LEAST SQUARES LINEAR REGRESSION

In regression problems, we are given training data

$$(x_1, y_1), \ldots, (x_n, y_n)$$

where

$$x_i \in \mathbb{R}^d$$
$$y_i \in \mathbb{R}.$$

We assume the (x_i, y_i) are realizations of a random pair (X, Y). The goal of regression is to predict the response y associated to a new input x.

A <u>regression model</u> posits

$$Y = f(X) + \epsilon$$

where

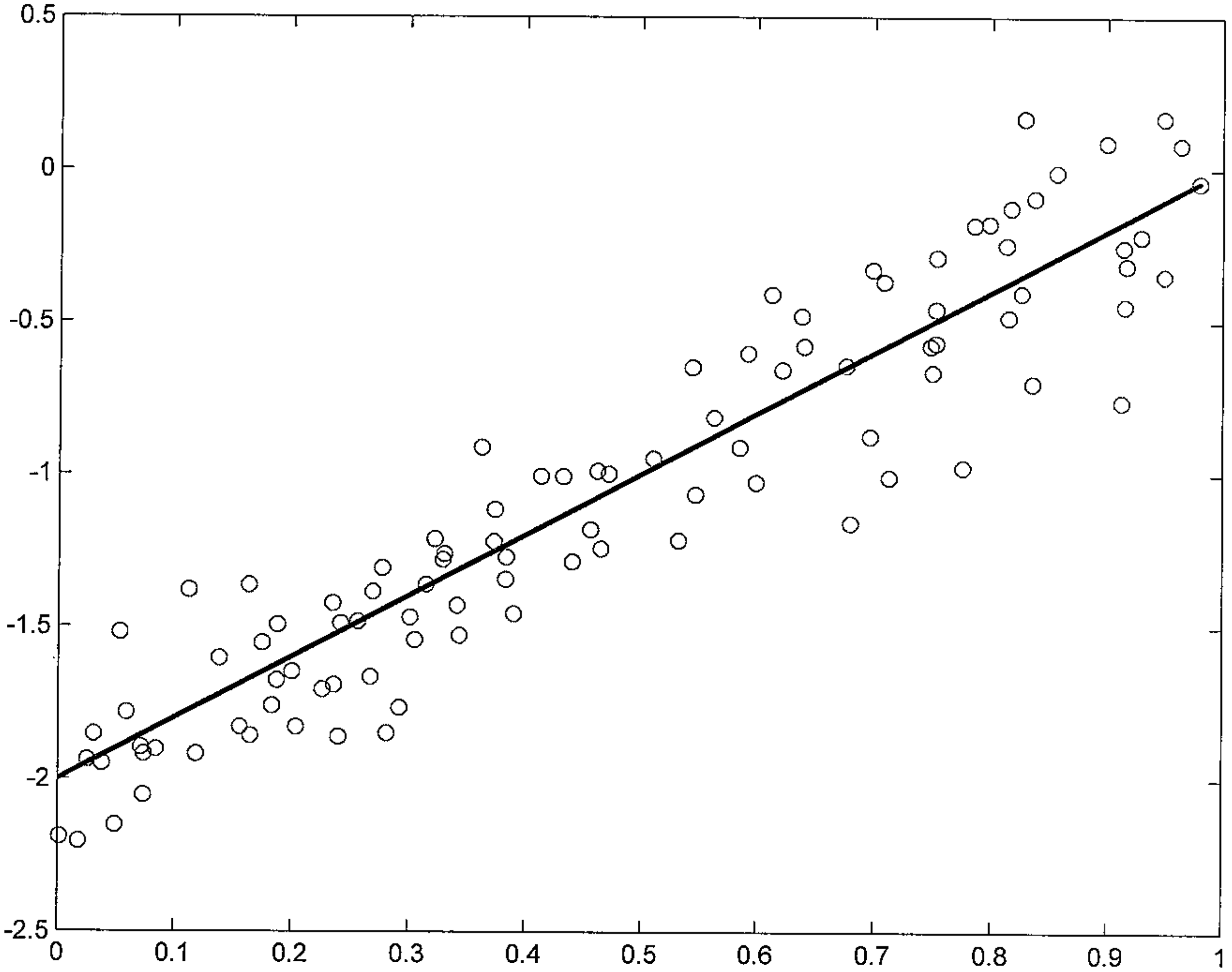
$$\epsilon = \text{noise}$$
$$f \in \text{some class of functions.}$$

In linear regression, we assume a linear model

$$f(x) = \beta^T x + \beta_0$$

where

$$\beta \in \mathbb{R}^d, \quad \beta_0 \in \mathbb{R}$$



The challenge in linear regression is to estimate the parameters β, β_0, from training data.

