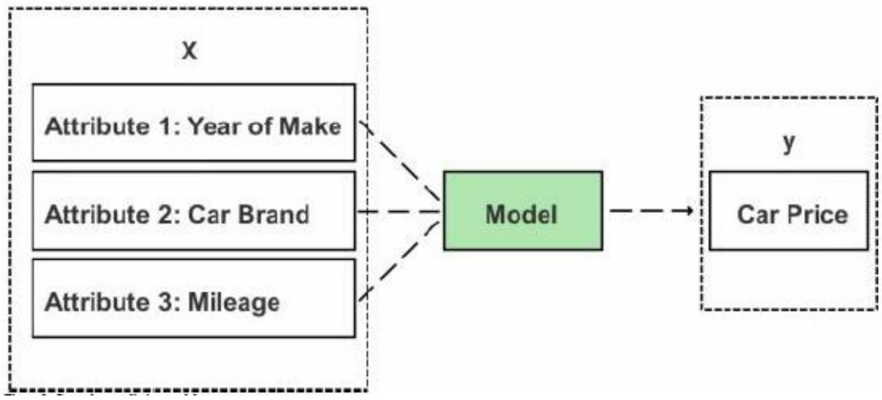
1.1. Simple Linear Regression

What is Regression Problem?

Any regression algorithm works by feeding the machine sample data with various features (represented as “X”) and a ***continuous or numeric value output*** (represented as “y”).

For instance, to predict the market rate for the purchase of a used car, a regression problem can be formulated to predict selling price of a used car by analyzing the relationship between car attributes (including the year of make, car brand, mileage, etc.) and the selling price of other cars sold based on historical data.



After the machine deciphers the parameter values from the data, it creates what is known as a model: *an algorithmic equation for producing an outcome with new data based on the rules derived from the training data*.

Once the model is prepared, it can be applied to new data and tested for accuracy. After the model has passed both the training and test data stages, it is ready to be applied and used in the real world.

Another simple example is to create a model for predicting house values where y is the actual house price and X are the variables that impact y, such as land size, location, and the number of rooms. Through supervised learning, we can create a rule to predict y (house value) based on the given values of various variables (X).

There are many types of Regression such Simple Linear Regression, Multiple Linear Regression, Polynomial Regression etc.

Simple Linear Regression

Simple linear regression is a type of regression analysis where the *number of independent variables is one* and there is a linear relationship between the independent (X) and dependent (Y) variable.

Independent variable, denoted by (X) is also known as the predictor, explanatory variable

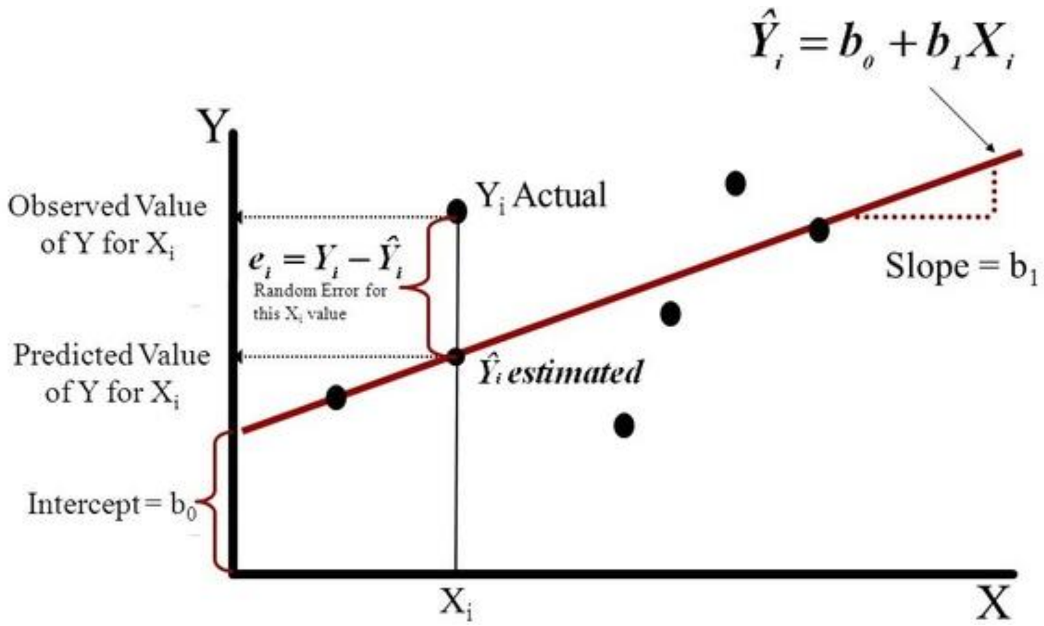
Dependent variable, denoted by (Y), is also known as the response, outcome, or target variable.

