

“Experimental Data Processing”

Topic 1

"Introduction to statistical analysis"

Tatiana Podladchikova

Term 1B, October 2017

t.podladchikova@skoltech.ru

The basis of statistical analysis

1 Least
Square Method
LSM

2 Linear
regression



Advantages



Limitations

Transform theoretical knowledge into useful skills

The basis of statistical analysis

1 Least
Square Method
LSM

2 Linear
regression

Advantages

Limitations

Based on needs
of practical
problems

Following
topics of
the course

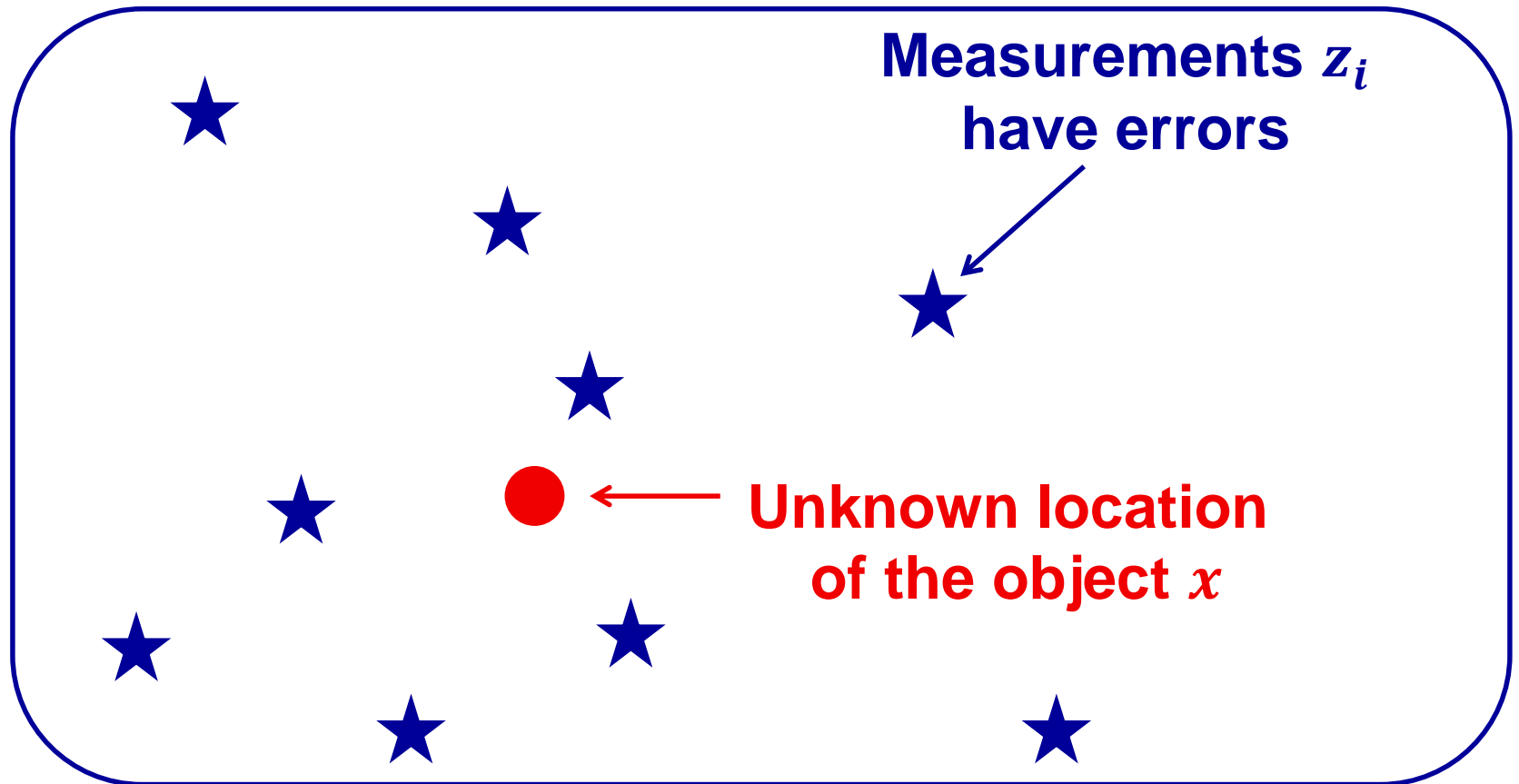
Combination of
strict techniques
with individual
approaches

Estimate the location of an unmoving object



LSM. Example 1

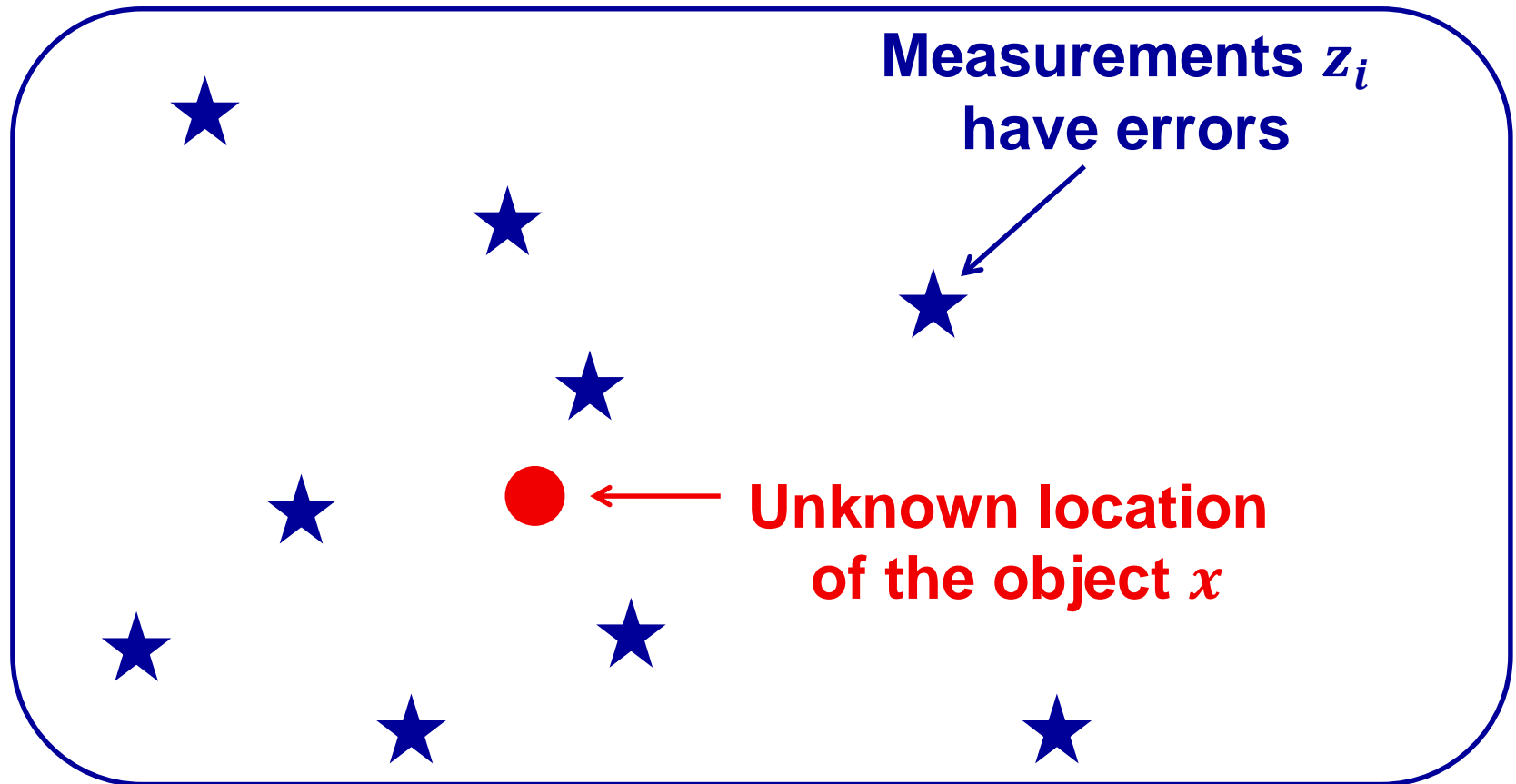
Estimate the location of an unmoving object



LSM. Example 1

Estimate the location of an unmoving object

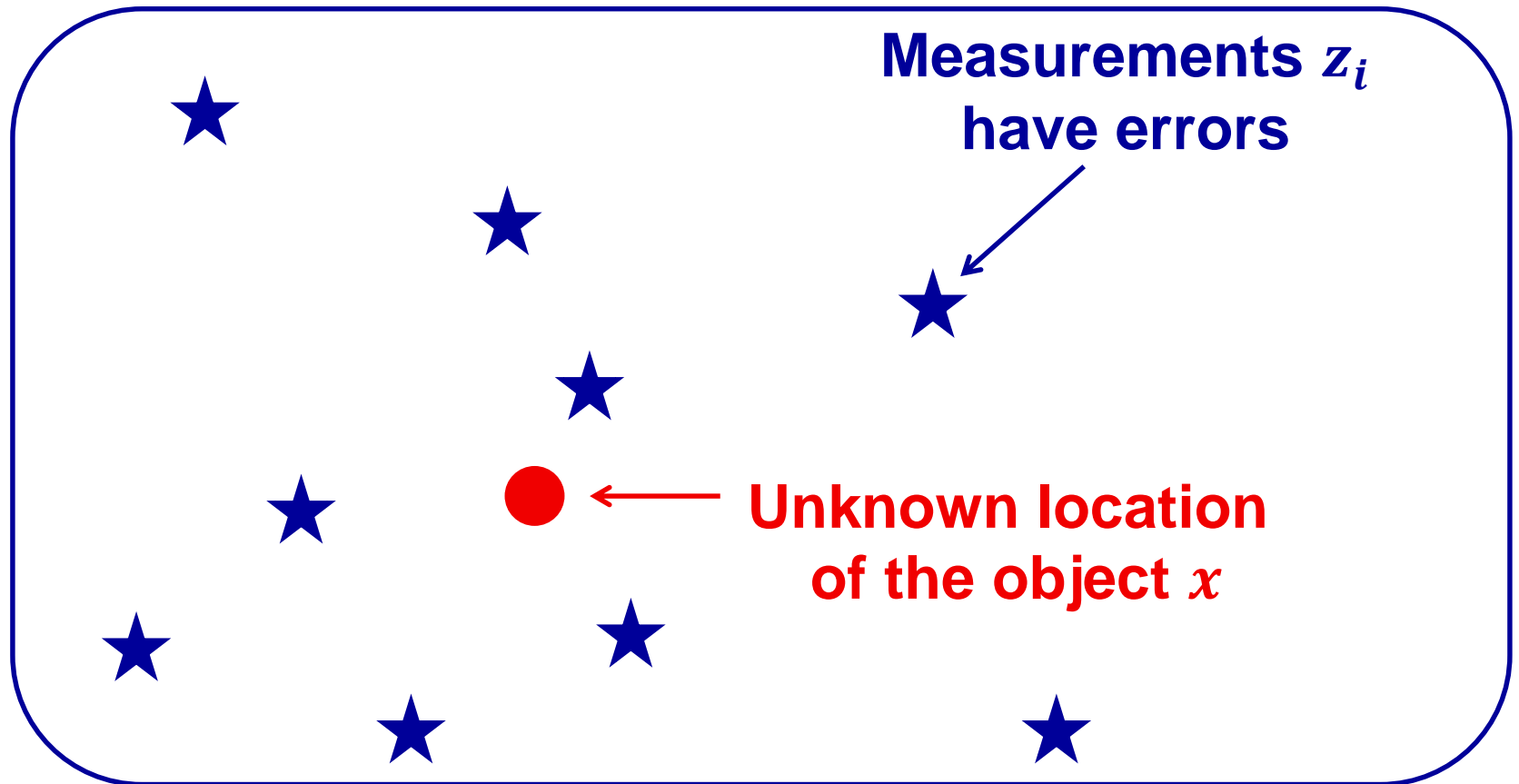
First use common sense!



