AI for biology

JxLi

September 15, 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 errorsε_i are independent and normally distributed:

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

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

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

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

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

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:

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2,$$

$$\frac{\partial f}{\partial \beta_0} = \frac{\partial f}{\partial \beta_0} = 0,$$

$$\sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i = 0, \sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\beta_0 = \hat{y}_i - \beta_1 \hat{x}_i, \sum_{i=1}^{n} x_i(y_i - \hat{y}_i - \beta_1(x_i - \hat{x}_i))$$

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \hat{y}_i)}{\sum_{i=1}^{n} x_i(x_i - \hat{x}_i)} = \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}$$

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:

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

[1] The 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."

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

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

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

Since $y^TX\beta$ is a scalar ,$y^TX\beta=\beta^TX^Ty$

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

To explain it from a higher perspective ,I want to introduce the matrix calculus rules . Firstly ,let's assume a vector $e=[e_1,e_2,\cdots,e_n]$

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

Then ,

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

$$\frac{\partial(\beta^TX^TX\beta)}{\partial\beta}=(\frac{\partial(X\beta)}{\partial\beta})^T\frac{\partial(\beta^TX^TX\beta)}{\partial(X\beta)}=2X^TX\beta \tag{1}$$

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 $x=[x_1,x_2,\cdots,x_n]^T$:

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

As for our case ,

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

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

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

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

if the X^TX is invertible

$$(X^TX)^{-1}X^TX\beta = (X^TX)^{-1}X^Ty$$

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

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

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

where $X^+ = (X^TX)^+X^T$.[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) $n - (p+1)$:

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

Why is it unbiased?

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

where $H = X(X^TX)^{-1}X^T$ 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 ,$(I-H)^T(I-H) = (I-H)^2 = I-H$

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

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

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

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

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

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

[3] We 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 $\mathrm{Cov}(\varepsilon) = \sigma^2 I_m, \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \mathbb{E}[(\varepsilon_i - \mathbb{E}\varepsilon_i)(\varepsilon_j - \mathbb{E}\varepsilon_j)] = \mathbb{E}[\varepsilon_i\varepsilon_j]$, 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:

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

□

$$\mathbb{E}[\varepsilon^T(I-H)\varepsilon] = \sigma^2 tr(I-H) = \sigma^2(n-(p+1))$$
$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}[\sigma^2]$$

This shows that $\hat{\sigma}^2$ is an unbiased estimator of the true error variance σ^2.
To compute the **R-squared** ,we firstly compute the **total sum of squares (TSS)** and the **residual sum of squares (RSS)**:

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 = ||y - \bar{y}\mathbf{1}||_2^2$$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = ||y - \hat{y}||_2^2$$

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{||y - \bar{y}\mathbf{1}||_2^2}{||y - \hat{y}||_2^2}$$

That is the normal way to get **R-squared** , if we take freedom degrees into thoughts ,the adjusted R^2 should be penalized as follows:

$$R^2_{\text{adj}} = 1 - \frac{RSS/(n-(p+1))}{TSS/(n-1)}$$

Then , we can test the significance of the predictor :

$$\hat{\beta} = (X^TX)^{-1}X^Ty = (X^TX)^{-1}X^T(X\beta + \varepsilon)$$

$$\mathrm{Cov}(\hat{\beta}) = \mathrm{Cov}((X^TX)^{-1}X^T\varepsilon) = (X^TX)^{-1}X^T\mathrm{Cov}(\varepsilon)((X^TX)^{-1}X^T)^T$$
$$= (X^TX)^{-1}X^T\sigma^2 I((X^TX)^{-1}X^T)^T = \sigma^2(X^TX)^{-1}X^TX(XX^T)^{-1} = \sigma^2(X^TX)^{-1}$$

And the variance of each component of $\hat{\beta}$ is given by the diagonal elements of the $\mathrm{Cov}(\hat{\beta})$:

$$\mathrm{Var}(\hat{\beta}_i) = \mathrm{Cov}(\hat{\beta})_{ii} = \sigma^2(X^TX)^{-1}_{ii}.$$

Thus the standard errors of $\hat{\beta}_i$:

$$SE(\hat{\beta}_i) = \sqrt{\sigma^2 (X^T X)^{-1}_{ii}}$$

To find that how many standard errors $\hat{\beta}$ away is from the hypothesized value ,we introduce the way of **t-statistics**:

$$t_j = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)}$$

where

- $\hat{\beta}_j$ is the estimated coefficient
- $\beta_{j,0}$ is the hypothesized true value under null hypothesis
- $\hat{\beta}_j$ is the estimated standard value