Least Squares

In <u>least squares</u> linear regression, we select β, β_0 to minimize the <u>sum of squared errors</u>,

$$SSE(\beta,\beta_0) := \sum_{i=1}^{n} (y_i - \beta^T x_i - \beta_0)^2 .$$

<u>Example</u> Suppose $d=1$, so x_i, β are scalars.

Ⓐ $$\frac{\partial SSE}{\partial \beta_0} = \qquad\qquad\qquad\qquad\qquad = 0$$

$$\Longrightarrow \quad \beta_0 =$$

$$\frac{\partial SSE}{\partial \beta} =$$

$$\Longrightarrow \quad \beta =$$

In matrix form,

$$\begin{bmatrix} \qquad\qquad\qquad\qquad \\ \\ \\ \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} = \begin{bmatrix} \qquad \\ \\ \\ \end{bmatrix}$$

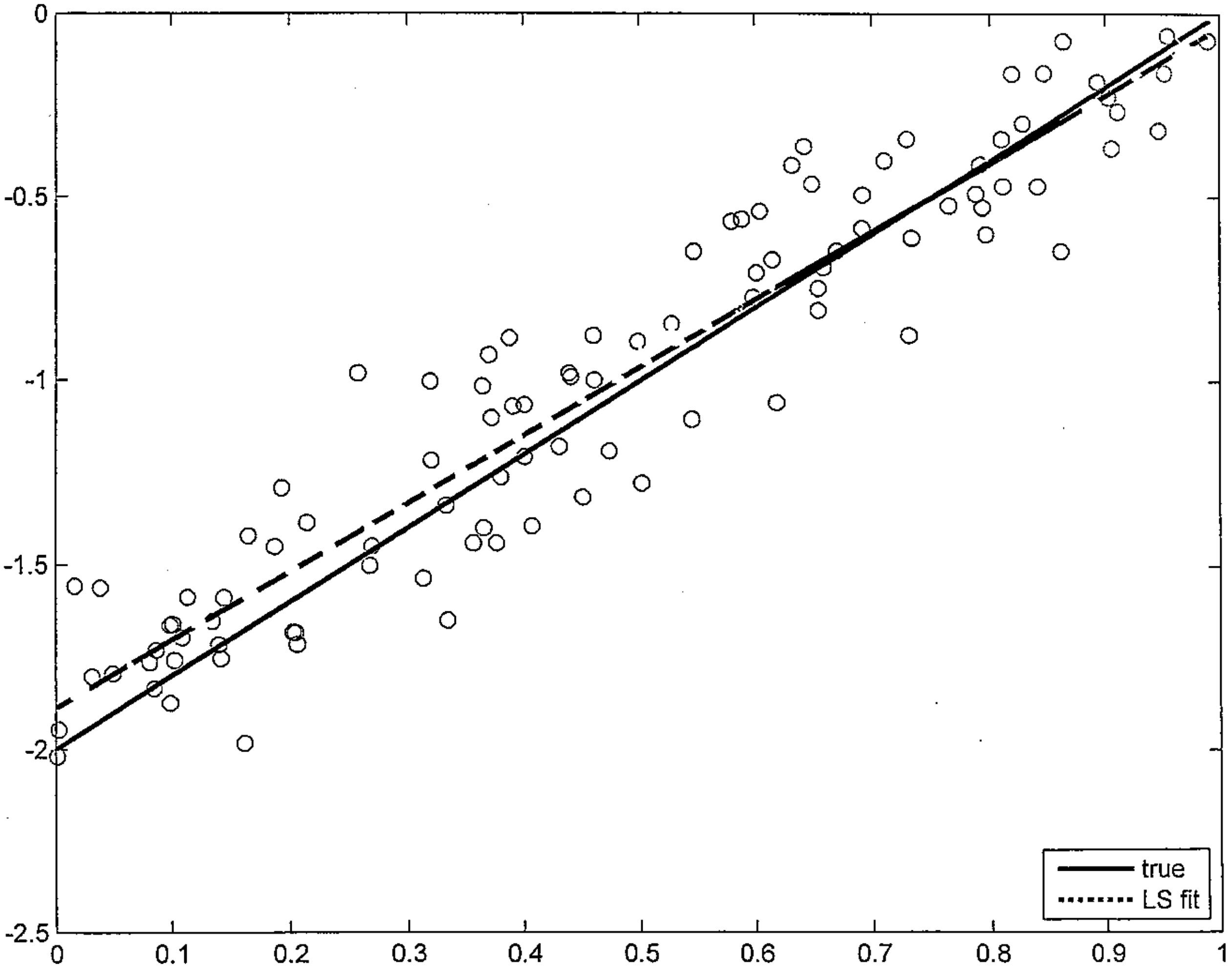
Inverting the matrix,

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{\bar{y}\left(\sum x_i^2\right) - \bar{x} \sum x_i y_i}{\sum x_i^2 - n\bar{x}^2} \\ \dfrac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \end{bmatrix}$$

where

$$\bar{x} = \frac{1}{n} \sum x_i \,, \quad \bar{y} = \frac{1}{n} \sum y_i$$



More generally, suppose d is arbitrary. Set

$$\theta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}$$

Then

