Machine Learning

Linear Regression

Hiroshi Shimodaira and Hao Tang

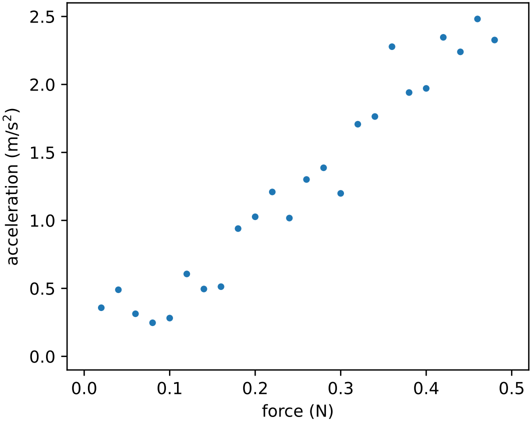
February 2026

Ver. 1.0

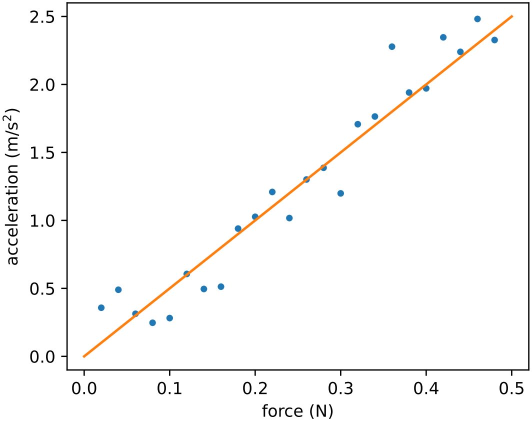
Topics

- Linear regression
- Geometry of linear regression
- Training of linear regression with MSE
- Normal equations
- Training of linear regression with MLE
- Terminology (multiple linear regression vs multivariate regression)

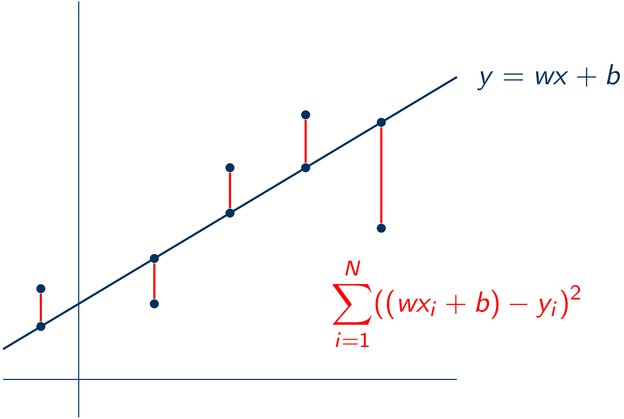
First example



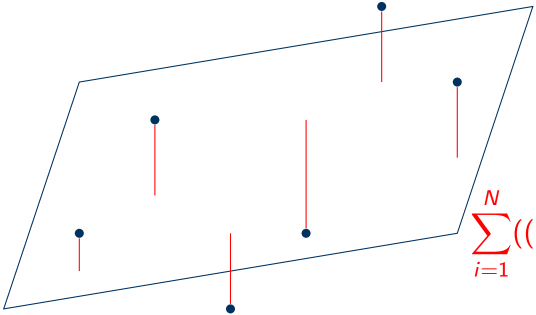
First example



Geometry of linear regression



Geometry of linear regression *(cont.)*



$$y = \boldsymbol{w}^\top \boldsymbol{x} + b$$

$$\sum_{i=1}^{N} ((\boldsymbol{w}^\top \boldsymbol{x}_i + b) - y_i)^2$$

Linear regression

- $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$: data set
  - $\boldsymbol{x}_i = \begin{bmatrix} x_{i1} & \cdots & x_{id} \end{bmatrix}^\top$: input, features, independent variables
  - $y_i \in \mathbb{R}$: target/dependent variable, ground truth, for $\boldsymbol{x}_i$.
- $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$: linear predictor, hyperplane
  - $\boldsymbol{w} = \begin{bmatrix} w_1 & \cdots & w_d \end{bmatrix}^\top$: weights
  - $b \in \mathbb{R}$: bias
  - $\{\boldsymbol{w}, b\}$: parameters $\cdots$ $\boldsymbol{\theta} = [b\ \boldsymbol{w}^\top]^\top$