LSM. Example 1

Estimate the location of an unmoving object

First use common sense!

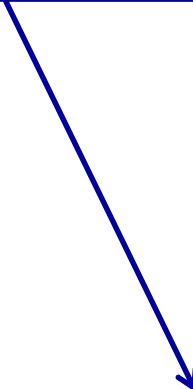


LSM. Example 1

**Solution using common sense:
to average measurements**

Estimate the location of an unmoving object

**Let's find the theoretical ground
of this solution using
least-square method (LSM)**

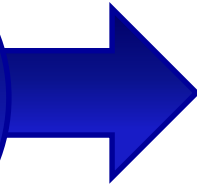


LSM. Example 1

**Solution using common sense:
to average measurements**

Estimate the location of an unmoving object

1
Let's
minimize
functional
 J



$$J = \sum_{i=1}^N (z_i - x)^2$$

z_i – measurements

x – object location

N – number of
measurements

LSM. Example 1

**Solution using common sense:
to average measurements**

Estimate the location of an unmoving object

1
Let's
minimize
functional
 J



$$J = \sum_{i=1}^N (z_i - x)^2$$

z_i – measurements
 x – object location
 N – number of
measurements

2
First
derivative
in respect
to x

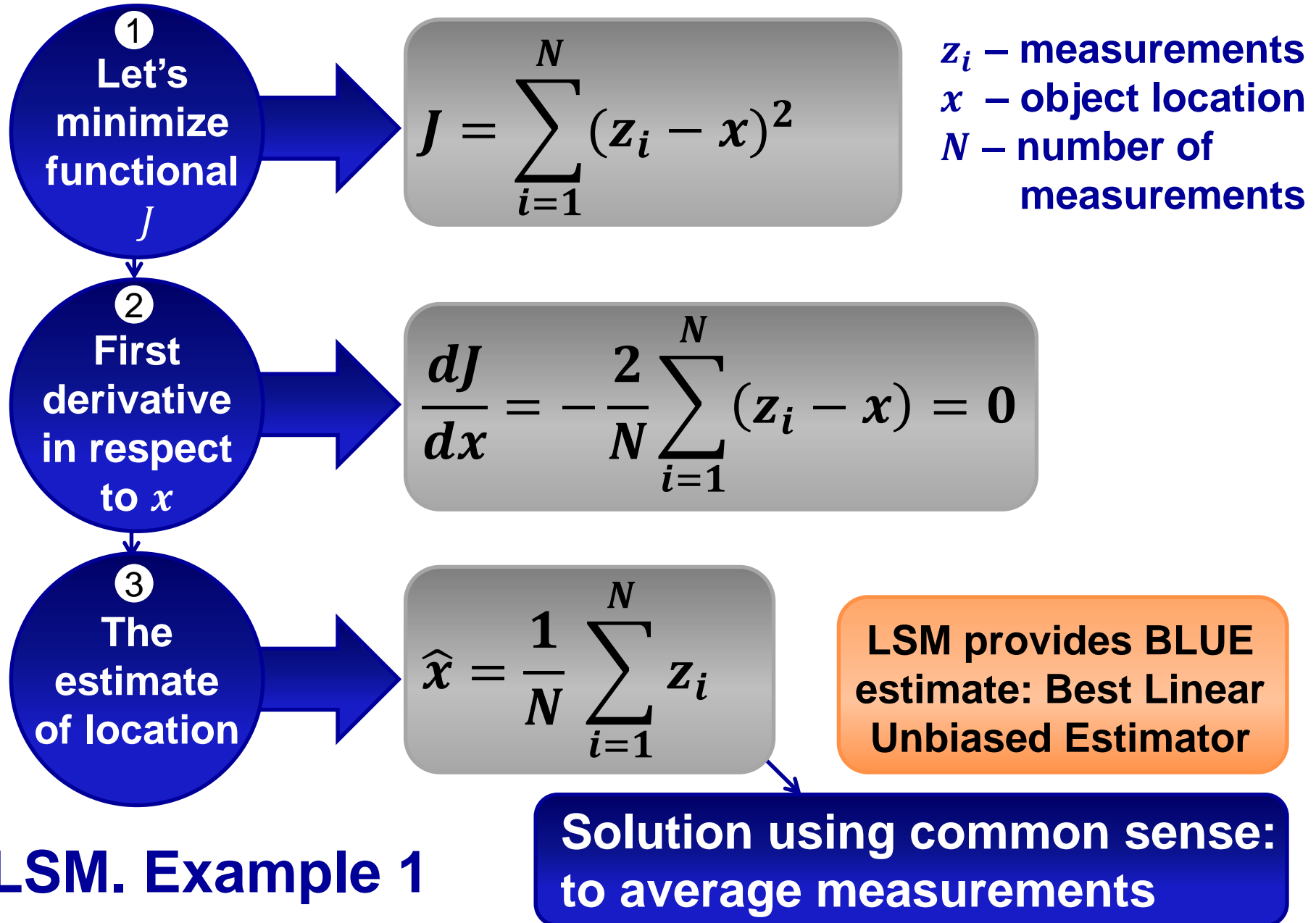


$$\frac{dJ}{dx} = -\frac{2}{N} \sum_{i=1}^N (z_i - x) = 0$$

LSM. Example 1

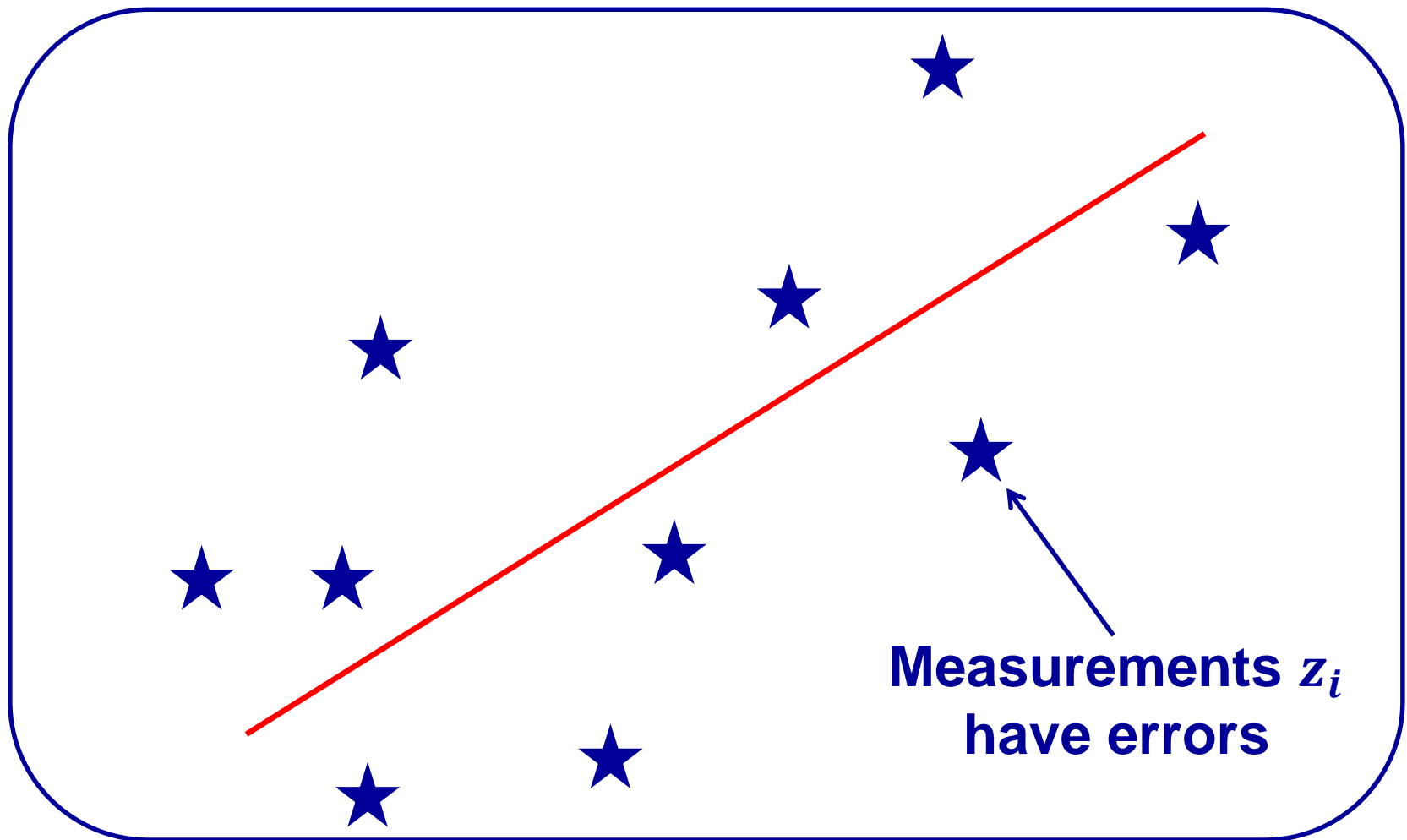
Solution using common sense:
to average measurements

