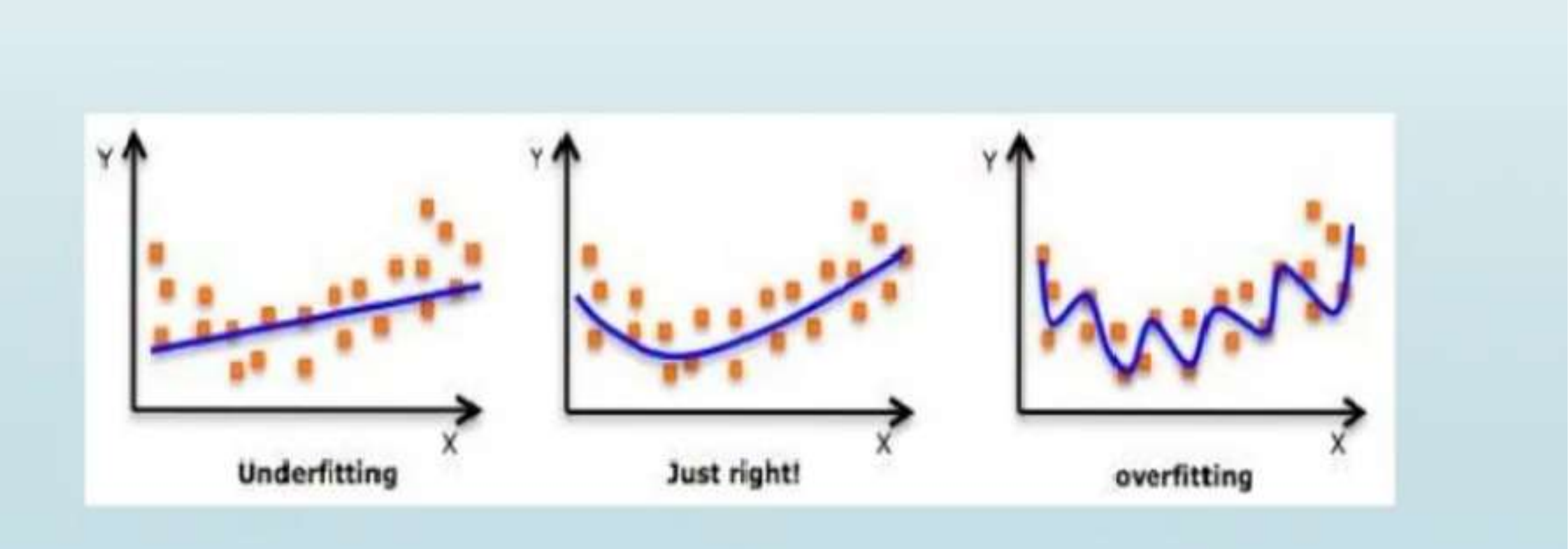
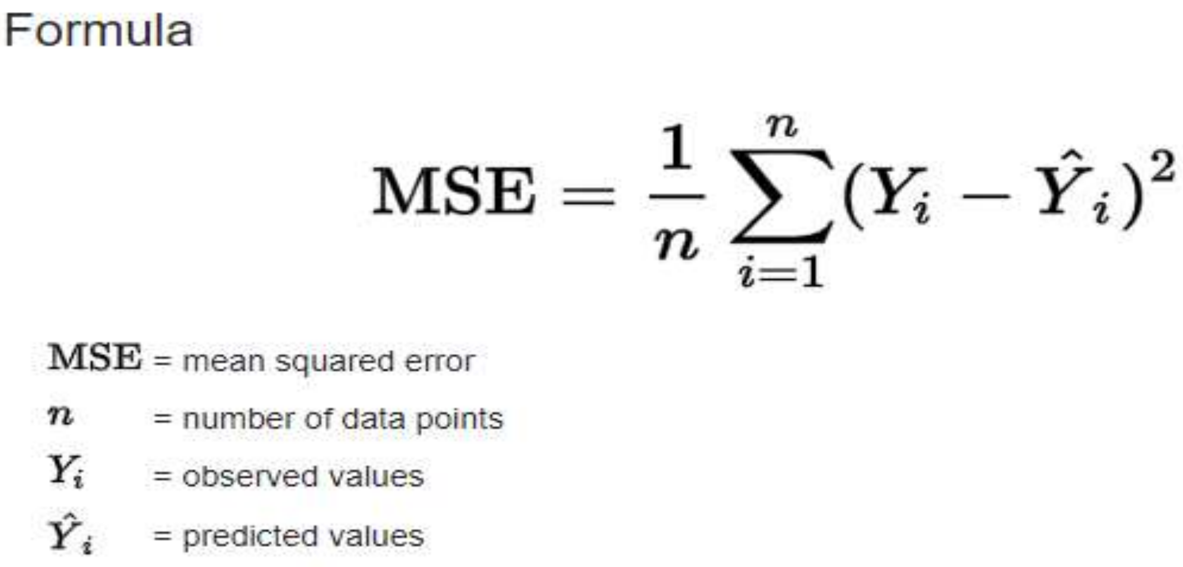
Overfitting and underfitting graphs



