Mathematics Behind Linear Regression Model

Supervised Algorithm

Dostdar Ali (Speaker)

Title Machine Learning and Neural Networks with Mathematical Modelling

Department of Mathematics
University of Baltistan, Skardu
Gilgit Baltistan.

May 22, 2024

Table of Contents

Linear Regression

Definition

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. It is widely used in predictive analysis and data modeling.

- Simple linear regression involves a single independent variable. The relationship is modeled with a straight line.

$$y = w_0 + w_1 x \tag{1}$$

- Where,
  y is the dependent variable.
  x is the independent variable.
  w_0 is the y-intercept of the regression line
  w_1 is the slope of the regression line.

Linear Regression

Definition

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. It is widely used in predictive analysis and data modeling.

- Simple linear regression involves a single independent variable. The relationship is modeled with a straight line.

$$y = w_0 + w_1 x \tag{1}$$

- Where,
  y is the dependent variable.
  x is the independent variable.
  w_0 is the y-intercept of the regression line
  w_1 is the slope of the regression line.

Linear Regression

Definition

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. It is widely used in predictive analysis and data modeling.

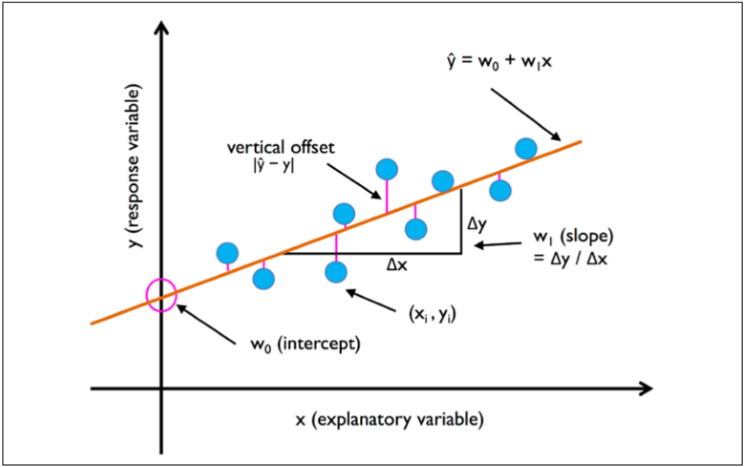
- Simple linear regression involves a single independent variable. The relationship is modeled with a straight line.

$$y = w_0 + w_1 x \tag{1}$$

- Where,
  y is the dependent variable.
  x is the independent variable.
  w_0 is the y-intercept of the regression line
  w_1 is the slope of the regression line.

Visual's of Linear regression Model

- Visual's representation of Simple Linear regression Model



Linear Regression Function

- The linear function is denote for i^{th} data points are:

$$\hat{y}_i = w_1 x_i + w_0 \tag{2}$$

- Now our objective, to find a approximated function that best fit the data points over all.
- Error Equations

$$r_i = y_i - \hat{y}_i \tag{3}$$

- The overall error can be calculated in different ways.
- Simply adding of Residual method
- Absolute error method
- Linear least square method

Linear Regression Function

- The linear function is denote for i^{th} data points are:

$$\hat{y}_i = w_1 x_i + w_0 \tag{2}$$

- Now our objective, to find a approximated function that best fit the data points over all.
- Error Equations

$$r_i = y_i - \hat{y}_i \tag{3}$$

- The overall error can be calculated in different ways.
- Simply adding of Residual method
- Absolute error method
- Linear least square method

Linear Regression Function

- The linear function is denote for i^{th} data points are:

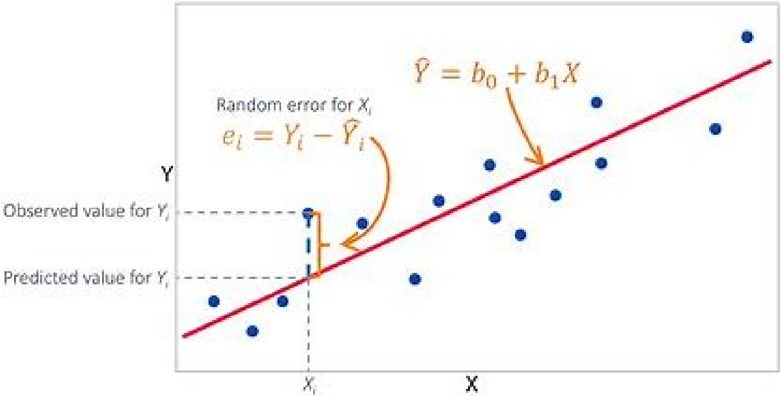
$$\hat{y}_i = w_1 x_i + w_0 \tag{2}$$

- Now our objective, to find a approximated function that best fit the data points over all.
- Error Equations

$$r_i = y_i - \hat{y}_i \tag{3}$$

- The overall error can be calculated in different ways.
- Simply adding of Residual method
- Absolute error method
- Linear least square method