Estimate the location of an unmoving object



LSM. Example 1

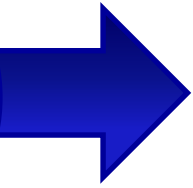
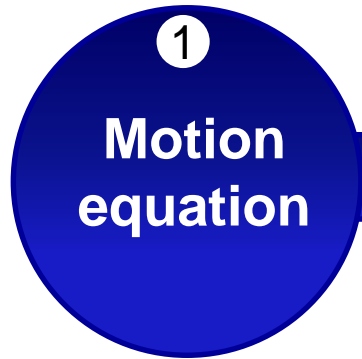
Estimate the location of a moving object



LSM. Example 2

Uniform and linear movement

Estimate the location of a moving object

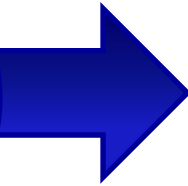


$$x_i = x_0 + V \cdot iT$$

x_0 – initial position
 V – velocity
 i – time step
 T – time interval
between
measurements

Estimate the location of a moving object

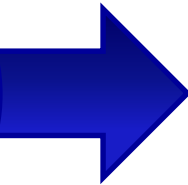
1
Motion
equation



$$x_i = x_0 + V \cdot iT$$

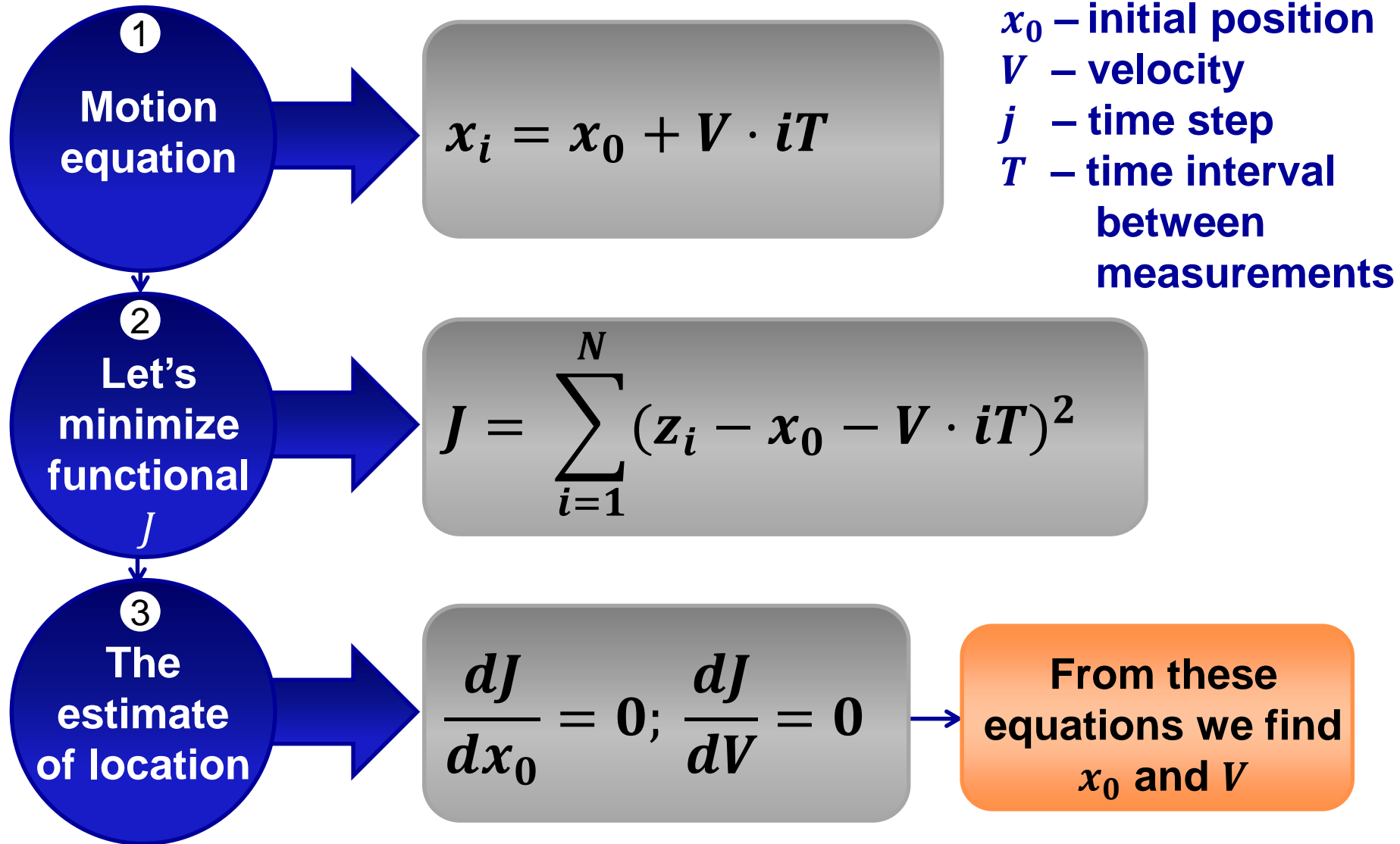
x_0 – initial position
 V – velocity
 i – time step
 T – time interval
between
measurements

2
Let's
minimize
functional
 J



$$J = \sum_{i=1}^N (z_i - x_0 - V \cdot iT)^2$$

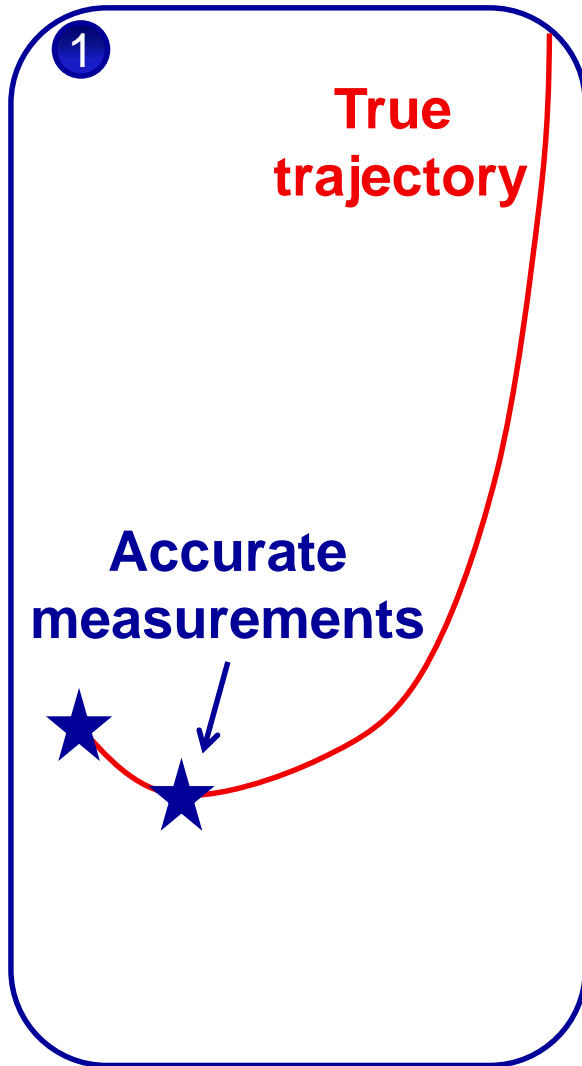
Estimate the location of a moving object



LSM. Example 2

Uniform and linear movement

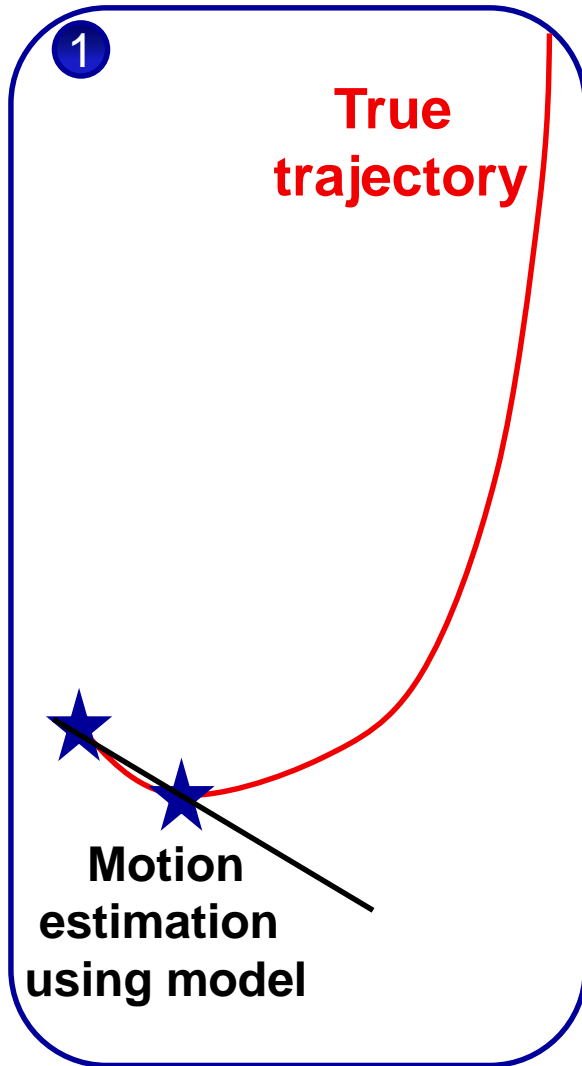
What happens if we assume the object to move along the line, but in fact it moves along the parabola?



