# Curve Fitting: Regression

Regression process selects a certain function that best fits for given set of experimental data (usually presented as a table of x and y values). The 'independent' variable x is usually called the regressor (there may be one or more of these), the 'dependent' variable y is the response variable. The function is also used to predict or forecast the value of dependent variable. The procedure of finding a function that establishes relationship between values is known as *curve fitting* or *regression analysis*.

Suppose the values of y for different values of x are given. If we want to know the effect of x on y, then we may write a functional relationship y = *f* (x) where the variable y is called the *dependent variable* or response variable and the variable x is called the *independent variable or regressor*. In curve fitting, the general problem is to find a mathematical relationship of the form y = *f* (x), suggested by the given values. The relationship may be either linear or non linear as shown in Figure 1.

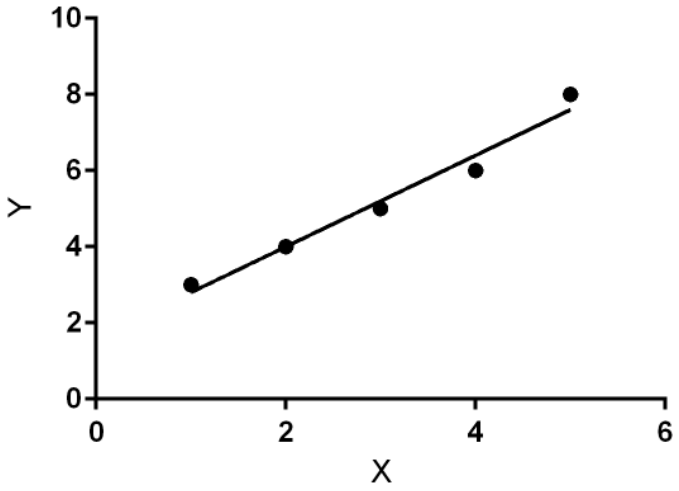


Figure 1

It is a standard practice to prepare a *scatter diagram* as shown in Figure 1 and try to determine the functional relationship needed to fit the points. The curve should best fit the plotted points. This means that the average error introduced by the assumed curve is minimum.

The class from which the functions are selected (the model) is usually one of the following types:

1. a linear function of x (i.e. y = a+ bx): simple (univariate) linear regression,
2. a linear function of $x_1$, $x_2$, ... $x_k$: multiple (multivariate) linear regression,
3. a polynomial function of x: polynomial regression,
4. any other type of function, with one or more parameters (e.g. $y = ae^{bx}$): nonlinear regression.

**Least Square Regression:**

Least square regression is a technique that fit the data as linear function, transcendental function or polynomial function.

Let us consider the mathematical equation for a straight line: y = f(x) = a + bx, to describe the data. We know that 'a' is the intercept of the line and 'b' is its slope.

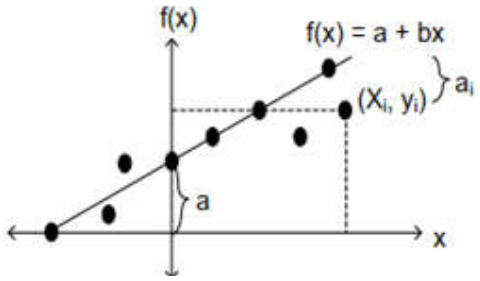


Figure 2: Least Square Regression

Consider a point $(x_i, y_i)$ as shown in figure. The vertical distance of this point from the line f(x) = a + bx is the error $a_i$.

Then $a_i = y_i - f(x_i) = y_i - a - bx_i$ ..... .....(1)

The best approach that could be tried for fitting a best line through the data is to minimize the sum of squares of errors.

$$i.e \sum a_i^2 = \sum (y_i - a - bx_i)^2$$

The technique of minimizing the sum of square of errors is known as least square regression.