- We have two values actual and approximated value, as shown below:



Visual of error function

Calculating Error Methods: Addition of residuals

One simple way is to add the residuals of all the points

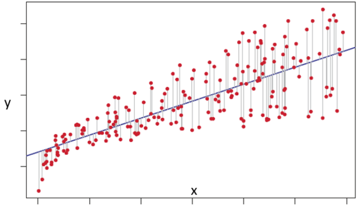
$$E = \sum_{i=1}^{n} r_i = \sum_{i=1}^{n} [y_i - (w_1 x_i + w_0)] \tag{4}$$

Residual

Residual –

The difference between residual the *i*th observed response value and the *i*th response value that is predicted by our linear model is known as residual

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



This is not a good choice to sum all : both +ve and -ve residuals both can be large zero or very close to zero.

Absolute, Linear least square method

- The next possibility: Absolute $||.||$

$$E = \sum_{i=1}^{n} r_i = \sum_{i=1}^{n} |y_i - (w_1 x_i + w_0)| \tag{5}$$

Now positive: (total error) but it cannot be used for determining the constants of the function that give the best fit. This is because the measure is not unique.

- Now we apply square on the all over error E is:

$$E = \sum_{i=1}^{n} r_i = \sum_{i=1}^{n} [y_i - (w_1 x_i + w_0)]^2 \tag{6}$$

Absolute, Linear least square method

- The next possibility: Absolute $||.||$

$$E = \sum_{i=1}^{n} r_i = \sum_{i=1}^{n} |y_i - (w_1 x_i + w_0)| \tag{5}$$

Now positive: (total error) but it cannot be used for determining the constants of the function that give the best fit. This is because the measure is not unique.

- Now we apply square on the all over error E is:

$$E = \sum_{i=1}^{n} r_i = \sum_{i=1}^{n} [y_i - (w_1 x_i + w_0)]^2 \tag{6}$$

Calculate w_1 and w_0

- $$w_1 = \frac{nS_{xy} - S_x S_y}{nS_{xx} - (S_x)^2} \tag{7}$$
  $$w_0 = \frac{S_{xx} S_y - S_{xy} S_x}{nS_{xx} - (S_x)^2} \tag{8}$$
- We have to calculate w_0 and w_1 using below formulas Where:

  w_1 : Coefficients for the respective independent variables
  w_0 : Intercept
  S_x : Sum of x
  S_y : Sum of y
  S_{xy} : Sum of x and y product
  S_{xx} : Sum of x product itself

Calculate w_1 and w_0

- $$w_1 = \frac{nS_{xy} - S_x S_y}{nS_{xx} - (S_x)^2} \tag{7}$$
  $$w_0 = \frac{S_{xx} S_y - S_{xy} S_x}{nS_{xx} - (S_x)^2} \tag{8}$$
- We have to calculate w_0 and w_1 using below formulas Where:

w_1 : Coefficients for the respective independent variables

w_0 : Intercept

S_x : Sum of x

S_y : Sum of y

S_{xy} : Sum of x and y product

S_{xx} : Sum of x product itself

Example for Linear Regression

Example

We want to predict the atmospheric pressure on given temperature and in between temperature the dataset are given below,

X_i	T(c^0)	0^0	30^0	70^0	100^0
y_i	P(atm)	0.94	1.05	1.17	1.28

$$S_x = \sum_{i=1}^{4} x_i = 0 + 30 + 70 + 100 = 200 \tag{9}$$

$$S_y = \sum_{i=1}^{4} y_i = 0.94 + 1.05 + 1.17 + 1.28 = 4.44 \tag{10}$$

Example For Linear Regression

- Similarly we have,

$$S_{xx} = \sum_{i=1}^{4} x_i^2 = 15800 \tag{11}$$

$$S_{xy} = \sum_{i=1}^{4} (x_i y_i) = 241.4 \tag{12}$$

- Putting these values in parameters equations we get,

$$w_1 = \frac{4(241.4) - 200 \times 4.44}{4(15800) - (200)^2} = 0.003345 \tag{13}$$

