PARAMETER ESTIMATION – 1

Parameter estimation of static models. Linear regression and its properties.

Katalin Hangos
Department of Electrical Engineering and Information Systems
University of Pannonia

Sept 2020

Contents
Lectures and tutorials

- Basic notions, Elements of random variables and mathematical statistics
- The properties of the estimates, Linear regression
- Stochastic processes, Discrete time stochastic dynamic models
- Least squares (LS) estimation by minimizing the prediction error, The properties of the LS estimation
- Special methods for LS estimation of dynamic model parameters: Instrumental variable (IV) method, Parameter estimation of dynamic nonlinear models
- Practical implementation of parameter estimation: Data checking and preparation, Evaluation of the results of parameter estimation

Lecture overview

1. Analysis of the properties of the estimates
   - Unbiased estimates
   - Confidence intervals, statistical hypothesis
2. Linear static models for parameter estimation
   - Simple linear scalar case
   - Linear models with vector valued parameters
3. Linear regression
   - The principle of LS estimation
   - The LS estimate
4. Properties of the LS estimate
   - Unbiasedness and covariance matrix
   - Evaluation of the residuals
5. Tutorial

Overview

Recall: Sample, statistics

Consider a (scalar valued) random variable ξ with probability density function $f_\xi(x)$.

- Sample
  is a collection (set) of n independent random variables

  $$S(\xi) = \{\xi_1, \xi_2, ..., \xi_n\}$$

  where every ξ_i has the same distribution as ξ.
  - the sample corresponds to a set of *measurements* about ξ
- Statistics
  is a (deterministic) function of the sample elements (a random variable itself)

  $$s(S) = F(\xi_1, \xi_2, ..., \xi_n)$$

  - a statistics is used to construct an *estimate*

Statistical properties of the sample mean – 1

Consider a **scalar valued** *random variable* ξ *with probability density function* $f_\xi(z)$ and a sample $S(\xi) = \{\xi_1, \xi_2, ..., \xi_n\}$.

Sample mean: a statistics for estimating the mean value

$$\mu(S) = \frac{1}{n}(\xi_1 + \xi_2 + ... + \xi_n)$$

Important

If the random variable ξ *has a normal or Gaussian distribution* $(\xi \sim \mathbb{N}(m, \sigma^2))$
then μ *has also a normal or Gaussian distribution.*
(For large n the distribution of μ *is approximately Gaussian).*

$$\mu \sim \mathbb{N}(m, \frac{\sigma^2}{n})$$

NOTES

It is important to notice that the **variance of the sample mean decreases with the increase of the number of measurements**. This is the reason, why we try to measure a random quantity more than once, as much times as we can.

Recall: Measured data set

Consider a **scalar valued** *random variable* ξ *with a sample* $S(\xi) = \{\xi_1, \xi_2, ..., \xi_n\}$.

Measured data set

is a collection (set) of n measurements of the sample elements $\{\xi_1, \xi_2, ..., \xi_n\}$

$$D(\xi, n) = \{x_1, x_2, ..., x_n\}$$

D is a realization of S.

Important

The measured data set contains an actual set of measurements about ξ *that are* **not** *random variables but deterministic values (a realization).*

Recall: Estimates

Consider a **scalar valued** *random variable* ξ *with a sample* $S(\xi) = \{\xi_1, \xi_2, ..., \xi_n\}$, and with a *measured data set*

$$D(\xi, n) = \{x_1, x_2, ..., x_n\}$$

Estimate
is a realization of a statistics $s(S) = F(\xi_1, \xi_2, ..., \xi_n)$

$$\hat{s}(D) = F(x_1, x_2, ..., x_n)$$

Important

an estimate is computed from the actual measurement values in the data set D

Unbiased estimates

Important (Unbiased estimate)

An estimate $\hat{s}(D)$ realizing a statistics $s(S)$ of a parameter p is **unbiased**, *if the mean value of its statistics is equal to the parameter, i.e.* $E\{s(S)\} = p$.

Important (Unbiasedness of the sample mean)

The sample mean

$$\hat{\mu}(D) = \frac{x_1 + ... + x_n}{n}$$

is an unbiased estimate of the mean value of the random variable ξ *underlying the sample* $S(\xi) = \{\xi_1, \xi_2, ..., \xi_n\}$.

