

AI for biology

JxLi

September 13, 2025

1 Supervised Learning

1.1 Linear Regression

1.1.1 simple linear regression

In simple linear regression, we model the relationship between a single variable x and a dependent variable y using a linear function:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where

- y is the dependent variable.
- x is the independent variable.
- β_0 is the intercept.
- β_1 is the slope coefficient.
- ε is the random error term.

The difference between the observed value y_i and the predicted value \hat{y}_i is called the residual:

$$e_i = y_i - \hat{y}_i$$

Suppose our regression model is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where we assume the errors ε_i are independent and normally distributed:

$$\varepsilon_i \sim N(0, \sigma^2).$$

This means each observed value y_i is a random variable with density:

$$p(y_i | x_i, \beta_0, \beta_1, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right).$$

The likelihood of observing all data points is:

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^n p(y_i | x_i, \beta_0, \beta_1, \sigma)$$

We need to find the $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}$, that maximize $L(\beta_0, \beta_1, \sigma)$, which means

$$\max \prod_{i=1}^n p(y_i | x_i, \beta_0, \beta_1, \sigma) \Leftrightarrow \max \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left(-\frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right)$$

For fixed σ , maximizing $L(\beta_0, \beta_1)$ is equivalent to minimizing $\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$, which is exactly the sum of squared residuals.

From a probability perspective, we prove that using the sum of squared residuals (SSR) to evaluate the loss of function is mathematically reasonable.

To figure out the right parameters, we try as follows:

$$\begin{aligned} f(\beta_0, \beta_1) &= \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2, \\ \frac{\partial f}{\partial \beta_0} &= \frac{\partial f}{\partial \beta_1} = 0, \\ \sum_{i=1}^n y_i - \beta_0 - \beta_1 x_i &= 0, \quad \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \beta_0 &= \hat{y}_i - \beta_1 \hat{x}_i, \quad \sum_{i=1}^n x_i (y_i - \hat{y}_i - \beta_1 (x_i - \hat{x}_i)) \\ \beta_1 &= \frac{\sum_{i=1}^n x_i (y_i - \hat{y}_i)}{\sum_{i=1}^n x_i (x_i - \hat{x}_i)} = \frac{\sum_{i=1}^n (x_i - \hat{x}_i) (y_i - \hat{y}_i)}{\sum_{i=1}^n (x_i - \hat{x}_i)^2} \end{aligned}$$

1.1.2 multiple linear regression

In matrix notation, the multiple linear regression can be written as follows:

$$y = X\beta + \varepsilon$$

where

- y is the $n \times 1$ vector
- X is the $n \times (p+1)$ matrix
- β is the $(p+1) \times 1$ vector of regression coefficients
- ε is the $n \times 1$ vector of random error terms, assumed to follow $\varepsilon \sim N(0, \sigma^2 I)$

Similarly, the least squares estimator of β is obtained by minimizing the sum of squared residuals:

$$\hat{\beta} = \arg \min_{\beta} \|y - X\beta\|^2.^1$$

which comes from as follows:

$$p(y|X, \beta, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{1}{2\sigma^2} \underbrace{(y - X\beta)^T (y - X\beta)}_{\|y - X\beta\|^2} \right)$$

¹The notation of *arg* means to find a argument that satisfy the requirements of the subsequent function. In the text, we are saying "Find the vector β that makes the sum of the squared residuals as small as possible."

$$L(\beta, \sigma) = \prod_{i=1}^n p(y_i | x_i^T, \beta, \sigma) = \prod_{i=1}^n p(y_i | X, \beta, \sigma) = p(y | x_i^T, \beta, \sigma).^2$$

For the same reason ,we find the targeted β only when function $(y - X\beta)^T(y - X\beta)$ get minimized.

$$(y - X\beta)^T(y - X\beta) = (y^T - \beta^T X^T)(y - X\beta) = y^T y + (X\beta)^T X\beta - \beta^T X^T y - y^T X\beta$$

Since $y^T X\beta$ is a scalar , $y^T X\beta = \beta^T X^T y$

$$(y - X\beta)^T(y - X\beta) = y^T y + (X\beta)^T X\beta - 2y^T X\beta.$$

To explain it from a higher perspective ,I want to introduce the matrix calculus rules .

