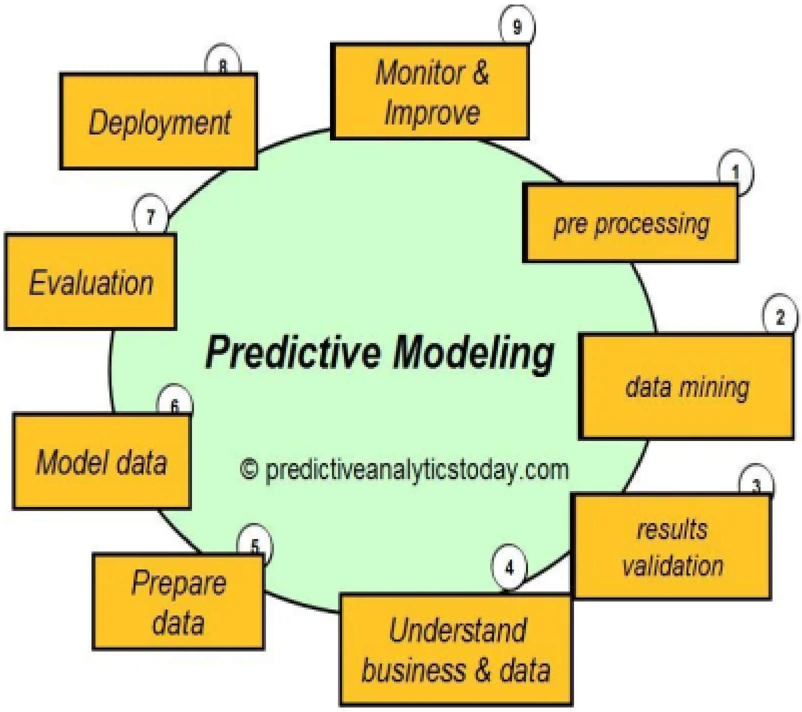
Predictive Modelling

The predictive modelling process starts with data collection, then a statistical model is formulated, predictions are made, and the model is revised as new data becomes available.



What is Predictive modelling?

In short, predictive modelling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data.

It works by analyzing current and historical data and projecting what it learns on a model generated to forecast likely outcomes. Predictive modelling can be used to predict just about anything, from TV ratings and a customer's next purchase to credit risks and corporate earnings.

A predictive model is not fixed; it is validated or revised regularly to incorporate changes in the underlying data. In other words, it's not a one-and-done prediction. Predictive models make assumptions based on what has happened in the past and what is happening now. If incoming, new data shows changes in what is happening now, the impact on the likely future outcome must be recalculated, too.

For example, a multi brand, multi branch store could model historical sales data against marketing expenditures across multiple regions to create a model for future revenue based on the impact of the marketing spend.

Most predictive models work fast and often complete their calculations in real time. That's why banks and retailers can calculate the risk of an online mortgage or credit card application and accept or decline the request almost instantly based on that prediction.

Some predictive models are more complex, such as those used in computational biology and quantum computing; the resulting outputs take longer to compute than a credit card application but are done much more quickly than was possible in the past thanks to advances in technological capabilities, including computing power.

Predictive Modelling Techniques

Predictive modelling is generally categorized as either parametric or nonparametric models. There are several different varieties of predictive analytics models which include Ordinary Least Squares, Generalized Linear Models, Linear Regression, Logistic Regression, Random Forests, Decision Trees and Neural Networks.

How to Make a Predictive Model

Regardless of the types of predictive models in place, the process of predictive model deployment follows the same steps:

- Clean up data by treating missing data and eliminating outliers
- Determine whether parametric or nonparametric predictive modelling is most effective
- Reprocess the data into a format appropriate for the modelling algorithm
- Specify a subset of data to be used for training the model
- Train model parameters from the training dataset
- Conduct predictive model performance monitoring tests to assess model efficacy
- Validate predictive modelling accuracy on data not used for calibrating the model
- Deploy the model for prediction

How to Evaluate a Predictive Model

A popular technique to employ in predictive model validation and evaluation is cross-validation. Datasets are split at random into training datasets, test datasets, and validation datasets. Training data is used to build the model, then the trained model is run against test data to evaluate performance, and the validation dataset ensures a neutral estimation of predictive model accuracy.

