Least squares estimators

Contents map

developed content units	taxonomy levels
least squares	u1, e1

prerequisite content units	taxonomy levels
dataset	u1, e1

Main ILO of sub-module "Least squares estimators"

Describe the concept of least squares in geometrical perspectives

Derive and use the normal equations for
solving separable least squares problems

Basic assumptions

data generation model: $y_t = f\left(u_t; \theta\right) + v_t$

dataset: $\mathcal{D} = \left\{\left(u_t, y_t\right)\right\}_{t=1,\ldots,N}$

hypothesis space: $\theta \in \Theta$

Basic assumptions

data generation model: $y_t = f(u_t; \theta) + v_t$

dataset: $\mathcal{D} = \{(u_t, y_t)\}_{t=1,\dots,N}$

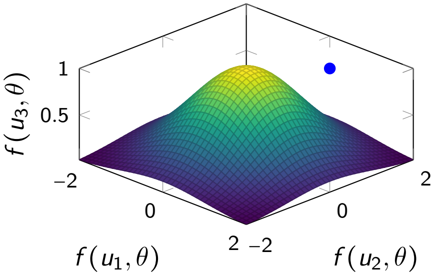
hypothesis space: $\theta \in \Theta$

Problem: *find a* $\widehat{\theta} \in \Theta$ *that "best explains"* $\mathcal{D}$

Geometrical interpretation

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{bmatrix} \text{fixed} \qquad \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \end{bmatrix} \text{fixed} \qquad \begin{bmatrix} f(u_1;\theta) \\ f(u_2;\theta) \\ f(u_3;\theta) \\ \vdots \end{bmatrix} \text{manifold in } \theta$$

example with $\theta \in \mathbb{R}^2$:



Question 1

Consider

$$\begin{bmatrix} f(u_1;\theta) \\ \vdots \\ f(u_N;\theta) \end{bmatrix};$$

varying $u_1, \ldots, u_N$ but keeping θ fixed corresponds in general to find:

Potential answers:

I: a scalar

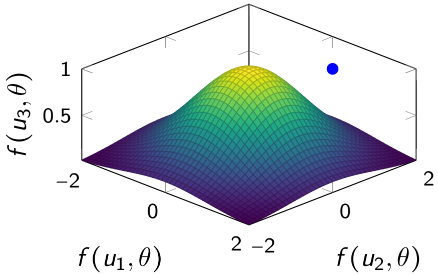
II: a vector

III: a manifold

IV: I do not know

mathematical formulation

Intuitions, towards a mathematical formulation



Mathematical formulation

$$y_t = f(u_t; \theta) + v_t \qquad \mathcal{D} = \{(u_t, y_t)\}_{t=1,\dots,N} \qquad \theta \in \Theta$$

Mathematical formulation

$$y_t = f(u_t; \theta) + v_t \qquad \mathcal{D} = \{(u_t, y_t)\}_{t=1,\dots,N} \qquad \theta \in \Theta$$

$$\widehat{\theta}_{\mathrm{LS}} = \arg\min_{\theta \in \Theta} \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} f(u_1; \theta) \\ \vdots \\ f(u_N; \theta) \end{bmatrix} \right\|^2$$

Mathematical formulation

$$y_t = f\left(u_t;\theta\right) + v_t \qquad \mathcal{D} = \left\{\left(u_t, y_t\right)\right\}_{t=1,\dots,N} \qquad \theta \in \Theta$$

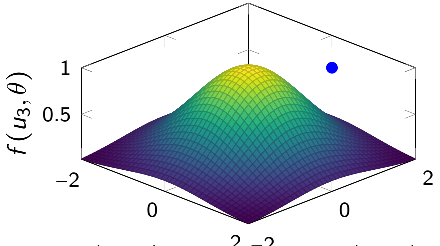
$$\widehat{\theta}_{\mathrm{LS}} = \arg\min_{\theta\in\Theta} \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} f\left(u_1;\theta\right) \\ \vdots \\ f\left(u_N;\theta\right) \end{bmatrix} \right\|^2 = \arg\min_{\theta\in\Theta} \sum_{t=1}^{N} \Big(y_t - f\left(u_t;\theta\right)\Big)^2$$

