Curve Fitting

Least Square Approximations

Applications of numerical techniques in science and engineering often involve curve fitting of experimental data. When data are derived from an experiment, there is some error in the measurement.

It is often the case that an experiment produces a set of data points

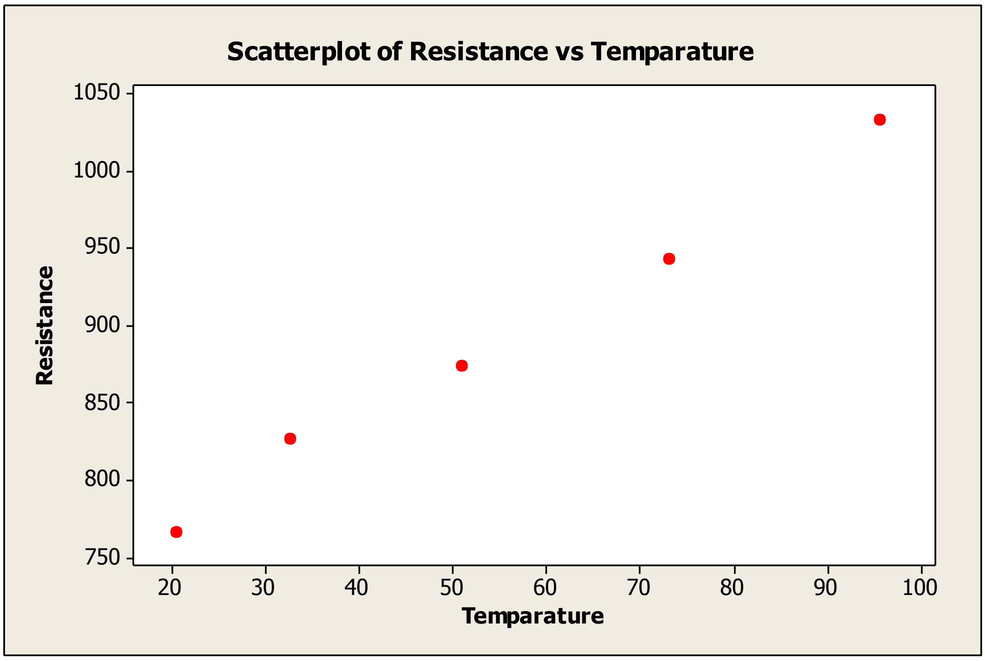
$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

- **One goal of numerical methods is to determine a formula y=f(x) that relates these variables.**
- **There are many type of function that can be used.**
- **Often there is underlying mathematical model, based on the physical situation, which will determine the form of the function.**

Example:

Some students are assigned to find the effect of temperature on the resistance of a metal wire. They have recorded the temperature and resistance values in a table and have plotted their findings.

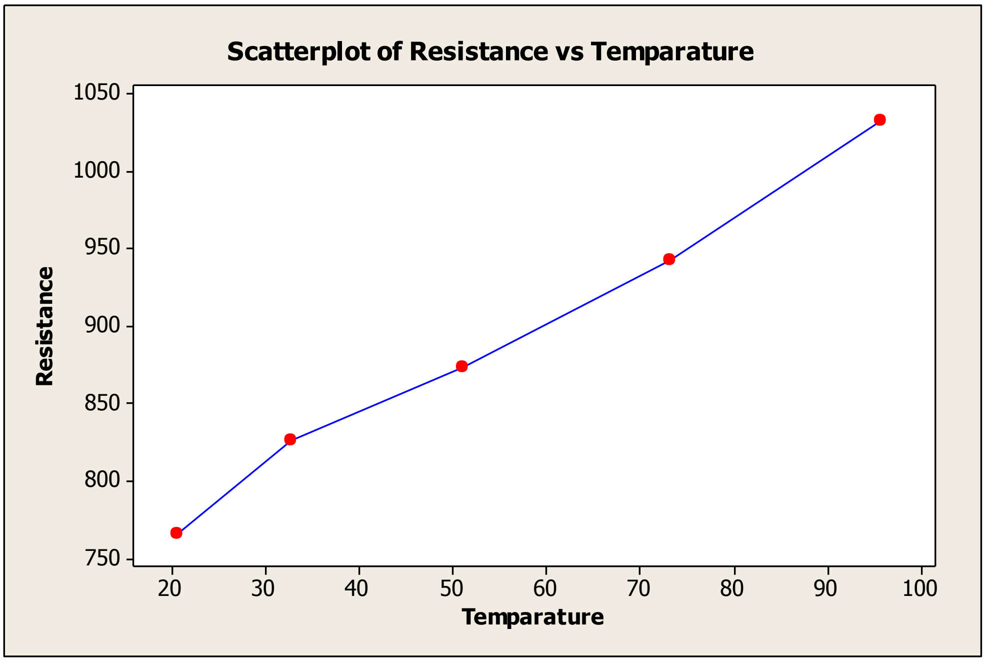
Temperature T C^0	Resistance R Ohms
20.5	765
32.7	826
51.0	873
73.2	942
95.7	1032