**Fitting a Straight Line**

Fitting a straight line is the simplest approach of regression analysis. Let us consider the mathematical equation for a straight line

$$y = f(x) = a + bx \quad \text{..... (1)}$$

Usually fitting a straight line means finding the values of the parameter a and b of the above equation, as well as actually constructing the line itself.

Consider a point $(x_i, y_i)$ as shown in Figure 2. The vertical distance of this point from the line $f(x) = a + bx$ is the error $a_i$. Then, $a_i = y_i - f(x_i) = y_i - a - bx_i$

There are many approaches that could be tried to fit a best line through the data points. They include:

a. Minimize the sum of errors i.e. minimize $\sum a_i = \sum (y_i - a - bx_i)$

b. Minimize the sum of absolute values of errors i.e. minimize $\sum |a_i| = \sum |(y_i - a - bx_i)|$

c. Minimize the sum of squares of errors i.e. minimize $\sum a_i^2 = \sum (y_i - a - bx_i)^2$

It can be easily verified that the first two strategies do not yield a unique line for a given set of data. The third strategy overcomes this problem and guarantees a unique line. The technique of minimizing the sum of squares of errors is known as *least square method*.

**Least Squares Method**

Let the sum of squares of individual errors be expressed as

$$Q = \sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

In the method of least squares, we choose a and b such that Q is minimum. The value of Q depends on a and b and a necessary condition for Q to be minimum is:

$$\frac{\partial Q}{\partial a} = 0 \text{ and } \frac{\partial Q}{\partial b} = 0$$

Then, $$\frac{\partial Q}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\frac{\partial Q}{\partial b} = -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0$$

Thus, $$-\sum y_i + na + b\sum x_i = 0$$

$$-\sum x_i y_i + a\sum x_i + b\sum x_i^2 = 0$$

Rearranging we get

$$n a + b \sum x_i = \sum y_i$$

$$a \sum x_i + b \sum x_i^2 = \sum x_i y_i$$

These are called *normal equations*. Solving for a an b, we get

$$\mathbf{a} = \frac{\mathbf{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}}{\mathbf{n \sum x_i^2 - (\sum x_i)^2}} \quad \text{and} \quad \mathbf{b} = \frac{\mathbf{n \sum x_i y_i - \sum x_i \sum y_i}}{\mathbf{n \sum x_i^2 - (\sum x_i)^2}} \quad \mathbf{or}$$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \quad \text{and} \quad a = \frac{1}{n}(\sum y - b\sum x)$$

***Example 1*:** Certain experimental values of x and y are given below in the following table:

| x | 0 | 2 | 5 | 7 |
|---|---|---|---|---|
| y | -1 | 5 | 12 | 20 |

If y = a + bx, find the value of a and b.

***Solution*:**

| x | Y | $x^2$ | Xy |
|---|---|---|---|
| 0 | -1 | 0 | 0 |
| 2 | 5 | 4 | 10 |
| 5 | 12 | 25 | 60 |
| 7 | 20 | 49 | 140 |
| ∑14 | ∑36 | ∑78 | ∑210 |