LSM. Example 3

Unknown object dynamics

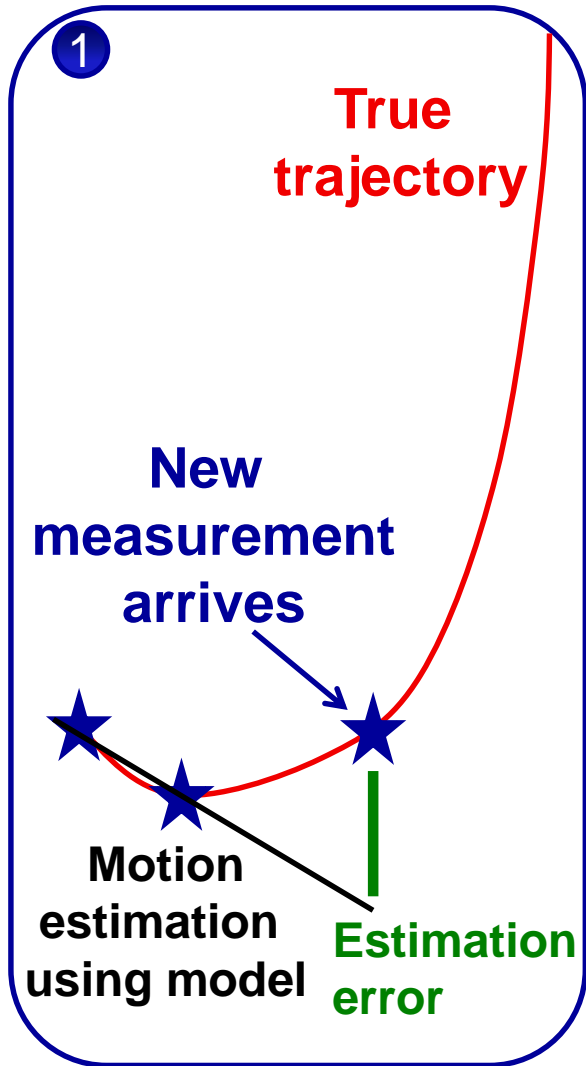
What happens if we assume the object to move along the line, but in fact it moves along the parabola?



LSM. Example 3

Unknown object dynamics

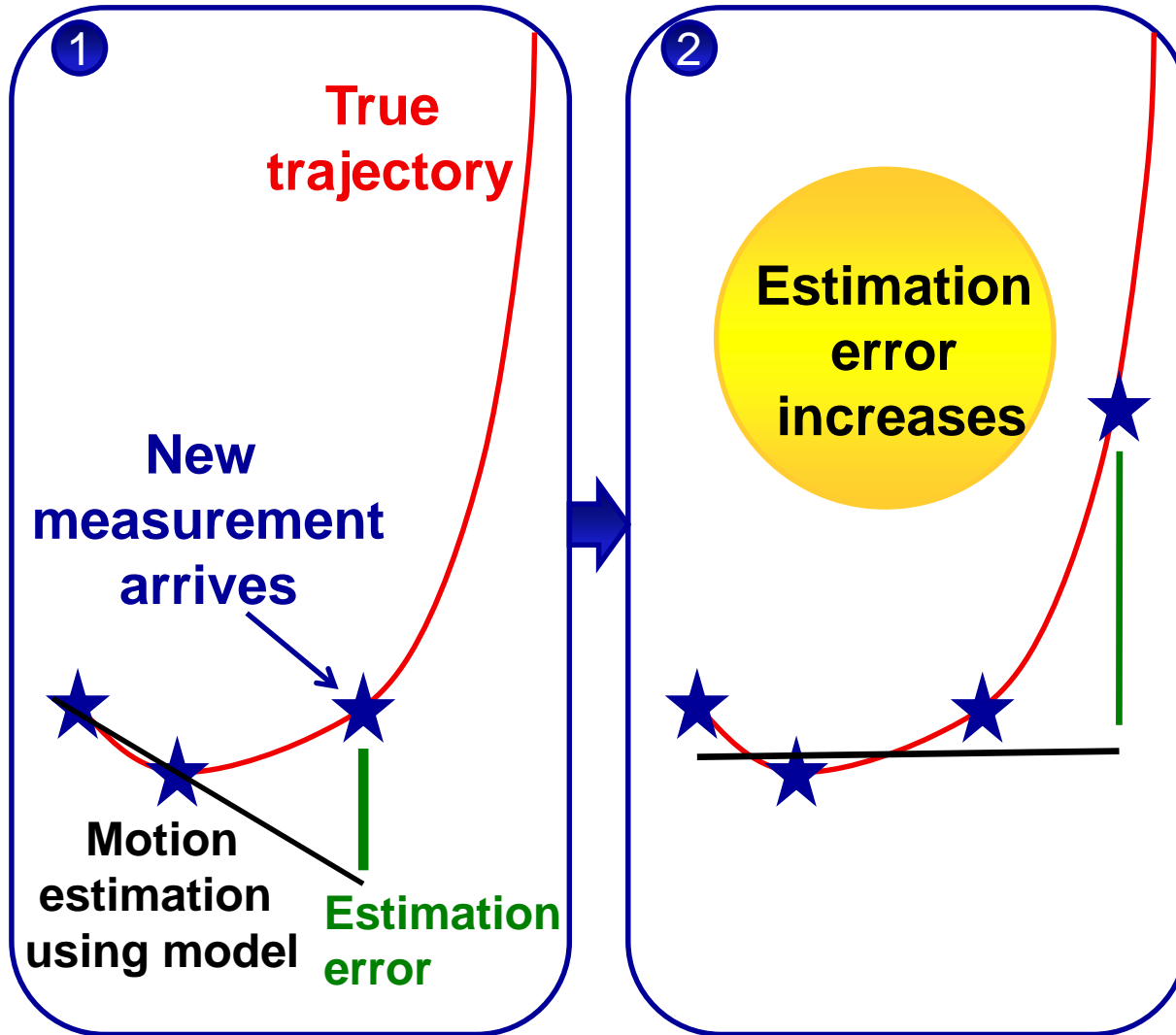
What happens if we assume the object to move along the line, but in fact it moves along the parabola?



LSM. Example 3

Unknown object dynamics

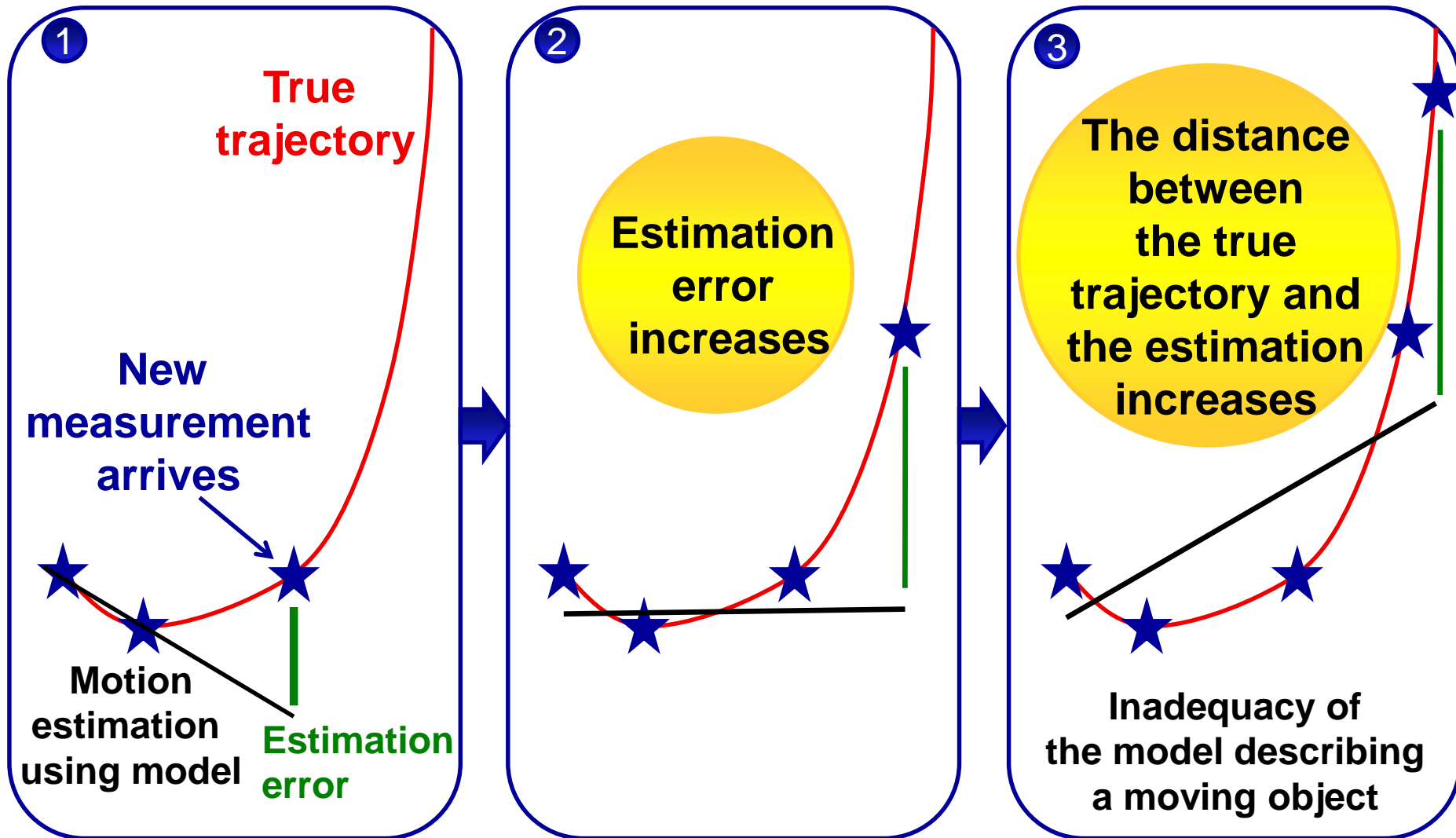
What happens if we assume the object to move along the line, but in fact it moves along the parabola?



LSM. Example 3

Unknown object dynamics

What happens if we assume the object to move along the line, but in fact it moves along the parabola?