Linear regression *(cont.)*

- Given $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$, find $\boldsymbol{\theta}$ such that the mean-squared error (MSE)

$$L = \frac{1}{N} \sum_{i=1}^{N} ((\boldsymbol{w}^\top \boldsymbol{x}_i + b) - y_i)^2 \tag{1}$$

  is minimised.th

- The act of finding $\boldsymbol{\theta}$ is called training.

- c.f. "least squares" – a parameter estimation method based on MSE or minimising the sum of squares of errors/residuals.

Linear regression: training with MSE

- The goal of linear regression is to solve

$$\min_{\boldsymbol{w},b} \quad \frac{1}{N}\sum_{i=1}^{N}((\boldsymbol{w}^\top \boldsymbol{x}_i + b) - y_i)^2. \tag{2}$$

- The optimal solution satisfies

$$\frac{\partial L}{\partial b} = 0\,, \qquad \frac{\partial L}{\partial \boldsymbol{w}} = \begin{bmatrix} \frac{\partial L}{\partial w_1} & \frac{\partial L}{\partial w_2} & \cdots & \frac{\partial L}{\partial w_d} \end{bmatrix}^\top = \mathbf{0}. \tag{3}$$

(Is this global optimal? More on this in lectures on optimisation.)

Linear regression: finding the bias b

$$\frac{\partial}{\partial b}\frac{1}{N}\sum_{i=1}^{N}((\boldsymbol{w}^\top \boldsymbol{x}_i + b) - y_i)^2 = \frac{2}{N}\sum_{i=1}^{N}((\boldsymbol{w}^\top \boldsymbol{x}_i + b) - y_i) \tag{4}$$

$$= 2b - \frac{2}{N}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) = 0 \tag{5}$$

$$b = \frac{1}{N}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) = \frac{1}{N}\sum_{i=1}^{N} y_i - \boldsymbol{w}^\top\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\right) = \bar{y} - \boldsymbol{w}^\top \bar{\boldsymbol{x}} \tag{6}$$

Linear regression: data centring (mean normalisation)

Using $b = \bar{y} - \boldsymbol{w}^\top \bar{\boldsymbol{x}}$,

$$L = \frac{1}{N}\sum_{i=1}^{N}\left((\boldsymbol{w}^\top \boldsymbol{x}_i + b) - y_i\right)^2 \tag{7}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\boldsymbol{w}^\top(\boldsymbol{x}_i - \bar{x}) - (y_i - \bar{y})\right]^2 \tag{8}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{w}^\top \dot{\boldsymbol{x}}_i - \dot{y}_i\right)^2 \tag{9}$$

where $\dot{\boldsymbol{x}}_i = \boldsymbol{x}_i - \bar{\boldsymbol{x}}, \quad \dot{y}_i = y_i - \bar{y}$

Linear regression: finding the weights w

$$\frac{\partial}{\partial \boldsymbol{w}} \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{w}^\top \dot{\boldsymbol{x}}_i - \dot{y}_i)^2 = \frac{2}{N} \sum_{i=1}^{N} (\boldsymbol{w}^\top \dot{\boldsymbol{x}}_i - \dot{y}_i)(\dot{\boldsymbol{x}}_i) \tag{10}$$

$$= \frac{2}{N} \sum_{i=1}^{N} ((\boldsymbol{w}^\top \dot{\boldsymbol{x}}_i)\dot{\boldsymbol{x}}_i - \dot{y}_i \dot{\boldsymbol{x}}_i) \tag{11}$$