Mathematical formulation

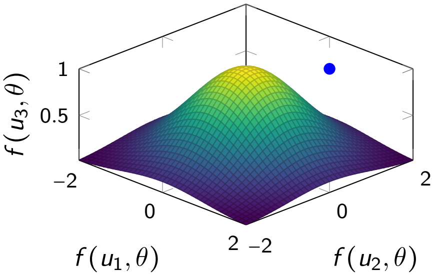
$$y_t = f(u_t;\theta) + v_t \qquad \mathcal{D} = \{(u_t, y_t)\}_{t=1,\dots,N} \qquad \theta \in \Theta$$

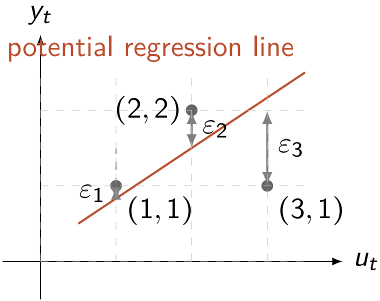
$$\widehat{\theta}_{\mathrm{LS}} = \arg\min_{\theta\in\Theta} \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} f(u_1;\theta) \\ \vdots \\ f(u_N;\theta) \end{bmatrix} \right\|^2 = \arg\min_{\theta\in\Theta} \sum_{t=1}^{N} \Big(y_t - f(u_t;\theta) \Big)^2$$

residual: $r_t(\theta) := y_t - f(u_t;\theta)$



Example: regression line



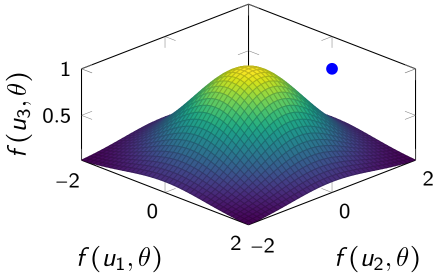


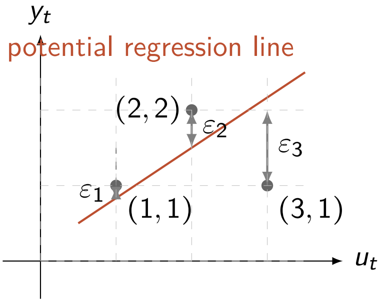
$y_t = \theta_1 + \theta_2 u_t + v_t$

$\mathcal{D} = \{(u_t, y_t)\}_{t=1}^{3} = \{(1,1),(2,2),(3,1)\}$

$\theta \in \mathbb{R}^2$

Example: regression line





$$y_t = \theta_1 + \theta_2 u_t + v_t \qquad \mathcal{D} = \{(u_t, y_t)\}_{t=1}^3 = \{(1,1),(2,2),(3,1)\} \qquad \theta \in \mathbb{R}^2$$

$$\widehat{\theta}_{\mathrm{LS}} = \left(\widehat{\theta}_{\mathrm{LS},1}, \widehat{\theta}_{\mathrm{LS},2}\right) = \arg\min_{\theta_1,\theta_2\in\mathbb{R}} \Big((1-\theta_1-\theta_2)^2 + (2-\theta_1-2\theta_2)^2 + (1-\theta_1-3\theta_2)^2 \Big)$$

Question 2

Consider

$$f(u;\theta) = \sum_{k=0}^{2} \theta_k u^k \qquad \mathcal{D} = \{(0,0),(1,1)\} \qquad \Theta = \mathbb{R}^2.$$

How many solutions will the LS problem have?

