Curve fitting: Regression

Prakash Kotecha
Debasis Maharana & Remya Kommadath
Department of Chemical Engineering
Indian Institute of Technology Guwahati

Outline

- Regression
- Linear Regression
  - Simple linear regression
  - Multiple linear regression
  - Polynomial regression
  - General linear least squares
- Non-linear Regression
- Transformations for Data Linearization
- MATLAB functions: regress, nlinfit, polyfit, polyval

Regression

- Fits a selected function to the general trend of data.
- An underlying mathematical model is selected, based on physical situation.
- Coefficients of the model are determined such that the error between model values and the given data is minimum.
- Applied when substantial error is associated with the data

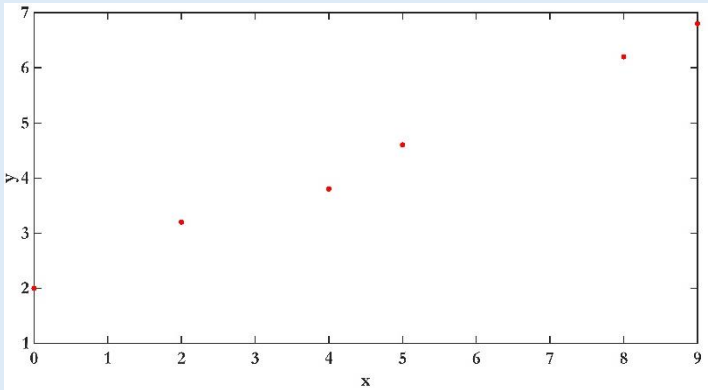
Let the approximating function be $y = f(x) = a_0x + a_1$

Value of the kth point can be obtained from $f(x_k) = y_k + e_k$

Error/deviation/residuals can be determined as $e_k = f(x_k) - y_k$

Average error: $E_a = \dfrac{\sum_{k=1}^{n} |f(x_k) - y_k|}{n}$

x	0	2	4	5	8	9
y	2	3.2	3.8	4.6	6.2	6.8



Regression

- Fits a selected function to the general trend of data.
- An underlying mathematical model is selected, based on physical situation.
- Coefficients of the model are determined such that the error between model values and the given data is minimum.
- Applied when substantial error is associated with the data

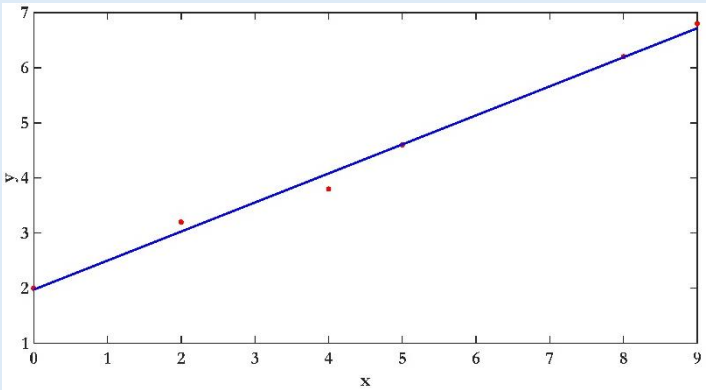
Let the approximating function be $y = f(x) = a_0x + a_1$

Value of the kth point can be obtained from $f(x_k) = y_k + e_k$

Error/deviation/residuals can be determined as $e_k = f(x_k) - y_k$

x	0	2	4	5	8	9
y	2	3.2	3.8	4.6	6.2	6.8

Average error: $E_a = \dfrac{\sum_{k=1}^{n} |f(x_k) - y_k|}{n}$



Types of regression analysis

- Linear regression: linear model is used to fit the data
  - Simple: linear model with one independent variable. $y = a_0 + a_1 x$
  - Multiple: linear model with two or more independent variables.

    Let number of independent variables be two, then $y = a_0 + a_1 x_1 + a_2 x_2$

    For m independent variables: $y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_m x_m$
  - Polynomial: model to fit the data is a higher order polynomial

    Let the order of the polynomial be two, then $y = a_0 + a_1 x + a_2 x^2$

    For an m^{th} order polynomial: $y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m$