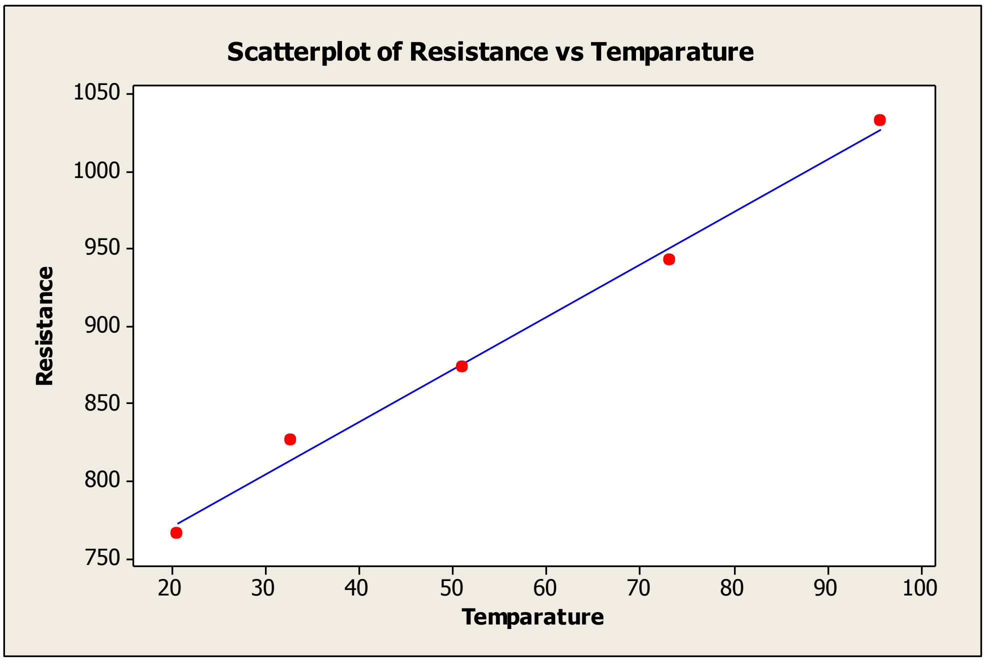
The graph suggests a linear relationship. If so, then

$$\mathbf{R = a + b\,T.}$$

Where a and b are parameters.



With Connected lines (Piecewise Linear Spline)



With Regression Line

Simple Linear Regression Model

A linear model that relates two variables, *x* and y. It can be written as:

$$y = \beta_0 + \beta_1 x + e$$

- ***y* and *x* may be referred in one of the following ways:**

x	y
Independent Variable **Explanatory Variable** **Control Variable** **Predictor Variable** **Regressor**	**Dependent Variable** **Explained Variable** **Response Variable** **Predicted Variable** **Regressand**

$$y = \beta_0 + \beta_1 x + e$$

- ***e* is referred to as the *error term* or *disturbance*. (Represents factors other than x that affect y. It is treated as unobservable.)**
- **β_0 is the *intercept.***
  (Gives the value of *y* when *x* = 0 and *u* = 0.)
- **β_1 is the *slope.***
  (Relates a change *in y* to a change in *x*)

Definition (Fitted Value):

The value of y when $x=x_i$ using estimates of the regression coefficients

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Definition (Residual):

The difference between the actual and fitted values of y_i

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Given a sample of data $\{(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ **, we want to find to estimates of the intercept and slope such that minimize the total amount of residuals.** ***Negative and positive residuals will cancel each other out, so look at the square of the residuals.***

$$Sum = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

The least square criterion requires that *Sum* be a <u>minimum.</u>
The first order conditions are:

$$\frac{\partial Sum}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0$$

$$\frac{\partial Sum}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0$$

Normal Equations

Dividing each of these equations by -2 and expanding the summation, we get the so called normal equations

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_0 n = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

Solving these equations simultaneously gives the estimations of intercept and slope.