Each time a subset of historical data is used as test data, remaining subsets are used as training data. As tests continue, a former test dataset will become one of the training datasets, and one of the former training datasets will become a test dataset, until every subset has been used as a test set. This allows the use of every data point in a historical dataset for both testing and training, which facilitates a less random and more effective, thorough method for evaluating data and testing model accuracy.

Some Applications of predictive Modelling

Predictive modelling, often associated with meteorology, is leveraged throughout a wide variety of disciplines. Some popular predictive modelling applications include:

- **Predictive modelling in healthcare**: identify high risk patients in poor health and inform clinical trial designs, predict optimized dosage of medicine and gain insights from patterns in patient data in order to develop effective treatment plans.
- **Predictive modelling in insurance**: turn data collected by insurers into actionable insights for pricing and risk selection purposes; identify customers at risk of cancellation; identify risk of fraud and outlier claims; anticipate the insured's needs, ultimately improving satisfaction and optimizing budget management and identify potential markets.
- **GIS predictive modelling**: describe spatial environment factors that constrain and influence the location of events by spatially correlating environmental factors that represent those constraints and influences with occurrences of historical geospatial locations.

Predictive Modelling and Data Analytics

Predictive modelling is also known as predictive analytics. Mostly the term "predictive modelling" is used in academic settings, while "predictive analytics" is the preferred term for commercial applications of predictive modelling.

Successful use of predictive analytics depends heavily on unconstrained access to sufficient volumes of accurate, clean and relevant data. While predictive models can be highly complex, such as those using decision trees and k-means clustering, the most complex part is always the neural network; that is, the model by which computers are trained to predict outcomes. Machine learning uses a neural network to find correlations in exceptionally large data sets and "to learn" and identify patterns within the data.

Benefits of Predictive Modelling

In a nutshell, predictive analytics reduce time, effort and costs in forecasting business outcomes. Variables such as environmental factors, competitive intelligence, regulation changes and market conditions can be factored into the mathematical calculation to render more complete views at relatively low costs.

Examples of specific types of forecasting that can benefit businesses include demand forecasting, headcount planning, competitive analysis, fleet and IT hardware maintenance and financial risks.

Challenges of Predictive Modelling

It’s essential to keep predictive analytics focused on producing useful business insights because not everything this technology digs up is useful. Some mined information is of value only in satisfying a curious mind and has few or no business implications.

Also, being able to use more data in predictive modelling is an advantage only to a point. Too much data can skew the calculation and lead to a meaningless or an erroneous outcome.

For example, more coats are sold as the outside temperature drops. But only to a point. People do not buy more coats when it’s 5 degrees outside than they do when it’s 20 degrees. At a certain point, more frigid temps no longer appreciably change that pattern.

And with the massive volumes of data involved in predictive modelling, maintaining security and privacy will also be a challenge. Further challenges rest in machine learning’s limitations.

Limitations of Predictive Modelling

1. **Errors in data labelling:** These can be overcome with reinforcement learning or generative adversarial networks (GANs).

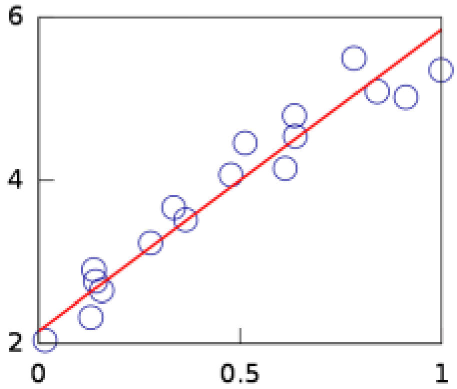
2. **Shortage of massive data sets needed to train machine learning:** A possible fix is “one-shot learning,” wherein a machine learns from a small number of demonstrations rather than on a massive data set.

