SOLUTION OF REGRESSION MODELS

Linear Regression

2007 *Fabian Farelo , Autar Kaw, Jamie Trahan*
University of South Florida
United States of America
kaw@eng.usf.edu
http://numericalmethods.eng.usf.edu

Introduction

Linear Regression is the most popular regression model. In this model we wish to predict response points to n data points (x_1,y_1), (x_2,y_2),..., (x_n,y_n) by a regression model given by:

$$y = a_0 + a_1 x \qquad (1.1)$$

where a_0 and a_1 are the constants of the regression model. A measure of goodness of fit, that is, how $a_0 + a_1x$ predicts the response variable y is the magnitude of the residual, ε_i , at each of the n data points

$$\varepsilon_i = \text{observed value at } x_i \text{ - predicted value at } x_i \qquad (1.2)$$

Ideally, if all of the residuals are zero, one may find an equation in which all the points lie on the model. Thus, minimization of the residual is an objective of obtaining regression coefficients. The most popular method to minimize the residual is the least squares method, where the estimates of the constants of the models are chosen such that the sum of the squared residuals, S_r , is minimized, that is minimize

$$S_r = \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 \cdot x_i)^2 \qquad (1.3)$$

Let us use the least squares criterion where we minimize the sum of the squared residual, S_r:

$$\frac{d}{da_0} S_r = 0 \qquad (1.4)$$

$$\frac{d}{da_1} S_r = 0 \qquad (1.5)$$

Once S_r is minimized with respect to the regression coefficients, a_0 and a_1 (see **Linear Regression** for complete derivation), the coefficients can be solved for:

$$a_1 = \frac{Sxy}{Sxx} \qquad (1.6)$$

$$a_0 = y_{ave} - a_1 * x_{ave} \qquad (1.7)$$

where S_{xy} and S_{xx} can be defined as:

$$Sxy = \sum_{i=1}^{n} \left(x_i \cdot y_i\right) - n \cdot x_{ave} \cdot y_{ave} \tag{1.8}$$

$$Sxx = \sum_{i=1}^{n} \left(x_i\right)^2 - n \cdot \left(x_{ave}\right)^2 \tag{1.9}$$

and the average values can be defined as:

$$x_{ave} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.10}$$

$$y_{ave} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{1.11}$$

Section 1: Input Data

Below are the input parameters to begin the simulation. This is the only section that requires user input. The user can change the values that are highlighted and Mathcad will calculate the Linear Regression Model for the data set.
Note: the origin has been set to one to redefine the starting index of all arrays. The user SHOULD NOT change this value. ORIGIN := 1

- Number of Data Points, *n*

$n := 5$

- Data Points

$i := 1 .. n$

$x_i :=$

1
2
3
4
5

$y_i :=$

1
4
9
16
25

Section 2: Linear Regression Method

Calculating average values using Equations (1.10) and (1.11):

$$xave := \frac{\sum_{i=1}^{n} x_i}{n}$$

$xave = 3$

$$yave := \frac{\sum_{i=1}^{n} y_i}{n}$$

$yave = 11$

Calculating *Sxy* and *Sxx* using Equations (1.8) and (1.9):

$$Sxy := \sum_{i=1}^{n} \left(x_i \cdot y_i\right) - n \cdot xave \cdot yave$$

$$Sxy = 60$$

$$Sxx := \sum_{i=1}^{n} \left(x_i\right)^2 - n \cdot xave^2$$

$$Sxx = 10$$

Calculating a_1 and a_0 using Equations (1.6) and (1.7):