From the normal equations of straight line we get,

$n\,a + b\,\sum x_i = \sum y_i$

$4a + 14b = 36$

And,

$a\,\sum x_i + b\,\sum x_i^2 = \sum x_i y_i$

$14a + 78b = 210$

Solving above two equations, we obtain

a = -1.1381 and b = 2.8966

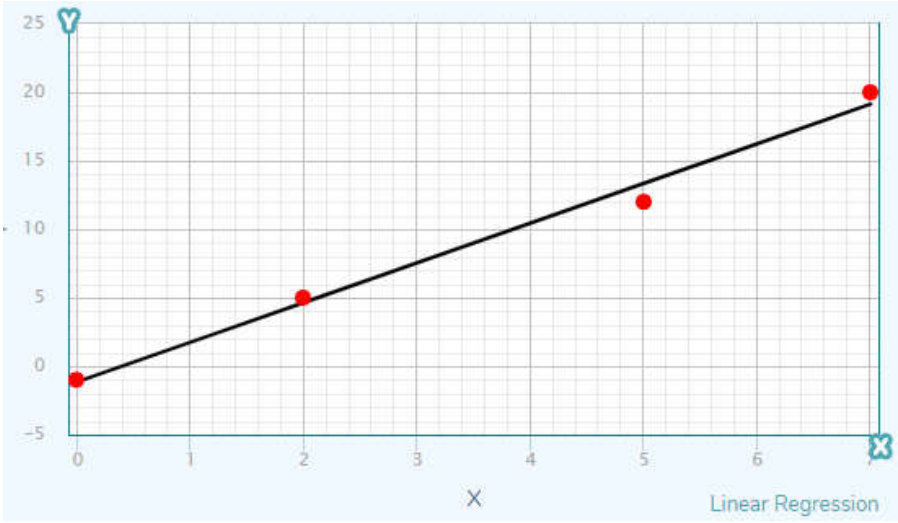


Figure 3: Linear Regression

**Nonlinear Curve Fitting: Fitting a Power Function**

Let $y = ax^b$ be the function to be fitted to the given data. Taking logarithms of both sides, we obtain the relation: **log y = log a + b log x,** which is the form of Y = A + B X.

Where Y = log y, A = log a, B = b and X = log x.

Hence, we can evaluate the value of A and B by the normal equations of straight line. Then a and b can be calculated from the formulae A = log a and B = b.

$$b = \frac{n\sum \ln x_i \ln y_i - \sum \ln x_i \sum \ln y_i}{n\sum (\ln x_i)^2 - (\sum \ln x_i)^2} \quad \text{and}$$

$$\ln a = R = \frac{1}{n}(\sum \ln y_i - b \sum \ln x_i) \quad \text{or,} \quad a = e^R$$

***Example 2*:**

Given the data table:

| x | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| y | 0.5 | 2 | 4.5 | 8 | 12.5 |

Fit a power function model of the form $y = ax^b$.

Solution:

| xi | yi | ln(xi) | ln(yi) | {ln(xi)}$^2$ | ln(xi).ln(yi) |
|---|---|---|---|---|---|
| 1 | 0.5 | 0 | -0.69 | 0 | 0 |
| 2 | 2 | 0.6931 | 0.693 | 0.4804 | 0.4804 |
| 3 | 4.5 | 1.0986 | 1.564 | 1.2069 | 1.6524 |
| 4 | 8 | 1.3863 | 2.079 | 1.9218 | 2.8827 |
| 5 | 12.5 | 1.6094 | 2.526 | 2.5901 | 4.0649 |
| ∑=15 | ∑=27.5 | ∑=4.7874 | ∑=6.1692 | ∑=6.1992 | ∑=9.0804 |

Now,

$$b = \frac{n\sum \ln x_i \ln y_i - \sum \ln x_i \sum \ln y_i}{n\sum (\ln x_i)^2 - (\sum \ln x_i)^2} \qquad b = \frac{5*9.0804-(4.7874*6.1692)}{5*6.1992-(4.7874)^2} = 2.2832$$