Potential answers:

I: 0

II: 1

III: $+\infty$

IV: I do not know

basic properties

Question 3

The concepts behind LS are simple, so it is simple to compute analytically $\widehat{\theta}_{\mathrm{LS}}$

Potential answers:

I: true

II: false

III: I do not know

Example: computing the LS may be numerically infeasible

$$u_t \in \mathbb{R}^{10^6}$$

$$f\left(u_t;\theta\right) \text{ extremely nonlinear}$$

$$\mathcal{D} = \left\{\left(u_t, y_t\right)_{t=1,\ldots,N}\right\}, \; N = 10^{12}$$

$$\theta \in \text{ very non-convex set}$$

Question 4

The LS estimate $\widehat{\theta}_{\mathrm{LS}}$ always exists

Potential answers:

I: true

II: false

III: I do not know

Example: the LS estimate may not exist

$$\widehat{\theta}_{\mathrm{LS}} = \arg\min_{\theta\in(0,1)} \sum_{t=1}^{100} r_t^2(\theta) \qquad\qquad r_t(\theta) = \frac{1}{\theta}$$

Question 5

When it exists, the LS estimate $\widehat{\theta}_{\mathrm{LS}}$ is unique

Potential answers:

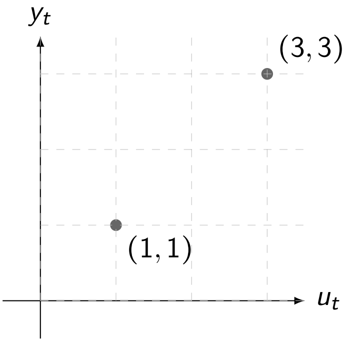
I: true

II: false

III: I do not know

Example: the LS estimate is not unique

How many quadratics fit perfectly this dataset?



linear least squares

If you don't remember how to do computations with matrices and vectors. . .

the matrix cookbook

Separable problems

$$y_t = \sum_{j=1}^{n} \theta_j \phi_j(u_t) + e_t$$

Separable problems

$$y_t = \sum_{j=1}^{n} \theta_j \phi_j(u_t) + e_t$$

$$\Downarrow$$

$$y_t = \begin{bmatrix} \phi_1(u_t) & \cdots & \phi_n(u_t) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} + e_t$$

Separable problems

$$y_t = \sum_{j=1}^{n} \theta_j \phi_j(u_t) + e_t$$

$$\Downarrow$$

$$y_t = \begin{bmatrix} \phi_1(u_t) & \cdots & \phi_n(u_t) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} + e_t$$

$$\Downarrow$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \phi_1(u_1) & \cdots & \phi_n(u_1) \\ \vdots & & \vdots \\ \phi_1(u_N) & \cdots & \phi_n(u_N) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}$$

Separable problems

$$y_t = \sum_{j=1}^{n} \theta_j \phi_j(u_t) + e_t$$

$$\Downarrow$$

$$y_t = \begin{bmatrix} \phi_1(u_t) & \cdots & \phi_n(u_t) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} + e_t$$

$$\Downarrow$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \phi_1(u_1) & \cdots & \phi_n(u_1) \\ \vdots & & \vdots \\ \phi_1(u_N) & \cdots & \phi_n(u_N) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}$$

$$\Downarrow$$

$$\boldsymbol{y} = \Phi(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{e}$$

LS for ***unconstrained*** separable problems $\implies$ normal equations

$$\boldsymbol{y} = \Phi(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{e}, \quad \boldsymbol{\theta} \in \mathbb{R}^n \qquad\qquad \widehat{\theta}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \left\|\boldsymbol{y} - \Phi(\boldsymbol{u})\boldsymbol{\theta}\right\|^2$$

LS for ***unconstrained*** separable problems $\implies$ normal equations

$$\boldsymbol{y} = \Phi(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{e}, \quad \boldsymbol{\theta} \in \mathbb{R}^n \qquad\qquad \widehat{\theta}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \left\| \boldsymbol{y} - \Phi(\boldsymbol{u})\boldsymbol{\theta} \right\|^2$$

Ideally, $\widehat{\theta}_{\mathrm{LS}}$ is s.t. $\Phi(\boldsymbol{u})\widehat{\theta}_{\mathrm{LS}} = \boldsymbol{y}$!