Types of regression analysis

- Nonlinear regression: non-linear model needs to be fit to the data

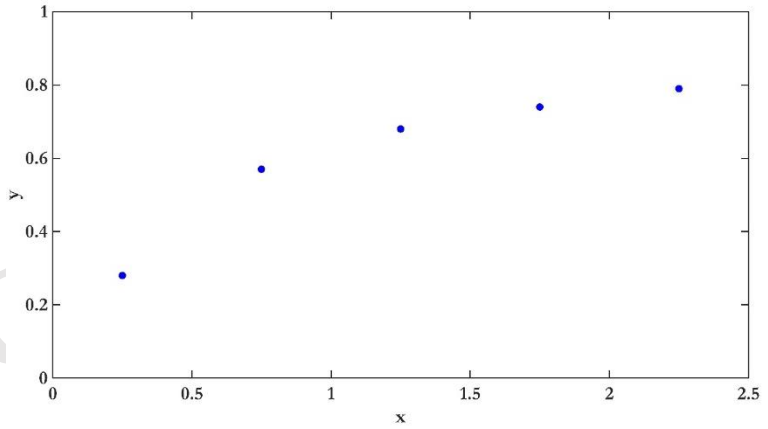
$$y = ae^{bx}$$

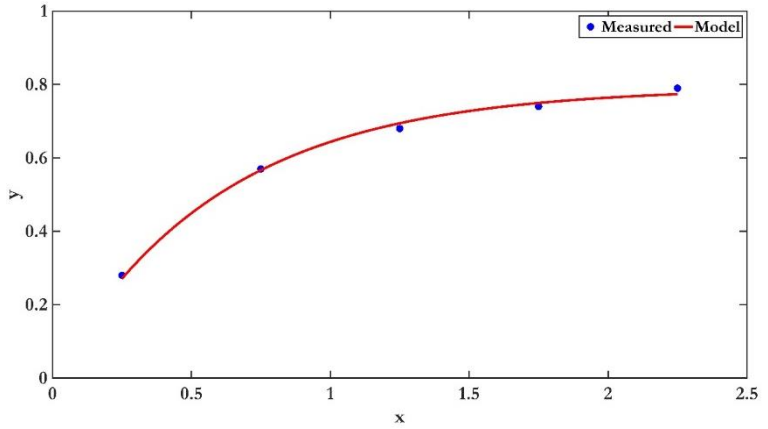
$$y = a\frac{x}{b+x}$$

$$y = (ax+b)^{-2}$$

$$y = a_0\left(1-e^{-a_1 x}\right)$$

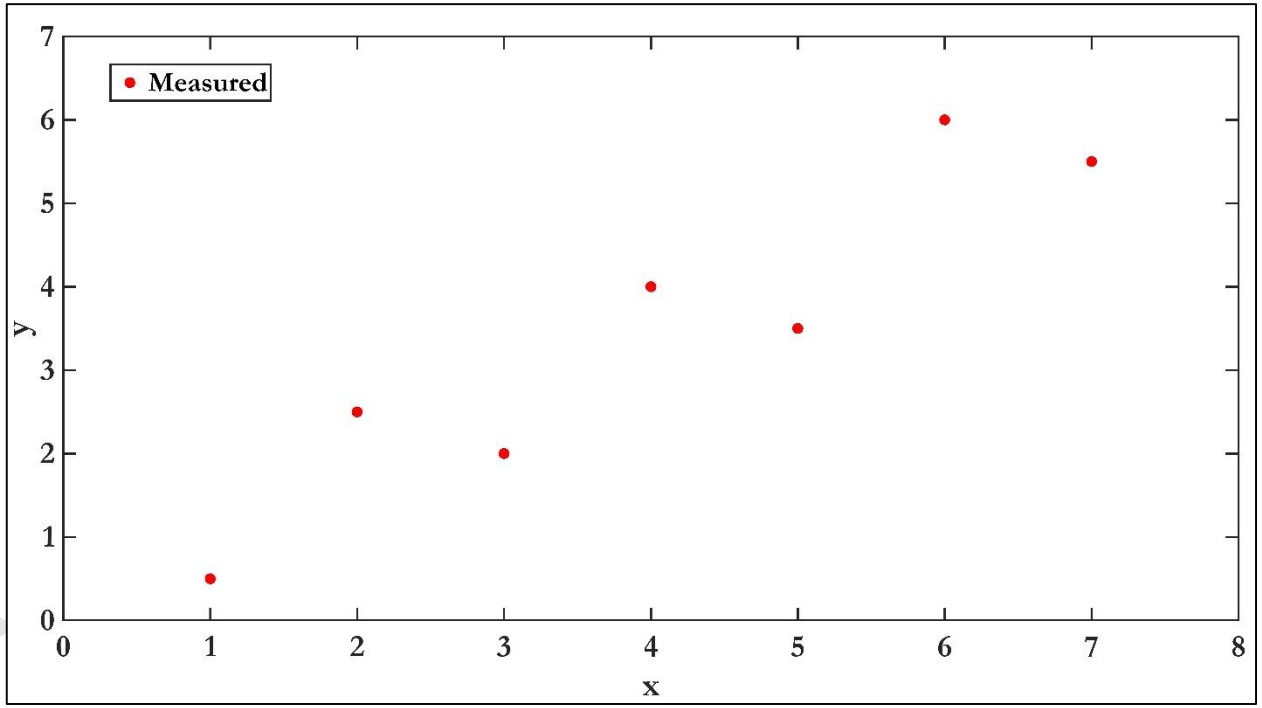
$$y = a_0 x_1^{a_1} x_2^{a_2}$$





Quantification of error in regression

x	y (measured)	y (model)
1	0.5	0.91
2	2.5	1.75
3	2	2.59
4	4	3.43
5	3.5	4.27
6	6	5.11
7	5.5	5.95



$$error(e) = (y_{measured} - y_{model})$$

$$y_{i,model} = a_0 + a_1 x_i + e$$

$$e_i = (y_i - a_0 - a_1 x_i)$$

Quantification of error in regression

x	y (measured)	y (model)
1	0.5	0.91
2	2.5	1.75
3	2	2.59
4	4	3.43
5	3.5	4.27
6	6	5.11
7	5.5	5.95

