Linear Regression

University of South Florida

United States of America

kaw@eng.usf.edu

Note: This worksheet demonstrates the use of Maple to illustrate the procedure to regress a given data set to a straight line.

Introduction

Linear Regression is the most popular regression model. In this model we wish to predict response points to n data points $(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$ 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 $a_0 + a_1 x$ predicts the response variable y is the magnitude of the residual, ε_i at each of the n data points

$$\varepsilon_i = (\text{observed value at } x_i - \text{predicted value at } x_i) = y - (a_0 + a_1 x) \qquad (1.2)$$

Ideally, if all the residuals, ε_i, are zero, one will 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

$Sr = Sum((\text{epsilon}[i])^2, i = 1..n)$

$$Sr = \sum_{i=1}^{n} \varepsilon_i^2 \qquad \textbf{(1.1)}$$

where

$Sum(\text{epsilon}[i]^2, i = 1..n) = Sum((y(i) - (a[0] + a[1] \cdot x[i]))^2, i = 1..n)$

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y(i) - a_0 - a_1 x_i\right)^2 \qquad \textbf{(1.2)}$$

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

$Diff((Sr), a0) = 0$

$$\frac{\partial}{\partial a0} Sr = 0 \tag{1.3}$$

$Diff((Sr), a1) = 0$

$$\frac{\partial}{\partial a1} Sr = 0 \tag{1.4}$$

Once Sr is minimized with respect to the regression coefficients, a0 and a1 (see Linear Regression notes for complete derivation), the coefficients can be solved for:

$a[0] = y[ave] - a[1] \cdot (x[ave])$

$$a_0 = y_{ave} - a_1 x_{ave} \tag{1.5}$$

$a[1] = \frac{Sxy}{Sxx}$

$$a_1 = \frac{Sxy}{Sxx} \tag{1.6}$$

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

$Sxy = Sum(x[i] \cdot y[i], i = 1..n) - n \cdot x[ave] \cdot y[ave]$

$$Sxy = \sum_{i=1}^{n} x_i y_i - n\, x_{ave} y_{ave} \tag{1.7}$$

$Sxx = Sum\left(x[i]^2, i = 1..n\right) - n \cdot (x[ave])^2$

$$Sxx = \sum_{i=1}^{n} x_i^2 - n\, x_{ave}^2 \tag{1.8}$$

and the average values can be defined as:

$x[ave] = \frac{Sum(x[i], i = 1..n)}{n}$

$$x_{ave} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.9}$$

$y[ave] = \frac{Sum(y[i], i = 1..n)}{n}$

$$y_{ave} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{1.10}$$

Section 1: Input Data

Below are the input parameters to begin the simulation. This is the only section that requires user input. Once the values are entered, Maple will calculate the Linear Regression model for the data set.

```
> x1:=[1,2,3,4,5];
  y1:=[1,4,9,16,25];
  n:=5;
```

$$x1 := [1, 2, 3, 4, 5]$$
$$y1 := [1, 4, 9, 16, 25]$$
$$n := 5 \quad \textbf{(2.1)}$$

Section 2: Results

Using Equations (1.5) and (1.6) to calculate x_{ave} and y_{ave}:

```
> xave:=add(x1[i],i=1..n)/n;
```

$$xave := 3 \quad \textbf{(3.1)}$$

```
> yave:=add(y1[i],i=1..n)/n;
```

$$yave := 11 \quad \textbf{(3.2)}$$

Using Equations (1.3) and (1.4) to calculate S_{xx} and S_{xy}:

```
> Sxy:=0:
  Sxx:=0:
  for i from 1 by 1 to n do
       Sxy:=Sxy+x1[i]*y1[i]-xave*yave:
       Sxx:=Sxx+((x1[i])^2)-(xave)^2:
  end do:
> Sxx;
```

$$10 \quad \textbf{(3.3)}$$

```
> Sxy;
```

$$60 \quad \textbf{(3.4)}$$

S_{xx} , S_{xy}, x_{ave}, and y_{ave} can be used to calculate the regression coefficients, a_0 and a_1 using Eq (1.7) and (1.8):

```
> a1:=evalf(Sxy/Sxx);
```

(3.5)

