Engineering Mathematics 2B

Module 20: Linear Regression.

Nick Polydorides

School of Engineering

THE UNIVERSITY *of* EDINBURGH

Module 20 contents

Motivation

Supervised Machine Learning

Theory

Linear regression

Objectives & conditions

Least squares

Goodness of fit

Outcomes

Motivation: Supervised machine learning

We want to make a software that can predict the selling prices of houses in Edinburgh:

We will use historical sales data. This includes information on:

- Postcode: EH x_1
- Kitchen area: x_2 m^2
- Age of the house: x_3 yr
- Average energy bill: £ x_4
- Number of bedrooms: x_5

We want to learn the function

$$y = f(\mathbf{x}), \quad \mathbf{x} := (x_1, x_2, x_3, x_4, x_5)$$

using training data $\{(y_i, \mathbf{x}_i)\}_{i=1}^{n}$ from previous sales in Edinburgh.

Supervised ML

The construction $f = f(\mathbf{x})$ tells us what variables (data) are involved in this prediction and in what form (=template):

Say,

$$y = f_{\boldsymbol{\theta}}(\mathbf{x}) := \theta_0 + \theta_1 e^{1-x_1} + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4^2 + \theta_5 x_5, \text{ or}$$

$$y = f_{\boldsymbol{\theta}}(\mathbf{x}) := \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5$$

SML uses regression to fix the model parameters $\boldsymbol{\theta} := (\theta_0, \ldots, \theta_p)$ from **many** historical data $\{(y, \mathbf{x})\}$ to ‘tune the model’.

Then once we have a new house (not in our training dataset) we extract its data ($\mathbf{x}_t := x_1, \ldots, x_5$) and make a prediction $\hat{y} = f(\mathbf{x}_t)$.

Linear regression

Let $\mathbf{x}_i = (x_{i_1}, \ldots, x_{i_p})$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ be real vectors and suppose that data $(y_1, \ldots, y_n)$ where

$$y_i \quad \text{is a linear combination of the parameters}\, \boldsymbol{\theta}$$

via a model construction

$$y_i = \mathbf{x}_i^\top \boldsymbol{\theta} + \epsilon_i, \quad i = 1, \ldots, n, \quad n > p,$$

for unknown **independent** **random** errors ϵ_i.

The pairs $(y_i, \mathbf{x}_i)$ are called (model) data, and $\boldsymbol{\theta}$ are the respective model parameters.

Linear regression

In $y_i = \mathbf{x}_i^\top \boldsymbol{\theta} + \epsilon_i$, for $i = 1, \dots, n$

- y_i is the i-th **dependent** (response) variable.
- $\mathbf{x}_i = \{x_{i_1}, \dots, x_{i_p}\}$ are the **independent** (explanatory) variables of the i-th data/observation.
- $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ are the unknown **model parameters**,
- p is the **number of samples**
- The samples $(y_i, \mathbf{x}_i)$ are assumed independent.

Known: $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$.

Unknown: Deterministic parameters $\boldsymbol{\theta}$ and random noise ϵ_i.

Underpinning assumption: y and $\boldsymbol{\theta}$ are linearly related.

Toy example

Consider the single variable, single parameter model

$$f_\theta(x) = x\theta, \quad f : \mathbb{R} \to \mathbb{R}$$

where both x and θ are scalar. We don't know θ, and want to estimate it from data