Linear regression: finding the weights w *(cont.)*

$$\frac{\partial}{\partial \boldsymbol{w}} \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{w}^\top \dot{\boldsymbol{x}}_i - \dot{y}_i)^2 = \frac{2}{N} \sum_{i=1}^{N} ((\boldsymbol{w}^\top \dot{\boldsymbol{x}}_i)\dot{\boldsymbol{x}}_i - \dot{y}_i \dot{\boldsymbol{x}}_i) \tag{12}$$

$$= \frac{2}{N} \left(\begin{bmatrix} \dot{\boldsymbol{x}}_1 & \cdots & \dot{\boldsymbol{x}}_N \end{bmatrix} \begin{bmatrix} \boldsymbol{w}^\top \dot{\boldsymbol{x}}_1 \\ \vdots \\ \boldsymbol{w}^\top \dot{\boldsymbol{x}}_N \end{bmatrix} - \begin{bmatrix} \dot{\boldsymbol{x}}_1 & \cdots & \dot{\boldsymbol{x}}_N \end{bmatrix} \begin{bmatrix} \dot{y}_1 \\ \vdots \\ \dot{y}_N \end{bmatrix} \right) \tag{13}$$

$$= \frac{2}{N} \left(\mathbf{X}^\top \mathbf{X} \boldsymbol{w} - \mathbf{X}^\top \dot{\boldsymbol{y}} \right) = \mathbf{0} \tag{14}$$

$$\longrightarrow \quad \mathbf{X}^\top \mathbf{X} \boldsymbol{w} = \mathbf{X}^\top \dot{\boldsymbol{y}} \qquad \cdots \text{ normal equations} \tag{15}$$

$$\boldsymbol{w} = \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \dot{\boldsymbol{y}} \tag{16}$$

$$\text{where} \quad \mathbf{X} = \begin{bmatrix} \dot{x}_{11} & \cdots & \dot{x}_{1d} \\ & \vdots & \\ \dot{x}_{N} & \cdots & \dot{x}_{Nd} \end{bmatrix} = \begin{bmatrix} \dot{\boldsymbol{x}}_1^\top \\ \vdots \\ \dot{\boldsymbol{x}}_N^\top \end{bmatrix}$$

Linear regression - training process

1. Centring

$$\dot{\boldsymbol{y}} = \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_N - \bar{y} \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} \dot{\boldsymbol{x}}_1^\top \\ \vdots \\ \dot{\boldsymbol{x}}_N^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^\top - \bar{\boldsymbol{x}}^\top \\ \vdots \\ \boldsymbol{x}_N^\top - \bar{\boldsymbol{x}}^\top \end{bmatrix} \tag{17}$$

Linear regression - training process

1. Centring

$$\dot{\boldsymbol{y}} = \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_N - \bar{y} \end{bmatrix}, \qquad \dot{\mathbf{X}} = \begin{bmatrix} \dot{\boldsymbol{x}}_1^\top \\ \vdots \\ \dot{\boldsymbol{x}}_N^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^\top - \bar{\boldsymbol{x}}^\top \\ \vdots \\ \boldsymbol{x}_N^\top - \bar{\boldsymbol{x}}^\top \end{bmatrix} \tag{17}$$

2. Computing the weights $\boldsymbol{w}$ and b

$$\boldsymbol{w} = (\dot{\mathbf{X}}^\top \dot{\mathbf{X}})^{-1} \dot{\mathbf{X}}^\top \dot{\boldsymbol{y}} \tag{18}$$

$$b = \bar{\boldsymbol{y}} - \boldsymbol{w}^\top \bar{\boldsymbol{x}} \tag{19}$$

NB: $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ is called a Moore-Penrose pseudoinverse of $\mathbf{X}$.