NOTES

Unbiasedness is a basic requirement for an estimate.
It ensures, that one obtains a reliable estimated value when the number of measurements is increasing.

Confidence intervals

Consider a **scalar valued** *statistics (i.e. a random variable)* $s(S)$ *of a parameter* p *with probability density function* $f_s(z)$ *and a confidence (significance) level* $(1-\pi)$ $(0 < \pi << 1)$.

Important (Confidence interval)

The **confidence interval**

$$[p_m(1-\pi), p_M(1-\pi)]$$

is an interval estimation of p *on the significance level* $(1-\pi)$ *if*

$$\int_{p_m(1-\pi)}^{p_M(1-\pi)} f_s(z)dz = (1-\pi)$$

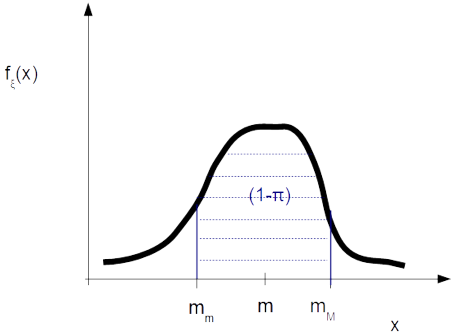
i.e. p *is in the interval* $[p_m(1-\pi), p_M(1-\pi)]$ *with probability* $(1-\pi)$

NOTES

The **notion of confidence intervals**. can be understood by recalling the meaning of a probability density function.
On the example when the parameter is the mean value of a Gaussian random variable with

$$m_m(1-\pi) \quad , \quad m_M(1-\pi)$$

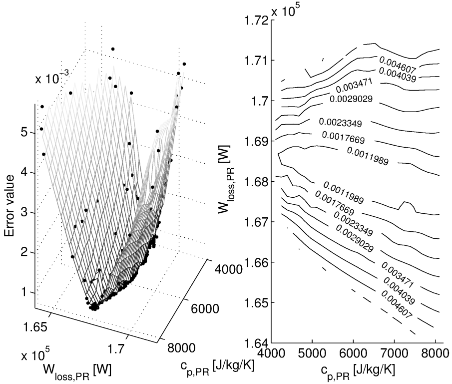
we can see the notion in the figure below:



Generalization to the vector valued case

If the parameter is vector valued, then it has a **confidence region** in the parameter space for a given significance level.

A two parameter example of confidence regions is seen in the figure below (right sub-figure).



Statistical hypothesis

Consider a **scalar valued** *statistics (i.e. a random variable) $s(S)$ and an estimate $\hat{s}(D)$ of a parameter p and a confidence (significance) level* $(1 - \pi)$ $(0 < \pi << 1)$

Important (Statistical hypothesis)

A **simple statistical hypothesis** *is a relation*

$$H_0 \;:\; p = p^*$$

for the parameter p with a given constant value p^* (we suggest that the value of p is p^*).

Testing a statistical hypothesis

Consider a (scalar valued) statistics (i.e. a random variable) $s(S)$ and an estimate $\hat{s}(D)$ of a parameter p and a confidence (significance) level $(1 - \pi)$ $(0 < \pi << 1)$ with a simple statistical hypothesis $H_0 \ : \ p = p^*$.

Important (Statistical hypothesis testing)

Hypothesis testing is to make a decision if we accept the hypothesis H_0 on the confidence (significance) level $(1 - \pi)$.

Hint: *if the estimate $\hat{s}(D)$ is within the confidence interval $[p_m^*(1 - \pi), p_M^*(1 - \pi)]$ for the parameter p then we accept the hypothesis H_0 on the confidence (significance) level $(1 - \pi)$.*

Overview

Recall: Model types

$$y = \mathcal{M}(x, p)$$

- **linear in parameters**

$$\mathcal{M}(x, p) = p^T \mathcal{F}(x)$$

  where $\mathcal{F}(x)$ is a possibly nonlinear function of the independent variable vector x

- *dynamic*
  discrete time index $k = 0, 1, ..., K, ...$ such that

$$y(k) = \mathcal{M}(x(k), x(k-1), ..., x(k-K); p) \quad , \quad k = K, K+1, ..., n$$

Simple linear scalar case: model form

Consider a (scalar valued) dependent variable y with a scalar independent variable x and a scalar parameter p.