$$w_0 = 0.9428 \tag{14}$$

- The Linear Equation is

$$\hat{y}_i = 0.003345 x_i + 0.9428. \tag{15}$$

- Similarly we have,

$$S_{xx} = \sum_{i=1}^{4} x_i^2 = 15800 \tag{11}$$

$$S_{xy} = \sum_{i=1}^{4} (x_i y_i) = 241.4 \tag{12}$$

- Putting these values in parameters equations we get,

$$w_1 = \frac{4(241.4) - 200 \times 4.44}{4(15800) - (200)^2} = 0.003345 \tag{13}$$

$$w_0 = 0.9428 \tag{14}$$

- The Linear Equation is

$$\hat{y}_i = 0.003345x_i + 0.9428. \tag{15}$$

Example For Linear Regression

- Similarly we have,

$$S_{xx} = \sum_{i=1}^{4} x_i^2 = 15800 \tag{11}$$

$$S_{xy} = \sum_{i=1}^{4} (x_i y_i) = 241.4 \tag{12}$$

- Putting these values in parameters equations we get,

$$w_1 = \frac{4(241.4) - 200 \times 4.44}{4(15800) - (200)^2} = 0.003345 \tag{13}$$

$$w_0 = 0.9428 \tag{14}$$

- The Linear Equation is

$$\hat{y}_i = 0.003345 x_i + 0.9428. \tag{15}$$

Example for Linear Regression

Data Overview

x_i	1	2	3	4	5
y_i	2	3	4	5	6

Calculating Means

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3$$
$$\bar{y} = \frac{2+3+5+4+6}{5} = 4$$

Calculating Slope (w_1)

$$w_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Calculation

$$
\begin{aligned}
&= \frac{(1-3)(2-4)+(2-3)(3-4)+(0)(5-4)+(4-3)(0)+(5-3)(2)}{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2} \\
&= \frac{(-2)(-2)+(-1)(-1)+(0)(1)+(1)(0)+(2)(2)}{(-2)^2+(-1)^2+(0)^2+(1)^2+(2)^2} \\
&= \frac{4+1+0+0+4}{4+1+0+1+4} \\
&= \frac{9}{10} = 0.9
\end{aligned}
$$

Example for Linear Regression

Calculating Intercept (w_0)

$$\begin{aligned} w_0 &= \bar{y} - w_1\bar{x} \\ &= 4 - 0.9 \times 3 \\ &= 4 - 2.7 \\ &= 1.3 \end{aligned}$$

Regression Equation

$$\hat{y} = 1.3 + 0.9x \tag{16}$$

Verification

$$\begin{aligned} \hat{y}_1 &= 1.3 + 0.9 \times 1 = 2.2 \\ \hat{y}_2 &= 1.3 + 0.9 \times 2 = 3.1 \\ \hat{y}_3 &= 1.3 + 0.9 \times 3 = 4.0 \\ \hat{y}_4 &= 1.3 + 0.9 \times 4 = 4.9 \\ \hat{y}_5 &= 1.3 + 0.9 \times 5 = 5.8 \end{aligned}$$

Example for Linear Regression

Sum of Squared Residuals (SSR)

$$
\begin{aligned}
e_1 &= 2 - 2.2 = -0.2 \\
e_2 &= 3 - 3.1 = -0.1 \\
e_3 &= 5 - 4.0 = 1.0 \\
e_4 &= 4 - 4.9 = -0.9 \\
e_5 &= 6 - 5.8 = 0.2 \\
SSR &= (-0.2)^2 + (-0.1)^2 + (1.0)^2 + (-0.9)^2 + (0.2)^2 \\
&= 0.04 + 0.01 + 1.00 + 0.81 + 0.04 \\
&= 1.90
\end{aligned}
$$

This show model fit

Model Evaluation

1. R-squared ($\mathbf{R}^2$)

The Total Sum of Squares (SST) is:

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 10$$

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} = 1 - \frac{1.90}{10} = 0.81$$

2. Mean Squared Error (MSE)

$$\text{MSE} = \frac{\text{SSR}}{n} = \frac{1.90}{5} = 0.38$$

3. Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{0.38} \approx 0.62$$

Conclusion

The linear regression model fits the data well:

- $R^2 = 0.81$ indicates that 81% of the variance in y is explained by x.
- MSE $= 0.38$ and RMSE $= 0.62$ indicate the average magnitude of prediction errors.

Questions and Answers

Search for Resource of Workshop