$$x \longrightarrow \boxed{\theta} \longrightarrow f_\theta(x) = x\theta$$

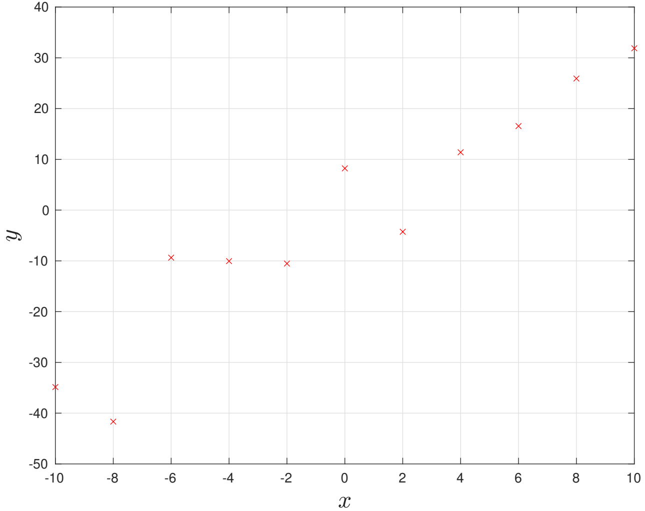
In reality however we can only have a noisy version of the data at different x_i points, for $i = 1, 2, \ldots$

$$x_i \longrightarrow \boxed{\theta} \longrightarrow f_\theta(x_i) \longrightarrow \boxed{\mathcal{N}(0, \sigma^2)} \longrightarrow y_i = f_\theta(x_i) + \epsilon_i$$

If we estimate the parameter θ then we can make predictions for any x.

Toy example

Data generated with $\theta = 3$ and $\sigma = 6$. What is $f(x = 1)$?



Line through the origin

Sometimes we know from physics not the data themselves that without the noise they belong to a **straight line through the origin**.

Take for example current and voltage data from Ohm's ($V = IR$) or force and extension data from Hooke's law ($F = Kx$).

We know the data correspond to a line and we know a point on that line. Here we assume the data satisfy the linear model

$$y_i = bx_i + \epsilon_i, \quad i = 1, \dots, n > p$$

$$p = 1 (= \text{number of parameters}), \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \theta = b$$

What is b? Find the line that goes through the origin and 'fits' the data the best.

Simple linear regression

Not all straight lines go through zero. Linear regression allows us to consider a more general linear model

$$y_i = a + bx_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n > p$$

or

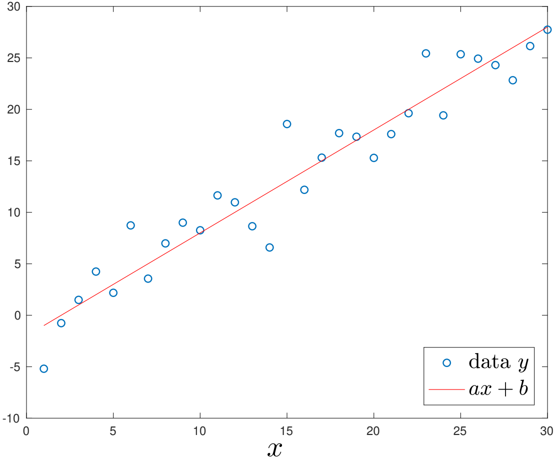
$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

where

$$p = 2, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} a \\ b \end{bmatrix}$$

What are a and b? Find the line that ‘fits’ the data the best.

Simple linear regression: Fitting $y = ax + b$ to noisy (x, y) data



Quadratic regression

Quadratic regression is based on the model

$$y_i = ax_i^2 + bx_i + c + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, ..., n > p$$

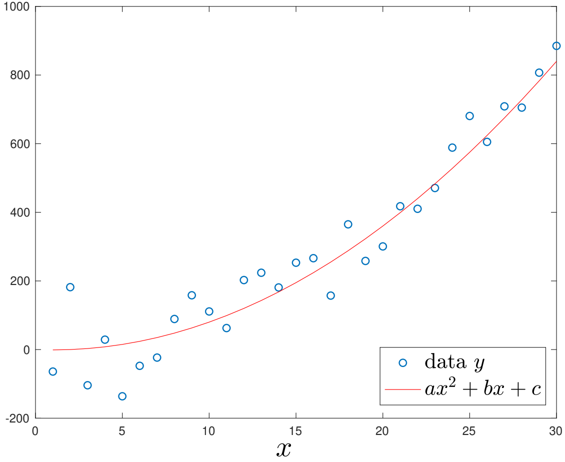
or $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ using data $\{(x_i, y_i)\}_{i=1}^n$. Here

$$p = 3, \quad \mathbf{X} = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

The function we are fitting is **quadratic** in the explanatory variables, but **linear** in the unknown coefficients.