LSM. Example 3

Unknown object dynamics

Conclusion. Advantages and limitations of LSM applications

**Process dynamics
is known**



Advantages

**LSM provides
BLUE estimate:
Best Linear
Unbiased
Estimator**

**Process dynamics
is unknown**



Limitations

**The LSM method
leads to divergence
and loses its
practical value**

Conclusion. Advantages and limitations of LSM applications

**Process dynamics
is known**



**Process dynamics
is unknown**

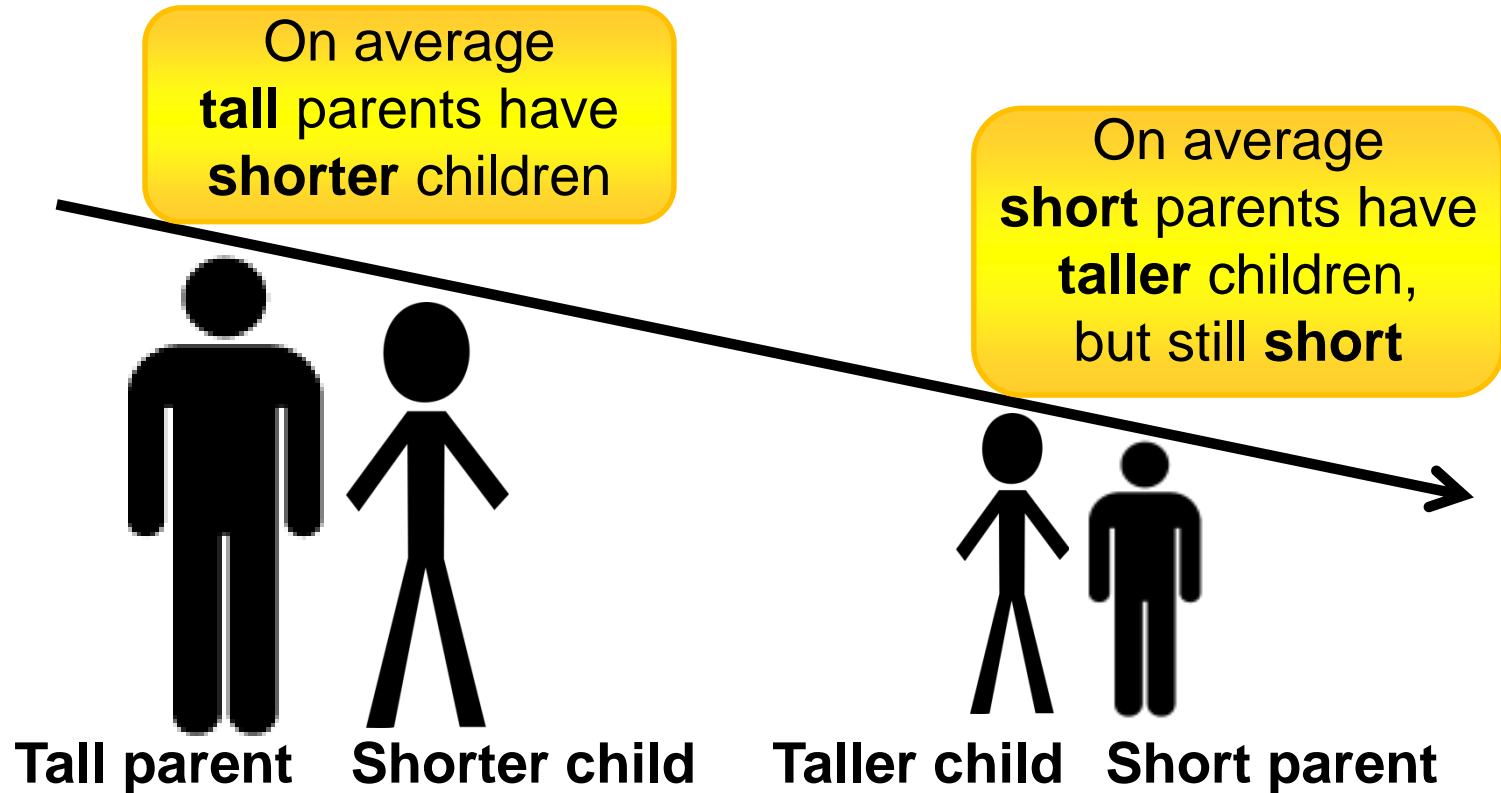


**Therefore in parallel
to optimal methods
the robust quasi optimal
methods are developed**

**Such methods can be
applied in conditions
of uncertainty
of process dynamics**

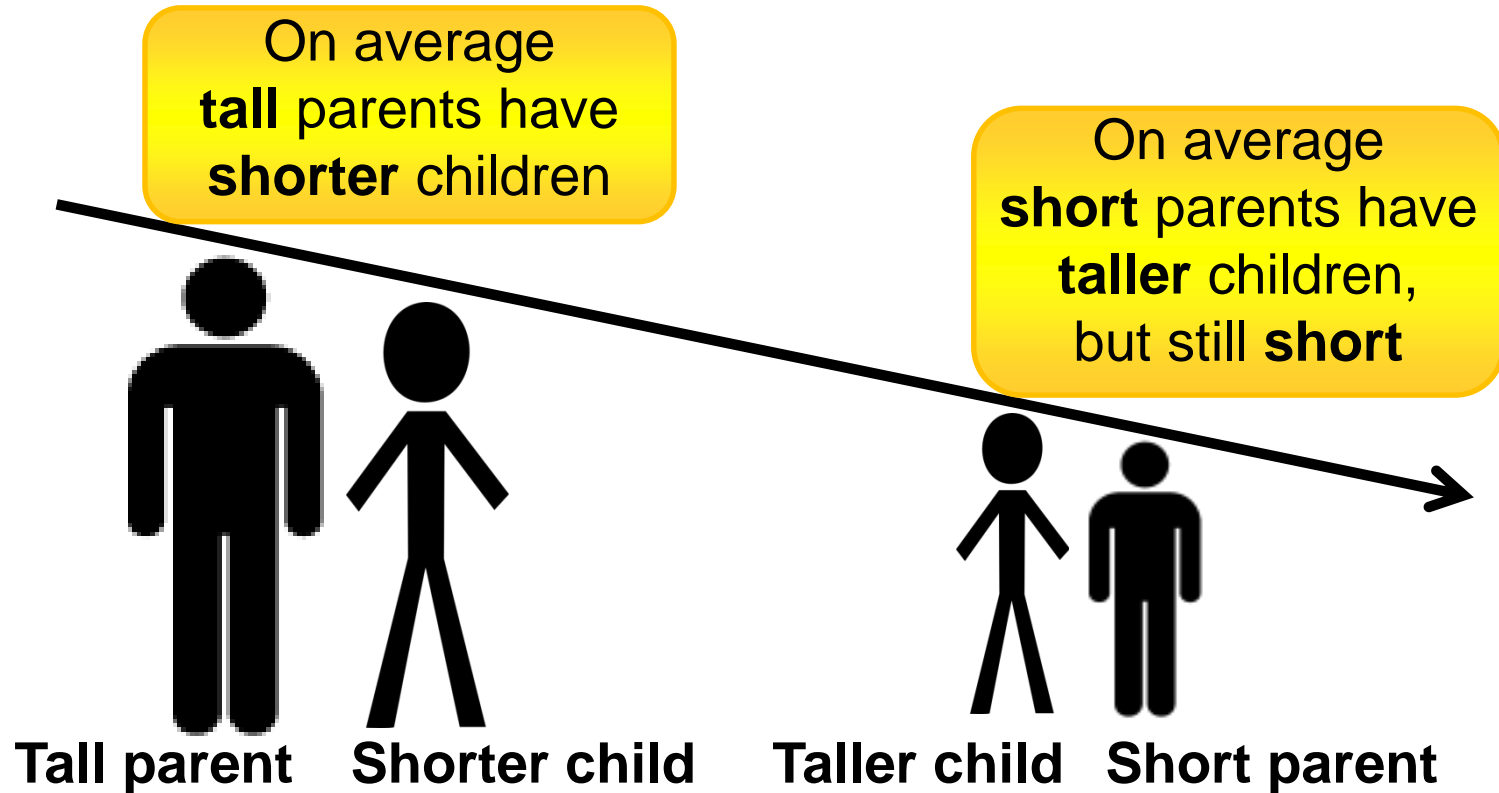
**Following
topics of
the course**

Is there any relationship between the height of parents and children?



Sir Francis Galton, 1822 –1911

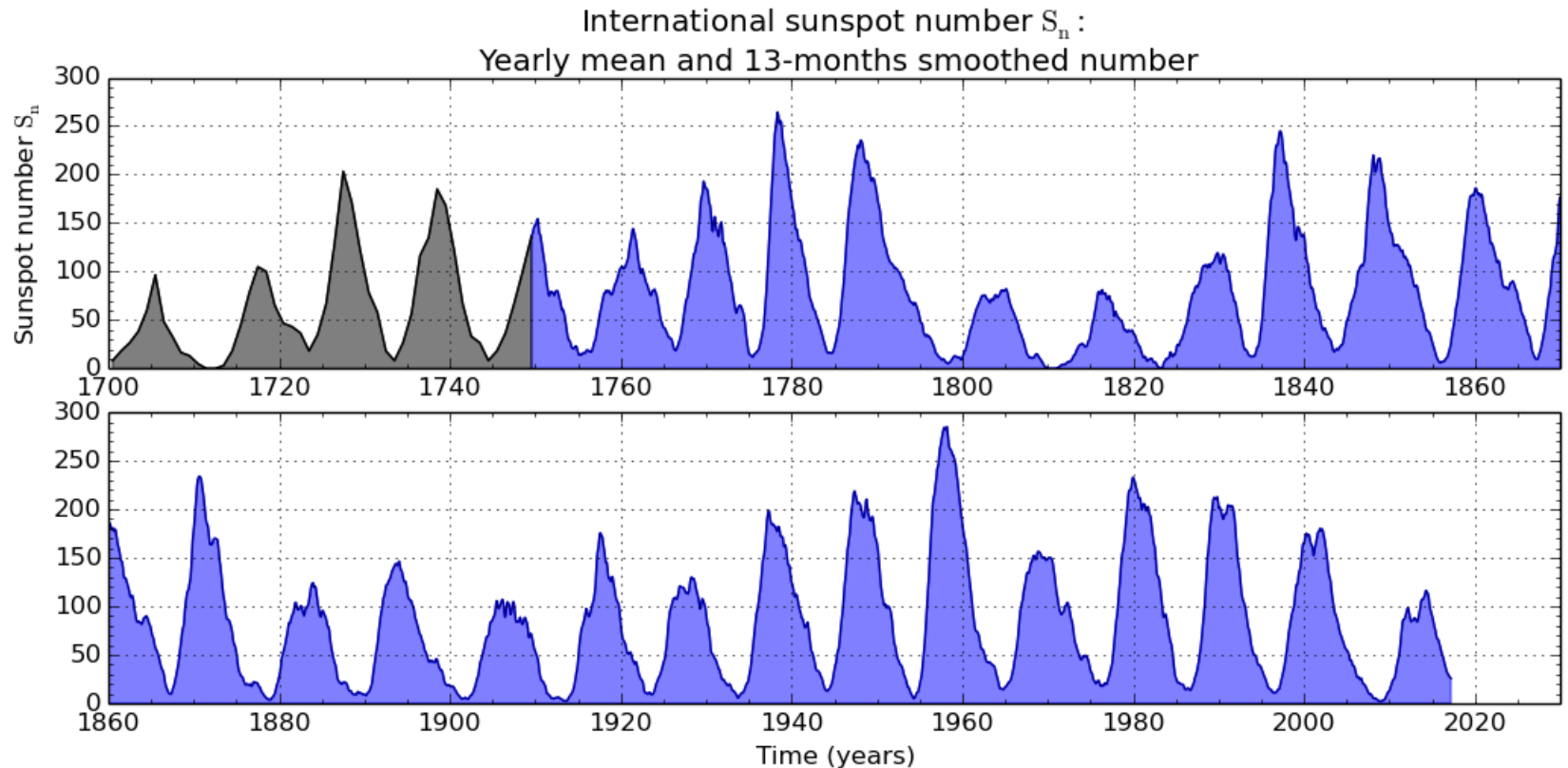
Is there any relationship between the height of parents and children?