$$\ln a = R = \frac{1}{n}(\sum \ln y_i - b \sum \ln x_i)$$

$R = 1/5(6.1692 - (2{,}2832*4.7874)) = \mathbf{-0.9641}$

$\mathbf{a = e^R = e^{-0.9641} = 0.381}$

Finally, the power form equation is: $y = 0.381.x^{2.2832}$

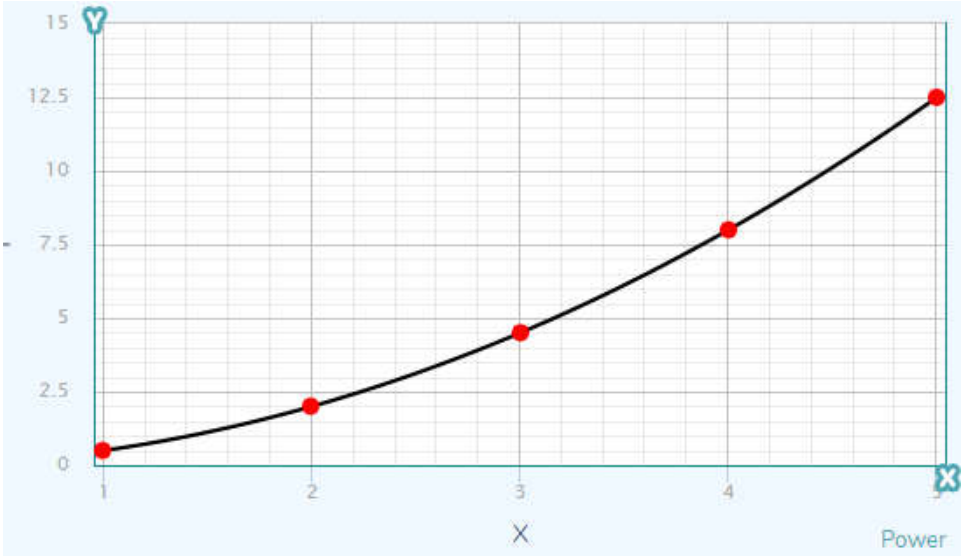


Figure 4: Power Curve Fitting

**Fitting a Polynomial of nth Degree**

When a given series of data does not appear to satisfy a linear equation, we can try a suitable polynomial as a regression to fit the data. The least square technique can be used to fit the data to a polynomial.

Let the curve be represented by the equation of degree m-1

$$y = f(x) = a_1 + a_2x + a_3x^2 + \ldots\ldots\ldots\ldots + a_mx^{m-1}$$

The sum of squares of individual errors be expressed as

$$Q = \sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2 = \sum_{i=1}^{n} (y_i - a_1 - a_2x_i - a_3x_i^2 - \ldots\ldots - a_mx_i^{m-1})^2$$

In the method of least squares, we have to choose $a_0, a_1, \ldots, a_n$ such that Q is minimum.

$$\frac{\partial Q}{\partial a_1} = 0, \frac{\partial Q}{\partial a_2} = 0, \frac{\partial Q}{\partial a_3} = 0, \ldots\ldots\ldots, \frac{\partial Q}{\partial a_m} = 0$$

Consider a general term,

$$\frac{\partial Q}{\partial a_j} = -2\sum_{i=1}^{n} [y_i - f(x_i)] \frac{\partial f(x_i)}{\partial a_j} = 0$$

$$\frac{\partial f(x_i)}{\partial a_j} = x_i^{j-1}$$

Thus, we have

$$\sum_{i=1}^{n} [y_i - f(x_i)]\ x_i^{j-1} = 0 \ ; j = 1, 2, \ldots\ldots, m$$

$$\sum_{i=1}^{n} [\, y_i \; x_i^{j-1} - x_i^{j-1} f(x_i)] = 0$$

Substituting for $f(x_i)$

$$\sum_{i=1}^{n} x_i^{j-1} \; (a_1 + a_2 x_i + a_3 x_i^2 + \ldots\ldots\ldots\ldots + a_m x_i^{m-1}) = \sum_{i=1}^{n} y_i \, x_i^{j-1}$$

Equating, as before, the first partial derivatives to zero and simplifying, we get the following normal equations

$$a_1 n + a_2 \sum x_i + a_3 \sum x_i^2 + \ldots\ldots\ldots + a_m \sum x_i^{m-1} = \sum y_i$$
$$a_1 \sum x_i + a_2 \sum x_i^2 + a_3 \sum x_i^3 + \ldots\ldots\ldots + a_m \sum x_i^{m} = \sum y_i x_i$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$a_1 \sum x_i^{m-1} + a_2 \sum x_i^{m} + a_3 \sum x_i^{m+1} + \ldots\ldots\ldots + a_m \sum x_i^{2m-2} = \sum y_i \, x_i^{m-1}$$