Regression metrics

Mean Squared Error - is a popular error metric for regression problems. It is also an important loss function for algorithms fit or optimized using the least squares framing of a regression problem. Here “least squares” refers to minimizing the mean squared error between predictions and expected values.



Formula

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE	= mean squared error
n	= number of data points
Y_i	= observed values
$\hat{Y}_i$	= predicted values

Root Mean Squared Error - RMSE, is an extension of the mean squared error. From the formula we can deduce that RMSE is square-root of MSE. Lower values of RMSE indicate better fit.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Mean Absolute Error - MAE, is a popular metric because the units of the error score match the units of the target value that is being predicted. The difference between the true and predicted value is called residual.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \hat{Y}_i|$$

R^2 coefficient

R^2 or Coefficient of Determination is a prevalent metric that uses two mean squared error calculations. While the former is the mean square of each real value versus the average of observations, the latter is the mean squared error of the actual value versu

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y_i})^2}$$

R^2 score ranges from -∞ to 1. The closest to 1 the R^2, the better the regression model is.

If R^2 is equal to 0, the model is not performing better than a random model.

If R^2 is negative, the regression model is erroneous.

Supervised Machine Learning Algorithms

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. Support Vector Machines
6. Naïve Bayes Classification
7. k-nearest neighbors (KNN) algorithm
8. Random Forest

Linear Regression in Machine Learning

- Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis.

- Prediction for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

- Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (x) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

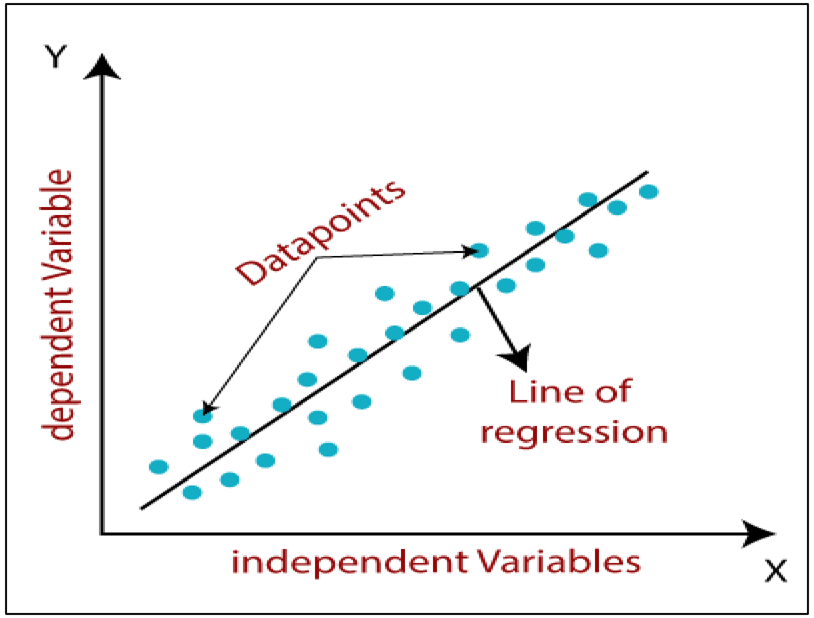
Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression (Cont..)