Height of children approaches
to average height of humans
REGRESSION

Sir Francis Galton, 1822 –1911

Is there any relationship between period of current sunspot cycle and the maximum of next sunspot cycle?

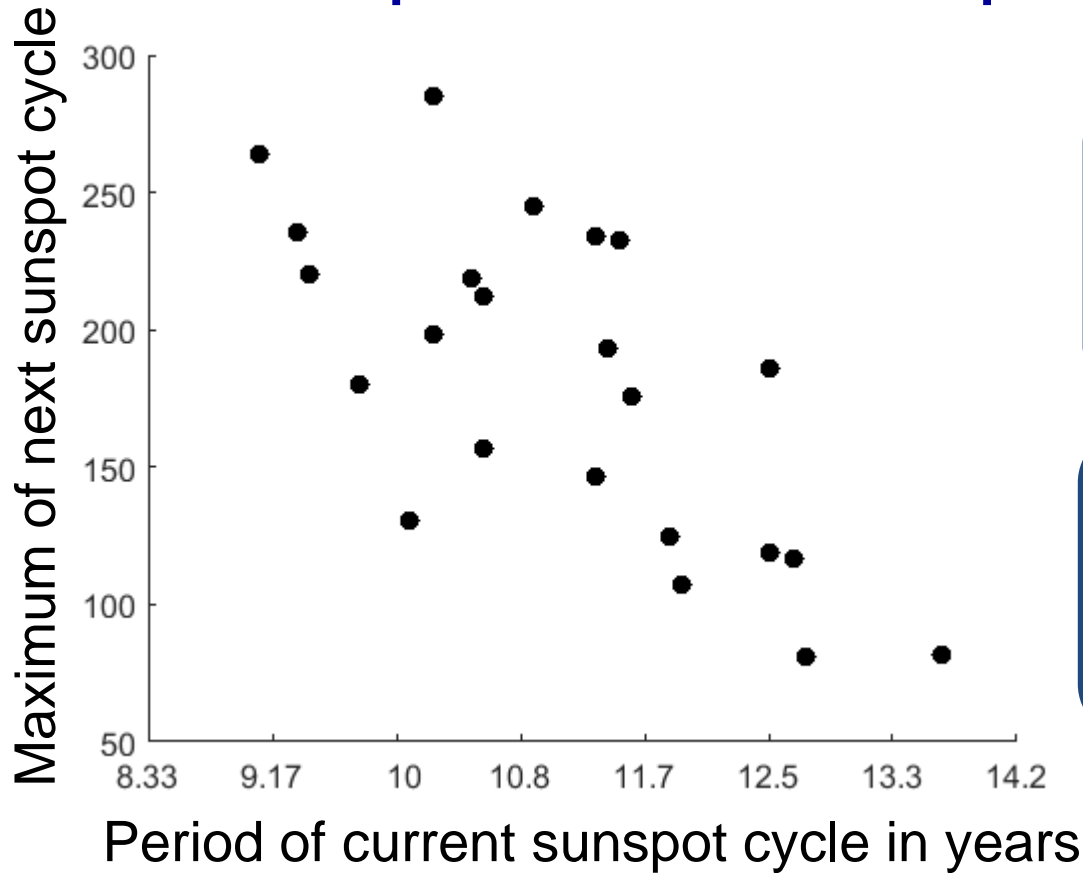


SILSO graphics (<http://sidc.be/silso>) Royal Observatory of Belgium 2017 October 1

On average a period of sunspot cycle is 11 years. It can be shorter or greater.

Is there any relationship between period of current sunspot cycle and the maximum of next sunspot cycle?

