Linear least squares

Volker Blobel – University of Hamburg
March 2005

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 Δy_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:

$$S = \sum_{i=1}^{n} \Delta y_i^2 \qquad S = \sum_{i=1}^{n} \left(\frac{\Delta y_i}{\sigma_i} \right)^2 \qquad S = \Delta \boldsymbol{y}^T \boldsymbol{V}^{-1} \Delta \boldsymbol{y}$$

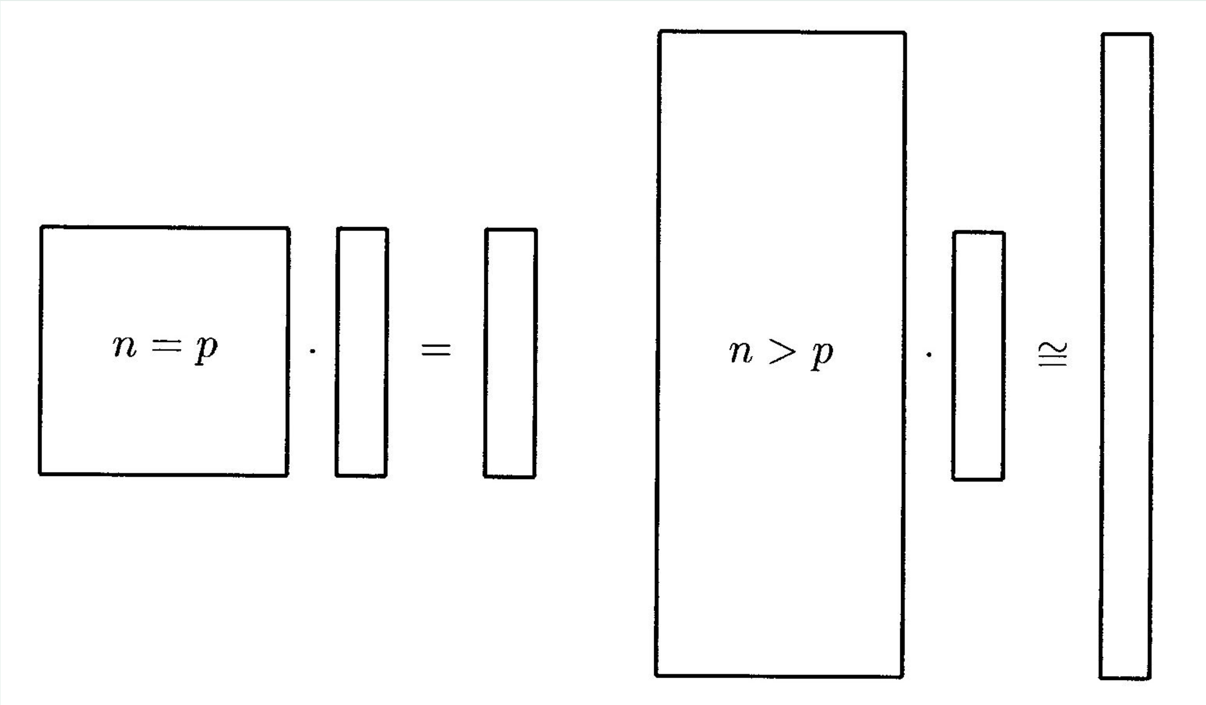
Example: mean value y of n measured values y_i:

$$S = \sum_{i=1}^{n} (y - y_i)^2 = \text{ minimum} \qquad \hat{y} = \sum_{i=1}^{n} y_i / n \quad \text{follows from } \operatorname{grad} S = 0$$

Systems of linear equations

Linear model: $\boldsymbol{A} \cdot \boldsymbol{a} = \boldsymbol{y}$ $\boldsymbol{A} \cdot \boldsymbol{a} \cong \boldsymbol{y}$

with n elements of the measured vector $\boldsymbol{y}$ and p elements of the parameter vector $\boldsymbol{a}$.



Linear least squares

The model of Linear Least Squares: $\boldsymbol{y} \cong \boldsymbol{A}\,\boldsymbol{a}$

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

Least Squares Principle: minimize the expression

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

with respect to $\boldsymbol{a}$.