Firstly ,let's assume a vector $e = e_1, e_2, \dots, e_n$

$$\frac{\partial e^T e}{\partial e} = \frac{\sum_{i=1}^n e_i^2}{\partial e} = 2e$$

Then ,

$$\begin{aligned} \frac{\partial (y - X\beta)^T (y - X\beta)}{\partial \beta} &= \frac{\partial (y^T y + (X\beta)^T X\beta - 2y^T X\beta)}{\partial \beta} = \frac{\partial (\beta^T X^T X\beta)}{\partial \beta} - 2X^T y \\ \frac{\partial (\beta^T X^T X\beta)}{\partial \beta} &= \left(\frac{\partial (X\beta)}{\partial \beta} \right)^T \frac{\partial (\beta^T X^T X\beta)}{\partial (X\beta)} = 2X^T X\beta \end{aligned} \quad (1)$$

Equation (1) comes from multivariable chain rule:

$$\nabla_{\beta} f = J_{v,\beta}^T \nabla_v f,$$

where

- $J_{v,\beta} = \frac{\partial v}{\partial \beta}$ is the Jacobian of v w.r.t. β
- $\nabla_v f$ is the gradient w.r.t. v

To review the concept of Jacobian,we take multivariable function f and column vector $x = [x_1, x_2, \dots, x_n]^T$:

$$J_{f,x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_2}{\partial x_3} & \dots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \frac{\partial f_m}{\partial x_3} & \dots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}, J_{f,x}(i; j) = \frac{\partial f_i}{\partial x_j}$$

As for our case ,

$$J_{v,\beta}^T = \left(\frac{\partial (X\beta)}{\partial \beta} \right)^T = X^T.$$

Now we can solve the $\hat{\beta}$ we want:

$$\frac{\partial (y - X\beta)^T (y - X\beta)}{\partial \beta} = 2X^T (X\beta - y) = 0$$

²Normally ,we take x_i as a column vector and x_i^T as a row vector.

if the $X^T X$ is invertible

$$(X^T X)^{-1} X^T X \beta = (X^T X)^{-1} X^T y$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

else ,like the X is not full ranked,we need to use the Moore-Penrose pseudoinverse:

$$\hat{\beta} = X^+ y$$

where $X^+ = (X^T X)^+ X^T$.³

1.1.3 post-estimation analysis

We always need a way to evaluate our function ,and it is called post-estimation analysis. Take multivariable linear regression for an example ,we do as follows:

First ,get our predicted values

$$\hat{y} = X \hat{\beta}.$$

Calculate the residual:

$$r = y - \hat{y}$$

To evaluate the residual without bias ,we divided the squared residual by degrees of freedom(DF) $n - (p + 1)$:

$$\hat{\sigma}^2 = \frac{r^T r}{n - (p + 1)}$$

Why is it unbiased?

$$r = y - X \hat{\beta} = y - X(X^T X)^{-1} X^T y = (I - H)y$$

where $H = X(X^T X)^{-1} X^T$ is the hat matrix ,which put a hat on y to transform it to \hat{y} as for hat matrix ,it has some good properties:

- Symmetric: $H = H^T$
- Idempotent: $H^2 = H$

$$r = (I - H)(X\beta + \varepsilon) = (I - H)X\beta + (I - H)\varepsilon$$

where $HX\beta = X\beta$

$$r = (I - H)\varepsilon$$

since $I - H$ is symmetric and idempotent , $(I - H)^T(I - H) = (I - H)^2 = I - H$

$$r^T r = \varepsilon^T (I - H)^T (I - H) \varepsilon = \varepsilon^T (I - H) \varepsilon$$

$$\mathbb{E}(r^T r) = \mathbb{E}(\varepsilon^T (I - H) \varepsilon)$$

Lemma 1 (Expectation of a quadratic form). *Let $\varepsilon \in \mathbb{R}^m$ be a random vector with*

$$\mathbb{E}[\varepsilon] = 0, \quad \text{Cov}(\varepsilon) = \sigma^2 I_m,$$

and let $A \in \mathbb{R}^{m \times m}$ be a symmetric matrix. Then

$$\mathbb{E}[\varepsilon^T A \varepsilon] = \sigma^2 \text{tr}(A).$$

³We will talk about it later if possible

Proof. Expand the quadratic form:

$$\varepsilon^\top A \varepsilon = \sum_{i=1}^m \sum_{j=1}^m A_{ij} \varepsilon_i \varepsilon_j.$$

Taking expectation:

$$\mathbb{E}[\varepsilon^\top A \varepsilon] = \sum_{i=1}^m \sum_{j=1}^m A_{ij} \mathbb{E}[\varepsilon_i \varepsilon_j].$$

Since $\text{Cov}(\varepsilon) = \sigma^2 I_m$, $\text{Cov}(\varepsilon_i, \varepsilon_j) = \mathbb{E}(\varepsilon_i - \mathbb{E}\varepsilon)(\varepsilon_j - \mathbb{E}\varepsilon) = \mathbb{E}(\varepsilon_i \varepsilon_j)$, we have $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$ for $i \neq j$ and $\mathbb{E}[\varepsilon_i^2] = \sigma^2$. Therefore, only diagonal terms remain:

$$\mathbb{E}[\varepsilon^\top A \varepsilon] = \sum_{i=1}^m A_{ii} \sigma^2 = \sigma^2 \text{tr}(A).$$

□

$$\mathbb{E}(\varepsilon^\top (I - H) \varepsilon) = \sigma^2 \text{tr}(I - H) = \sigma^2 (n - (p + 1))$$

$$\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}(\sigma^2)$$

This shows that $\hat{\sigma}^2$ is an unbiased estimator of the true error variance σ^2 .