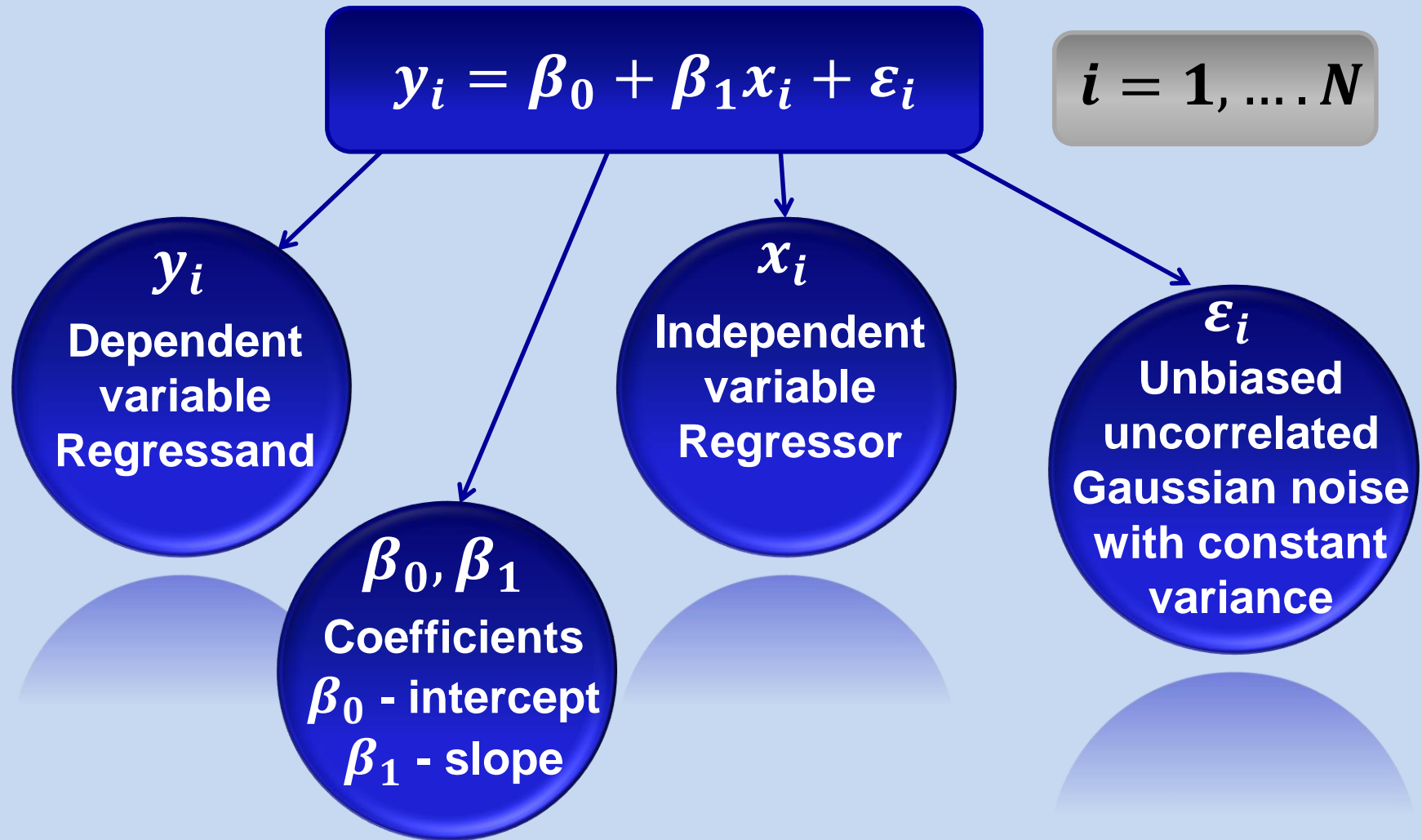
Scatter plot maximum versus period



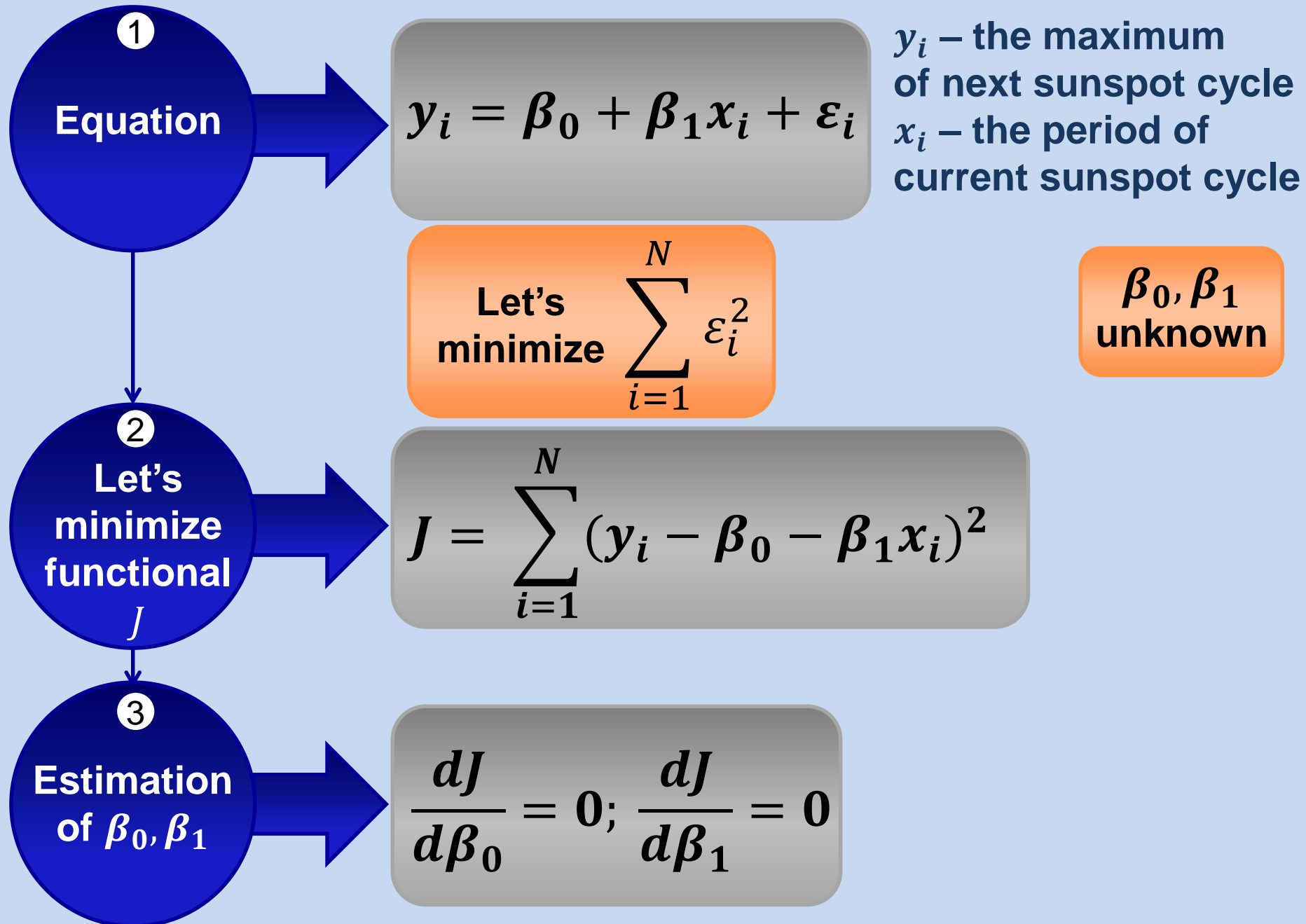
**Correlation
coefficient
-0.7**

**Greater the period
of cycle – less the
cycle maximum**

One-dimensional linear regression



One-dimensional linear regression



Determine coefficients β_0 and β_1 using LSM

Let's
minimize
functional
 J

$$J = \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i)^2$$

1

$$\frac{dJ}{d\beta_0} = -2 \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i) = 0$$

1

$$\frac{dJ}{d\beta_0} = \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i) = 0$$

2

$$\frac{dJ}{d\beta_1} = -2 \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

2

$$\frac{dJ}{d\beta_1} = \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

Excursus: solution

Determine coefficients β_0 and β_1 using LSM

$$\frac{dJ}{d\beta_0} = \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i) = 0 \quad 1.1$$

$$\frac{dJ}{d\beta_0} = \sum_{i=1}^N y_i - \sum_{i=1}^N \beta_0 - \sum_{i=1}^N \beta_1 x_i = 0 \quad 1.2$$

$$\frac{dJ}{d\beta_0} = \sum_{i=1}^N y_i - N\beta_0 - \beta_1 \sum_{i=1}^N x_i = 0 \quad 1.3$$

$$\beta_0 = \frac{1}{N} \left(\sum_{i=1}^N y_i - \beta_1 \sum_{i=1}^N x_i \right) \quad 1.4$$

Excursus: solution

Determine coefficients β_0 and β_1 using LSM

$$\frac{dJ}{d\beta_1} = \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \quad 2.1$$

$$\frac{dJ}{d\beta_1} = \sum_{i=1}^N y_i x_i - \sum_{i=1}^N \beta_0 x_i - \sum_{i=1}^N \beta_1 x_i^2 = 0 \quad 2.2$$

$$\frac{dJ}{d\beta_1} = \sum_{i=1}^N y_i x_i - \beta_0 \sum_{i=1}^N x_i - \beta_1 \sum_{i=1}^N x_i^2 = 0 \quad 2.3$$

Let's substitute value of β_0 from Equation 1.4 to Equation 2.3

Excursus: solution

Determine coefficients β_0 and β_1 using LSM

$$\sum_{i=1}^N y_i x_i - \frac{1}{N} \left(\sum_{i=1}^N y_i - \beta_1 \sum_{i=1}^N x_i \right) \sum_{i=1}^N x_i - \beta_1 \sum_{i=1}^N x_i^2 = 0 \quad 2.4$$

$$\sum_{i=1}^N y_i x_i - \frac{1}{N} \sum_{i=1}^N y_i \sum_{i=1}^N x_i + \frac{1}{N} \beta_1 \left(\sum_{i=1}^N x_i \right)^2 - \beta_1 \sum_{i=1}^N x_i^2 = 0 \quad 2.5$$

$$\beta_1 \left[\frac{1}{N} \left(\sum_{i=1}^N x_i \right)^2 - \sum_{i=1}^N x_i^2 \right] = \frac{1}{N} \sum_{i=1}^N y_i \sum_{i=1}^N x_i - \sum_{i=1}^N y_i x_i \quad 2.6$$

$$\beta_1 = \left[\frac{1}{N} \sum_{i=1}^N y_i \sum_{i=1}^N x_i - \sum_{i=1}^N y_i x_i \right] / \left[\frac{1}{N} \left(\sum_{i=1}^N x_i \right)^2 - \sum_{i=1}^N x_i^2 \right] \quad 2.7$$

Excursus: solution

Determine coefficients β_0 and β_1 using LSM

$$\beta_1 = \left[\frac{1}{N} \sum_{i=1}^N y_i \sum_{i=1}^N x_i - \sum_{i=1}^N y_i x_i \right] / \left[\frac{1}{N} \left(\sum_{i=1}^N x_i \right)^2 - \sum_{i=1}^N x_i^2 \right]$$

$$\beta_0 = \frac{1}{N} \left(\sum_{i=1}^N y_i - \beta_1 \sum_{i=1}^N x_i \right)$$

$$\beta_0 = 551.4$$

$$\beta_1 = -2.8$$

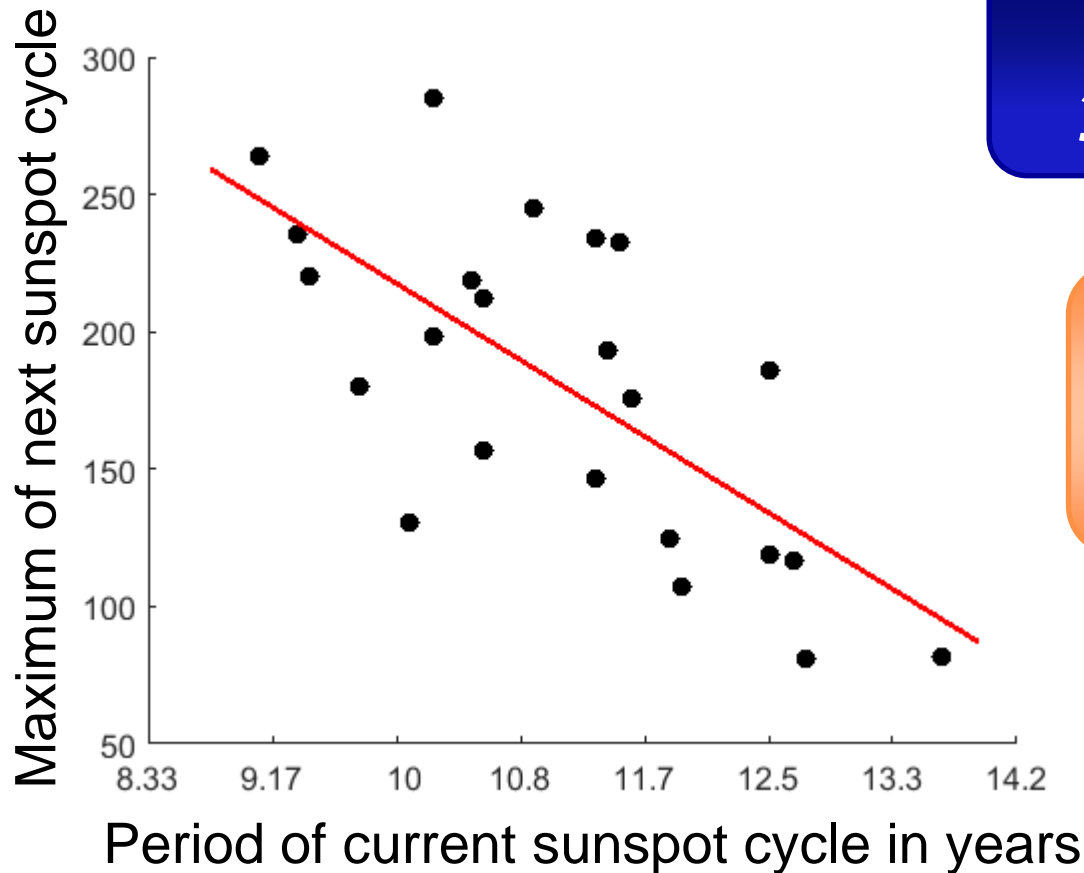
Excursus: solution

Is there any relationship between period of current sunspot cycle and the maximum of next sunspot cycle?

Linear regression

$$y_i = 551.4 - 2.8x_i$$

Thus, current cycle contains information on future sunspot activity



Precursor techniques to predict the next 11-year sunspot cycle strength

Extraction of useful knowledge from current sunspot cycle to predict future sunspot activity

Example

Regressand:
Next sunspot cycle maximum

Regressors:

Cycle
minimum

Geomagnetic
index around
its minimum

Number
of geomagnetic
disturbed days

Reversed
polar field

Example of inadequate regression model

True
trajectory

Motion
estimation
using model

Measurements

$$z_i = x_i + \varepsilon_i$$

$$x_i = x_{i-1} + VT$$

We assume
the object
to move
uniformly

$$x_i = x_{i-1} + V_i T + \frac{aT^2}{2}$$

But in
fact it is
uniformly
accelerated
motion

$$z_i = x_0 + V \cdot iT + \varepsilon_i$$

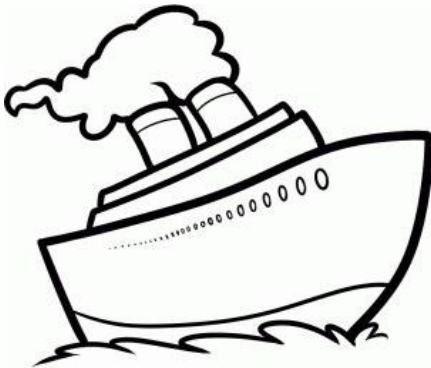
Thus,
regression
model is
inadequate

No practical value

How to take into account the unintentional maneuver?