In practice, we find the solution $\boldsymbol{w}$ without calculating $(\mathbf{X}^\top \mathbf{X})^{-1}$

Linear regression - training process

1. Centring

$$\dot{\boldsymbol{y}} = \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_N - \bar{y} \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} \dot{\boldsymbol{x}}_1^\top \\ \vdots \\ \dot{\boldsymbol{x}}_N^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^\top - \bar{\boldsymbol{x}}^\top \\ \vdots \\ \boldsymbol{x}_N^\top - \bar{\boldsymbol{x}}^\top \end{bmatrix} \tag{17}$$

2. Computing the weights $\boldsymbol{w}$ and b

$$\boldsymbol{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \dot{\boldsymbol{y}} \tag{18}$$

$$b = \bar{\boldsymbol{y}} - \boldsymbol{w}^\top \bar{\boldsymbol{x}} \tag{19}$$

NB: $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ is called a Moore-Penrose pseudoinverse of $\mathbf{X}$.

In practice, we find the solution $\boldsymbol{w}$ without calculating $(\mathbf{X}^\top \mathbf{X})^{-1}$

Question: Are the above $\{\boldsymbol{w}, b\}$ globally optimal?

What is $\mathbf{X}^\top\mathbf{X}$?

- $\mathbf{X}^\top\mathbf{X}$ is a $d \times d$ *symmetric matrix*, where $\mathbf{X} = \begin{bmatrix} \dot{\boldsymbol{x}}_1 & \cdots & \dot{\boldsymbol{x}}_N \end{bmatrix}^\top$ is a $N \times d$ matrix.
- $\mathbf{X}^\top\mathbf{X}$ is *positive semi-definite*, i.e. $\boldsymbol{z}^\top(\mathbf{X}^\top\mathbf{X})\boldsymbol{z} \geq 0$ for any $\boldsymbol{z} \in \mathbb{R}^d$

  NB: Eigen values of a positive semi-definite matrix are non-negative, i.e. $\lambda_i \geq 0$ for $i = 1, \ldots, d$
- $\boldsymbol{\Sigma} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$ is called an (empirical) **covariance matrix**
  - $\boldsymbol{\Sigma} = (\sigma_{ij})$: σ_{ii} is the variance of i-th dimension of data, σ_{ij} is the covariance between i-th and j-th dimensions of data.
  - used in many areas, e.g. multivariate normal distributions, principal component analysis (PCA)
- $\det(\boldsymbol{\Sigma}) = \displaystyle\prod_{i=1}^{d} \lambda_i$ and $tr(\boldsymbol{\Sigma}) = \sum_{i=1}^{d} \lambda_i$, where λ_i is the i-the eigenvalue of $\boldsymbol{\Sigma}$
- $\det(\boldsymbol{\Sigma}) = 0$ and $\text{rank}(\boldsymbol{\Sigma}) < d$ if dimensions are not linearly independent (e.g. $N \leq d$)

A probabilistic interpretation

- Assume we cannot get a perfect fit because of noise.
- In particular, we assume the noise is additive and Gaussian.
- In other words, $y_i = \boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
- If $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, then $y_i \sim \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_i + b, \sigma^2)$.
- The likelihood of $\boldsymbol{\theta}$: $p(y_1, \ldots, y_N \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\theta})$

$$= \prod_{i=1}^{N} p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid \boldsymbol{w}^\top \boldsymbol{x}_i + b, \sigma^2) \tag{20}$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - (\boldsymbol{w}^\top \boldsymbol{x}_i + b))^2\right) \tag{21}$$

A probabilistic interpretation *(cont.)*

- The log-likelihood of $\boldsymbol{\theta}$:

$$\log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - (\boldsymbol{w}^\top \boldsymbol{x}_i + b))^2\right) \tag{22}$$

$$= \sum_{i=1}^{N} \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - (\boldsymbol{w}^\top \boldsymbol{x}_i + b))^2\right] \tag{23}$$

- Note that the mean-squared error is given as

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - (\boldsymbol{w}^\top \boldsymbol{x}_i + b))^2 \tag{24}$$

- The maximum likelihood estimator is the optimal solution for MSE.