From the first normal equation we obtain

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

(Estimator of β_0)

From the second normal equation we obtain

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

(Estimator of β_1)

Example:

Temperature T C^0	Resistance R Ohms
20.5	765
32.7	826
51.0	873
73.2	942
95.7	1032

i	x_i	y_i	x_i^2	$x_i y_i$
1	20.5	765	420.25	15682.5
2	32.7	826	1069.29	27010.2
3	51.0	873	2601.00	44523.0
4	73.2	942	5358.24	68954.4
5	95.7	1032	9158.49	98762.4
Total	273.1	4438	18607.3	254933

Normal equations

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_0 n = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

$$273.1\hat{\beta}_1 \ + 5\hat{\beta}_0 = 4438$$

$$18607.3\hat{\beta}_1 + 273.1\hat{\beta}_0 = 254933$$

From these we find $\hat{\beta}_1$ = 3.395, $\hat{\beta}_0$ = 702.2, and hence we write estimated linear equation as

$$\hat{y} = 702.2 + 3.395x$$

(Estimated Regression Line)

MATLAB "polyfit" Command

MATLAB gets a least-squares polynomial with its "polyfit" command, the same one that fits an interpolating polynomial to the data defined in vectors x and. y.

```
>> x=[20.5 32.7 51.0 73.2 95.7]

x =   20.5000   32.7000   51.0000   73.2000   95.7000

>> y=[765 826 873 942 1032]

y =        765        826        873        942        1032

>> equation=polyfit(x,y,1)

equation =

   3.3949  702.1721
```