Linear model

$$y^{(M)} = p \cdot x$$

Measurements (independent!):

$$Y = \{y_1, y_2, ..., y_m\} \text{ for fixed } X = \{x_1, x_2, ..., x_m\}$$

such that $y_j = p \cdot x_j + \varepsilon_j$, where ε_j, $j = 1, ..., m$ are independent identically distributed random variables with p.d.f. $f_\varepsilon(z)$.

Important

Sample for the measurement error

$$S(\varepsilon) = \{(y_1 - px_1), \ ..., (y_m - px_m)\}$$

Simple linear scalar case: residuals

Consider a **scalar valued dependent variable** y *with a* **scalar independent variable** x *and a* **scalar parameter** p

$$y^{(M)} = p \cdot x$$

and with independent measurements such that $y_j = p \cdot x_j + \varepsilon_j$, $\varepsilon_j, \ j = 1, ..., m$ are *independent identically distributed random variables*

Measured data set*:* $D_m = \{(y_j; x_j) \mid j = 1, ..., m\}$

Residuals*:*

$$r_j = y_j - y_j^{(M)} = y_j - p \cdot x_j$$

Sample *for the estimation of the residual properties:*
$S(\varepsilon) = \{r_1, r_2, ..., r_m\}$ *where every* r_i *has the same distribution as* ε*.*

NOTES

Residual: characterizes the deviation of the measured values from the "ideal", model-predicted values.
Residuals play a major role in parameter estimation. One needs to have a so called **predictive model** to be able to compute the residual values.

$$y^{(M)} = \mathcal{M}(x, p) \quad , \quad r = y - y^{(M)}$$

Parameter estimation based on residuals: aims to find such a parameter value $\hat{p}$ that minimizes the magnitude of the residuals (i.e. the deviation from the "ideal", model-predicted values.

Static linear models: vector valued parameter – 1

Static linear model
that is linear in parameters $p \in \mathbb{R}^n$ and also in independent variables $x \in \mathbb{R}^n$ but has a single dependent variable y

$$y^{(M)} = x^T p = \sum_{i=1}^{n} x_i p_i$$

Measured data: *m independent measurements*

$$y_j = \sum_{i=1}^{n} x_{ji} p_i + \varepsilon_j \quad , \quad D_m = \{(y_j; x_{j1}, ..., x_{jn}) \mid j = 1, ..., m\}$$

with fixed independent variable values x_{ji}, $j = 1, ..., m$; $i = 1, ..., n$ *and independent identically distributed measurement errors* ε_j.

Static linear models: vector valued parameter – 2

Static linear model

that is linear in parameters $p \in \mathbb{R}^n$ and also in independent variables $p \in \mathbb{R}^n$ but has a single dependent variable y

$$y^{(M)} = x^T p = \sum_{i=1}^{n} x_i p_i$$

$$y_j = \sum_{i=1}^{n} x_{ji} p_i + \varepsilon_j \quad , \quad D_m = \{(y_j; x_{j1}, ..., x_{jn}) \mid j = 1, ..., m\}$$

Residuals

$$r_j = y_j - y_j^{(M)} = y_j - x^{(j)T} p$$

where $x^{(j)}$ is the jth fixed independent variable set.

Overview

Simple linear scalar case: LS estimation – 1

Linear model

$$y^{(M)} = p \cdot x$$

Residuals

$$r_j = y_j - y_j^{(M)} = y_j - p \cdot x_j \quad , \quad j = 1, ..., m$$

Loss function: *squared deviation from the model*

$$V(p; X) = \sum_{j=1}^{m} r_j^2 = \sum_{j=1}^{m} (y_j - p \cdot x_j)^2$$

with fixed X.

Important (LS principle)

The least squares (LS) estimation principle: choose the parameter estimate $\hat{p}$ such that the quadratic function $V(p)$ is minimal.

NOTES

Least squares principle

In the general case, the parameter estimation procedure using the residuals tries to find such a parameter estimate $\hat{p}$ that minimizes the magnitude of the residuals.

We can consider the residuals

$$r_j = y_j - y_j^{(M)} = y_j - p \cdot x_j \quad , \quad j = 1, \ldots, m$$

as entries of a vector, then its magnitude is measured using a suitable **vector norm**.