LS for ***unconstrained*** separable problems $\Longrightarrow$ normal equations

$$\boldsymbol{y} = \Phi(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{e}, \quad \boldsymbol{\theta} \in \mathbb{R}^n \qquad\qquad \widehat{\theta}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \left\| \boldsymbol{y} - \Phi(\boldsymbol{u})\boldsymbol{\theta} \right\|^2$$

Ideally, $\widehat{\theta}_{\mathrm{LS}}$ is s.t. $\Phi(\boldsymbol{u})\widehat{\theta}_{\mathrm{LS}} = \boldsymbol{y}$!

normal equations:

LS for ***unconstrained*** separable problems $\Longrightarrow$ normal equations

$$\boldsymbol{y} = \Phi(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{e}, \quad \boldsymbol{\theta} \in \mathbb{R}^n \qquad\qquad \widehat{\theta}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \left\| \boldsymbol{y} - \Phi(\boldsymbol{u})\boldsymbol{\theta} \right\|^2$$

Ideally, $\widehat{\theta}_{\mathrm{LS}}$ is s.t. $\Phi(\boldsymbol{u})\widehat{\theta}_{\mathrm{LS}} = \boldsymbol{y}$!

normal equations: $\Phi(\boldsymbol{u})^T \Phi(\boldsymbol{u})\widehat{\theta}_{\mathrm{LS}} = \Phi(\boldsymbol{u})^T \boldsymbol{y}$

Exercise

Compute the solution of

$$\arg\min_{\theta\in\mathbb{R}^n} \left(\boldsymbol{y} - \Phi(\boldsymbol{u})\boldsymbol{\theta}\right)^T W \left(\boldsymbol{y} - \Phi(\boldsymbol{u})\boldsymbol{\theta}\right)$$

Question 6

Starting from

$$\Phi(\boldsymbol{u})^T \Phi(\boldsymbol{u}) \widehat{\theta}_{\mathrm{LS}} = \Phi(\boldsymbol{u})^T \boldsymbol{y}$$

we can always set

$$\widehat{\theta}_{\mathrm{LS}} = \left(\Phi(\boldsymbol{u})^T \Phi(\boldsymbol{u})\right)^{-1} \Phi(\boldsymbol{u})^T \boldsymbol{y}$$

Potential answers:

I: true

II: false

III: I do not know

Using the pseudoinverse when necessary

what if $\Phi(\boldsymbol{u})^T\Phi(\boldsymbol{u})$ does not have an inverse?

Using the pseudoinverse when necessary

what if $\Phi(\boldsymbol{u})^T\Phi(\boldsymbol{u})$ does not have an inverse?

Definition (Moore-Penrose pseudoinverse of a matrix)

Given $A \in \mathbb{R}^{m\times n}$, $A^\dagger$ is its pseudoinverse if

$$AA^\dagger A = A$$

$$A^\dagger A A^\dagger = A^\dagger$$

$$\left(AA^\dagger\right)^H = AA^\dagger$$

$$\left(A^\dagger A\right)^H = A^\dagger A$$

Using the pseudoinverse when necessary

what if $\Phi(\boldsymbol{u})^T \Phi(\boldsymbol{u})$ does not have an inverse?

Definition (Moore-Penrose pseudoinverse of a matrix)

Given $A \in \mathbb{R}^{m \times n}$, $A^\dagger$ is its pseudoinverse if

$$AA^\dagger A = A$$

$$A^\dagger A A^\dagger = A^\dagger$$

$$\left(AA^\dagger\right)^H = AA^\dagger$$

$$\left(A^\dagger A\right)^H = A^\dagger A$$

more in `http://www.math.ucla.edu/~laub/33a.2.12s/mppseudoinverse.pdf`

Using the pseudoinverse when necessary

$$\boldsymbol{y} = \Phi(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{e}, \quad \boldsymbol{\theta} \in \mathbb{R}^n \qquad\qquad \widehat{\theta}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \left\| \boldsymbol{y} - \Phi(\boldsymbol{u})\boldsymbol{\theta} \right\|^2$$

$$\implies \qquad \widehat{\theta}_{\mathrm{LS}} = \Phi(\boldsymbol{u})^{\dagger}\boldsymbol{y}$$

Using the pseudoinverse when necessary

$$\boldsymbol{y} = \Phi(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{e}, \quad \boldsymbol{\theta} \in \mathbb{R}^n \qquad \widehat{\theta}_{\text{LS}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \left\| \boldsymbol{y} - \Phi(\boldsymbol{u})\boldsymbol{\theta} \right\|^2$$

$$\Longrightarrow \qquad \widehat{\theta}_{\text{LS}} = \Phi(\boldsymbol{u})^{\dagger}\boldsymbol{y}$$

strong connections with singular values decompositions!