***Example 3*:** Fit a second order polynomial to the data in the table below:

| x | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| y | 6 | 11 | 18 | 27 |

***Solution:***

$$a_1 n + a_2 \sum x_i + a_3 \sum x_i^2 = \sum y_i$$
$$a_1 \sum x_i + a_2 \sum x_i^2 + a_3 \sum x_i^3 = \sum y_i x_i$$
$$a_1 \sum x_i^2 + a_2 \sum x_i^3 + a_3 \sum x_i^4 = \sum y_i x_i^2$$

| **x** | **y** | $\mathbf{x^2}$ | $\mathbf{x^3}$ | $\mathbf{x^4}$ | **yx** | $\mathbf{yx^2}$ |
|---|---|---|---|---|---|---|
| 1 | 6 | 1 | 1 | 1 | 6 | 6 |
| 2 | 11 | 4 | 8 | 16 | 22 | 44 |
| 3 | 18 | 9 | 27 | 81 | 54 | 162 |
| 4 | 27 | 16 | 64 | 256 | 108 | 432 |
| $\sum$ 10 | $\sum$ 62 | $\sum$ 30 | $\sum$ 100 | $\sum$ 354 | $\sum$ 190 | $\sum$ 644 |

Substituting these values, we get

$$4a_1 + 10a_2 + 30a_3 = 62$$
$$10a_1 + 30a_2 + 100a_3 = 190$$
$$30a_1 + 100a_2 + 354a_3 = 644$$

Solve $a_1$, $a_2$ and $a_3$ using cramer's rule:

$$\begin{matrix} 4 & 10 & 30 & 62 \\ 10 & 30 & 100 & \vdots\, 190 \\ 30 & 100 & 354 & 644 \end{matrix}$$

Determinate $D = \begin{bmatrix} 4 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{bmatrix} = 4*(30*354 - 100*100) - 10\,(10*354 - 100*30) + 30(10*100 - 30*30) = 80$

$D_x = \begin{bmatrix} 62 & 10 & 30 \\ 190 & 30 & 100 \\ 644 & 100 & 354 \end{bmatrix} = 240$

$D_y = \begin{bmatrix} 4 & 62 & 30 \\ 10 & 190 & 100 \\ 30 & 644 & 354 \end{bmatrix} = 160$

$D_z = \begin{bmatrix} 4 & 10 & 62 \\ 10 & 30 & 190 \\ 30 & 100 & 644 \end{bmatrix} = 80$