- The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematical Representation of Linear Regression

Mathematically, we can represent a linear regression as:

$y= a_0+a_1x+ \varepsilon$

Here,

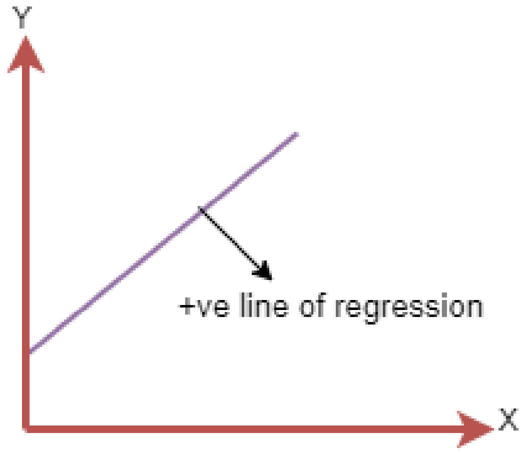
y= Dependent Variable (Target Variable)
x= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Linear Relationship

1. Positive Linear Relationship:

If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.

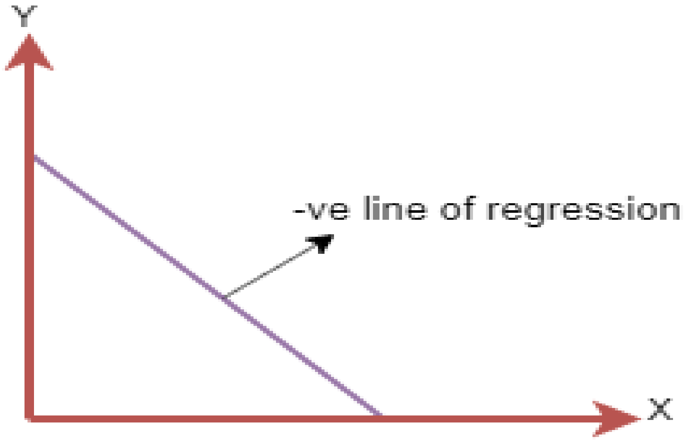


The line equation will be: **$Y= a_0+a_1x$**

Linear Relationship (Cont..)

- **Negative Linear Relationship:**

  If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: **$Y= -a_0+a_1x$**

Mean Squared error

- For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

$$MSE= 1\frac{1}{N}\sum_{i=1}^{n}(y_i \;\; - (a_1 x_i + a_0))^2$$

Where,

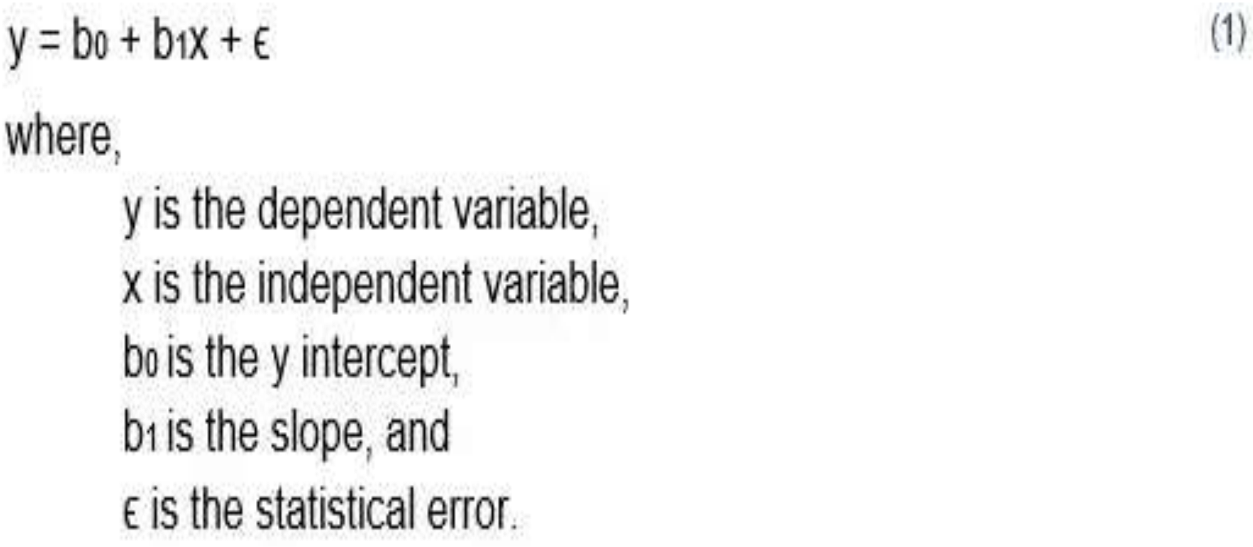
N=Total number of observation

Yi = Actual value

($a1x_i$+a_0)= Predicted value.