$$\begin{aligned} SSE(\theta) &= \sum_{i=1}^{n} \left(y_i - \beta^T x_i - \beta_0 \right)^2 \\ &= \left\| \underline{y} - A\theta \right\|^2 \end{aligned}$$

where

Ⓑ $$\underline{y} = \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}, \quad A = \underset{n \times (d+1)}{\begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}}$$

The minimizer $\hat{\theta}$ of this quadratic objective function is

$$\hat{\theta} = (A^T A)^{-1} A^T \underline{y}$$

provided

To see this, write

Ⓒ $$\|\underline{y} - A\theta\|^2 =$$

$$=$$

Linear regression is easy and works well when the true f is indeed linear. But often it is not. What can we do?

Sometimes f is linear in variables $\phi_1(x), \ldots, \phi_K(x)$, where ϕ_k is possibly nonlinear. In such a case we can model

$$f(x) = \sum_{k=1}^{K} \beta_k \phi_k(x) + \beta_0$$

and estimate $\beta_0, \beta_1, \ldots, \beta_K$ using least squares applied to the "data"

$$(\phi(x_1), y_1), \ldots, (\phi(x_n), y_n).$$

Exercise | Suppose $f(x)$ is a cubic polynomial. Determine the least-squares estimate of f given $(x_1, y_1), \dots, (x_n, y_n)$.