Evaluation measures

- RSS (residual sum of squares)

$$\mathrm{RSS} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{25}$$

- RMSE (root mean squared error)

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2} = \sqrt{\frac{1}{N} \mathrm{RSS}} \tag{26}$$

- R^2 – coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (\bar{y} - y_i)^2} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{Var}\left(\hat{y} - y\right)}{\mathrm{Var}\left(y\right)} \tag{27}$$

where TSS (total sum of squares) $= \sum_{i=1}^{N} (\bar{y} - y_i)^2$.

Example

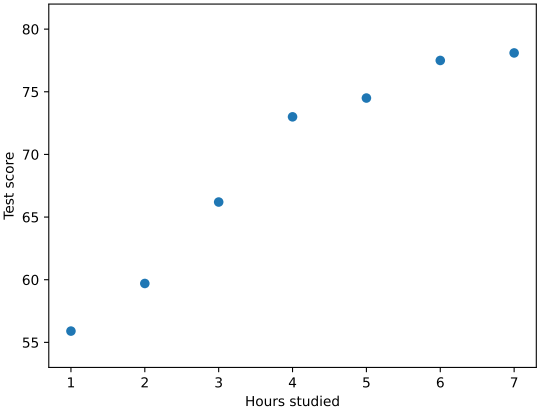
Let's say you're trying to figure out if studying more hours actually prepares your better for the exam, or if it's just the coffee. You decide to collect some data on how many hours you studied and what mark you got.

Hours Studied (x)	Test Score (y)
1	55.9
2	59.7
3	66.2
4	73.0
5	74.5
6	79.5
7	88.1

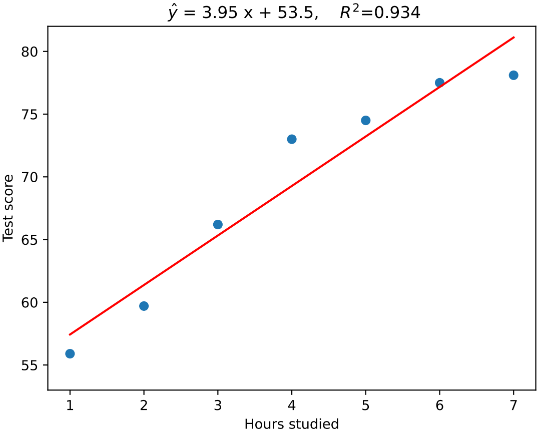
Fit a linear regression model to this data in the form $y = w_1 x + b$.

Adapted from Kia Nazarpour's slides on linear regression

Example *(cont.)*



Example *(cont.)*



Terminology

- Simple linear regression

$$y = wx + b, \quad \{y, x, w, b\} \in \mathbb{R}$$

- Multiple linear regression

$$y = \boldsymbol{w}^\top \boldsymbol{x} + b, \qquad y \in \mathbb{R},\ \boldsymbol{x} \in \mathbb{R}^d,\ \boldsymbol{w} \in \mathbb{R}^d,\ b \in \mathbb{R}$$

- Multivariate linear regression

$$\boldsymbol{y} = \mathbf{W}^\top \boldsymbol{x} + \boldsymbol{b}, \qquad \boldsymbol{y} \in \mathbb{R}^m,\ \boldsymbol{x} \in \mathbb{R}^d,\ \mathbf{W} \in \mathbb{R}^{d \times m},\ \boldsymbol{b} \in \mathbb{R}^m$$

Quizzes

1. What is the number of dimensions of the hyperplane formed by linear regression?

2. Give detailed derivations for Eqs. (13) and (14).

3. Find the Hessian of L defined in Eq. (9).

4. Show that $\mathbf{X}^\top \mathbf{X}$ that appears in Eq. (18) is positive semi-definite.