**Motion
model**

$$x_i = x_{i-1} + VT$$

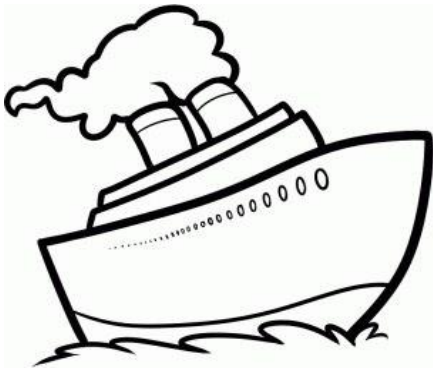


**However how to take into account
unintentional maneuver to track moving
object (i.e., airplane turbulence,
ship pitching or undercurrents)?**

How to take into account the unintentional maneuver?

**Motion
model**

$$x_i = x_{i-1} + V_i T + \frac{a_i T^2}{2}$$



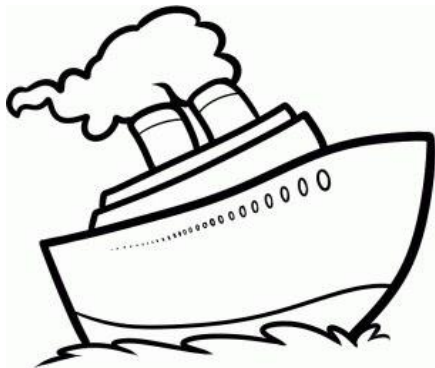
**However how to take into account
unintentional maneuver to track moving
object (i.e., airplane turbulence,
ship pitching or undercurrents)?**

**Unintentional maneuver can be
described by random acceleration a**

Process noise should not be filtered

**Motion
model**

$$x_i = x_{i-1} + V_i T + \frac{a_i T^2}{2}$$



**However how to take into account
unintentional maneuver to track moving
object (i.e., airplane turbulence,
ship pitching or undercurrents)?**

**Unintentional maneuver can be
described by random acceleration a**

$$\frac{a_i T^2}{2}$$

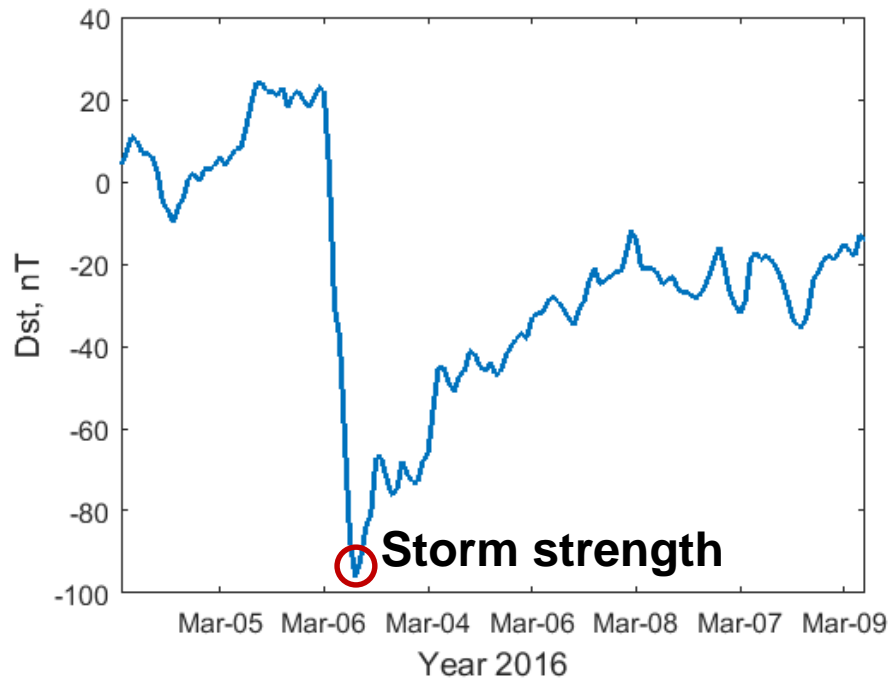
**Noise intrinsic
to the process
itself that should
not be filtered**

**However linear regression
doesn't separate process noise
and measurement noise and thus
loses its practical value**

Multi-dimensional linear regression

Dependence on
multiple regressors

Example: Disturbance storm
time (Dst) geomagnetic index



Dst is sensitive to:

① Solar wind speed

② Southward component
of Interplanetary
magnetic field (IMF)

③ Previous Dst values

④ Solar wind density
and pressure

Multi-dimensional linear regression

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{k-1} x_{i,k-1} + \varepsilon_i$$

$$i = 1, \dots, N$$

y_i

**Dependent
variable
Regressand**

β_j

**Coefficients
of regression**

$x_{i,j}$

**Independent
variable
Regressor**

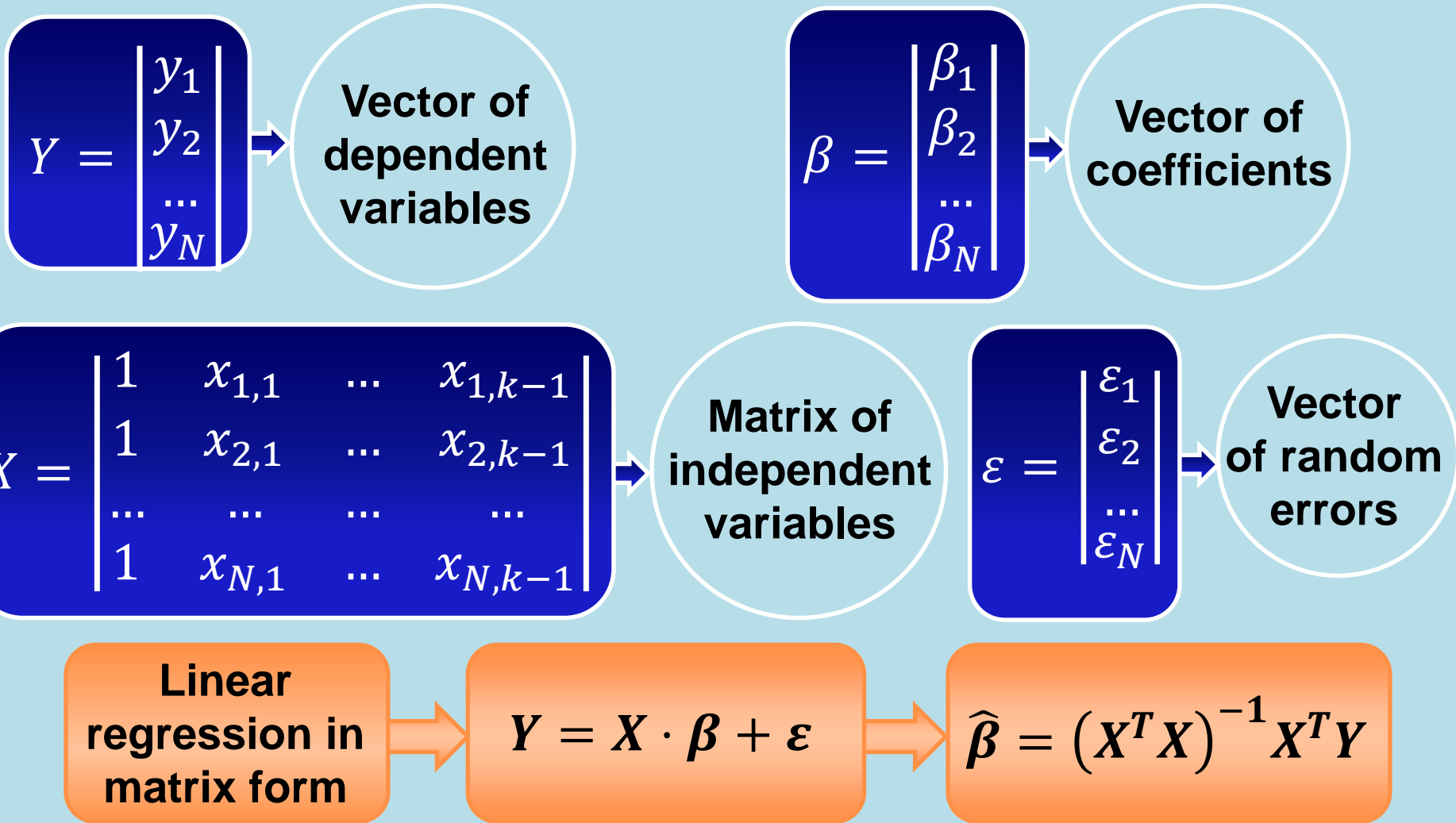
ε_i

**Unbiased
uncorrelated
Gaussian noise
with constant
variance**

**Coefficients
 β_j are
determined
by LSM**

$$\sum_{i=1}^N \varepsilon_i^2 \rightarrow \min$$

Multi-dimensional linear regression



Linear Regression Analysis,
G.A.F. Seber and J. Lee, Wiley, N.Y., 2003

Estimation error of coefficients β

**Covariance
matrix of
estimation
error**

$$\text{cov}(\hat{\beta}) = \sigma_{\varepsilon}^2 (X^T X)^{-1}$$

**Variance
estimation
of random
noise**

$$\sigma_{\varepsilon}^2 = \frac{1}{N - k} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

**Weighted
LSM**

$$\sum_{i=1}^N \frac{\varepsilon_i^2}{\sigma_{\varepsilon_i}^2} \rightarrow \min$$

Main problems of applying LSM and linear regression

LSM

If process dynamics
is unknown



The LSM method
leads to divergence
and loses its
practical value

Linear regression



No separation
between process
noise and
measurement noise
and thus no
practical value

Following topics of
the course overcome
these problems