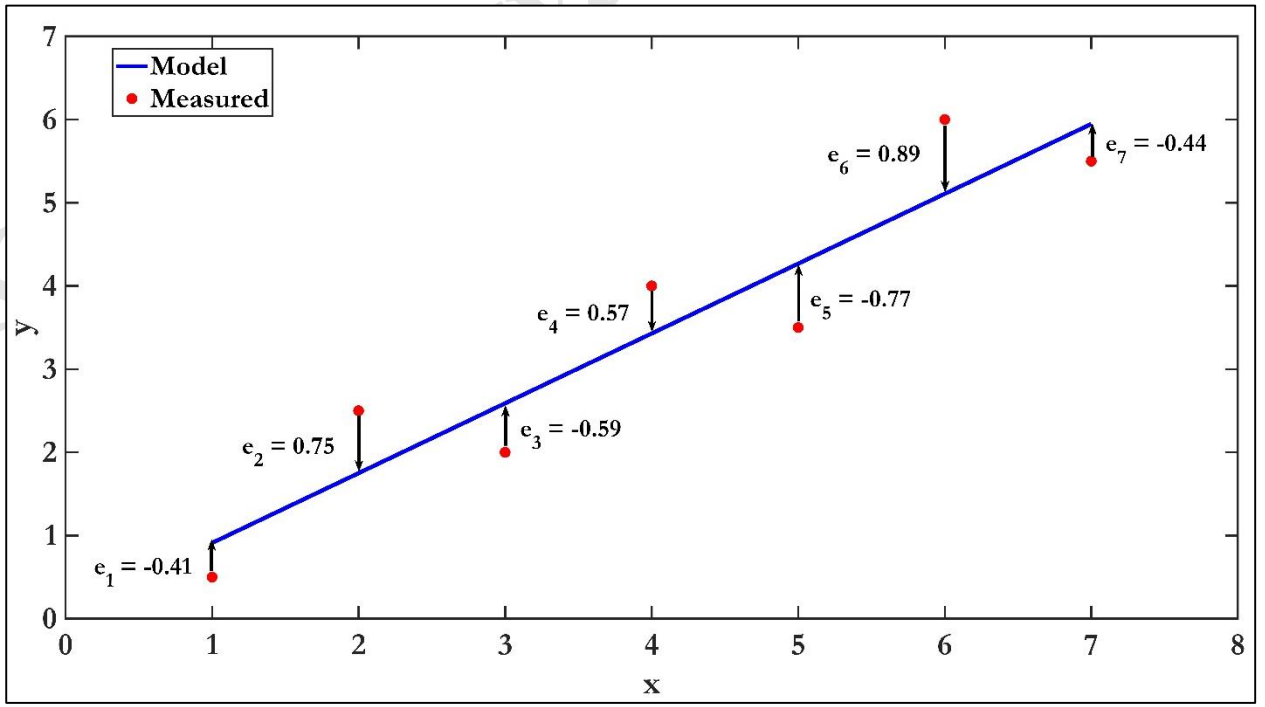
Regression model: $y = a_0 + a_1 x + e$

$a_0 = 0.07, a_1 = 0.84$

$$error(e) = (y_{measured} - y_{model})$$

$$y_{i,model} = a_0 + a_1 x_i + e$$

$$e_i = (y_i - a_0 - a_1 x_i)$$



Linear regression ($y = a_0 + a_1 x$)

- Sum of squares of residuals for n data points

$$Min\ S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_{i,measured} - y_{i,model} \right)^2$$

$$= \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i \right)^2$$

- Differentiating S_r equation with respect to each unknown coefficients of polynomial

$$\frac{\partial S_r}{\partial a_o} = -2\sum \left(y_i - a_o - a_1 x_i \right) = 0$$

$$-\sum y_i + \sum a_o + \sum a_1 x_i = 0$$

$$\sum a_o + \sum a_1 x_i = \sum y_i$$

$$n a_0 + a_1 \sum x_i = \sum y_i$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum x_i \left(y_i - a_o - a_1 x_i \right) = 0$$

$$-\sum x_i y_i + \sum x_i a_o + \sum a_1 x_i^2 = 0$$

$$a_0 \sum x_i + a_1 \sum x_i^2 = \sum x_i y_i$$

$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i \right)^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