3. **Generalizability of learning**: Unlike humans, machines have difficulty carrying what they’ve learned forward. In other words, they have trouble applying what they’ve learned to a new set of circumstances. Whatever it has learned is applicable to one use case only. For predictive modelling using machine learning to be reusable—that is, useful in more than one use case—a possible way is transfer learning.

4. **Bias in data and algorithms:** non-representation can skew outcomes and lead to mistreatment of large groups of humans. Further, built-in biases in data are difficult to find. In other words, biases tend to self-perpetuate. This is a moving target, and no clear fix has yet been identified.

Regression

In statistical modelling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.



Ex: predicting the price of a house

Price (Y) depends on the following variables:

- Size of the house in square feet (X1)
- Number of rooms in the house (X2)
- City where the house is located (X3)
- Age of the house (X4)

…and many others

In such a case, the regression analysis problem tries to answer the following question – given the above variables, what would be the predicted price of a house?

Clearly, as the size of the house increases, price increases. Same with the number of rooms. However, with an increase in age, the price decreases. A house in Gachibowli would be far more expensive than a house in Nadergul. Regression tries to model the underlying correlation between these variables to develop a framework to predict the price of a house.

Two commonly used types of Regression Models:

- Linear Regression
  - Simple Linear Regression
  - Multiple linear regression
- Logistic Regression

Simple Linear Regression

Simple linear regression is a statistical method that allows us to summarize and study relationships between *two* continuous (quantitative) variables.

One variable, denoted by ‘x’ is regarded as the predictor or independent variable.
The other variable, denoted by ‘y’ is regarded as the outcome, or dependent variable.

The relationship shown by a Simple Linear Regression model is linear or a sloped straight line, hence it is called Simple Linear Regression.

The key point in Simple Linear Regression is that the dependent variable must be a continuous/real value. However, the independent variable can be measured on continuous or categorical values.

Simple Linear regression algorithm has mainly two objectives:

- Model the relationship between the two variables. Such as the relationship between Income and expenditure, experience and Salary, etc.
- Forecasting new observations. Such as Weather forecasting according to temperature, Revenue of a company according to the investments in a year, etc.

The Simple Linear Regression model can be represented using the below equation:
(With simple linear regression we want to model our data as follows:)

$$y = B0 + B1 * x$$

This is a line where y is the output variable we want to predict, x is the input variable we know and B0 and B1 are coefficients that we need to estimate that move the line around.

Technically, B0 is called the intercept because it determines where the line intercepts the y-axis. In machine learning we call this the bias, because it is added to offset all predictions that we make.

The B1 term is called the slope because it defines the slope of the line or how x translates into a y value before we add our bias.

The goal is to find the best estimates for the coefficients to minimize the errors in predicting y from x.

Rather than having to search for values by trial and error or calculate them analytically using more advanced linear algebra, we can estimate them directly from our data.

We can start off by estimating the value for B1 as:

$$\boxed{\text{B1 = sum((xi-mean(x)) * (yi-mean(y))) / sum((xi – mean(x)) \^{}2)}}$$

Where mean() is the average value for the variable in our dataset. The xi and yi refer to the fact that we need to repeat these calculations across all values in our dataset and i refers to the i’th value of x or y.

We can calculate B0 using B1 and some statistics from our dataset, as follows:

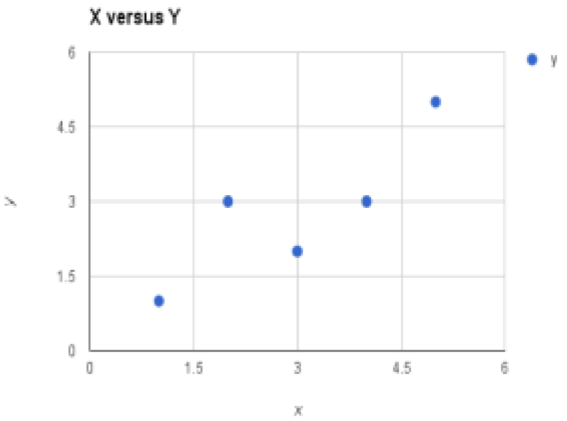
$$\boxed{\text{B0 = mean(y) – B1 * mean(x)}}$$

Ex:

x	y
1	1
2	3
4	3
3	2
5	5

The attribute x is the input variable and y is the output variable that we are trying to predict. If we got more data, we would only have x values and we would be interested in predicting y values.

