Linear Regression

Linear Regression is a statistical modeling technique that is used to estimate the relationship between variables. For example, a linear regression model can be used to predict house prices based on variables such as number of rooms, age of the house, size of the house, and economy. This article provides an explanation of simple linear regression using an example to illustrate the steps in calculating the regression parameters. The same is then calculated using matrix algebra.

Simple Linear Regression

In simple linear regression, a relationship is established between two variables, an independent or predictor variable x and a dependent or response variable y. Lets create a regression model where we predict house prices based on a single variable, age of house (# of years since the house was built). Equation 1 defines this linear relationship between age (x) and price (y) of a house.

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

where,

y is the dependent variable,
x is the independent variable,
β_0 is the y intercept,
β_1 is the slope, and
ϵ is the statistical error.

The statistical error accounts for error that results from the inaccuracy in modeling and measuring the relationship between x and y.

It is important to note that the parameters β_0 and β_1, and the error term ϵ are unknown. These are population parameters which are theoretical values and cannot be determined. We estimate these parameters from a sample of data collected on x and y.

We will explain these concepts using an example with a few (x, y) observations as shown in table 1.

Age of House (x)	Price of House ($,000) (y)
10	350
15	250
20	300
20	240

25	225

Table 1: Data on age and price

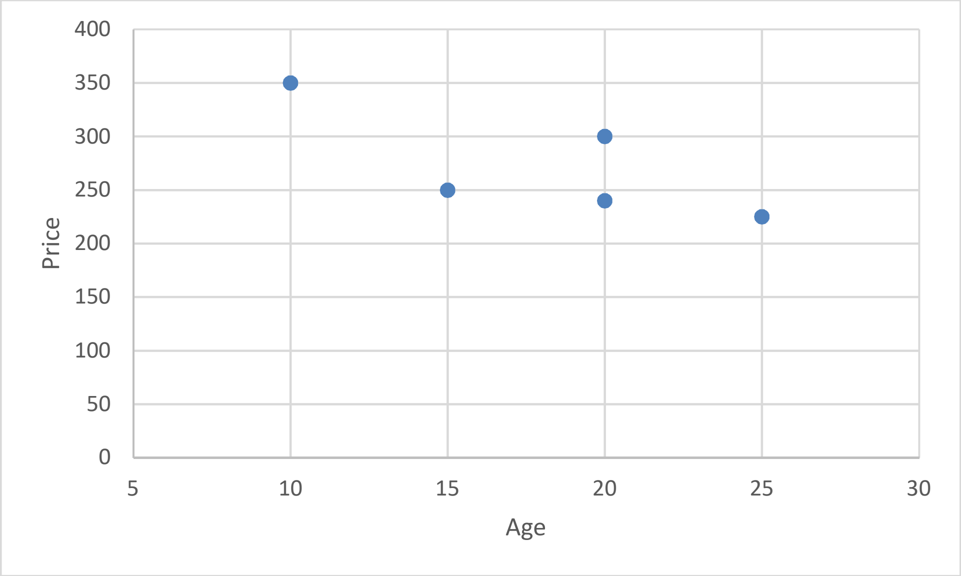

Figure 1. Scatter plot of age and price

Data from table 1 is plotted in figure 1. The scatter plot shows a negative relationship between age and price of house. For newer houses, prices will be higher than for older houses.

Estimating Regression Parameters

The most common method used to estimate the parameters β_0 and β_1 is the method of ***least squares***. According to this method, the regression parameters are estimated by minimizing the sum of squared errors, the vertical distance of each observed response from the regression line. The least square method yields equations 2 and 3 which are used to find the estimated parameters $\hat{\beta}_1$and $\hat{\beta}_0$for β_1 and β_0, respectively.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (2)$$

Where,

$\bar{x}$ is the mean of x, and

$\bar{y}$ is the mean of y

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \tag{3}$$

The regression line using the estimated parameters is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{4}$$

Table 2 shows the steps to calculate $\hat{\beta}_1$and $\hat{\beta}_0$.