$\bar{x}, \bar{y}$: means of x, y

- Simultaneous linear equations:

$$n a_0 + a_1 \sum x_i = \sum y_i$$

$$a_0 \sum x_i + a_1 \sum x_i^2 = \sum x_i y_i$$

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$A \qquad x \qquad b$

Example: Linear regression ($y = a_0 + a_1 x$)

	x	y	x^2	xy
1	1	4	1	4
2	3	5	9	15
3	5	6	25	30
4	7	5	49	35
5	10	8	100	80
6	12	7	144	84
7	13	6	169	78
8	16	9	256	144
9	18	12	324	216
$\sum$				

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$\begin{bmatrix} 9 & 85 \\ 85 & 1077 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 62 \\ 686 \end{bmatrix}$$

$$a_0 = 3.43$$
$$a_1 = 0.37$$

Coefficient of determination (r^2)

- Quantifies the 'goodness' of a fit.
- S_t: Magnitude of the residual error associated with the dependent variable with respect to the mean.

$$S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad \bar{y} = \text{mean}(y)$$

- S_r: The sum of the squares of the residuals around the regression line.

$$S_r = \sum_{i=1}^{n} (y_i - y_{i,model})^2$$

- $(S_t - S_r)$: Quantifies the improvement or error reduction due to describing data in terms of a straight line rather than as an average value.
- Coefficient of determination:

$$r^2 = \left[\frac{S_t - S_r}{S_t} \right]$$

Coefficient of determination (r^2)

- **Case 1:** Perfect fit ($r^2 = 1$)
  - The line explains 100 percent of the variability of the data.

- **Case 2:** No improvement ($r^2 = 0$)
  - No reduction of error by describing the data in a straight line.

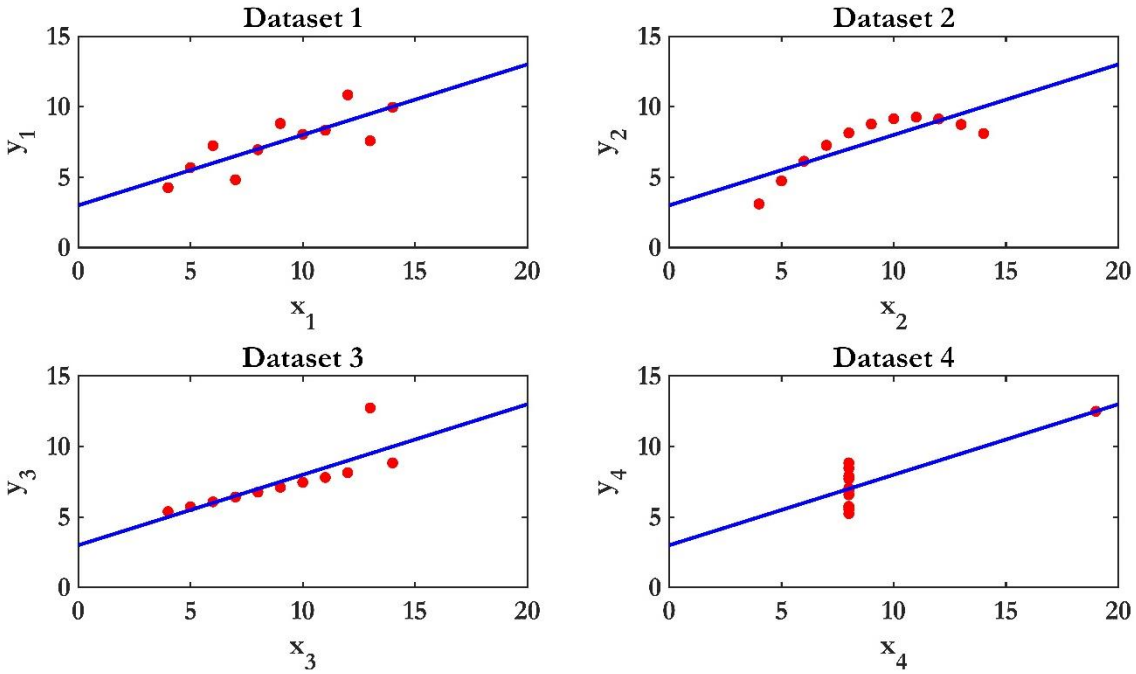
- r^2 close to 1 does not mean the fit is necessarily good.