Below is a simple scatter plot of x versus y.



We can see the relationship between x and y looks kind of linear. As in, we could probably draw a line somewhere diagonally from the bottom left of the plot to the top right to generally describe the relationship between the data.

First, calculate the mean value of our x and y variables:

mean(x) = 3 and mean(y) = 2.8

Now we need to calculate the error of each variable from the mean.

for x:

x	**mean(x)**	**x - mean(x)**
1	3	-2
2	3	-1
4	3	1
3	3	0
5	3	2

for y:

y	**mean(y)**	**y - mean(y)**
1	2.8	-1.8
3	2.8	0.2
3	2.8	0.2
2	2.8	-0.8
5	2.8	2.2

We now have the parts for calculating the numerator. All we need to do is multiply the error for each x with the error for each y and calculate the numerator as follows:

x - mean(x)	**y - mean(y)**	**Multiplication (Numerator)**
-2	-1.8	3.6
-1	0.2	-0.2
1	0.2	0.2
0	-0.8	0
2	2.2	4.4
	Total:	**8**

Now we need to calculate the bottom part of the equation or the denominator for calculating B1. This is calculated as the sum of the squared differences of each x value from the mean.

x - mean(x)	**squared**
-2	4
-1	1
1	1
0	0
2	4
sum	**10**

Now we can calculate the value of our slope(B1):

B1 = 8 / 10 ==> B1 = 0.8

Estimating The Intercept (B0):

*Using the equation: B0 = mean(y) – B1 * mean(x)*

B0 = 2.8 – 0.8 * 3

B0 = 0.4

Making Predictions:

We now have the coefficients for our simple linear regression equation.

y = B0 + B1 * x

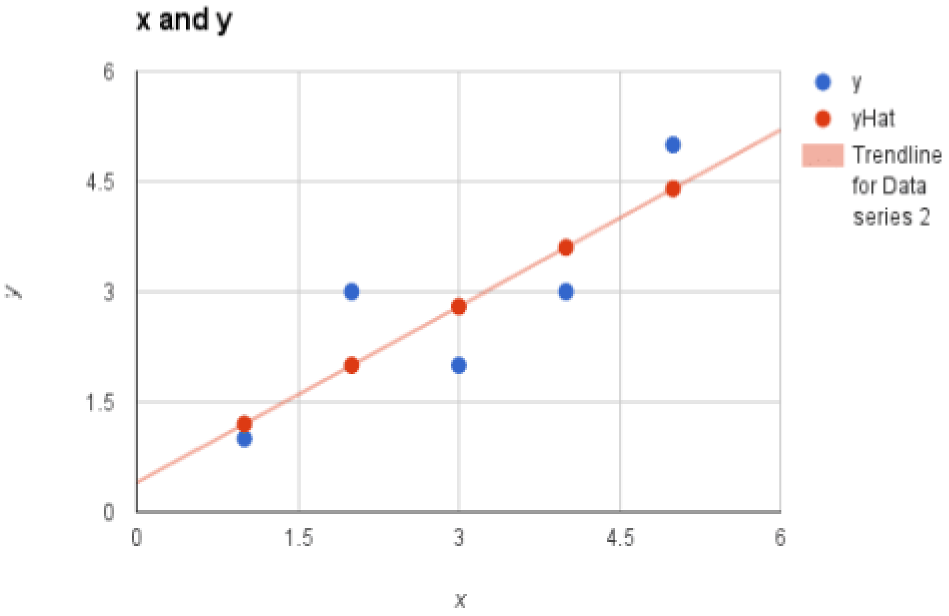
or

y = 0.4 + 0.8 * x

Now, Let’s try out the model by making predictions for our training data.

x	y	predicted y
1	1	1.2
2	3	2
4	3	3.6
3	2	2.8
5	5	4.4

We can plot these predictions as a line with our data. This gives us a visual idea of how well the line models our data.



y—original Values
yHat—Values predicted by model.