$$a1 := 6. \quad (3.5)$$

```
> a0:=yave-a1*xave;
```

$$a0 := -7. \quad (3.6)$$

The Linear model is described as

```
> y:=a0+a1*'x';
```

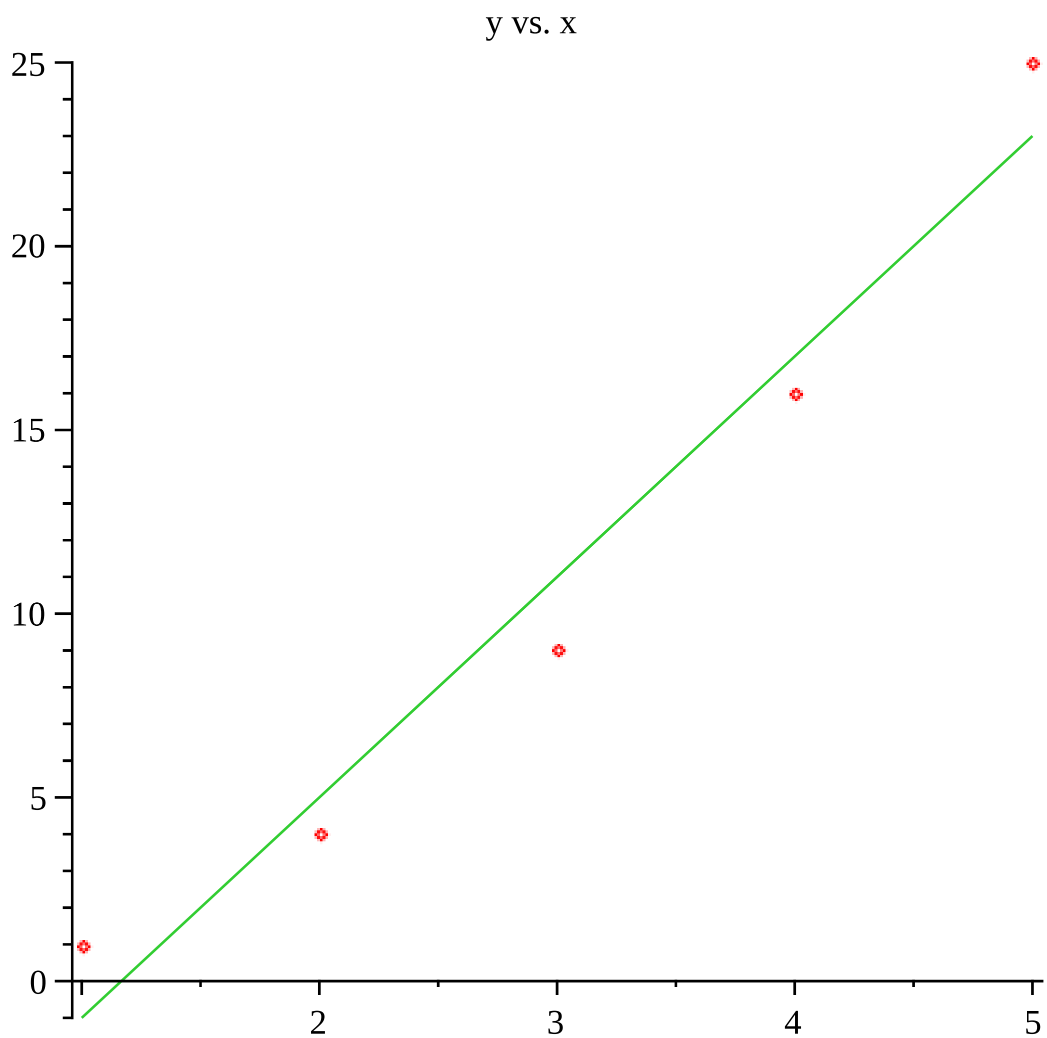
$$y := -7. + 6.\,x \quad (3.7)$$

The following plot demonstrates the data points as well as the least squares regression line:

```
> data:=[seq([x1[i],y1[i]],i=1..n)];
```

$$data := [[1, 1], [2, 4], [3, 9], [4, 16], [5, 25]] \quad (3.8)$$

```
> with(plots):
  plot([data,[seq([x1[i],a0+a1*x1[i]],i=1..n)]],style=[point,
  line],title="y vs. x");
```



```
>
```

Section 3: Coefficient of determination

One of the major indicators of how well least squares characterizes or predicts the whole data is a quantity called the coefficient of determination, r^2

$$r^2 = \frac{St-Sr}{St} \tag{4.1}$$

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 the two parameters (St-Sr) 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 value of $r^2 = 1$ is obtained, while a minimum value of $r^2 = 0$ is obtained when there is only one data point that falls on the line or the straight line model is a constant line. (Note that $0 < r^2 < 1$)

Calculation of the coefficient of determination:

```
> Sr:=0:
  St:=0:
  for i from 1 by 1 to n do
       Sr:=Sr+(y1[i]-a0-a1*x1[i])^2:
       St:=St+(y1[i]-yave)^2:
  end do:
  r2:=(St-Sr)/St:
> St;
```

$$374 \tag{4.1}$$

```
> Sr;
```

$$14. \tag{4.2}$$

```
> r2;
```

$$0.9625668449 \tag{4.3}$$

References

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

Conclusion

Using Maple, we were able to use the Linear Regression method to regress a given data set to a straight line. The accuracy of the regression model was then determined by calculating the coefficient of determination.

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

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

Temperature (*F*)	Instantaneous Expansion E-06 (*in/inF*)
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 unidirectional composite material.

Strain (%)	Stress (*MPa*)
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
1.56	2896

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