Example of Linear Regression

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.



$$y = b_0 + b_1x + \epsilon \qquad (1)$$

where,

y is the dependent variable,

x is the independent variable,

b_0 is the y intercept,

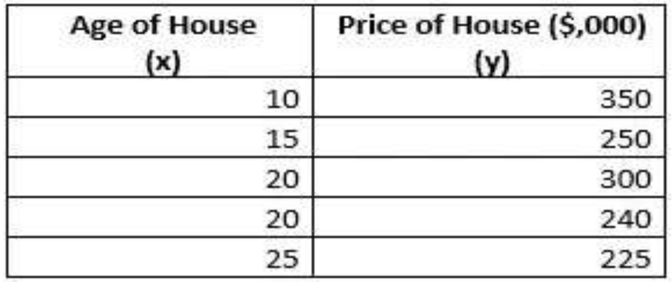
b_1 is the slope, and

ϵ is the statistical error.

Example…(Cont)

It is important to note that the parameters b0 and b1, 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

Scatter Diagram

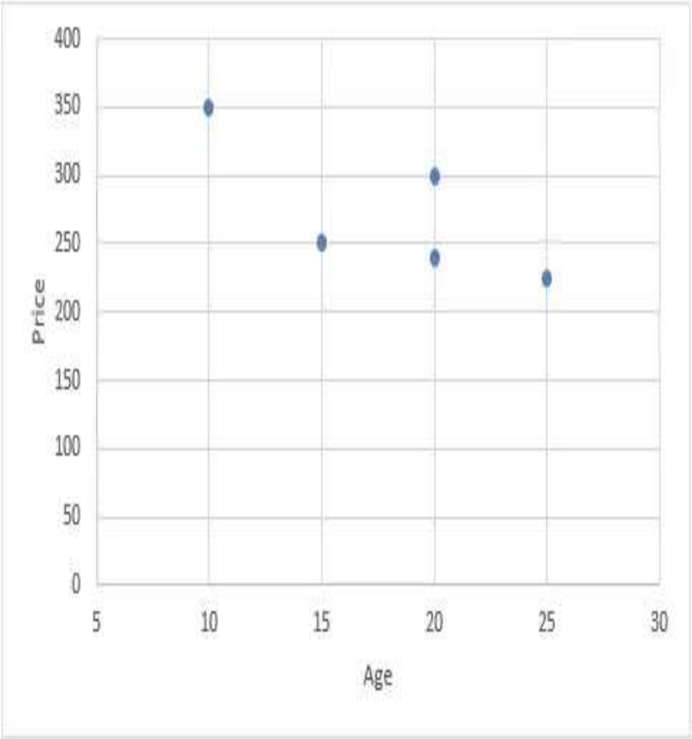


Figure 1. Scatter plot of age and price

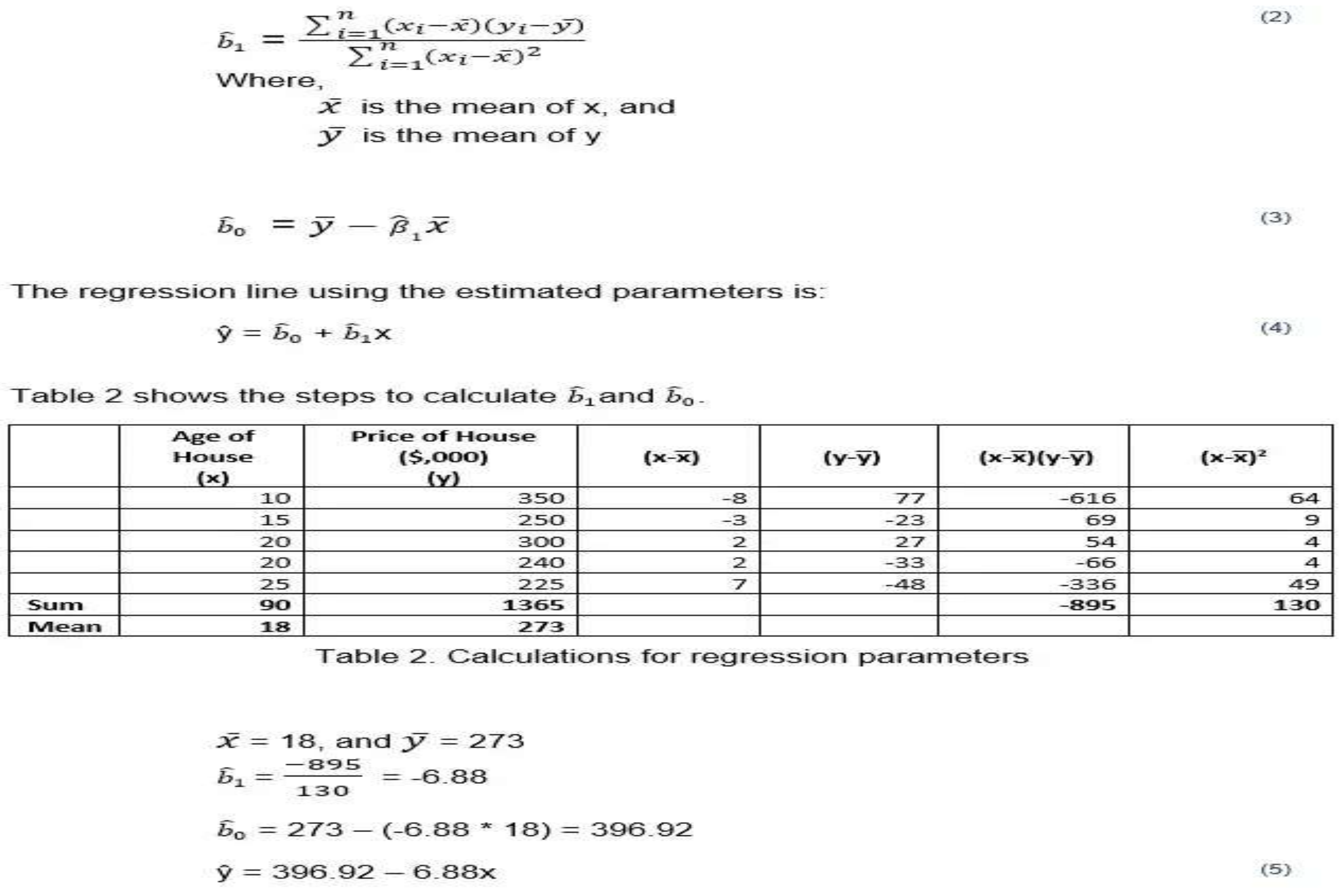
Data from table 1 is plotted in figure 1. The scatter plot shows a negative relationship between age and price of house. For newer houses, prices will be higher than for older houses.

Estimating Regression Parameters

The most common method used to estimate the parameters b0 and b1 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 $\hat{b}1$ and $\hat{b}0$ for b1 and b0, respectively.

Estimating Regression Parameters



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

Where,

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

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

$$\hat{b}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (3)$$

The regression line using the estimated parameters is:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x \qquad (4)$$

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

	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

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

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

$$\hat{y} = 396.92 - 6.88x \qquad (5)$$

Computation of Prediction Error

- 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.