The loss function is then nothing else, but the value of this norm.

The least squares (LS) estimation principle is obtained when we use the 2-norm of a vector, i.e.:

$$V(p; X) = \sum_{j=1}^{m} r_j^2 = \sum_{j=1}^{m} (y_j - p \cdot x_j)^2$$

This choice is practical, because

- it punishes the deviations both with positive and negative sides,
- it is an analytically tractable smooth (quadratic) function of the unknown parameters.

Simple linear scalar case: LS estimation – 2

Loss function: squared deviation from the model

$$V(p;X) = \sum_{j=1}^{m} r_j^2 = \sum_{j=1}^{m} (y_j - p \cdot x_j)^2$$

with *fixed* X.

Choose the parameter estimate $\hat{p}$ such that the quadratic function $V(p)$ is minimal.

Important

Solution: *using optimization*

$$\frac{dV(p)}{dp} = -2 \cdot \sum_{j=1}^{m} x_j (y_j - p \cdot x_j) = 0 \quad \Rightarrow \quad \hat{p} = \frac{1}{\sum_{j=1}^{m} x_j x_j} \cdot \sum_{j=1}^{m} x_j y_j$$

The estimate is a linear function of the measured y_j-s.

Parameter estimation of linear static models – 1

Problem statement

Given:

- A model that is linear in parameters $p \in \mathbb{R}^n$

$$y^{(M)} = x^T p = \sum_{i=1}^{n} x_i p_i$$

  where $x \in \mathbb{R}^n$ are deterministic independent variables (measured and set) and $y^{(M)} \in \mathbb{R}$ is the model output, measured value y is a random variable with **measurement error**.

- From m $(m \geq n)$ measurements we form

$$y = \begin{bmatrix} y_1 \\ y_2 \\ .. \\ y_m \end{bmatrix} \quad , \quad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

Parameter estimation of linear static models – 2

- Consider a weighted **quadratic loss function** V

$$V(p;X) = r^T W r = \sum_{i=1}^{m} \sum_{j=1}^{m} r_i W_{ij} r_j$$

$$r_j = y_j - y_j^{(M)} = y_j - x^{(j)T} p \quad , \quad j = 1, ..., m$$

where r is the *residual vector* and W is a weighting matrix (often $W = I$)

Important (Least squares (LS) estimate)

The LS estimate $\hat{p}$ of the parameters p minimizes V.

The minimum of V is at $\frac{\partial V}{\partial p} = 0$ with $W = I$

$$\hat{p} = (X^T X)^{-1} X^T y$$

The estimate is a linear function of the measured y_j-s.

NOTES

LS estimate with $W = I$
If one chooses the weighting matrix W to be the unit matrix, then the quadratic loss function specializes to

$$\begin{aligned} V(p;X) &= r^T W r = \sum_{i=1}^{m} \sum_{j=1}^{m} r_i W_{ij} r_j = \sum_{i=1}^{m} r_i^2 \\ r_j &= y_j - y_j^{(M)} = y_j - x^{(j)T} p \quad , \quad j = 1, ..., m \end{aligned}$$

that is simply the sum of squares of the individual residual entries.

Overview

Recall: Vector-valued random variables

Given a vector valued random variable ξ

$$\xi \ : \ \ \xi(\omega), \quad \omega \in \Omega, \quad \xi(\omega) \in \mathbb{R}^{\mu}$$

Its **mean value** $m \in R^{\mu}$ *is a real vector.*

Its **variance** $COV\{\xi\}$ is a square real matrix, the *covariance matrix*:

$$COV\{\xi\} = E\{(\xi - E\{\xi\})(\xi - E\{\xi\})^T\}$$

Covariance matrices are positive definite symmetric matrices:

$$z^T COV\{\xi\} z \geq 0 \quad , \quad \forall z \in \mathbb{R}^{\mu}$$

Recall: Linearly transformed random variables

Let us transform the vector-valued random variable $\xi(\omega) \in R^n$ using the non-singular square transformation matrix $T \in R^{n \times n}$:

$$\eta = T\xi$$

The properties of the vector-valued random variable η:

$$E\{\eta\} = TE\{\xi\} \quad , \quad COV\{\eta\} = TCOV\{\xi\}T^T$$

If the random variable ξ has a Gaussian distribution $N(m_\xi, \Delta_\xi)$ with mean value m_ξ and covariance matrix Δ_ξ, then the transformed random variable η will also be Gaussian $N(m_\eta, \Delta_\eta)$, where

$$m_\eta = Tm_\xi \quad , \quad \Delta_\eta = T\Delta_\xi T^T$$

The distribution of the LS estimate

The LS estimate

$$\hat{p} = (X^T X)^{-1} X^T y$$

with X being a fixed independent variable value matrix, and the measured dependent variable vector y is

$$y = X \cdot p + \varepsilon$$

where the measurement errors $\varepsilon_j,\ j = 1, ..., m$ are independent identically distributed random variables with p.d.f. $f_\varepsilon(z)$ and zero mean $E\{\varepsilon\} = 0$.

$$\hat{p} = (X^T X)^{-1} X^T (X \cdot p + \varepsilon) = p + (X^T X)^{-1} X^T \varepsilon$$

Important (Unbiasedness of the LS estimate)

The LS estimate is unbiased, because $E\{\hat{p}\} = p$.

The covariance matrix of the LS estimate

The LS estimate

$$\hat{p} = (X^T X)^{-1} X^T y = p + (X^T X)^{-1} X^T \varepsilon$$

with X being a fixed independent variable value matrix resulting in the transformation matrix $T = (X^T X)^{-1} X^T$ (from ε to $\hat{p}$).

The covariance matrix of the estimate is

$$COV\{\hat{p}\} = (X^T X)^{-1} \sigma_\varepsilon^2$$

where σ_ε^2 is the variance of the measurement errors.

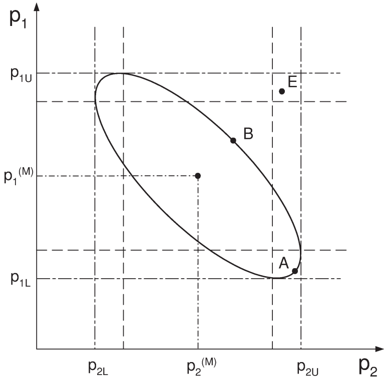
Important (Experiment design)

We can influence the covariance matrix of the estimate by choosing the fixed values of the independent variables properly.

The confidence region of the parameters in the vector valued case

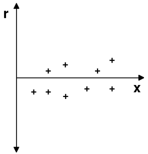
If the parameter is vector valued with a given mean value vector and covariance matrix, then it has a **confidence region of ellipsoidal shape** in the parameter space for a given significance level.

A two parameter example of a confidence region is seen in the figure below.

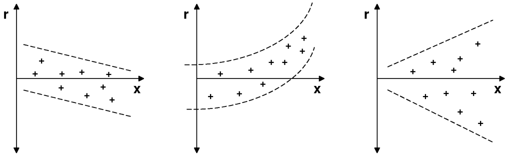


Evaluation of the residuals

For unbiased estimates the residuals should be realizations of independent identically distributed random variables with zero mean.



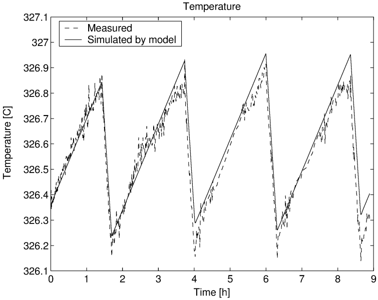
random



non random

Evaluation of the residuals a real industrial example

An example of a controlled pressure signal (both measured and model-predicted) is seen below.



The residuals are not everywhere independent zero-mean random variables. It indicates a modelling problem.