Coefficient of determination (r^2)

- Anscombe's (1973) four data sets ($n = 11$)

Dataset I		Dataset II		Dataset III		Dataset IV	
x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

Property	Value
Mean of x	9
Mean of y	7.5
Linear regression line	y = 3.00 + 0.500x
Coefficient of determination of the linear regression	0.67

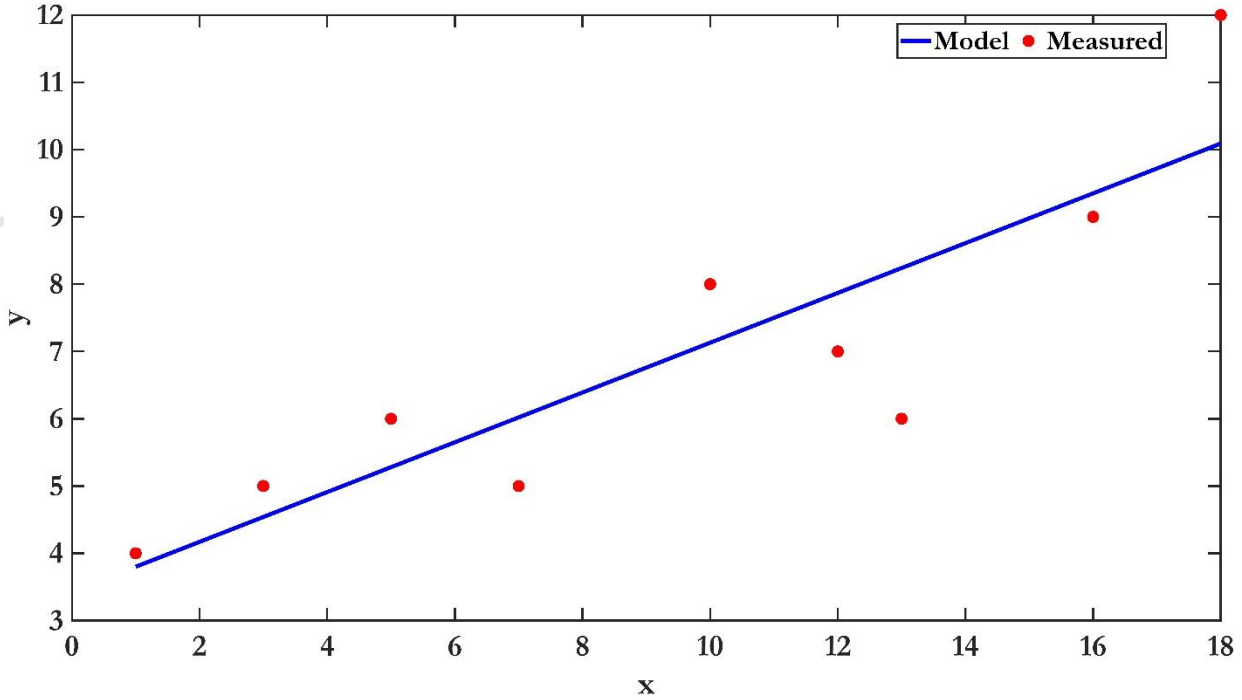


Linear regression: Coefficient of determination

x	y	y_{model}	$(y-y_{mean})^2$	$(y-y_{model})^2$
1	4	3.8	8.35	0.04
3	5	4.54	3.57	0.21
5	6	5.28	0.79	0.52
7	5	6.02	3.57	1.04
10	8	7.13	1.23	0.76
12	7	7.87	0.01	0.76
13	6	8.24	0.79	5.02
16	9	9.35	4.45	0.12
18	12	10.09	26.11	3.65
	$\bar{y}=6.89$		$S_t=48.87$	$S_r=12.12$

$$S_t=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 \qquad S_r=\sum_{i=1}^{n}\left(y_i-y_{i,model}\right)^2$$

$$r^2=\left[\frac{S_t-S_r}{S_t}\right] \qquad r^2=0.75$$



Multiple linear regression

- Extension of simple linear regression: y is a function of two or more independent variables.

x_1	x_2	y
0	0	5
2	1	10
2.5	2	9
1	3	0
4	6	3
9	2	27
8	4	15

Multiple linear regression model equation with x_1 and x_2 as independent variables

$$y = a_0 + a_1 x_1 + a_2 x_2 + e$$

- General equation of multiple linear regression model with ***m*** independent variables

$$y = a_0 + \sum_{i=1}^{m} a_i x_i + e \qquad y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m + e$$

- Sum of squares of the residuals for two independent variables and ***n*** data points