Least Squares solution

Derivatives of expression $S(\boldsymbol{a})$:

$$\frac{1}{2}\,\mathrm{grad}\,S = \frac{1}{2}\frac{\partial S}{\partial \boldsymbol{a}} = \left(-\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{y} + \left(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\right)\boldsymbol{a}\right)$$

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

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

$$-\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{y} + \left(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\right)\boldsymbol{a} = 0$$

is linear transformation of the data vector $\boldsymbol{y}$:

$$\hat{\boldsymbol{a}} = \left(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\right)^{-1}\ \boldsymbol{A}^T\boldsymbol{W}\,\boldsymbol{y} \qquad = \boldsymbol{B}\boldsymbol{y}$$

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

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

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

Solution vector $\boldsymbol{a}$ and covariance matrix $\boldsymbol{V}[\boldsymbol{y}]$ 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: $E[\boldsymbol{y}] = \boldsymbol{A}\,\boldsymbol{a}_{\text{true}}$ ($\boldsymbol{a}_{\text{true}}$ = true parameter vector)
- Covariance matrix $\boldsymbol{V}[\boldsymbol{y}]$ is known (correct) and finite

Properties:

- Estimated parameters are unbiased:

$$E[\hat{\boldsymbol{a}}] = \left(\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A}\right)^{-1} \boldsymbol{A}^T \boldsymbol{W}\, E[\boldsymbol{y}] = \boldsymbol{a}_{\text{true}}$$

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

Properties are not valid, if conditions violated.

Simplification for independent (=uncorrelated) data

…assuming same variance σ^2 for all data.

Covariance matrix and weight matrix are diagonal:

$$\boldsymbol{V}(\boldsymbol{y}) = \sigma^2 \boldsymbol{I}_n \qquad\qquad \boldsymbol{W} = \frac{1}{\sigma^2}\boldsymbol{I}_n$$

($\boldsymbol{I}_n$ is n-by-n unit matrix).

$$\begin{aligned} \text{solution} \quad \hat{\boldsymbol{a}} &= \boldsymbol{C}^{-1}\boldsymbol{A}^T\boldsymbol{y} \qquad \text{with} \quad \boldsymbol{C} = \boldsymbol{A}^T\boldsymbol{A} \\ \text{covariance matrix} \quad \boldsymbol{V}(\hat{\boldsymbol{a}}) &= \sigma^2 \boldsymbol{C}^{-1} \end{aligned}$$

Note: the solution $\hat{\boldsymbol{a}}$ does not depend on σ^2, but the covariance matrix is **proportional** to σ^2.

Properties of least square estimates

Basic assumptions on the properties of the data:

1. the data are unbiased: $E[\boldsymbol{y}] = \boldsymbol{A}\boldsymbol{a}_{\text{true}}$ or $E[\boldsymbol{y} - \boldsymbol{A}\boldsymbol{a}_{\text{true}}] = 0$
2. the variances are all the same: $\boldsymbol{V}[\boldsymbol{y} - \boldsymbol{A}\boldsymbol{a}_{\text{true}}] = \sigma^2 \boldsymbol{I}_n$

(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:

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

First property: Least square estimates are unbiased.

Proof:

$$E[\hat{\boldsymbol{a}}] = \boldsymbol{C}^{-1}\boldsymbol{A}^T E[\boldsymbol{y}] = \boldsymbol{C}^{-1}\boldsymbol{A}^T\boldsymbol{A}\,\boldsymbol{a}_{\text{true}} = \boldsymbol{a}_{\text{true}}$$

Gauß-Markoff Theorem

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

$$E[\boldsymbol{a}^*] = \boldsymbol{U} E[\boldsymbol{y}] = \underbrace{\boldsymbol{U}\boldsymbol{A}}_{=\boldsymbol{I}_p} \boldsymbol{a}_{\text{true}} = \boldsymbol{a}_{\text{true}} \qquad \boldsymbol{V}[\boldsymbol{a}^*] = \sigma^2 \boldsymbol{U}\boldsymbol{U}^T$$

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

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

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

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

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

$$\begin{aligned} \boldsymbol{U}\boldsymbol{U}^T &= \boldsymbol{C}^{-1} + (\boldsymbol{U} - \boldsymbol{C}^{-1}\boldsymbol{A}^T)(\boldsymbol{U} - \boldsymbol{C}^{-1}\boldsymbol{A}^T)^T \\ &= \boldsymbol{C}^{-1} + \boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}\boldsymbol{A}\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{A}^T\boldsymbol{U}^T + \boldsymbol{C}^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{C}^{-1} \end{aligned}$$

For the covariance matrix follows:

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

Product on the right has diagonal elements ≥ 0 $\quad$ ($\rightarrow$ 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 $\hat{\boldsymbol{a}}$:

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

This equation is rewritten in terms of $\boldsymbol{a}_{\text{true}}$ (instead of $\hat{\boldsymbol{a}}$) using the matrix $\boldsymbol{U} = \boldsymbol{I}_n - \boldsymbol{A}\boldsymbol{C}^{-1}\boldsymbol{A}^T$ and the vector $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{a}_{\text{true}}$.

$$\hat{S} = (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{a}_{\text{true}})^T\boldsymbol{U}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{a}_{\text{true}}) = \boldsymbol{z}^T\boldsymbol{U}\boldsymbol{z}$$

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

Properties of $\boldsymbol{z}$: $E[\boldsymbol{z}] = 0$ and covariance matrix

$$\boldsymbol{V}[\boldsymbol{z}] = \sigma^2\boldsymbol{I}_n \quad \text{i.e.} \quad V[z_i] = E\left[z_i^2\right] = \sigma^2 \quad \text{and} \quad E[z_i z_k] = 0 \; .$$

$$\hat{S} = \sum_{i=1}^{n} \sum_{k=1}^{n} U_{ik} \; z_i \; z_k \qquad E\left[\hat{S}\right] = \sum_{i=1}^{n} U_{ii} \; E\left[z_i^2\right] = \sigma^2 \; \sum_{i=1}^{n} U_{ii} = \sigma^2 \; \text{trace}(\boldsymbol{U})$$

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

$$\begin{aligned} \text{trace}(\boldsymbol{U}) &= \text{trace}(\boldsymbol{I}_n - \boldsymbol{A}\boldsymbol{C}^{-1}\boldsymbol{A}^T) = \text{trace}(\boldsymbol{I}_n) - \text{trace}(\boldsymbol{A}\boldsymbol{C}^{-1}\boldsymbol{A}^T) \\ &= \text{trace}(\boldsymbol{I}_n) - \text{trace}(\boldsymbol{C}^{-1}\boldsymbol{A}^T\boldsymbol{A}) \\ &= \text{trace}(\boldsymbol{I}_n) - \text{trace}(\boldsymbol{I}_p) = n - p. \qquad \rightarrow \text{Proof} \end{aligned}$$

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

Special case of **Gaussian distributed measurement errors**:

$\hat{S}/\sigma^2$ distributed according to the χ^2_{n-p} distribution

to be used for goodness-of-fit test.

Summary of properties

Distribution-free properties of least squares estimates in linear problems:

1. Least square estimates are unbiased.
2. The least square estimate $\hat{\boldsymbol{a}}$ has the property

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

   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 $\boldsymbol{V}(\boldsymbol{y})$ and the weight matrix $\boldsymbol{W}$ are *diagonal*.

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

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

which are necessary for the solution

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

Note: the parameters $\boldsymbol{a}$ will be **correlated** through the model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{a}$ and the covariance matrix $\boldsymbol{V}(\hat{\boldsymbol{a}})$ will be non-diagonal.

Normal equations for independent data

The diagonal elements of the weight matrix $\boldsymbol{W}$ 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 $\boldsymbol{A}_i$, with

$$i\text{-th row of } \boldsymbol{A} \qquad \boldsymbol{A}_i = (d_1,\, d_2,\, \ldots,\, d_p) \qquad y = d_1 a_1 + d_2 a_2 + \ldots + d_p a_p$$

the contributions of this row to $\boldsymbol{C}$ and $\boldsymbol{b}$ can be written as the $p \times p$-matrix $w_i \boldsymbol{A}_i^T \cdot \boldsymbol{A}_i$ and the p-vector $w_i \boldsymbol{A}_i^T \cdot y_i$.

The **contributions of a single row** are:

	d_1	d_2	$\ldots$	d_p
d_1	$w_i d_1^2$	$w_i d_1 d_2$	$\ldots$	$w_i d_1 d_p$
d_2		$w_i d_2^2$	$\ldots$	$w_i d_2 d_p$
$\ldots$			$\ldots$	$\ldots$
d_p				$w_i d_p^2$

	y_i
d_1	$w_i d_1 y_i$
d_2	$w_i d_2 y_i$
$\ldots$	$\ldots$
d_p	$w_i d_p y_i$

,

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

Contributions from an arbitrary number of rows from $\boldsymbol{A}$ can be accumulated in $\boldsymbol{C}$ and $\boldsymbol{b}$ (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 $\boldsymbol{A}$ and parameter vector $\boldsymbol{a}$

$$\boldsymbol{A} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \boldsymbol{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$$