	Age of House (x)	Price of House ($,000) (y)	(x-x̄)	(y-ȳ)	(x-x̄)(y-ȳ)	(x-x̄)²
	10	350	-8	77	-616	64
	15	250	-3	-23	69	9
	20	300	2	27	54	4
	20	240	2	-33	-66	4
	25	225	7	-48	-336	49
Sum	**90**	**1365**			**-895**	**130**
Mean	**18**	**273**				

Table 2. Calculations for regression parameters

$\bar{x}$ = 18, and $\bar{y}$ = 273

$\hat{\beta}_1 = \frac{-895}{130}$ = -6.88

$\hat{\beta}_0$ = 273 – (-6.88 * 18) = 396.92

$$\hat{y} = 396.92 - 6.88x \tag{5}$$

Equation 5 is the regression line that is used to estimate y for given values of x. The regression line is plotted in figure 2. The line gives ŷ (pronounced y-hat), the predicted values of y, for different values of x. Some of the observed values of y are above the regression line and some are below. The difference (y - ŷ) is the prediction error called the residual. The regression line is the line of “**best fit**” as this line minimizes the sum of squared errors of prediction.

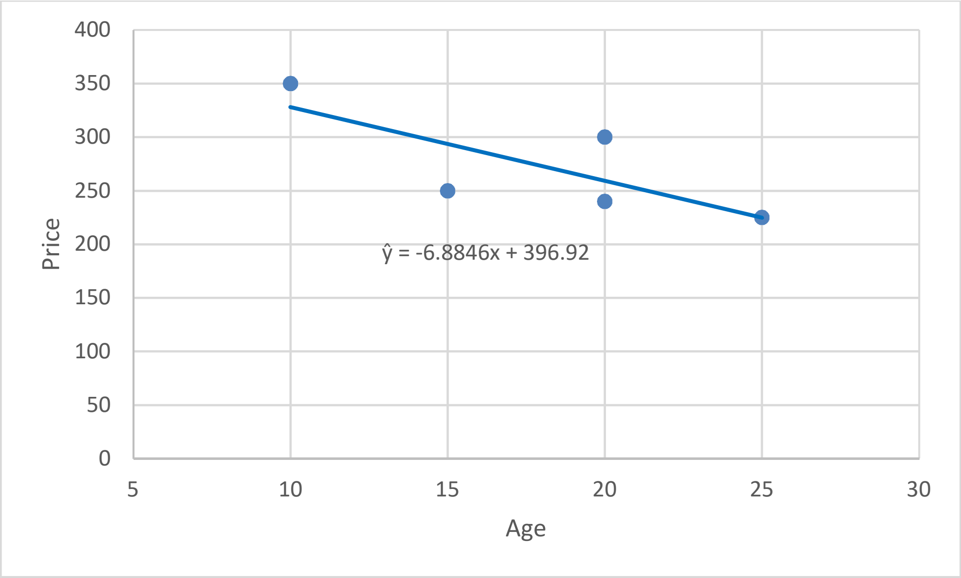

Figure 2. Regression Line (line of best fit)

Table 3 shows the prediction error for each observation.

Age of House (x)	Price of House ($,000) (y)	Predicted Price ($\hat{y}$)	Residual ($e = y - \hat{y}$)	Residual^2
10	350	328	22	481
15	250	294	-44	1905
20	300	259	41	1662
20	240	259	-19	370
25	225	225	0	0
Sum of Squared Errors				**4418**

Table 3. Prediction and residual values for line of best fit

Any line with different values for the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$will give a sum of squared errors that will be larger than what is achieved from the line of best fit. Table 4 and Figure 3 illustrate this by using a different value for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Age of House (x)	Price of House ($,000) (y)	Predicted Price ($\hat{y}$)	Residual (e)	Residual^2
10	350	260	90	8100
15	250	190	60	3600
20	300	120	180	32400
20	240	120	120	14400
25	225	50	175	30625
Sum of Squared Errors				**89125**

Table 4. Prediction and residual values for $\hat{\beta}_0$ = 400 and $\hat{\beta}_1$ = -14