Simple linear regression gets its adjective "simple," because it concerns the study of only one predictor variable. In contrast, multiple linear regression, which we study later in this course, gets its adjective "multiple," because it concerns the study of two or more predictor variables.

Simple linear regression uses traditional slope-intercept form,

$$\boldsymbol{y = b_0 + b_1 x}$$

where b_0 (the intercept) and b_1 (the slope) are the parameters our algorithm will try to "learn" to produce the most accurate predictions.



Example:

Let's consider a scenario where we want to determine the linear relationship between: How much a company spends on Radio advertising each year and its annual Sales in terms of units sold. We are trying to develop an equation that will let us to predict units sold based on how much a company spends on radio advertising. Given Below is a screenshot of dataset

radio	sales
37.8	22.1
39.3	10.4
45.9	9.3
41.3	18.5
10.8	12.9
48.9	7.2
[illegible]	[illegible]

A Simple Linear Regression model can be formulated between sales (dependent variable) & radio (independent variable)

$$\boldsymbol{sales = b_0 + b_1 * radio}$$

Our algorithm will try to learn the correct values for $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ using:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(X,Y)}{Var(X)} \qquad\qquad b_0 = \bar{y} - b_1\bar{x}$$

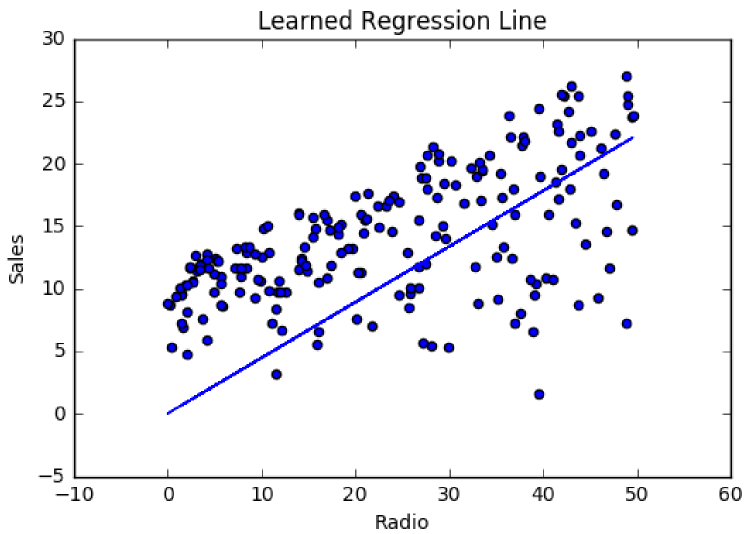
By the end of our training we get:

the intercept $b_0 = 9.31$ and the slope $b_1 = 0.2$

And now our equation:

$$\boldsymbol{sales = 9.31 + 0.2 * radio}$$

will approximate the line of best fit.



Our prediction model can now output an estimate of sales given a new value of company's radio advertising expenditure. For example, if the company spends $ 50 on radio advertisement then it can expect an average sale of:

$$\mathbf{9.31 + 0.2 * 50 = 19.31}\ \boldsymbol{units} \sim \mathbf{19}\ \boldsymbol{units}$$

To evaluate model performance, we can use any of the following error metric:

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

Or coefficient of determination i.e., $\boldsymbol{R^2}$