	1	x_i
1	w_i	$w_i x_i$
x_i		$w_i x_i^2$

	y_i
1	$w_i y_i$
x_i	$w_i x_i y_i$

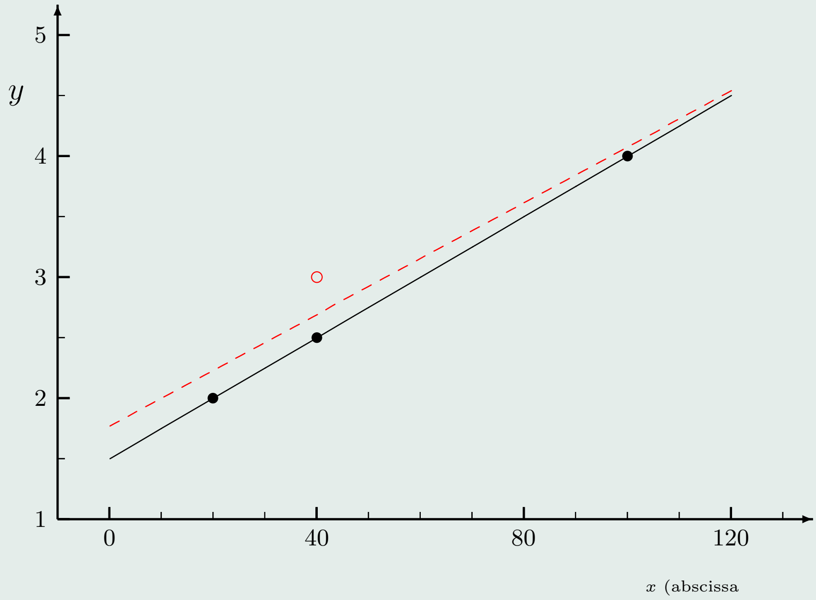
,

Weight matrix is diagonal (*independent* measurements):

$$\boldsymbol{C} = \boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A} = \begin{pmatrix} \sum w_i & \sum w_i x_i \\ \sum w_i x_i & \sum w_i x_i^2 \end{pmatrix} \qquad \boldsymbol{b} = \boldsymbol{A}^T \boldsymbol{W} \boldsymbol{y} = \begin{pmatrix} \sum w_i y_i \\ \sum w_i x_i y_i \end{pmatrix}$$

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

- parameters biased, and usually χ^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^*:

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

  The factor c regulates the amount of robustness, a goid 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

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

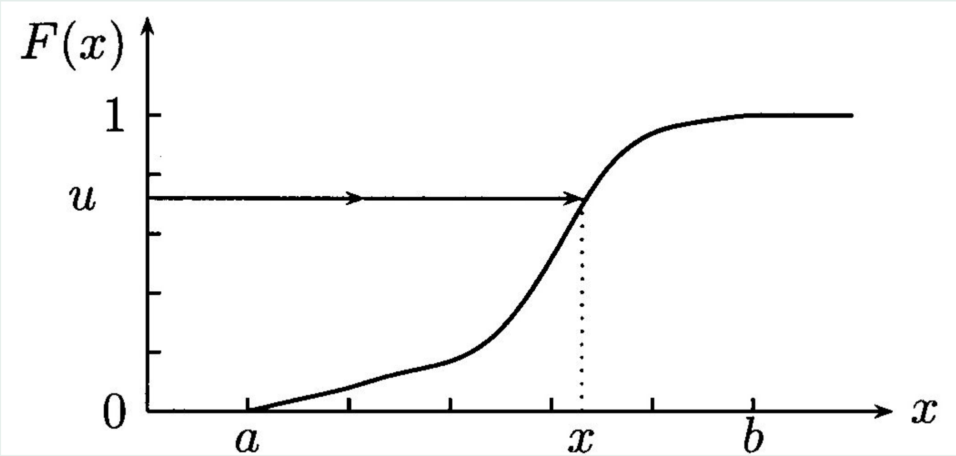
Maximizing the $\mathcal{L}(a_0, a_1)$ w.r.t. a_0, a_1 is equivalent to minimizing 2 times the negative logarithm

$$-2 \ln \mathcal{L}(a_0, a_1) = \sum_{i=1}^{n} \frac{(y_i - a_0 - a_1 x_i)^2}{\sigma_i^2} + \text{const.}$$

Relation between χ^2 and P-value

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

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



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

Linear least squares