What values of the parameters a, b and c fit best the quadratic function to the data?

Quadratic regression: Given data (x, y) find the best-fit parabola



Objectives in regression analysis

1. **To fit the data** considering the error margin. Typically via the method of Least Squares (LS) but there are other alternatives, e.g. least absolute sums, generalised LS.

2. **Robust parameters**. How does the response y change when an explanatory variable x changes?

3. **Good predictions**. What response can we expect to get under a new set of experimental conditions?

4. An indication of the **uncertainty** underlying 1-3 above using hypothesis tests and confidence intervals.

5. Developing a simple and functioning model. This is usually the result of an iterative refinement process.

Linear regression necessary conditions

Fitting a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ is meaningful when:

1. The data $\mathbf{y}$ are representative and meaningful, i.e. data are a random sample from the **same population**.
2. The regression equation is correct: $\mathbb{E}[\epsilon_i] = 0 \ \forall i$, i.e., **data are unbiased**.
3. Data errors are **uncorrelated**: $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \ \forall i \neq j$.
4. Matrix $\mathbf{X}$ is known **exactly**, i.e. its elements are exact.
5. **Homoscedasticity**. Data errors have the same, constant variance: $\mathbb{E}[\epsilon_i^2] = \sigma^2, \ \forall i$.
6. The errors $(\epsilon_1, \ldots, \epsilon_n)$ belong to a **joint normal distribution** (e.g. when the CLT applies). In effect, data $\mathbf{y}$ belong to a joint normal with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\theta}$ and $\boldsymbol{\Sigma} = \sigma^2 I$.

Estimating the parameters in linear regression

Consider the linear model where $\mathbf{y} \in \mathbb{R}^n$, $\boldsymbol{\theta} \in \mathbb{R}^p$ with $n \gg p$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \text{where} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I).$$

The **Least Squares Estimator** (LS) of the p parameters in $\boldsymbol{\theta}$, denoted as $\hat{\boldsymbol{\theta}}_{\text{LS}}$ is the one that has minimum variance and zero bias.

To compute it we seek the argument that minimises the sum of squared residuals loss function

$$\hat{\boldsymbol{\theta}}_{\text{LS}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \ell(\boldsymbol{\theta}) \doteq \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2,$$

recalling that $\|\epsilon\|_2^2 = \sum_{i=1}^n \epsilon_i^2 = \epsilon^\top \epsilon$ for any vector $\epsilon \in \mathbb{R}^n$.

Computing the Least Squares estimator

Since the loss $\ell(\boldsymbol{\theta})$ is quadratic in $\boldsymbol{\theta}$, we can find the argument where the unique minimum is attained by setting its gradient to zero

$$\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = 0$$

This mounts to solving the system of **normal equations**

$$\mathbf{X}^\top\mathbf{X}\,\boldsymbol{\theta} = \mathbf{X}^\top\mathbf{y},$$

which is a system of p linear equations in p unknowns. If there are p linearly independent observations/rows in $\mathbf{X}$ then matrix $\mathbf{X}^\top\mathbf{X}$ has full rank implying it has an inverse, hence

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}, \qquad \mathrm{Cov}(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1},$$

where $\mathrm{Cov}(\hat{\boldsymbol{\theta}}_{\mathrm{LS}})$ is a matrix with (i,j)-th entry $\mathrm{Cov}(\hat{\theta}_{i\,\mathrm{LS}}, \hat{\theta}_{j\,\mathrm{LS}})$.

Covariance proof

If $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a multivariate random variable with zero mean and covariance matrix $\Sigma = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \sigma^2 I$, then $\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ why is $\mathrm{Cov}(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$?