$$a1 := \frac{Sxy}{Sxx}$$

$$a1 = 6$$

$$a0 := yave - a1 \cdot xave$$

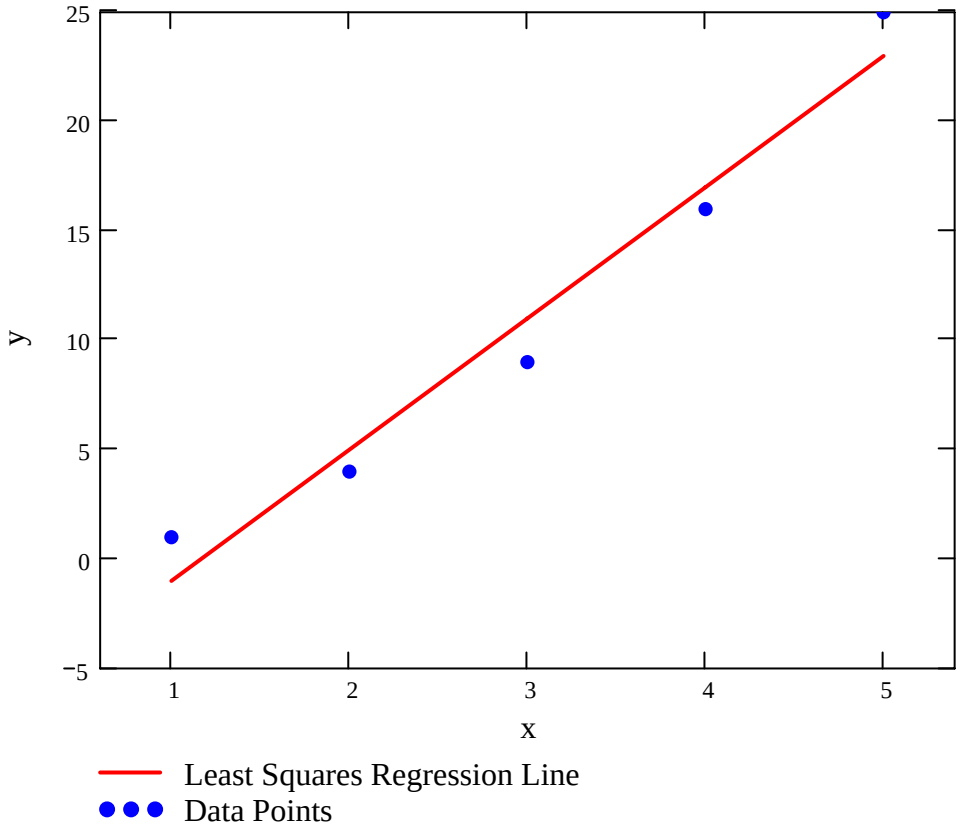
$$a0 = -7$$

Results

The regression coefficients a_0 and a_1 can now be used to describe the regression model. The figure shows the data points as well as the least squares regression line.

$$y1 := a0 + a1 \cdot x$$

Figure 1: Linear Regression Model shown with data points



Coefficient of determination

A good indicator of how well least squares characterizes or predicts the whole data is a quantity called the **coefficient of determination**, r^2 ,

$$r^2 = \frac{St-Sr}{St} \qquad (2.1)$$

where
S_r = the sum of the squares of the residuals (a value that quantifies the spread around the regression line)
and
S_t = the sum of the squares of deviation from the mean (a value that measures the spread between the data and its mean).

The objective of least squares method is to obtain a compact equation that best describes all data points. The mean can also be used to describe only data points. The magnitude of the sum of squares of deviation from the mean or from the least squares line is therefore a good indicator of how well the mean or least squares characterizes the whole data.

The difference between these two parameters, (S_t-S_r), measures the error due to describing or characterizing the data in one form instead of the other. A relative comparison of this difference with the sum of squares deviation associated with the mean, (i.e. r^2), describes the proportion of variation in the response data that is explained by the regression model. When all the points in a data set lie on the regression model, the largest possible value of $r^2 = 1$ is obtained, while a minimum possible value of $r^2 = 0$ is obtained when there is only one data point or if the straight line regression model is a constant line. (Note that $0 \leq r^2 \leq 1$)

Calculation of the coefficient of determination:

Sum of the square of the difference between observed value and average value, St:

$$St := \sum_{i=1}^{n} \left(y_i - yave\right)^2$$

$$St = 374$$

Sum of the square of the residuals, Sr:

$$Sr := \sum_{i=1}^{n} \left(y_i - y1_i\right)^2$$

$$Sr = 14$$

Coefficient of determination, r^2:

$$r2 := \frac{St - Sr}{St}$$

$$r2 = 0.96257$$

As r^2 approaches unity, the regression model describes the data more accurately and the model is more apt to predict the response variable.

References

[1] Autar Kaw, *Holistic Numerical Methods Institute, http://numericalmethods.eng.usf.edu*, See **Linear Regression**

Conclusion

Mathcad helped us apply our knowledge of linear regression method to regress a given d set to a straight line.

Question 1: In the table below is given the instantaneous thermal expansion coefficient a function of temperature. Find the linear regression model that relates the Instantaneous Thermal Expansion α, as a function of temperature. What is the coefficient of determinat of the model?

Table 1: Instantaneous thermal expansion coefficient as a function of temperature.

$i := 1 .. 22$

Temperature *T, (F)*	Instantaneous Expansion E-06 *in/(inF)*
$T_i :=$	$\alpha_i :=$
80	6.47
60	6.36
40	6.24
20	6.12
0	6.00
–20	5.86
–40	5.72
–60	5.58
–80	5.43
–100	5.28
–120	5.09
–140	4.91
–160	4.72
–180	4.52
–200	4.30
–220	4.08
–240	3.83
–260	3.58
–280	3.33
–300	3.07
–320	2.76
–340	2.45

Question 2: In the table below is given the stress-strain data for a tensile test of a unidirect composite material.

$i := 12$

Strain (%)	Stress (MPa)
$\varepsilon_i :=$	$\sigma_i :=$
0	0
0.183	306
0.36	612
0.5324	917
0.702	1223
0.867	1529
1.0244	1835
1.1774	2140
1.329	2446
1.479	2752
1.5	2767
1.563	2896

Find the straight line regression model that finds the relationship between the stress and s
Note that the straight line has no intercept. The slope of the straight line is the longitudin
Young's Modulus of the composite material.