Regression Line of Best Fit

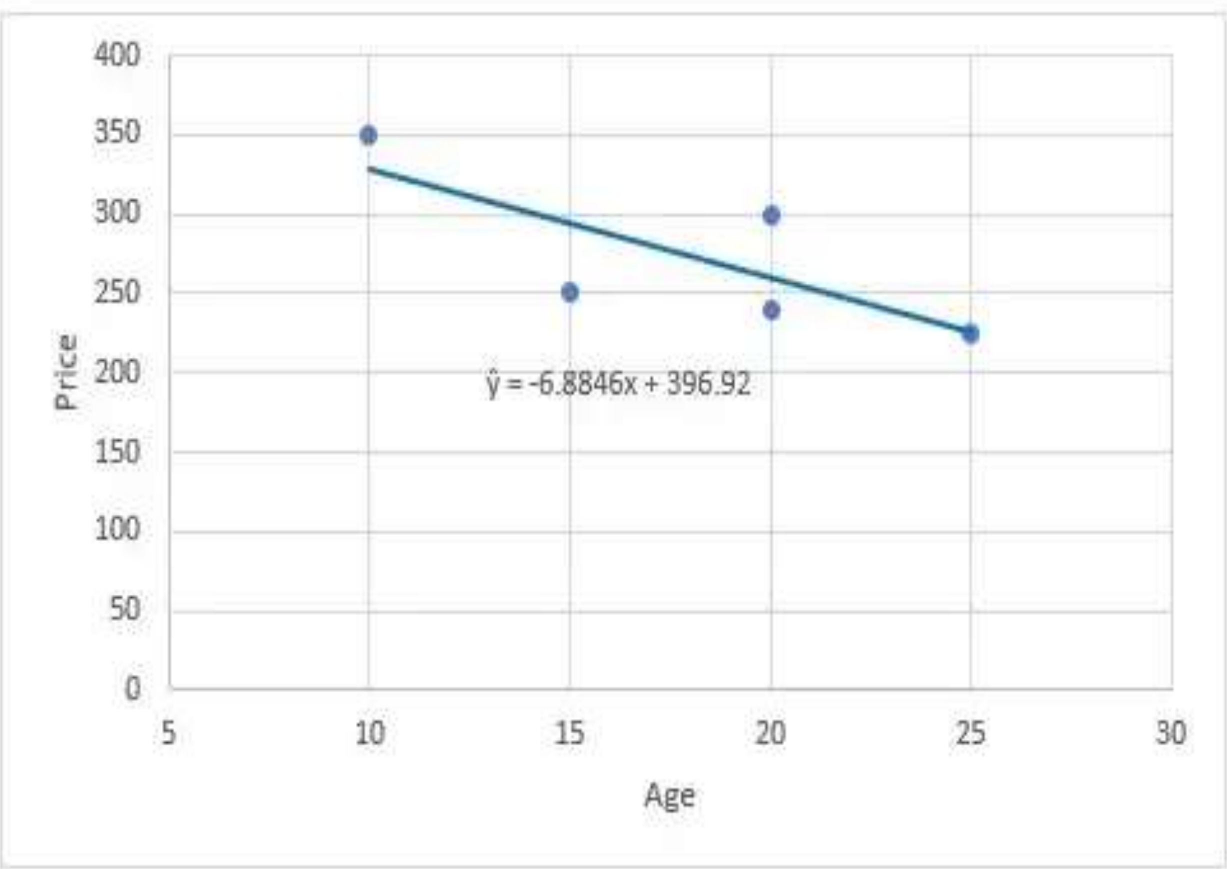
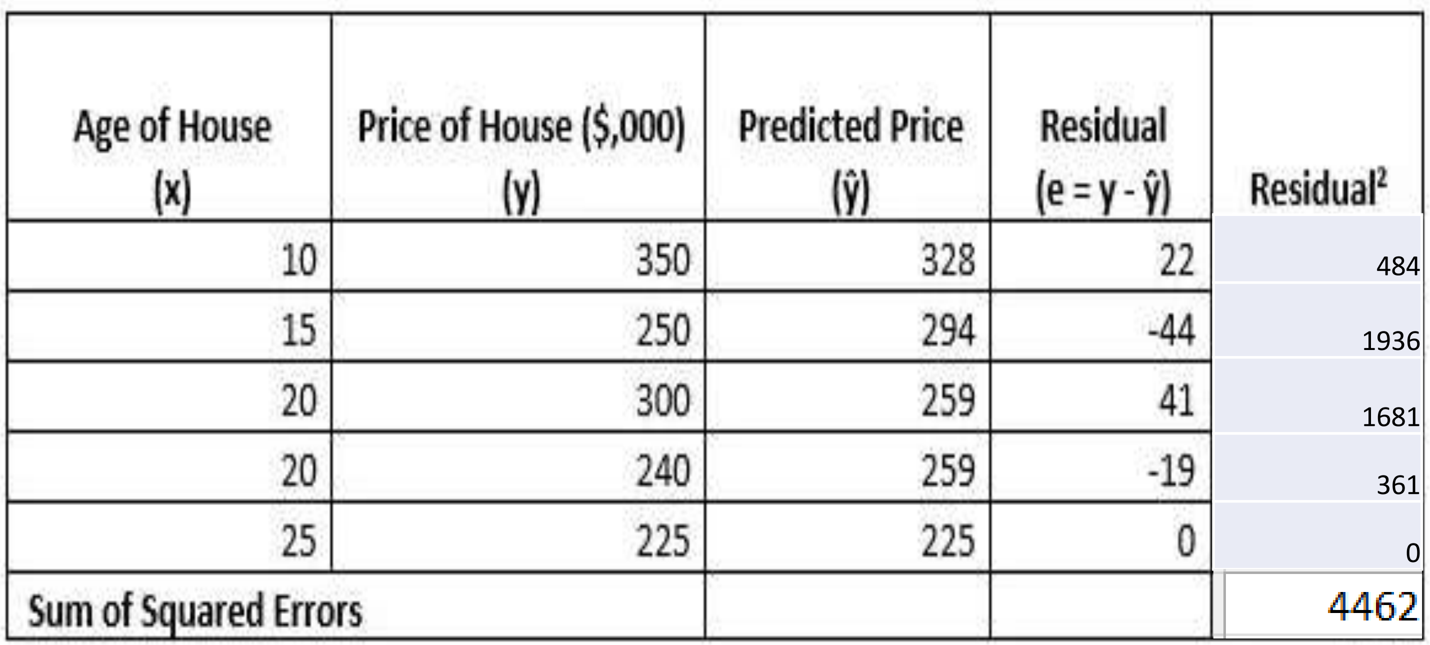