$$\hat{y} = 702.2 + 3.395x$$

MINITAB "regress" Command

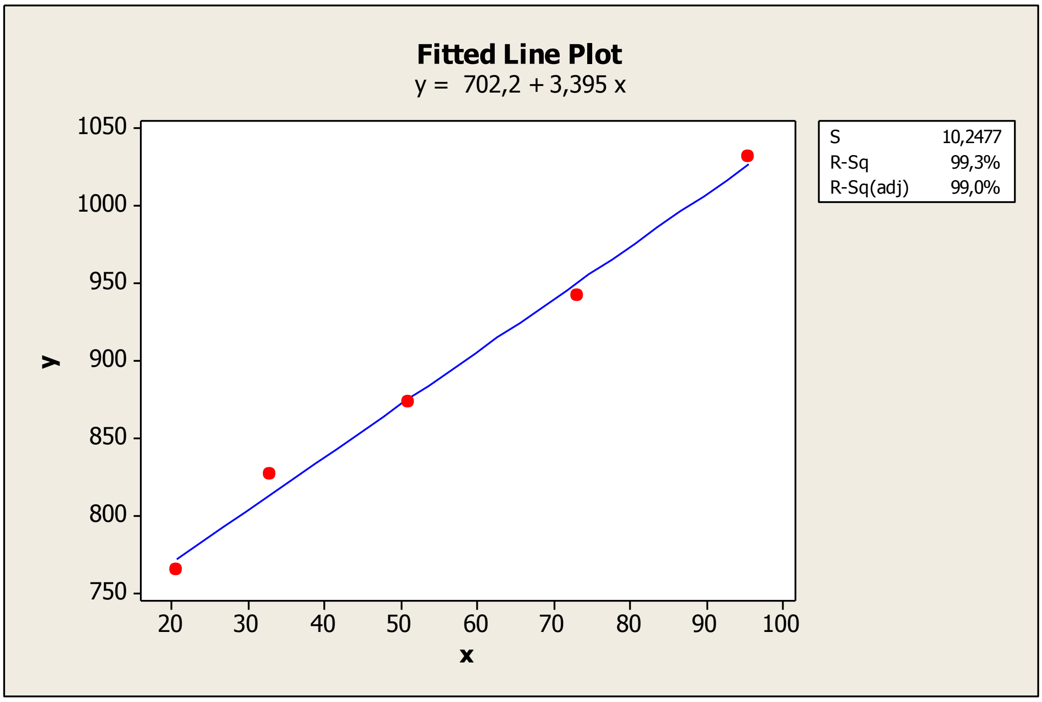
```
MTB > regress c2 1 c1
```

Regression Analysis: y versus x

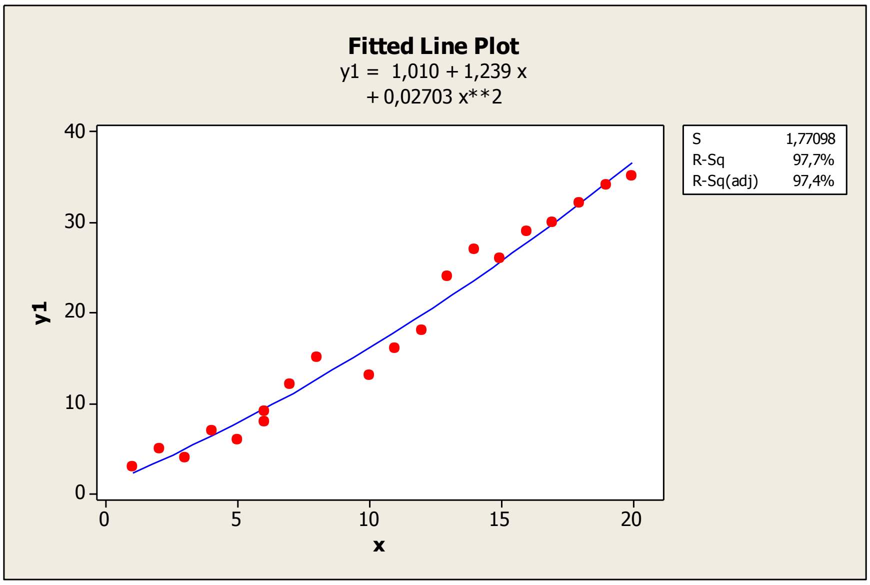
```
The regression equation is
y = 702 + 3.39 x
```