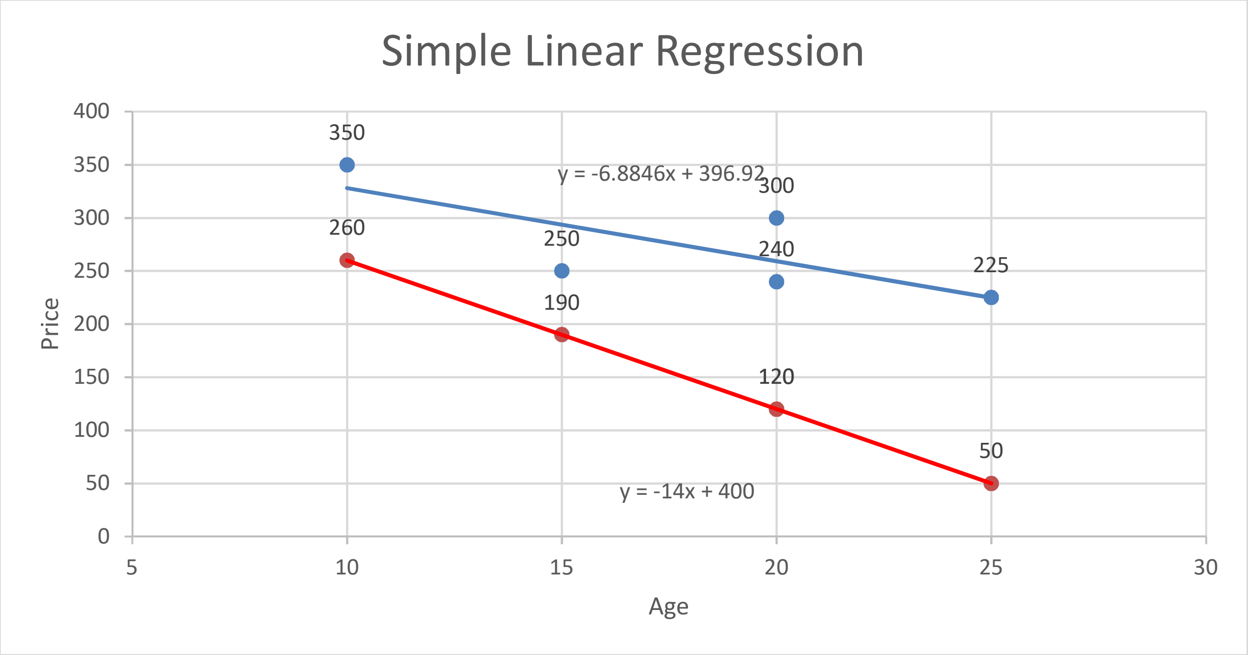

Figure 3. Regression line comparison

In figure 3, the regression line in blue is the line of best fit. The line in red is the line with $\hat{\beta}_0$ = 400 and $\hat{\beta}_1$ = -14. This line has a residual sum of squared errors of 89125 compared to 4418 for the line of best fit. You can use the link to download an Excel spreadsheet and try different values of $\hat{\beta}_0$ and $\hat{\beta}_1$ and observe the change in the regression line and the corresponding sum of squared errors.

Simple Linear Regression Using Matrix Algebra

Here, we will use matrix algebra to determine the line of best fit. A simple linear regression is expressed as:

$y = \beta_0 + \beta_1 X + \epsilon$

Our objective is to estimate the coefficients β_0 and β_1 by using matrix algebra to minimize the residual sum of squared errors.

A set of n observations (x, y) can be written in the form:

$y_1 = \beta_0 1 + \beta_1 x_1 + \epsilon_1$
$y_2 = \beta_0 1 + \beta_1 x_2 + \epsilon_2$
:
$y_n = \beta_0 1 + \beta_1 x_n + \epsilon_n$

Write the above equations in matrix form as:

$$y = \begin{bmatrix} y1 \\ y2 \\ \vdots \\ yn \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x1 \\ 1 & x2 \\ \vdots & \vdots \\ 1 & xn \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta 0 \\ \beta 1 \end{bmatrix} \qquad \epsilon = \begin{bmatrix} \epsilon 1 \\ \epsilon 2 \\ \vdots \\ \epsilon n \end{bmatrix}$$

y is a (n x 1) dimension vector that represents the prices of n houses.