Solving these equations gives,

$a_1 = D_x/D = 3$, $a_2 = D_y/D = 2$, and $D_z/D = 1$

So the least squares quadratic polynomial is: $y = 3 + 2x + x^2$

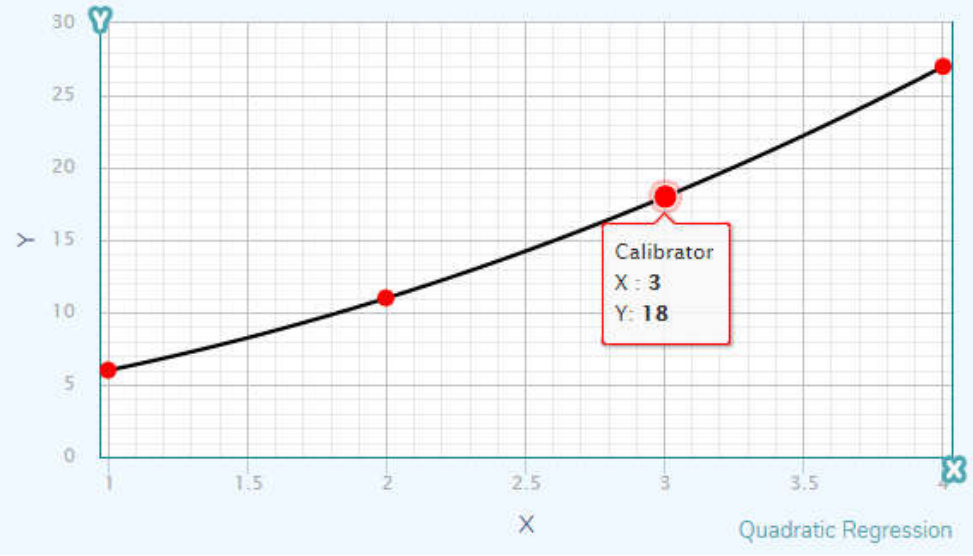


Figure 4: Polynomial Regression

**MULTIPLE LINEAR REGRESSIONS:**

There are number of situations, where the dependent variable is a function of two or more variables. For example, the salary of a sales person may be expressed as: y = 500 + 5$x_1$ + 8$x_2$, where $x_1$ and $x_2$ are the number of unit sold of product 1 and 2 respectively. We shall discuss an approach to fit the experimental data, where the variable under consideration is a linear function of two independent variables. Let us consider a two variables linear function as follows:

$y = f(x, z) = a_1 + a_2 x_i + a_3 z_i$ ……(1)

Then, the sum of square of errors is given by:

$Q = \sum_{i=1}^{n}\{y_i - f(x,z)\}^2 = \sum_{i=1}^{n}\{y_i - (a_1 + a_2 x_i + a_3 z_i)\}^2$ ……(2)

Differentiation equation … (ii) w. r. t. $a_1$, $a_2$, $a_3$ and equating them to zero, we will get the condition for minimum error.

$$\frac{dq}{da_1} = -2\sum\{y_i - (a_1 + a_2 x_i + a_3 z_i)\} = 0$$

$$\frac{dq}{da_2} = -2\sum\{y_i - (a_1 + a_2 x_i + a_3 z_i)x_i\} = 0$$

$$\frac{dq}{da_3} = -2\sum\{y_i - (a_1 + a_2 x_i + a_3 z_i)z_i\} = 0$$

With these conditions, we will get the following simultaneous equation:

$a_1 n + a_2\sum x_i + a_3\sum z_i = \sum y_i$

$a_1\sum x_i + a_2\sum \boldsymbol{x_i^2} + a_3\sum x_i z_i = \sum x_i y_i$

$a_1\sum \boldsymbol{z_i} + a_2\sum x_i z_i + a_3\sum \boldsymbol{z_i^2} = \sum z_i y_i$

***Example 4:***

Given the following table of data find the value of $a_1$, $a_2$ and $a_3$

| **x** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **z** | **0** | **1** | **2** | **3** |
| **y** | **12** | **18** | **24** | **30** |

*Solution:*

| **x** | **z** | **Y** | **$x_2$** | **$z_2$** | **z.x** | **x.y** | **y.z** |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 12 | 1 | 0 | 0 | 12 | 0 |
| 2 | 1 | 18 | 4 | 1 | 2 | 36 | 18 |
| 3 | 2 | 24 | 9 | 4 | 6 | 72 | 48 |
| 4 | 3 | 30 | 16 | 9 | 12 | 120 | 90 |
| ∑10 | ∑6 | ∑84 | ∑30 | ∑14 | ∑20 | ∑240 | ∑156 |

So, three simultaneous equations will be:

$4a_1 + 10a_2 + 6a_3 = 84$ …….. (i)

$10a_1 + 30a_2 + 20a_3 = 240$ … ……(ii)

$6a_1 + 20a_2 + 14a_3 = 156$ … ……..(iii)

This pdf is based on Text Book Numerical Methods by E Balagurushamy; Publisher: Tata McGraw-Hill Publishing Company Limited

On solving we can get the values of unknowns.