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2$$

Multiple linear regression

- Differentiate S_r with respect to each unknown coefficient of the polynomial

$$\frac{\partial S_r}{\partial a_0} = 0$$

$$-2\left[\sum\left(y_i - a_0 - a_1 x_{1i} - a_2 x_{2i}\right)\right] = 0$$

$$-\sum y_i + \sum a_0 + \sum a_1 x_{1i} + \sum a_2 x_{2i} = 0$$

$$n a_0 + a_1 \sum x_{1i} + a_2 \sum x_{2i} = \sum y_i$$

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2$$

Multiple linear regression

- Differentiate S_r with respect to each unknown coefficient of the polynomial

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2$$

$$a_0 n + a_1 \sum x_{1i} + a_2 \sum x_{2i} = \sum y_i$$

Multiple linear regression

- Differentiate S_r with respect to each unknown coefficient of the polynomial

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2$$

$$\frac{\partial S_r}{\partial a_1} = 0$$

$$-2\sum x_{1i}\left(y_i - a_0 - a_1 x_{1i} - a_2 x_{2i}\right) = 0$$

$$-\sum y_i x_{1i} + \sum a_0 x_{1i} + \sum a_1 x_{1i}^2 + \sum a_2 x_{1i} x_{2i} = 0$$

$$a_0 \sum x_{1i} + a_1 \sum x_{1i}^2 + a_2 \sum x_{1i} x_{2i} = \sum y_i x_{1i}$$

Multiple linear regression

- Differentiate S_r with respect to each unknown coefficient of the polynomial

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2$$

$$a_0 \sum x_{1i} + a_1 \sum x_{1i}^2 + a_2 \sum x_{1i} x_{2i} = \sum y_i x_{1i}$$

Multiple linear regression

- Differentiate S_r with respect to each unknown coefficient of the polynomial

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2$$

$$\frac{\partial S_r}{\partial a_2} = 0$$

$$-2\sum x_{2i}\left(y_i - a_0 - a_1 x_{1i} - a_2 x_{2i}\right) = 0$$

$$-\sum y_i x_{2i} + \sum a_0 x_{2i} + \sum a_1 x_{1i} x_{2i} + \sum a_2 x_{2i}^2 = 0$$

$$a_0 \sum x_{2i} + a_1 \sum x_{1i} x_{2i} + a_2 \sum x_{2i}^2 = \sum y_i x_{2i}$$

Multiple linear regression

- Differentiate S_r with respect to each unknown coefficient of the polynomial

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2$$

$$a_0 n + a_1 \sum x_{1i} + a_2 \sum x_{2i} = \sum y_i$$

$$a_0 \sum x_{1i} + a_1 \sum x_{1i}^2 + a_2 \sum x_{1i} x_{2i} = \sum y_i x_{1i}$$

$$a_0 \sum x_{2i} + a_1 \sum x_{1i} x_{2i} + a_2 \sum x_{2i}^2 = \sum y_i x_{2i}$$

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i} x_{2i} \\ \sum x_{2i} & \sum x_{2i} x_{1i} & \sum x_{2i}^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_{1i} \\ \sum y_i x_{2i} \end{bmatrix}$$

Multiple linear regression

For two independent variables and **n** data points

$$y_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + e$$

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2$$

$$a_0 n + a_1 \sum x_{1i} + a_2 \sum x_{2i} = \sum y_i$$
$$a_0 \sum x_{1i} + a_1 \sum x_{1i}^2 + a_2 \sum x_{1i} x_{2i} = \sum y_i x_{1i}$$
$$a_0 \sum x_{2i} + a_1 \sum x_{1i} x_{2i} + a_2 \sum x_{2i}^2 = \sum y_i x_{2i}$$

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i} x_{2i} \\ \sum x_{2i} & \sum x_{2i} x_{1i} & \sum x_{2i}^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_{1i} \\ \sum y_i x_{2i} \end{bmatrix}$$

For **m** independent variables and **n** data points

$$y_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + ... + a_m x_{mi} + e$$

$$S_r = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_{1i} - a_2 x_{2i} - \cdots - a_m x_{mi} \right)^2$$

$$a_0 n + a_1 \sum x_{1i} + a_2 \sum x_{2i} \cdots + a_m \sum x_{mi} = \sum y_i$$
$$a_0 \sum x_{1i} + a_1 \sum x_{1i}^2 + a_2 \sum x_{2i} x_{1i} \cdots + a_m \sum x_{mi} x_{1i} = \sum y_i x_{1i}$$
$$a_0 \sum x_{2i} + a_1 \sum x_{1i} x_{2i} + a_2 \sum x_{2i}^2 \cdots + a_m \sum x_{mi} x_{2i} = \sum y_i x_{2i}$$
$$\vdots$$
$$a_0 \sum x_{mi} + a_1 \sum x_{1i} x_{mi} + a_2 \sum x_{2i} x_{mi} \cdots + a_m \sum x_{mi}^2 = \sum y_i x_{mi}$$

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{mi} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{2i} x_{1i} & \cdots & \sum x_{mi} x_{1i} \\ \sum x_{2i} & \sum x_{1i} x_{2i} & \sum x_{2i}^2 & \cdots & \sum x_{mi} x_{2i} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum x_{mi} & \sum x_{1i} x_{mi} & \sum x_{2i} x_{mi} & \cdots & \sum x_{mi}^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{Bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_{1i} \\ \sum y_i x_{2i} \\ \vdots \\ \sum y_i x_{mi} \end{bmatrix}$$