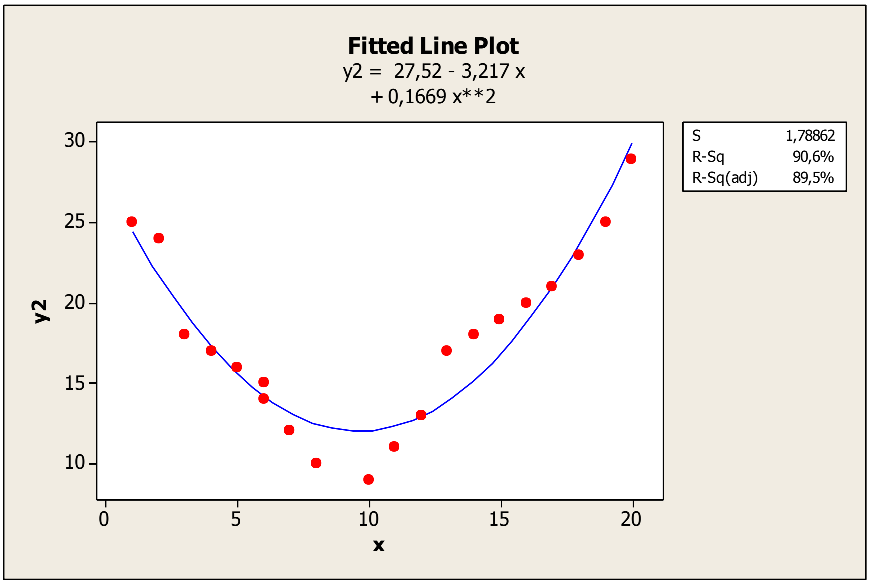
$$\hat{y} = 702.2 + 3.395x$$

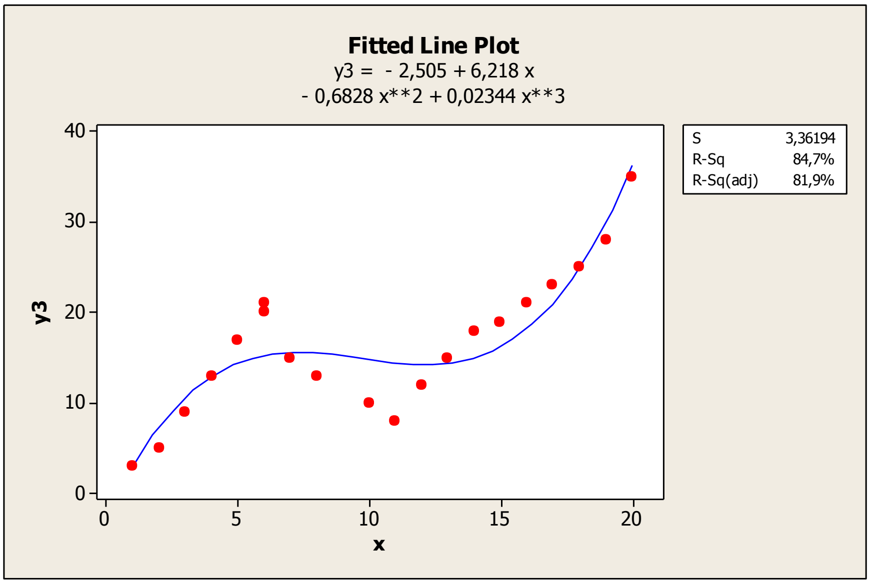
```
Predictor      Coef   SE Coef      T      P
Constant     702,17     10,29  68,23  0,000
x            3,3949    0,1687  20,13  0,000

S = 10,2477   R-Sq = 99,3%   R-Sq(adj) = 99,0%

Analysis of Variance

Source          DF     SS     MS       F      P
Regression       1  42534  42534  405,03  0,000
Residual Error   3    315    105
Total            4  42849
```



Least Square Polynomials


Fitted Line Plot
y1 = 1,010 + 1,239 x
+ 0,02703 x**2
S 1,77098
R-Sq 97,7%
R-Sq(adj) 97,4%
y1
x


Fitted Line Plot
y2 = 27,52 - 3,217 x
+ 0,1669 x**2
S 1,78862
R-Sq 90,6%
R-Sq(adj) 89,5%
y2
x



Polynomial regression model (kth degree)

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + e$$

Because polynomials can be readily manipulated, fitting such functions to data that do not plot linearly is common.

Obviously, if n=k+1, (n: number of observation, k: degrees of polynomial) the polynomial passes exactly through each point and the methods discussed earlier, so we will always have n> k+1.

At the minimum, all of the partial derivatives vanish. Writing the partial derivative equations and rearranging them gives k+1 normal equation to be solved simultaneously.

$$Sum = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \ldots \hat{\beta}_k x_i^k)^2$$

Normal Equations for Polynomials:

$$\hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_i^k \quad = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_i^3 + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_i^{k+1} \quad = \sum_{i=1}^{n} x_i y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_1 \sum_{i=1}^{n} x_i^3 + \hat{\beta}_2 \sum_{i=1}^{n} x_i^4 + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_i^{k+2} \quad = \sum_{i=1}^{n} x_i^2 y_i$$

$$\vdots$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i^k + \hat{\beta}_1 \sum_{i=1}^{n} x_i^{k+1} + \hat{\beta}_2 \sum_{i=1}^{n} x_i^{k+2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_i^{2k} \quad = \sum_{i=1}^{n} x_i^k y_i$$

Putting these equations in matrix form shows an interesting pattern in the coefficient matrix.

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 & \dots & \sum x_i^k \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \dots & \sum x_i^{k+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \dots & \sum x_i^{k+2} \\ & & & & \\ \sum x_i^k & \sum x_i^{k+1} & \sum x_i^{k+2} & \dots & \sum x_i^{2k} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \\ \\ \sum x_i^k y_i \end{bmatrix}$$

The above coefficient matrix is called normal matrix for the least square problem.

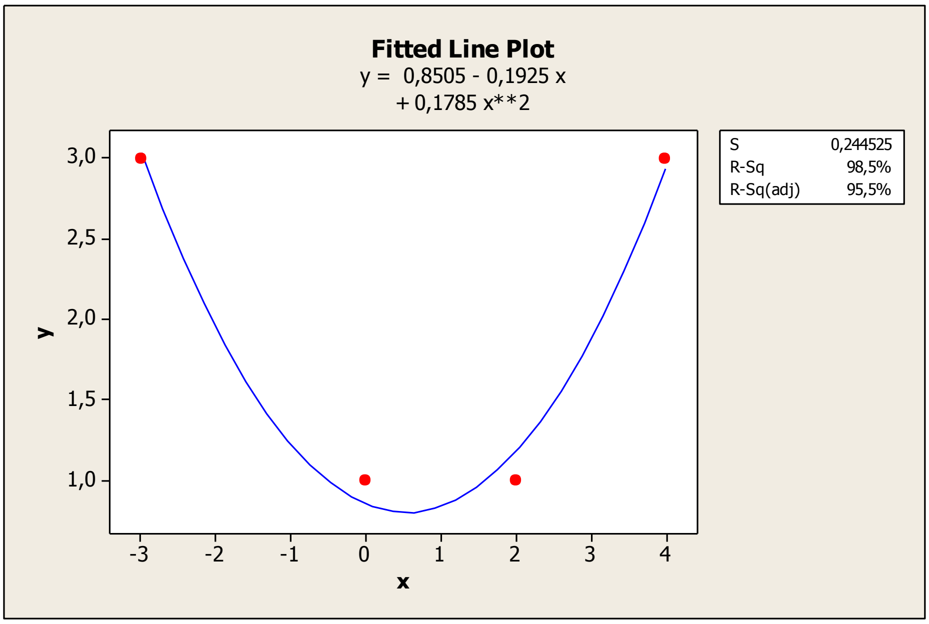
2nd degree-Polynomial Regression Model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 e$$