Combining the two formulas above gives

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta}+\boldsymbol{\epsilon} = \boldsymbol{\theta}+(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}$$

hence $\hat{\boldsymbol{\theta}}_{\mathrm{LS}} - \boldsymbol{\theta}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}$ Here $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ is a random variable with $\mathbb{E}[\hat{\boldsymbol{\theta}}_{\mathrm{LS}}] = \boldsymbol{\theta}$, hence

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}) = \mathrm{Cov}(\hat{a}_{\mathrm{LS}}, \hat{b}_{\mathrm{LS}}) &= \mathbb{E}\big[\big(\hat{\boldsymbol{\theta}}_{\mathrm{LS}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_{\mathrm{LS}} - \boldsymbol{\theta}\big)^\top\big] \\
&= \mathbb{E}\Big[\Big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}\Big)\Big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}\Big)^\top\Big] \\
&= \mathbb{E}\Big[\Big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}\Big)\boldsymbol{\epsilon}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Big) \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top]\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}
\end{aligned}$$

where we moved the scalar σ^2 at the front of the matrix expression and used the matrix id $(\mathbf{X}\mathbf{Y})^\top = \mathbf{Y}^\top\mathbf{X}^\top$.

Example

Compute the LS estimator for the linear regression model

$$y = a + bx + \epsilon,$$

and data

$$(x, y) = \Big\{(1, 7.2725), (2, 6.7041), (3, 5.2231), (4, 3.6351), (5, 3.9816)\Big\}$$

Tabulating the data leads to

$$\mathbf{y} = \begin{bmatrix} 7.2725 \\ 6.7041 \\ 5.2231 \\ 3.6351 \\ 3.9816 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\theta} = \begin{bmatrix} a \\ b \end{bmatrix}$$

(Data were simulated with $\epsilon_i \sim \mathcal{N}(0, 0.25)$, $a = 8$, $b = -1$.)

Example

Forming the normal equations yields

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}, \quad \text{and} \quad \mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 26.8165 \\ 70.7987 \end{bmatrix}$$

The determinant $|\mathbf{X}^\top \mathbf{X}| = 50$ so the matrix is invertible[1] and therefore

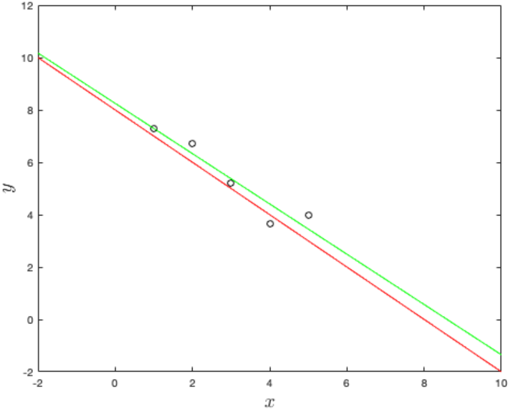
$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 8.25 \\ -0.96 \end{bmatrix}$$

thus

$$\hat{a}_{\mathrm{LS}} = 8.25, \quad \hat{b}_{\mathrm{LS}} = -0.96$$

[1] Recall that singular matrices have zero determinant, and are not invertable.

Example graphical



Good fit for a small data sample (∘). In red is the true line for $(a, b) = (8, -1)$ and in green the LS estimated at $(\hat{a}_{\mathrm{LS}}, \hat{b}_{\mathrm{LS}}) = (8.25, -0.96)$.

Formulas for simple linear regression

Instead of computing gradients or inverting matrices, we can get $\hat{a}$ and $\hat{b}$ in the linear model $y = a + bx + \epsilon$ using formulas. Like before we are given $n > 2$ data points $\{(x_i, y_i)\}_{i=1}^{n}$.

Due to the zero-mean Gaussian noise assumption, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ we effectively have $\big(y_i - (a + bx_i)\big) \sim \mathcal{N}(0, \sigma^2)$, which means that our underlying assumption on the data points we have is

$$p(\mathbf{y} \,|\, \mathbf{x}, a, b) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - a - bx_i)^2\Big\}$$

Note that this assumes that the x_i points are exact.