Estimating Error:

We can calculate the error for our predictions called the **Root Mean Squared Error** or

RMSE as follows:

RMSE = sqrt(sum((pi – yi)^2)/n)

Where sqrt() is the square root function, ‘p’ is the predicted value and ‘y’ is the actual value, i is the index for a specific instance, ‘n’ is the number of predictions, because we must calculate the error across all predicted values.

First, we must calculate the difference between each model prediction and the actual y values.

predicted-y	Actual-y	Error
1.2	1	0.2
2	3	-1
3.6	3	0.6
2.8	2	0.8
4.4	5	-0.6

We can easily calculate the square of each of these error values (error*error or error^2).

error	squared error
0.2	0.04
-1	1
0.6	0.36
0.8	0.64
-0.6	0.36
Sum	**2.4**

Dividing **Sum** by n and taking the square root gives us:

RMSE = 0.692

Or, *each prediction is on average wrong by about 0.692 units.*

Multiple Linear Regression

When we want to understand the relationship between a single predictor(independent) variable and a response(dependent) variable, we often use simple linear regression.

However, if we’d like to understand the relationship between multiple predictor variables and a single response variable then we can instead use multiple linear regression.

If we have p predictor variables, then a multiple linear regression model takes the form:

$$\boxed{Y = \beta0 + \beta1X1 + \beta2X2 + \ldots + \beta pXp + \varepsilon}$$

where:
Y: The response variable
Xj: The jth predictor variable
βj: The average effect on Y of a one unit increase in Xj, holding all other predictors fixed
ε: The error term

The values for $\beta0$, $\beta1$, $B2$, … , βp are chosen using the least square method, which minimizes the sum of squared residuals (RSS):

$$\boxed{RSS = \Sigma(yi - \hat{y}i)2}$$

where:
Σ: A greek symbol that means sum
yi: The actual response value for the ith observation
$\hat{y}i$: The predicted response value based on the multiple linear regression model

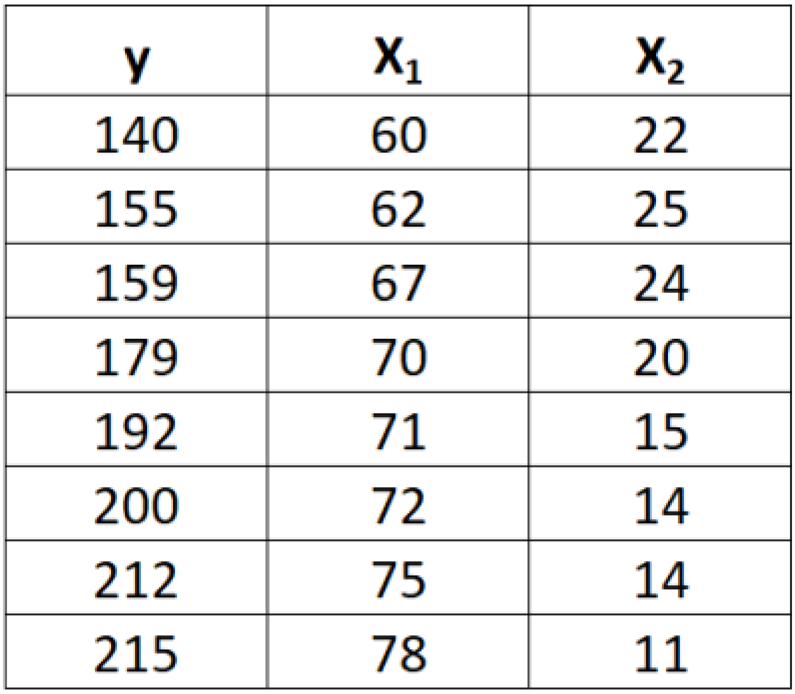
The method used to find these coefficient estimates relies on matrix algebra.