X is a (n x 2) dimension matrix where each column represents the values for each predictor and each row represents the values of predictors for a house.

β is a (2, 1) dimension vector of parameters.

ϵ is a (n x 1) dimension vector of errors.

The linear regression model can now be written as:

$y = X\beta + \epsilon$

Estimating Regression Parameters

As explained in the previous section, we will use the method of least squares to estimate regression parameters. The least squares method finds the vector β by minimizing the sum of squared errors:

$$\sum_{i=1}^{n} \epsilon_i^2 = \epsilon^T \epsilon$$

This gives a vector $\hat{\beta}$ which is an estimate of β.

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

Using the observations on house age and price, the following illustrates the steps to calculate the parameters by using matrix algebra.

Age of House (x)	Price of House ($,000) (y)
10	350
15	250
20	300
20	240
25	225

The linear regression model is:

y = $\beta_0 + \beta_1$ x + ϵ

where,

$$y = \begin{bmatrix} 350 \\ 250 \\ 300 \\ 240 \\ 225 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 10 \\ 1 & 15 \\ 1 & 20 \\ 1 & 20 \\ 1 & 25 \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta 0 \\ \beta 1 \end{bmatrix}$$

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

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 10 & 15 & 20 & 20 & 25 \end{bmatrix} \begin{bmatrix} 1 & 10 \\ 1 & 15 \\ 1 & 20 \\ 1 & 20 \\ 1 & 25 \end{bmatrix} = \begin{bmatrix} 5 & 90 \\ 90 & 1750 \end{bmatrix}$$

To find inverse,

$$\text{If } X = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\text{Then } X^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

So,

$$(X^TX)^{-1} = \frac{1}{(8750-8100)}\begin{bmatrix}1750 & -90\\ -90 & 5\end{bmatrix} = \frac{1}{650}\begin{bmatrix}1750 & -90\\ -90 & 5\end{bmatrix}$$

and $$(X^Ty) = \begin{bmatrix}1 & 1 & 1 & 1 & 1\\ 10 & 15 & 20 & 20 & 25\end{bmatrix}\begin{bmatrix}350\\ 250\\ 300\\ 240\\ 225\end{bmatrix} = \begin{bmatrix}1365\\ 23675\end{bmatrix}$$

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

$$= \frac{1}{650}\begin{bmatrix}1750 & -90\\ -90 & 5\end{bmatrix}\begin{bmatrix}1365\\ 23675\end{bmatrix}$$

$$= \frac{1}{650}\begin{bmatrix}258000\\ 591875\end{bmatrix}$$

$$= \begin{bmatrix}396.92\\ -6.88\end{bmatrix}$$

Therefore,

$\hat{\beta}_0$ = 396.92, and

$\hat{\beta}_1$ = -6.88

The regression equation is:

ŷ = 396.92 – 6.88x

Using this equation, the error values for each observation can be determined as shown previously.

Age of House (x)	**Price of House ($,000) (y)**	**Predicted Price (ŷ)**	**Residual (e = y - ŷ)**
10	350	328	22
15	250	294	-44
20	300	259	41
20	240	259	-19
25	225	225	0

In matrix notation, error is written as:

E = $\begin{bmatrix} 22 \\ -44 \\ 41 \\ -19 \\ 0 \end{bmatrix}$

The sum of squared errors (SSE) = E^TE

= $\begin{bmatrix} 22 & -44 & 41 & -19 & 0 \end{bmatrix} \begin{bmatrix} 22 \\ -44 \\ 41 \\ -19 \\ 0 \end{bmatrix}$

= [4418]

There are several statistical software packages available which can provide the complete regression analysis including several statistical tests to determine the accuracy and significance of the model. Here, an attempt is made to illustrate the steps in calculating the parameters using equations as well as matrix algebra. Today, linear regression using least square estimates is widely being used in machine learning. Having a good understanding of the basic concepts of linear regression will be very helpful in performing advanced machine learning modeling.