Example: Multiple linear regression ($y = a_0 + a_1x_1 + a_2x_2$)

	x_1	x_2	y	x_1^2	x_1x_2	x_1y	x_2^2	x_2y
1	0	0	14	0	0	0	0	0
2	0	2	21	0	0	0	4	42
3	1	2	11	1	2	11	4	22
4	2	4	12	4	8	24	16	48
5	0	4	23	0	0	0	16	92
6	1	6	23	1	6	23	36	138
$\sum$								

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{2i}x_{1i} & \sum x_{2i}^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_{1i} \\ \sum y_i x_{2i} \end{bmatrix}$$

$$\begin{bmatrix} 6 & 4 & 18 \\ 4 & 6 & 16 \\ 18 & 16 & 76 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{bmatrix} 104 \\ 58 \\ 342 \end{bmatrix}$$

$$a_0 = 14.02$$
$$a_1 = -6.44$$
$$a_2 = 2.53$$

Multiple linear regression: Coefficient of determination

$y = a_0 + a_1 x_1 + a_2 x_2$

$a_0 = 14.02 \, , \, a_1 = -6.44 \, , \, a_2 = 2.53$

x_1	x_2	y	y_{model}	$(y - y_{mean})^2$	$(y - y_{model})^2$
0	0	14	14.02	11.09	0.00
0	2	21	19.08	13.47	3.69
1	2	11	12.64	40.07	2.69
2	4	12	11.26	28.41	0.55
0	4	23	24.14	32.15	1.3
1	6	23	22.76	32.15	0.06
	$\bar{y} = 17.33$			$S_t = 157.34$	$S_r = 8.29$

$$S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad S_r = \sum_{i=1}^{n} (y_i - y_{i,model})^2$$

$$r^2 = \left[\frac{S_t - S_r}{S_t} \right] \qquad r^2 = 0.95$$

Multiple linear regression without constant

With constant coefficient

$$y_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + ... + a_m x_{mi} + e$$

$$S_r = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_{1i} - a_2 x_{2i} - \cdots - a_m x_{mi} \right)^2$$

$$\begin{aligned}
&a_0 n + a_1 \sum x_{1i} + a_2 \sum x_{2i} \cdots + a_m \sum x_{mi} = \sum y_i \\
&a_0 \sum x_{1i} + a_1 \sum x_{1i}^2 + a_2 \sum x_{2i} x_{1i} \cdots + a_m \sum x_{mi} x_{1i} = \sum y_i x_{1i} \\
&a_0 \sum x_{2i} + a_1 \sum x_{1i} x_{2i} + a_2 \sum x_{2i}^2 \cdots + a_m \sum x_{mi} x_{2i} = \sum y_i x_{2i} \\
&\qquad\qquad \vdots \\
&a_0 \sum x_{mi} + a_1 \sum x_{1i} x_{mi} + a_2 \sum x_{2i} x_{mi} \cdots + a_m \sum x_{mi}^2 = \sum y_i x_{mi}
\end{aligned}$$

$$\begin{bmatrix}
n & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{mi} \\
\sum x_{1i} & \sum x_{1i}^2 & \sum x_{2i} x_{1i} & \cdots & \sum x_{mi} x_{1i} \\
\sum x_{2i} & \sum x_{1i} x_{2i} & \sum x_{2i}^2 & \cdots & \sum x_{mi} x_{2i} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\sum x_{mi} & \sum x_{1i} x_{mi} & \sum x_{2i} x_{mi} & \cdots & \sum x_{mi}^2
\end{bmatrix}
\begin{Bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{Bmatrix}
=
\begin{bmatrix} \sum y_i \\ \sum y_i x_{1i} \\ \sum y_i x_{2i} \\ \vdots \\ \sum y_i x_{mi} \end{bmatrix}$$

Without constant coefficient

$$y = a_1 x_{1i} + a_2 x_{2i} + ... + a_m x_{mi} + e$$

$$S_r = \sum_{i=1}^{n} \left(y_i - a_1 x_{1i} - a_2 x_{2i} - \cdots - a_m x_{mi} \right)^2$$

$$\begin{aligned}
&a_1 \sum x_{1i}^2 + a_2 \sum x_{2i} x_{1i} \cdots + a_m \sum x_{mi} x_{1i} = \sum y_i x_{1i} \\
&a_1 \sum x_{1i} x_{2i} + a_2 \sum x_{2i}^2 \cdots + a_m \sum x_{mi} x_{2i} = \sum y_i x_{2i} \\
&\qquad\qquad \vdots \\
&a_1 \sum x_{1i} x_{mi} + a_2 \sum x_{2i} x_{mi} \cdots + a_m \sum x_{mi}^2 = \sum y_i x_{mi}
\end{aligned}$$

$$\begin{bmatrix}
\sum x_{1i}^2 & \sum x_{2i} x_{1i} & \cdots & \sum x_{mi} x_{1i} \\
\sum x_{1i} x_{2i} & \sum x_{2i}^2 & \cdots & \sum x_{mi} x_{2i} \\
\vdots & \vdots & \vdots & \vdots \\
\sum x_{1i} x_{mi} & \sum x_{2i} x_{mi} & \cdots & \sum x_{mi}^2
\end{bmatrix}
\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}
=
\begin{bmatrix} \sum y_i x_{1i} \\ \sum y_i x_{2i} \\ \vdots \\ \sum y_i x_{mi} \end{bmatrix}$$