Multiple Linear Regression by Hand (Step-by-Step)

Multiple linear regression is a method we can use to quantify the relationship between two or more predictor variables and a response variable.

Example: Manually calculating coefficients of Multiple Linear Regression.

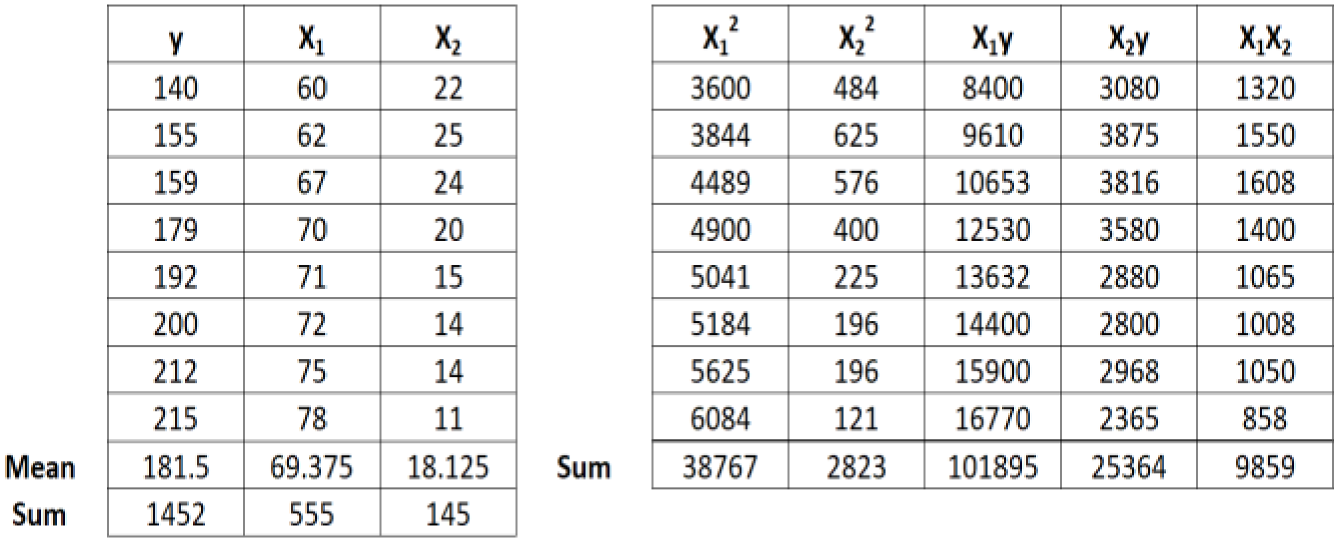
Suppose we have the following dataset with one response variable y and two predictor variables X1 and X2:



y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

Use the following steps to fit a multiple linear regression model to this dataset.

Step 1: Calculate X1^2, X2^2, X_1y, X_2y and X_1X_2.



	y	X_1	X_2
	140	60	22
	155	62	25
	159	67	24
	179	70	20
	192	71	15
	200	72	14
	212	75	14
	215	78	11
Mean	181.5	69.375	18.125
Sum	1452	555	145

	X_1^2	X_2^2	X_1y	X_2y	X_1X_2
	3600	484	8400	3080	1320
	3844	625	9610	3875	1550
	4489	576	10653	3816	1608
	4900	400	12530	3580	1400
	5041	225	13632	2880	1065
	5184	196	14400	2800	1008
	5625	196	15900	2968	1050
	6084	121	16770	2365	858
Sum	38767	2823	101895	25364	9859

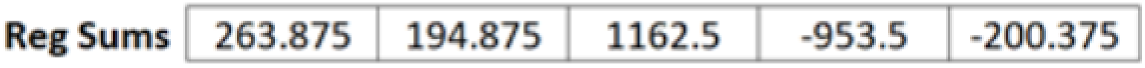
Step 2: Calculate Regression Sums

Next, make the following regression sum calculations:

1. $\Sigma X_1^2 – (\Sigma X_1)^2 / n =$ 38,767 – $(555)^2$ / 8 = **263.875**
2. $\Sigma X_2^2 – (\Sigma X_2)^2 / n =$ 2,823 – $(145)^2$ / 8 = **194.875**
3. $\Sigma X_1 y – (\Sigma X_1 \Sigma y) / n =$ 101,895 – (555*1,452) / 8 = **1,162.5**

4. $\Sigma X_2y - (\Sigma X_2\Sigma y) / n = 25,364 - (145*1,452) / 8 =$ **-953.5**

5. $\Sigma X_1X_2 - (\Sigma X_1\Sigma X_2) / n = 9,859 - (555*145) / 8 =$ **-200.375**



Reg Sums	263.875	194.875	1162.5	-953.5	-200.375

Step 3: Calculate b0, b1, and b2

The formula to calculate b1 is:

$$\boxed{[(\Sigma x_2^2)(\Sigma x_1y) - (\Sigma x_1x_2)(\Sigma x_2y)] \ / \ [(\Sigma x_1^2)\ (\Sigma x_2^2) - (\Sigma x_1x_2)^2]}$$

Thus, b1 = [(194.875) (1162.5) – (-200.375)(-953.5)] / [(263.875) (194.875) – (-200.375)2] = **3.148**

The formula to calculate b2 is:

$$\boxed{[(\Sigma x_1^2)(\Sigma x_2y) - (\Sigma x_1x_2)(\Sigma x_1y)] \ / \ [(\Sigma x_1^2)\ (\Sigma x_2^2) - (\Sigma x_1x_2)^2]}$$

Thus, b2 = [(263.875)(-953.5) – (-200.375)(1152.5)] / [(263.875) (194.875) – (-200.375)2] = -1.656

$$\boxed{y - b_1X_1 - b_2X_2}$$

The formula to calculate b0 is:

Thus, b0 = 181.5 – 3.148(69.375) – (-1.656)(18.125) = -6.867

Step 4: Place b0, b1, and b2 in the estimated linear regression equation

The estimated linear regression equation is:

$$\boxed{\hat{y} = b_0 + b_1{*}x_1 + b_2{*}x_2}$$

In our example, it is $\mathbf{\hat{y} = -6.867 + 3.148x_1 - 1.656x_2}$

Interpreting a Multiple Linear Regression Equation

Here is how to interpret this estimated linear regression equation:

$$\hat{y} = -6.867 + 3.148x_1 - 1.656x_2$$

b0 = -6.867 indicates that when both predictor variables are equal to zero, the mean value for y is -6.867.

b1 = 3.148 indicates that one unit increase in x1 is associated with a 3.148 unit increase in

y, on average, assuming x2 is held constant.
b2 = -1.656 indicates that one unit increase in x2 is associated with a 1.656 unit decrease in y, on average, assuming x1 is held constant.

To make sure your model fits the data use the r^2 value both for Linear Regression and Multiple Linear Regression.

The r^2 value (also called the coefficient of determination) states the portion of change in the data set predicted by the model.

The value will range from 0 to 1, with 0 stating that the model has no ability to predict the result and 1 stating that the model predicts the result perfectly. You should expect the r^2 value of any model you create to be between those two values. If it isn't, retrace your steps because you've made a mistake somewhere.

r^2 = 1 — (Sum of squared errors) / (Total sum of squares)

You can calculate the coefficient of determination for a model using the following equations:

Where,

(Total sum of squares) = Sum(y_i — mean(y))^2

(Sum of squared errors) = sum((Actual_i — Prediction_i)^2)

Simple Logistic Regression

Simple logistic regression computes the probability of some outcome given a single predictor variable as