Formulas for simple linear regression

Compute from the data

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \quad SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

and then the regression line is $y = \hat{a}_{\text{LS}} + \hat{b}_{\text{LS}} x$ where

$$\hat{b}_{\text{LS}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{SS_x}, \quad \hat{a}_{\text{LS}} = \bar{y} - \hat{b}_{\text{LS}}\bar{x}.$$

Note that a_{LS} and b_{LS} are correlated unless $\bar{x} = 0$. This means errors in computing the one, contaminate the computation of the other. This hints on how to collect the data!

Linear regression coefficients derivation

From $y_i = a + bx_i + \epsilon_i$ for $i = 1, \dots, n$ and $\hat{\boldsymbol{\theta}}_{\text{LS}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$, recalling that $\|\epsilon\|_2^2 = \sum_{i=1}^n \epsilon_i^2$ then

$$\begin{aligned}(\hat{a}_{\text{LS}}, \hat{b}_{\text{LS}}) &= \arg\min \sum_{i=1}^n \big(y_i - (a + bx_i)\big)^2 := \ell(a,b) \\ &= \arg\min \sum_{i=1}^n y_i^2 + b^2 x_i^2 + 2abx_i - 2ay_i + a^2 - 2bx_iy_i\end{aligned}$$

Then setting $\frac{\partial \ell(a,b)}{\partial a} = 0 = \sum_{i=1}^n 2bx_i - 2y_i + 2a$ which by rearrangement gives

$$na = \sum_{i=1}^n y_i - b\sum_{i=1}^n x_i, \quad \Rightarrow \quad a = \bar{y} - b\bar{x}.$$

derivation cont

Similarly, from $\frac{\partial \ell(a,b)}{\partial b} = 0$ we have

$$0 = \sum_{i=1}^{n} x_i \big(y_i - a - b x_i \big) = \sum_{i=1}^{n} x_i \big(y_i - \bar{y} + b\bar{x} - b x_i \big)$$

hence

$$b = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}$$

To see why the above fits the form given in the previous slide notice that since $\sum_{i=1}^{n}(x_i - \bar{x}) = 0 = \sum_{i=1}^{n}(y_i - \bar{y})$ we can subtract zero from the numerator to yield

$$\begin{aligned} \sum_{i=1}^{n} x_i (y_i - \bar{y}) - \bar{x} \sum_{i=1}^{n} (y_i - \bar{y}) &= \sum_{i=1}^{n} x_i (y_i - \bar{y}) - \sum_{i=1}^{n} \bar{x} (y_i - \bar{y}) \\ &= \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}), \end{aligned}$$

and similarly for the denominator.

Normality of the regressed coefficients

From the linear model

$$y_i = a + bx_i + \epsilon_i, \quad i = 1, \ldots, n$$

we see that for exact x_i

$$y_i \sim \mathcal{N}(a + bx_i, \sigma^2), \quad \text{since} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

In turn $\hat{a}$ and $\hat{b}$ are also **normal**, although typically **correlated** (unless $\bar{x} = 0$) centred at their exact values

$$\hat{a}_{\text{LS}} \sim \mathcal{N}\left(a, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)\right), \quad \mathbb{E}[\hat{a}] = a, \quad \text{and}$$

$$\hat{b}_{\text{LS}} \sim \mathcal{N}\left(b, \frac{\sigma^2}{SS_x}\right), \quad \mathbb{E}[\hat{b}] = b,$$

where $SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Distribution of LS estimators

If a_{LS} and b_{LS} are normal with known means and variances, what can be said about their joint distribution?

From the template of the bivariate Gaussian, the only bit of information we are missing is the correlation ρ.

$$\rho(\hat{a}_{\text{LS}}, \hat{b}_{\text{LS}}) = \frac{\text{Cov}(\hat{a}_{\text{LS}}, \hat{b}_{\text{LS}})}{\sqrt{\text{Var}(\hat{a}_{\text{LS}}) \cdot \text{Var}(\hat{b}_{\text{LS}})}} = -\frac{\bar{x}\sqrt{n}}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

Also the covariance is

$$\text{Cov}(\hat{a}_{\text{LS}}, \hat{b}_{\text{LS}}) = -\frac{\bar{x}\sigma^2}{SS_x}$$

Data prediction with regression

Once we estimated the coefficients we can predict new data $\hat{y}_*$ for new x_* points as

$$\hat{y}_* = \hat{a}_{\mathrm{LS}} + \hat{b}_{\mathrm{LS}} x_*,$$

an operation often referred to as **data prediction**.