Question 7

We can always solve the normal equations for every unconstrained separable LS problem

Potential answers:

I: true
II: false
III: I do not know

Question 8

We can always solve the normal equations for every separable LS problem, even for constrained ones

Potential answers:

I: true

II: false

III: I do not know

LS for constrained separable problems $\not\Longrightarrow$ normal equations

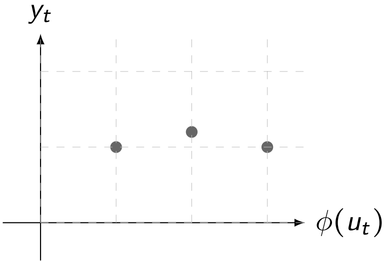
$$\boldsymbol{y} = \Phi(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{e}, \quad \boldsymbol{\theta} \in \Theta \qquad\qquad \widehat{\theta}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{y} - \Phi(\boldsymbol{u})\boldsymbol{\theta}\|^2$$

LS for constrained separable problems $\not\Longrightarrow$ normal equations

$$\boldsymbol{y} = \Phi(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{e}, \quad \boldsymbol{\theta} \in \Theta \qquad\qquad \widehat{\theta}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \left\| \boldsymbol{y} - \Phi(\boldsymbol{u})\boldsymbol{\theta} \right\|^2$$

Ideally, looking for $\boldsymbol{\theta}^*$ s.t. $\Phi(\boldsymbol{u})\boldsymbol{\theta}^* - \boldsymbol{y} = \boldsymbol{0}$, but it may happen that $\boldsymbol{\theta}^* \notin \Theta$!

Example = fitting a convex quadratic here:



Summarizing

Describe the concept of least squares in geometrical perspectives

Derive and use the normal equations for
solving separable least squares problems

- visualize the dataset in an opportune multidimensional plot
- if we have separability, we can use linear algebra to arrive at $X^T X \widehat{\theta} = Xy$

Most important python code for this sub-module

Illustrative example

https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html

Self-assessment material

Question 9

In the geometric interpretation of least squares, what does the vector $\boldsymbol{y}$ represent?

Potential answers:

I: The model parameters to be estimated
II: The fixed vector of measured output values
III: The manifold of all possible model predictions
IV: The noise affecting the measurements
V: I do not know

Question 10

What is the fundamental assumption required to derive the normal equations for least squares?

Potential answers:

I: The noise must be Gaussian distributed
II: The model must be nonlinear in parameters
III: The problem must be linear in parameters (separable)
IV: The hypothesis space must be constrained
V: I do not know

Question 11

When is the Moore-Penrose pseudoinverse required in least squares problems?

Potential answers:

I: When dealing with nonlinear models
II: When the measurements are noisy
III: When $\Phi^T\Phi$ is not invertible
IV: When the hypothesis space is constrained
V: I do not know

Question 12

What guarantees the existence of a unique least squares solution?

Potential answers:

I: Having more parameters than measurements
II: Φ having full column rank and unconstrained parameters
III: The hypothesis space being compact
IV: The noise being normally distributed
V: I do not know

Question 13

What is a key difference between constrained and unconstrained least squares problems?

Potential answers:

I: Constrained problems always have unique solutions
II: The normal equations may give solutions outside the constraint set
III: Only unconstrained problems can use the pseudoinverse
IV: Constrained problems require nonlinear optimization
V: I do not know

Recap of sub-module “linear least squares”

- Least squares aims to minimize the squared residuals between model predictions and observed data
- The geometric interpretation views system identification as finding the closest point on a model manifold to measurement vectors
- Normal equations provide an analytical solution for unconstrained linear least squares problems through $\Phi^T \Phi \theta = \Phi^T y$
- The pseudoinverse generalizes solutions for rank-deficient systems and connects with singular value decomposition
- Existence and uniqueness of LS solutions depend on hypothesis space topology and model structure identifiability
- Constrained LS problems require different approaches than normal equations when parameters must satisfy domain restrictions

?