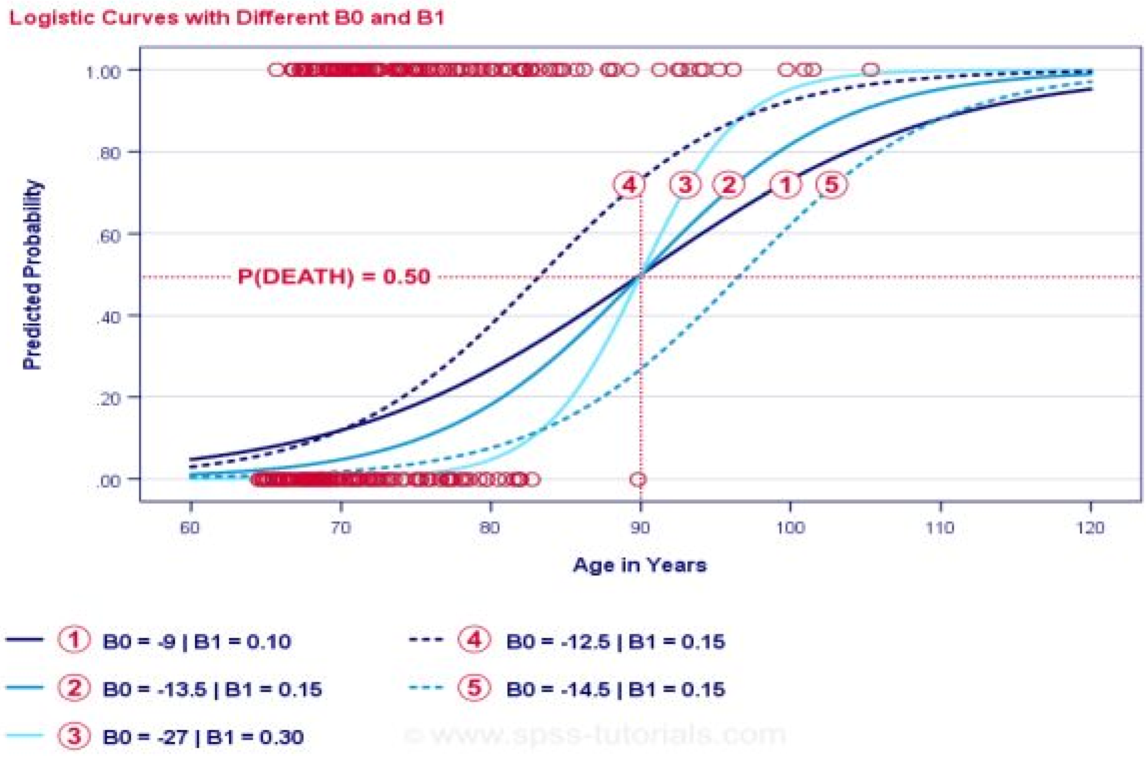
$$P(Y_i) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i})}}$$

where

- $P(Y_i)$ is the predicted probability that Y is true for case i
- e is a mathematical constant of roughly 2.72
- b_0 is a constant estimated from the data
- b_1 is a b-coefficient estimated from the data
- X_i is the observed score on variable X for case i

The very essence of logistic regression is estimating b_0 and b_1. These 2 numbers allow us to compute the probability.

Logistic Regression Example Curves



Logistic regression analysis requires the following assumptions:

- Independent observations
- Correct model specification
- Errorless measurement of outcome variable and all predictors

Multiple Logistic Regression

Simple logistic regression uses only one predictor. The model can easily extend with additional predictors, resulting in multiple logistic regression:

$$P(Y_i) = \frac{1}{1 + e^{-\,(b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki})}}$$

where

$P(Y_i)$ is the predicted probability that Y is true for case i;

e is a mathematical constant of roughly 2.72;
b0 is a constant estimated from the data;
b1, b2, ... ,bk are the b-coefficient for predictors 1, 2, ... ,k;
$X1_i$, $X2_i$, ... ,Xk_i are observed scores on predictors X1, X2, ... ,Xk for case i.