathbf{C}^{-1} \mathbf{A}^T \mathbf{A} \mathbf{C}^{-1}$$

For the covariance matrix follows:

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

Product on the right has diagonal elements $\geq 0$ (→ Theorem).
Sum of squares of residuals

Third property: The expectation of the sum of squares of the residuals is $\hat{S} = \sigma^2(n - p)$.

Definition of $\hat{S}$ in terms of the fitted vector $\hat{a}$:

$$\hat{S} = (y - A\hat{a})^T(y - A\hat{a}) = y^Ty - y^TA\hat{a}$$

This equation is rewritten in terms of $a_{\text{true}}$ (instead of $\hat{a}$) using the matrix $U = I_n - AC^{-1}A^T$ and the vector $z = y - Aa_{\text{true}}$.

$$\hat{S} = (y - Aa_{\text{true}})^TU(y - Aa_{\text{true}}) = z^TUz$$

(check the agreement with $\hat{S}$ above by multiplication).

Properties of $z$: $E[z] = 0$ and covariance matrix

$$V[z] = \sigma^2I_n \quad \text{i.e.} \quad V[z_i] = E[z_i^2] = \sigma^2 \quad \text{and} \quad E[z_i z_k] = 0.$$
\[ \hat{S} = \sum_{i=1}^{n} \sum_{k=1}^{n} U_{ik} \, z_i \, z_k \quad E \left[ \hat{S} \right] = \sum_{i=1}^{n} U_{ii} \quad E \left[ z_i^2 \right] = \sigma^2 \sum_{i=1}^{n} U_{ii} = \sigma^2 \text{ trace}(U) \]

(the trace of a square matrix is the sum of the diagonal elements). Calculation of the trace of \( U \):

\[
\text{trace}(U) = \text{trace}(\mathbf{I}_n - \mathbf{A} \mathbf{C}^{-1} \mathbf{A}^T) = \text{trace}(\mathbf{I}_n) - \text{trace}(\mathbf{A} \mathbf{C}^{-1} \mathbf{A}^T) \\
= \text{trace}(\mathbf{I}_n) - \text{trace}(\mathbf{C}^{-1} \mathbf{A}^T \mathbf{A}) \\
= \text{trace}(\mathbf{I}_n) - \text{trace}(\mathbf{I}_p) = n - p. \quad \rightarrow \text{ Proof}
\]

Application: estimate data variance (for \( n \gg p \)) by \( \hat{\sigma}^2 = \hat{S}/(n - p) \)

Special case of Gaussian distributed measurement errors:

\[ \frac{\hat{S}}{\sigma^2} \text{ distributed according to the } \chi^2_{n-p} \text{ distribution} \]

to be used for goodness-of-fit test.
Distribution-free properties of least squares estimates in linear problems:

1. Least square estimates are unbiased.

2. The least square estimate $\hat{a}$ has the property

\[ V[a^*]_{jj} \geq V[\hat{a}]_{jj} \text{ for all } j, \]

i.e., the least squares estimate has the smallest possible error. (Gauß-Markoff Theorem)

3. The expectation of the sum of squares of the residuals is $\hat{S} = \sigma^2(n - p)$.

Valid under the condition that the data are unbiased!
Independent data

Often the direct measurements, which are input to a least squares problem, are independent, i.e. the covariance matrix $V(y)$ and the weight matrix $W$ are diagonal.

This property, which is assumed here, simplifies the computation of the matrix products

$$C = A^T W A \quad \text{and} \quad b = A^T W y$$

which are necessary for the solution

$$\hat{a} = C^{-1} b \quad \quad V(\hat{a}) = C^{-1}$$

Note: the parameters $a$ will be correlated through the model $y = Aa$ and the covariance matrix $V(\hat{a})$ will be non-diagonal.
Normal equations for independent data

The diagonal elements of the weight matrix $W$ are denoted by $w_i$, with $w_i = 1/\sigma_i^2$. Each data value $y_i$ with its weight $w_i$ makes an independent contribution to the final matrix products. Calling the $i$-th row $A_i$, with

$$i\text{-th row of } A \quad A_i = (d_1, d_2, \ldots, d_p) \quad y = d_1a_1 + d_2a_2 + \ldots + d_pa_p$$

the contributions of this row to $C$ and $b$ can be written as the $p \times p$-matrix $w_iA_i^T \cdot A_i$ and the $p$-vector $w_iA_i^T \cdot y_i$.

