

Lecture Notes - 03: Linear Regression

Dihui Lai

September 29, 2020

Contents

1	Linear Regression	1
1.1	Single Variate Linear Regression	1
1.2	Multivariate Linear Regression	3
2	Likelihood Function	4
2.1	Definition	4
2.2	Example: Likelihood of Bernoulli Distribution	5
2.3	Example: Likelihood of Gaussian Distribution	5
3	Maximum Likelihood Estimator of the Multivariate Linear Model	5

1 Linear Regression

1.1 Single Variate Linear Regression

A simple linear model can be formulated by assuming the target variable is dependent only on one predictor i.e.

$$\hat{y} = \beta_0 + \beta_1 x$$

In order to have our estimate as close as to the actual value of y , we want to find the β s that minimize the sum squared error function

$$\epsilon = \sum_{i=1}^n (y^i - \hat{y}^i)^2 = \sum_{i=1}^n (y^i - \beta_0 - \beta_1 x^i)^2 \quad (1)$$

i.e.

$$\begin{cases} \frac{\partial \epsilon}{\partial \beta_1} = 0 \\ \frac{\partial \epsilon}{\partial \beta_0} = 0 \end{cases} \quad (2)$$

$$\Rightarrow \begin{cases} \sum_{i=1}^n (y^i - \beta_0 - \beta_1 x^i) x^i = 0 \\ \sum_{i=1}^n (y^i - \beta_0 - \beta_1 x^i) = 0 \end{cases} \quad (3)$$

Sorting the equations to get the solution for β_0 and β_1

$$\begin{aligned} \Rightarrow \begin{cases} \beta_0 \sum_{i=1}^n x^i = \sum_{i=1}^n y^i x^i - \beta_1 \sum_{i=1}^n x^i x^i \\ \sum_{i=1}^n \beta_0 = \sum_{i=1}^n y^i - \beta_1 \sum_{i=1}^n x^i \end{cases} \\ \Rightarrow \begin{cases} \beta_1 = \frac{\sum_{i=1}^n y^i x^i - \beta_0 \sum_{i=1}^n x^i}{\sum_{i=1}^n x^i x^i} = \frac{\frac{1}{n} \sum_{i=1}^n y^i x^i - \beta_0 \bar{x}}{\frac{1}{n} \sum_{i=1}^n x^i x^i} \\ \beta_0 = \frac{1}{n} (\sum_{i=1}^n y^i - \beta_1 \sum_{i=1}^n x^i) = (\bar{y} - \beta_1 \bar{x}) \end{cases} \end{aligned} \quad (4)$$

Substitute β_0 in to the first equation in equation set (4). We have

$$\beta_1 = \frac{\frac{1}{n} \sum_{i=1}^n y^i x^i - (\bar{y} \bar{x} - \beta_1 \bar{x} \bar{x})}{\frac{1}{n} \sum_{i=1}^n x^i x^i}$$

Solving for β_1 we have

$$\begin{aligned} \beta_1 \frac{1}{n} \sum_{i=1}^n x^i x^i &= \frac{1}{n} \sum_{i=1}^n y^i x^i - \bar{y} \bar{x} + \beta_1 \bar{x} \bar{x} \\ \Rightarrow \beta_1 \frac{1}{n} \sum_{i=1}^n x^i x^i &= \frac{1}{n} \sum_{i=1}^n y^i x^i - \bar{y} \bar{x} + \beta_1 \bar{x} \bar{x} \end{aligned}$$

$$\Rightarrow \beta_1 \left(\frac{1}{n} \sum_{i=1}^n x^i x^i - \bar{x}\bar{x} \right) = \frac{1}{n} \sum_{i=1}^n y^i x^i - \bar{y}\bar{x}$$

Thus we get the solution of β_1

$$\beta_1 = \frac{\frac{1}{n} \sum_{i=1}^n y^i x^i - \bar{y}\bar{x}}{\frac{1}{n} \sum_{i=1}^n x^i x^i - \bar{x}\bar{x}} \quad (5)$$

Take a close look it is not hard to find that the numerator is the co-variance of X and Y . The denominator is the variance of X .

Therefore β_1 can also be written as

$$\beta_1 = \frac{Cov(X, Y)}{Var(X)} = \rho_{XY} \frac{\sigma_X}{\sigma_Y} \quad (6)$$

1.2 Multivariate Linear Regression

Assume y is a linear superposition of multiple x s, the model for y is then formulated as

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$$

or simply

$$\hat{y} = \sum_{j=1}^m \beta_j x_j = \vec{\beta} \cdot \vec{x}$$

To estimate β s that best fit the data, we need to minimize the error

$$\epsilon = \sum_{i=1}^n (y^i - \hat{y}^i)^2 \quad (7)$$

$$= \sum_{i=1}^n (y^i - \vec{x}^i \cdot \vec{\beta})^2 \quad (8)$$

Writing in matrix notation, we have

$$\begin{aligned} \epsilon &= (y - \hat{y})^T (y - \hat{y}) \\ &= (y - X\beta)^T (y - X\beta) \end{aligned}$$

here β is a $m \times 1$ matrix defined as

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$$

To minimize the error ϵ we want to the β s satisfy the following equation set:

$$\frac{\partial \epsilon}{\partial \beta_j} = 0, j = 1, 2, 3, 4 \dots m$$

Using equation (7), we have

$$\sum_{i=1}^n \frac{\partial}{\partial \beta_j} (y^i - \hat{y}^i) = 0$$

$$\sum_{i=1}^n (y^i - \hat{y}^i) \frac{\partial \hat{y}^i}{\partial \beta_j} = 0 \quad (9)$$

$$\Rightarrow \sum_{i=1}^n (y^i - \hat{y}^i) x_j^i = 0 \quad (10)$$

Going from equation (9) to equation (10), we use the fact that $\hat{y} = \vec{x} \cdot \vec{\beta} = \sum_{l=1}^n x_l^i \beta_l$ and

$$\frac{\partial \hat{y}^i}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \sum_{l=1}^n x_l^i \beta_l = x_j^i \quad (11)$$

Write equation (10) in matrix format we have

$$(y - X\beta)^T X = 0$$

or after transposing

$$X^T y - X^T X \beta = 0$$

Therefore the β that minimize ϵ has to satisfy the following equation

$$\beta = (X^T X)^{-1} X^T y \quad (12)$$

Loosely speaking, we can interpreting equation (12) as composed two components the covariance related term $X^T y$ and a term that is related to the variance of X i.e. $X^T X$

2 Likelihood Function

2.1 Definition

If a set of random variables $Y_1, Y_2 \dots Y_n$ has a joint probability distribution density/mass $f(y_1, y_2, \dots y_n; \vec{\theta})$, where $\vec{\theta}$ is a set of parameters, the likelihood function is defined as

$$L(\vec{\theta}) = f(y_1, y_2, \dots y_n; \vec{\theta}) \quad (13)$$

Where $y_1, y_2, \dots y_n$ are values drawn from the random distribution $Y_1, Y_2 \dots, Y_n$

2.2 Example: Likelihood of Bernoulli Distribution

Assuming an event has two possible outcomes $y = 1$ or $y = 0$, with probability p of being 1, i.e. the outcome follows a Bernoulli distribution. As we learned in lecture 2, the probability mass function is

$$f(y; p) = \begin{cases} p, & y = 1 \\ 1 - p, & y = 0 \end{cases}$$

Or if you express this in a single formula it is

$$f(y; p) = p^y(1 - p)^{1-y}$$

you can verify that when $y = 1$, $f(y, p) = p$ and when $y = 0$, $f(y, p) = 1 - p$

When you draw a data point from this distribution, the likelihood that the value being 1 is $f(1, p) = p$ and the value being 0 is $f(0, p) = 1 - p$

2.3 Example: Likelihood of Gaussian Distribution

The probability density function (PDF) for a standard Gaussian distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

When you draw a data point from this distribution, the likelihood that the value being 1 is

$$f(1, p) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}}$$

The likelihood that the value being 0.5 is

$$f(0.5, p) = \frac{1}{\sqrt{2\pi}} e^{-\frac{0.5^2}{2}}$$

3 Maximum Likelihood Estimator of the Multivariate Linear Model

Assume the target variable y in a dataset is drawn from a Gaussian distribution and for the i th data point, the Gaussian distribution has a mean μ_i ; the variance for all the data points are the same i.e. σ^2 . The Gaussian distribution is therefore

$$f(y, \mu_i, \sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu_i)^2}{2\sigma^2}}, i = 1, 2, 3, \dots, n$$

Moreover, assume that the parameter μ_i can be estimated from the predictors of the i th data point: x_j^i , where $j = 1, 2, 3, \dots, m$

$$\mu_i = \sum_{j=1}^m \beta_j x_j^i = \vec{\beta} \cdot \vec{x}^i \quad (14)$$

The likelihood of i th data point being y^i is therefore

$$f(y^i, \mu_i, \sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y^i - \mu_i)^2}{2\sigma^2}}$$

Assuming all data points are independent we would end up having the total likelihood to be

$$L = \prod_{i=1}^n f(y^i, \mu_i, \sigma) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^n \frac{(y^i - \mu_i)^2}{2\sigma^2}}$$

The corresponding loglikelihood function is

$$\ell = \log(L) = -\frac{n}{2} \log(2\pi) - \sum_{i=1}^n \frac{(y^i - \mu_i)^2}{2\sigma^2}$$

The β s that maximize the loglikelihood function need to satisfy the following conditions

$$\frac{\partial \ell}{\partial \beta_j} = 0, j = 1, 2, \dots, m$$

Because $-\frac{n}{2} \log(2\pi)$ is constant, substitute μ_i with equation (14) we end up having an equation set

$$\frac{\partial}{\partial \beta_j} \sum_{i=1}^n \frac{(y^i - \vec{\beta} \cdot \vec{x}^i)^2}{2\sigma^2} = 0, j = 1, 2, \dots, m$$

Because σ^2 is constant and independent of β , we have

$$\frac{\partial}{\partial \beta_j} \sum_{i=1}^n (y^i - \vec{\beta} \cdot \vec{x}^i)^2 = 0, j = 1, 2, \dots, m$$

Notice the summation term is exactly the same as the sum-square-error ϵ . **Hence, in this case, the maximum likelihood estimator (MLE) is equivalent to the OLS estimator!**