The predicted data $\hat{\mathbf{y}}$ for the original points x used in to estimate the regression line, these relate to the original data $\mathbf{y}$ via the orthogonal projection matrix $\mathbf{H}$ as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\theta}} = \underbrace{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}_{\mathbf{H}}\mathbf{y}$$

that allows to compute the Residual Sum of Squares in the data RSS that tells us how well, overall, can we predict the original data given the linear model we estimated, as

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{H}\mathbf{y}$$

MLE & LS

Since we know that

$$y_i \sim \mathcal{N}(a + bx_i, \sigma^2), \quad \text{since} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

then the PDF of the i-th datum is

$$p_{Y_i}(y_i|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - \mathbf{X}_{(i)}\boldsymbol{\theta}\right)^2\right\}$$

where $\mathbf{X}_{(i)}$ is the i-th row of $\mathbf{X}$.

If $\{y_1, \ldots, y_n\} := \mathbf{y}$ are iid then the likelihood function of all these data, conditioned on the unknown parameters $\boldsymbol{\theta}$ and the variance σ^2 is

$$L(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p_{Y_i}(y_i|\boldsymbol{\theta})$$

Maximum likelihood & Least squares

Taking the product of n Gaussian PDFs yields the joint likelihood of $\mathbf{y}$ data vector

$$\begin{aligned} L(\mathbf{y}|\boldsymbol{\theta}) &= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\exp\Big\{-\frac{1}{2\sigma^2}\Big[(y_1 - \mathbf{X}_{(1)}\boldsymbol{\theta})^2 + \dots \\ &\qquad \dots + (y_n - \mathbf{X}_{(n)}\boldsymbol{\theta})^2\Big]\Big\} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\exp\Big\{-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2\Big\} \end{aligned}$$

Thus the **maximum likelihood estimator** (MLE) for $\hat{\boldsymbol{\theta}}$ coincides with the least squares estimator

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} L(\mathbf{y}|\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 = \hat{\boldsymbol{\theta}}_{\text{LS}}.$$

How good is the LS fit?

The **goodness of fit** is typically measured by the R^2 statistic, the coefficient of determination

$$R^2 = 1 - \frac{RSS}{SS_\epsilon}$$

where the sum of squared deviations (aka Total Sum of Squares) in the data is

$$SS_\epsilon = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

R^2 is the fraction of the variance in the data y that is captured by the regressed model, that is how much of SS_ϵ can be explained (=predicted) by the model.

A perfect fit corresponds to $R^2 = 1$, but this is unrealistic.

Formulas

- For $y_i = a + bx_i + \epsilon_i$, with $i = 1, \ldots, n$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ then $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and the LS line is

  $$y = \hat{a}_{\mathrm{LS}} + \hat{b}_{\mathrm{LS}}\, x, \quad \hat{b}_{\mathrm{LS}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{a}_{\mathrm{LS}} = \bar{y} - \hat{b}_{\mathrm{LS}}\bar{x}.$$

- $$\hat{a}_{\mathrm{LS}} \sim \mathcal{N}\left(a, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)\right),\ \hat{b}_{\mathrm{LS}} \sim \mathcal{N}\left(b, \frac{\sigma^2}{SS_x}\right),$$

  where $SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$

- $$R^2 = 1 - \frac{RSS}{SS_\epsilon}, \quad SS_\epsilon = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  where $\hat{y}_i = \hat{a}_{\mathrm{LS}} + \hat{a}_{\mathrm{LS}} x_i$.

Main outcomes of module 20

You MUST know:

1. The simple linear regression models.
2. The conditions for SLR.
3. To compute the line parameters using least squares in matrix and formula forms.
4. How to compute the goodness of fit.