Tutorial problems – Linear regression

- A. Scalar valued parameter
- B. Vector valued parameter

Tutorial problems – A

Example (Linear regression for scalar parameter – 1)

Consider the following model that is linear in parameters:

$$y^{(M)} = px \quad (1)$$

- *How many parameters does this model have?*
  1 (scalar parameter)
- *Consider a measured data set consisting of $(y_j; x_j)$ pairs*

  $$D_5 = \{(0.5; 1.0),\ (0.2; 1.0),\ (0.0; 1.0),\ (-0.5; 1.0),\ (-0.2; 1.0)\}$$

  Compute an estimate of p if possible with its mean value and variance.
  $\hat{p} = 0.0 \quad , \quad \hat{\sigma}_p^2 = 0.0825$

Tutorial problems – A

Example (Linear regression for scalar parameter – 2)

Consider the following model that is linear in parameters:

$$y^{(M)} = px \quad (2)$$

- *Consider a measured data set consisting of* $(y_j; x_j)$ *pairs*

$$D_5 = \{(0.5; 1.0),\ (0.2; 1.0),\ (0.0; 1.0),\ (-0.2; 1.0),\ (-0.5; 1.0)\}$$

Compute an estimate of p if possible with its mean value and variance.

$\hat{p} = 0.0 \quad , \quad \hat{\sigma}_p^2 = 0.0825$

Tutorial problems – A

Example (Linear regression for scalar parameter – 2)

Consider the following model that is linear in parameters:

$$y^{(M)} = px \quad (2)$$

- *Consider a measured data set consisting of* $(y_j; x_j)$ *pairs*

  $$D_5 = \{(0.5; 1.0),\ (0.2; 1.0),\ (0.0; 1.0),\ (-0.2; 1.0),\ (-0.5; 1.0)\}$$

  Compute an estimate of p if possible with its mean value and variance.
  $\hat{p} = 0.0 \quad , \quad \hat{\sigma}_p^2 = 0.0825$

- *Evaluate the properties of the residuals (mean value, variance)*
  may not be independent – slow drift

Tutorial problems – B

Example (Linear regression for vector valued parameter – 1.1)

Consider the modified model that is linear in parameters:

$$y^{(M)} = ax + b$$

where a and b are unknown scalar parameters.

How many parameters does this model have? Construct the parameter vector p.

$2;\ \ p = [\ a\ ,\ \ b\ ]^T$

Tutorial problems – B

Example (Linear regression for vector valued parameter – 1.2)

Consider the modified model that is linear in parameters:

$$y^{(M)} = ax + b$$

where a and b are unknown scalar parameters.

Consider a measured data set consisting of $(y_j; x_j)$ *pairs*

$$D_5 = \{(0.5; 1.0),\ (0.6; 1.0),\ (0.3; 1.0),\ (-0.2; 1.0),\ (0.5; 1.0)\}$$

Construct the matrix X and the vector y needed for the estimation. Comment on the solvability of the estimation problem.

$$y = \begin{bmatrix} 0.5 \\ 0.6 \\ 0.3 \\ -0.2 \end{bmatrix} \quad , \quad X = \begin{bmatrix} 1.0 & 1.0 \\ 1.0 & 1.0 \\ 1.0 & 1.0 \\ 1.0 & 1.0 \end{bmatrix}$$

Tutorial problems – B

Example (Linear regression for vector valued parameter – 2)

Consider the model that is linear in parameters:

$$y^{(M)} = ax + b$$

where a and b are unknown scalar parameters.

Consider a **modified measured data set** *consisting of $(y_j; x_j)$ pairs*

$$D_4 = \{(0.5; 1.0),\ (0.6; 1.0),\ (0.3; 0.5),\ (0.2; 0.5)\}$$

Construct the matrix X and the vector y needed for the estimation. Comment on the solvability of the estimation problem.

$$y = \begin{bmatrix} 0.5 \\ 0.6 \\ 0.3 \\ 0.2 \end{bmatrix} \quad , \quad X = \begin{bmatrix} 1.0 & 1.0 \\ 1.0 & 1.0 \\ 0.5 & 1.0 \\ 0.5 & 1.0 \end{bmatrix}$$

HOMEWORK

Consider the following model that is linear in parameters:

$$y^{(M)} = \sum_{i+1}^{2} p_i x_i + b$$

where the unknown model parameters are p_1, p_2 and b.

- Consider a measured data set consisting of $(y_j; x_{j1}, x_{j2})$ values

  $$D_4 = \{(0.5; 1.0, 1.0),\ (0.6; 1.0, 0.9),\ (0.3; 1.0, 0.5),\ (0.2; 0.5, 1.0)\}$$

  Compute an estimate of p if possible with its mean value and covariance matrix.

- Evaluate the properties of the residuals (mean value, variance).