$$Sum = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2)^2$$

Example: Find the least squares parabola for the following data

x	y
-3	3
0	1
2	1
4	3



$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix}$$

n	x	y	x^2	x^3	x^4	xy	x^2y
1	-3	3	9	-27	81	-9	27
2	0	1	0	0	0	0	0
3	2	1	4	8	16	2	4
4	4	3	16	64	256	12	48
Sum	3	8	29	45	353	5	79

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix}$$

$$\begin{bmatrix} 4 & 3 & 29 \\ 3 & 29 & 45 \\ 29 & 45 & 353 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 5 \\ 79 \end{bmatrix}$$

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 4 & 3 & 29 \\ 3 & 29 & 45 \\ 29 & 45 & 353 \end{bmatrix}^{-1} \begin{bmatrix} 8 \\ 5 \\ 79 \end{bmatrix}$$

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 0.626297 & 0.0187614 & -0.0538438 \\ 0.018761 & 0.0435479 & -0.0070927 \\ -0.053844 & -0.0070927 & 0.0081605 \end{bmatrix} \begin{bmatrix} 8 \\ 5 \\ 79 \end{bmatrix}$$

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 0.850519 \\ -0.192495 \\ 0.178462 \end{bmatrix}$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

$$\hat{y} = 0.851 - 0.192x + 0.178x^2$$

MATLAB SOLUTION

Example: Find the least squares parabola for the following data using MATLAB M file

x	y
-3	3
0	1
2	1
4	3

```
>> X=[-3 0 2 4]
X =
   -3     0     2     4
>> Y=[3 1 1 3]
Y =
    3     1     1     3
>> lspoly(X,Y,2)
ans =
   0.1785
  -0.1925
   0.8505
```

$$y = 0,851 - 0,192\, x + 0,178\, x^2$$

SOLUTION IN MINITAB

y	x	x^2
3	-3	9
1	0	0
1	2	4
3	4	16

```
MTB > regress y 2 x x²
```

Regression Analysis: y versus x; x2

```
The regression equation is
y = 0,851 - 0,192 x + 0,178 x²
```
same result

$$\hat{y} = 0.851 - 0.192x + 0.178x^2$$

```
Predictor       Coef   SE Coef      T      P
Constant      0,8505    0,1935   4,40  0,142
x           -0,19250   0,05103  -3,77  0,165
x2           0,17846   0,02209   8,08  0,078

S = 0,244525   R-Sq = 98,5%   R-Sq(adj) = 95,5%

Analysis of Variance

Source           DF       SS       MS       F
P
Regression        2   3,9402   1,9701   32,95
0,122
Residual Error    1   0,0598   0,0598
Total             3   4,0000

Source   DF   Seq SS
x         1   0,0374
x2        1   3,9028
```

MATLAB M FILE (Least Square Polynomial)

```
function C = lspoly(X,Y,M)
%Input    - X is the 1xn abscissa vector
%         - Y is the 1xn ordinate vector
%         - M is the degree of the least-squares polynomial
% Output - C is the coefficient list for the polynomial
n=length(X);
B=zeros(1:M+1);
F=zeros(n,M+1);
%Fill the columns of F with the powers of X
for k=1:M+1
  F(:,k)=X'.^(k-1);
end
%Solve the linear system from (25)
A=F'*F;
B=F'*Y';
C=A\B;
C=flipud(C);
```

```
>> X=[-3 0 2 4]
X = -3     0     2     4
>> Y=[3 1 1 3]
Y =     3     1     1     3
>> lspoly(X,Y,2)
ans =
    0.1785
   -0.1925
    0.8505
```

$$\hat{y} = 0.851 - 0.192x + 0.178x^2$$