Figure 2. Regression Line (line of best fit)

Prediction and Residual Values

Table 3 shows the prediction error for each observation.

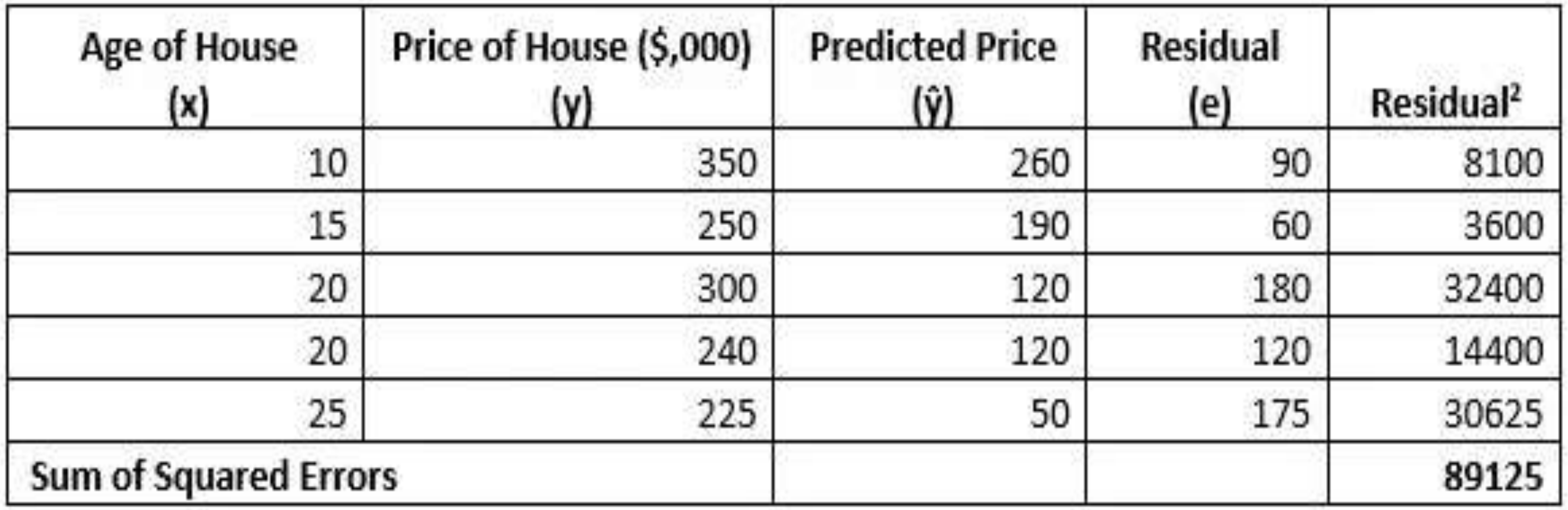


Age of House (x)	Price of House ($,000) (y)	Predicted Price ($\hat{y}$)	Residual ($e = y - \hat{y}$)	Residual^2
10	350	328	22	484
15	250	294	-44	1936
20	300	259	41	1681
20	240	259	-19	361
25	225	225	0	0
Sum of Squared Errors				4462

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

Prediction and Residual Values for Different Value of Parameters

Any line with different values for the parameters b̂0 and b̂1 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 b̂0 and b̂1.