Solution

$$f(x) = \beta_3 x^3 + \beta_2 x^2 + \beta_1 x + \beta_0$$

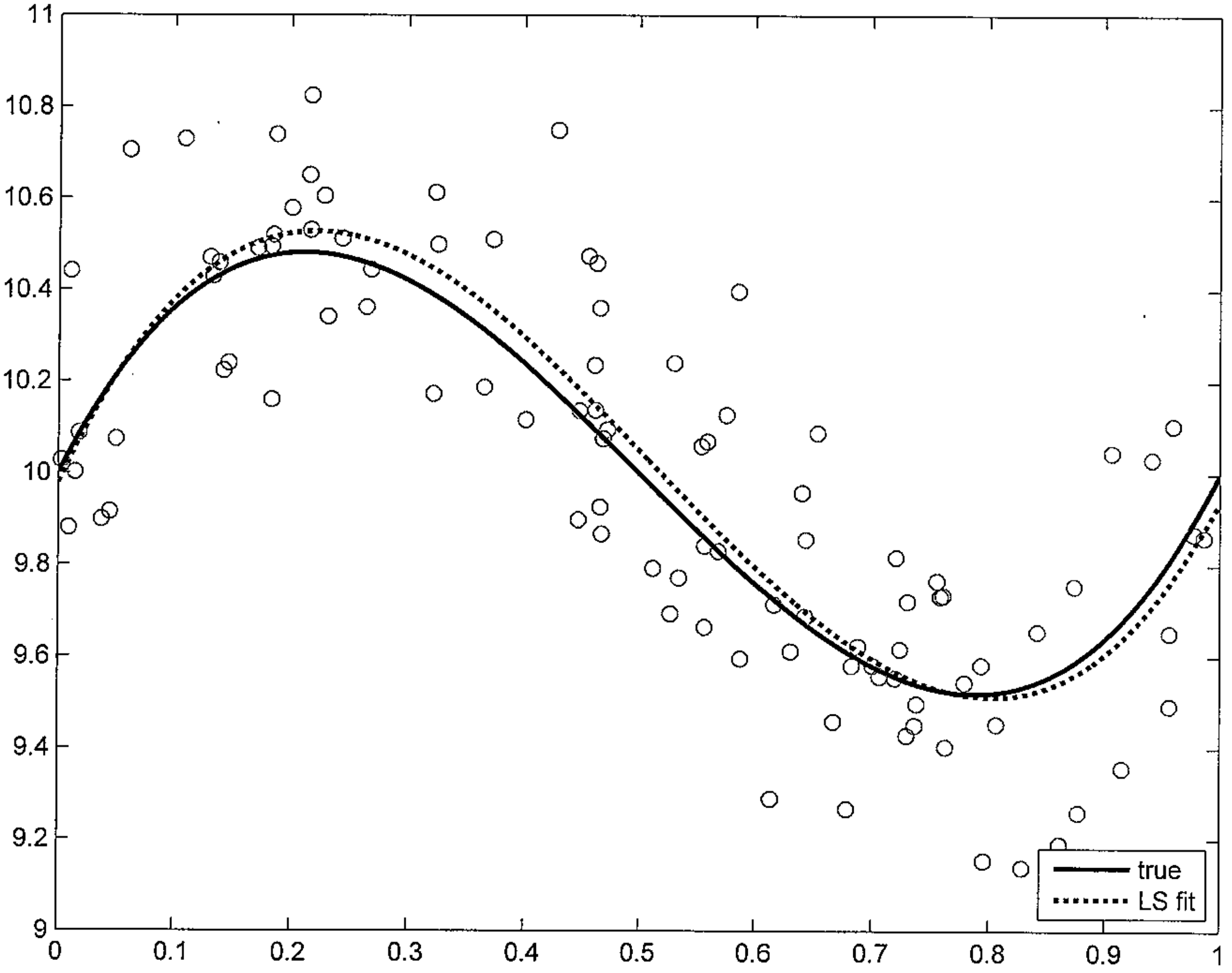
$$\Longrightarrow \quad \phi_k(x) = x^k$$

Ⓓ $$\Longrightarrow \quad A = \begin{bmatrix} \quad & & & \quad \\ & & & \\ & & & \end{bmatrix}$$

$$\Longrightarrow \quad \hat{\theta} = (A^T A)^{-1} A^T \underline{y}$$

gives the LS cubic polynomial fit.

What if a polynomial model is also not appropriate, or the degree is unknown? We'll address these and other issues later in the course.


11
10.8
10.6
10.4
10.2
10
9.8
9.6
9.4
9.2
9
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
true
LS fit

Key

A. $$\frac{\partial\, SSE}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta x_i - \beta_0) = 0$$

$$\Rightarrow \quad \beta_0 = \frac{1}{n} \sum_i (y_i - \beta x_i)$$

$$\frac{\partial\, SSE}{\partial \beta} = -2 \sum_i x_i (y_i - \beta x_i - \beta_0) = 0$$

$$\Rightarrow \quad \beta = \frac{\sum x_i (y_i - \beta_0)}{\sum x_i^2}$$

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

B. $$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix}$$

$$\hat{\theta} = (A^T A)^{-1} A^T \underline{y}$$

provided $A^T A$ is nonsingular

C. $\|\underline{y} - A\theta\|^2 = (\underline{y} - A\underline{\theta})^T(\underline{y} - A\theta)$

$$= \theta^T A^T A \theta - 2\underline{y}^T A \theta + \underline{y}^T\underline{y}$$

proof 1 : $\frac{\partial}{\partial\theta}(\swarrow) = 2A^TA\theta - 2A^T\underline{y} = 0$

$$\Longrightarrow \hat{\theta} = (A^TA)^{-1}A^T\underline{y}$$

proof 2 : $\theta^T B \theta + \underline{c}^T\theta + d$

$$= (\theta + \tfrac{1}{2}B^{-1}\underline{c})^T B (\theta + \tfrac{1}{2}B^{-1}\underline{c}) + (d - \tfrac{1}{4}\underline{c}^T B^{-1}\underline{c})$$

If B is positive definite, then the unique minimizer is

$$\theta = -\tfrac{1}{2}B^{-1}\underline{c}.$$

Apply this with $B = A^TA$, $\underline{c} = -2A^T\underline{y}$, $d = \underline{y}^T\underline{y}$

D.

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix}$$