The contributions of a single row are:

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$\ldots$</th>
<th>$d_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$</td>
<td>$w_id_1^2$</td>
<td>$w_id_1d_2$</td>
<td>$\ldots$</td>
<td>$w_id_1d_p$</td>
</tr>
<tr>
<td>$d_2$</td>
<td>$w_id_2^2$</td>
<td>$w_id_2d_2$</td>
<td>$\ldots$</td>
<td>$w_id_2d_p$</td>
</tr>
<tr>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td></td>
</tr>
<tr>
<td>$d_p$</td>
<td>$w_id_p^2$</td>
<td>$w_id_p^2$</td>
<td>$\ldots$</td>
<td>$w_id_1d_p$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$</td>
<td>$w_id_1y_i$</td>
</tr>
<tr>
<td>$d_2$</td>
<td>$w_id_2y_i$</td>
</tr>
<tr>
<td>$\ldots$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$d_p$</td>
<td>$w_id_py_i$</td>
</tr>
</tbody>
</table>

where the symmetric elements in the lower half are not shown.

Contributions from an arbitrary number of rows from $A$ can be accumulated in $C$ and $b$ (use Double precision words, if number of rows is large).
Straight line fit

Example: track fit of \( y \) (measured) vs. abscissa \( x \)

\[ y_i = a_0 + a_1 \cdot x_i \]

Matrix \( A \) and parameter vector \( a \)

\[
A = \begin{pmatrix}
1 & x_1 \\
1 & x_2 \\
\vdots & \vdots \\
1 & x_n
\end{pmatrix} \quad a = \begin{pmatrix}
a_0 \\
a_1
\end{pmatrix}
\]

Weight matrix is diagonal \((independent)\) measurements:

\[
C = A^T W A = \left( \frac{\sum w_i}{\sum w_i x_i} \right) \quad b = A^T W y = \left( \frac{\sum w_i y_i}{\sum w_i x_i y_i} \right)
\]

If one measured \( y_i \)-value is shifted (biased), then

- parameters biased, and usually \( \chi^2 \)-value very high
The full line is a straight line fit to three well aligned data points (black dots).
The dashed curve is the straight line fit, if the middle point is "badly aligned" (circle).
Recipe for robust least square fit

Assume estimate for the standard error of $y_i$ (or of $r_i$) to be $s_i$.
Do least square fit on observations $y_i$, yielding fitted values $\hat{y}_i$, and residuals $r_i = y_i - \hat{y}_i$.

- ”Clean” the data by pulling outliers towards their fitted values: winsorize the observations $y_i$ and replace them by pseudo-observations $y_i^*$:

$$y_i^* = y_i, \quad \text{if } |r_i| \leq c s_i,$$
$$= \hat{y}_i - c s_i, \quad \text{if } r_i < -c s_i,$$
$$= \hat{y}_i + c s_i, \quad \text{if } r_i > +c s_i.$$

The factor $c$ regulates the amount of robustness, a good choice is $c = 1.5$.

- Refit iteratively: the pseudo-observations $y_i^*$ are used to calculate new parameters and new fitted values $\hat{y}_i$. 
Least squares and Maximum Likelihood method

Example: straight line fit of \( y \) (measured data) vs. abscissa \( x \)

\[
y_i = a_0 + a_1 \cdot x_i.
\]

In the Maximum Likelihood method, assuming a Gaussian distribution of the data:

\[
p(x_i|a_0, a_1) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left( - \frac{(y_i - a_0 - a_1 x_i)^2}{2\sigma_i^2} \right),
\]

the Likelihood function is

\[
\mathcal{L}(a_0, a_1) = p(x_1|a_0, a_1) \cdot p(x_2|a_0, a_1) \cdots p(x_n|a_0, a_1) = \prod_{i=1}^{n} p(x_i|a_0, a_1).
\]

Maximizing the \( \mathcal{L}(a_0, a_1) \) w.r.t. \( a_0, a_1 \) is equivalent to minimizing 2 times the negative logarithm

\[
-2 \ln \mathcal{L}(a_0, a_1) = \sum_{i=1}^{n} \frac{(y_i - a_0 - a_1 x_i)^2}{\sigma_i^2} + \text{const}.
\]
Relation between $\chi^2$ and P-value

Assume $x$ follows the density $f(x)$. The cumulative probability $F(x)$ is defined as integral:

$$\int_{-\infty}^{x} f(x') \, dx' = F(x) = u.$$

If the random variable $x$ is transformed to the random variable $u$, then the random variable $u$ (and also $1-u$) will follow the uniform distribution $U(0,1)$.

For the $\chi^2$ distribution: probability $P = 1 - F_n(\chi^2)$ should follow a uniform distribution ($n =$ number of degrees of freedom).
Linear least squares

The least squares principle 2
   Systems of linear equations 3

Linear least squares 4
   Least Squares solution 5
   Properties of solution 6
   Simplification for independent (=uncorrelated) data 7

Properties of least square estimates 8
   Gauß-Markoff Theorem 9
   Sum of squares of residuals 10
   Summary of properties 12

Independent data 13
   Normal equations for independent data 14
   Straight line fit 15
   Recipe for robust least square fit 17
   Least squares and Maximum Likelihood method 18
   Relation between $\chi^2$ and P-value 19