Age of House (x)	Price of House ($,000) (y)	Predicted Price (ŷ)	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 $\hat{b}_0$ = 400 and $\hat{b}_1$ = -14

Regression Line Comparison

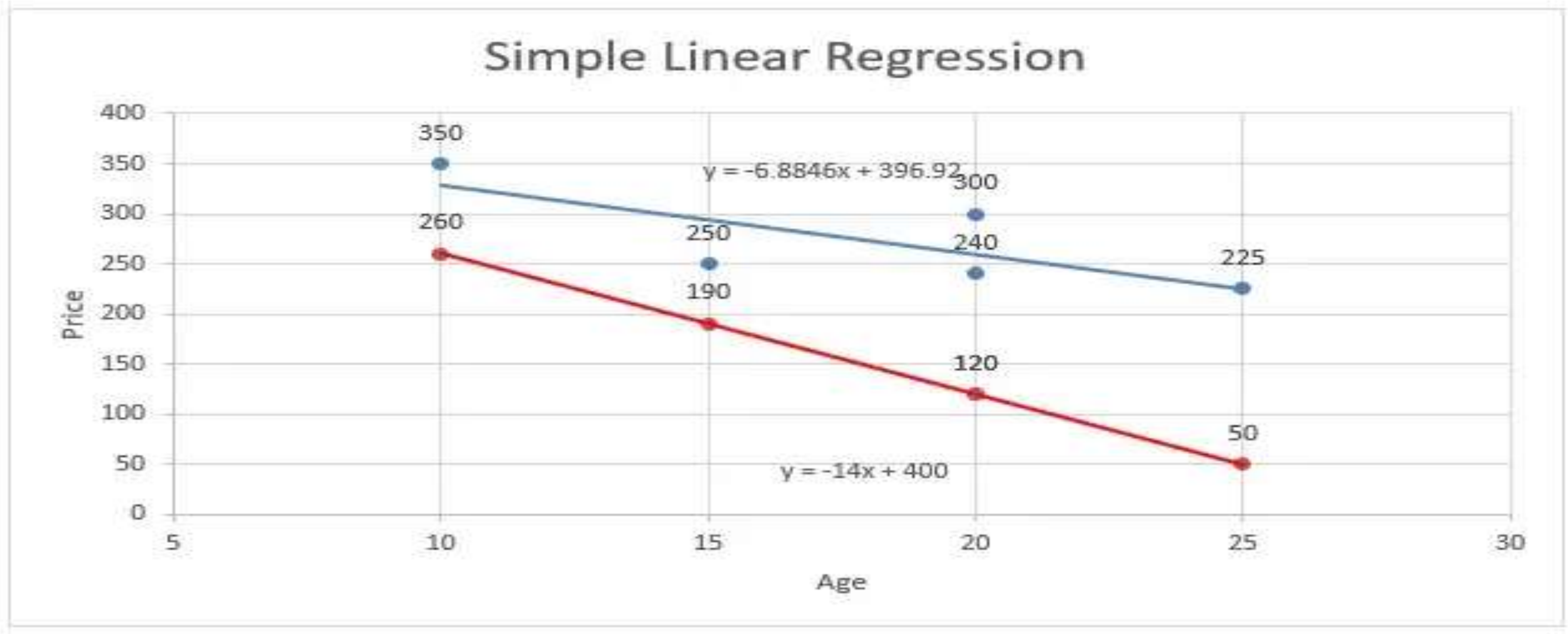


Figure 3. Regression line comparison

In figure 3, the regression line in blue is the line of best fit. The line in red is the line with $\hat{b}0 = 400$ and $\hat{b}1 = -14$. This line has a residual sum of squared errors of 89125 compared to 4462 for the line of best fit.