Polynomial regression

- For the cases where a curve is better suited for the data

- General equation for polynomial regression with ***m*** independent variables

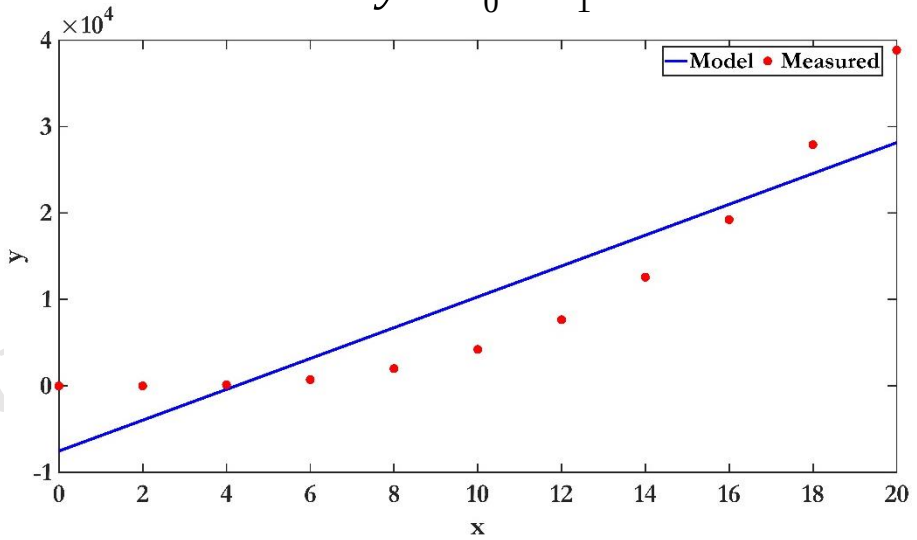
$$y = a_0 + \sum_{j=1}^{m} a_j x^j + e$$

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m + e$$

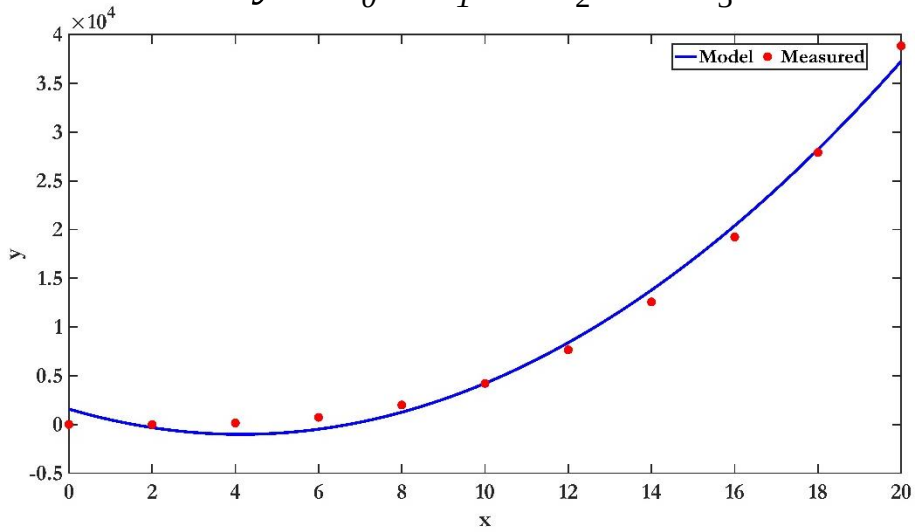
- Sum of the squares of the residuals with two independent variables and ***n*** data points

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$

$y = a_0 + a_1 x$



$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$



Polynomial regression

- Differentiating S_r equation with respect to each unknown coefficients of the polynomial

$$\frac{\partial S_r}{\partial a_o} = 0$$

$$-2\sum\left(y_i - a_o - a_1 x_i - a_2 x_{2i}\right) = 0$$

$$-\sum y_i + \sum a_0 + \sum a_1 x_i + \sum a_2 x_i^2 = 0$$

$$n a_0 + a_1 \sum x_i + a_2 \sum x_i^2 = \sum y_i$$

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$

Polynomial regression

- Differentiating S_r equation with respect to each unknown coefficients of the polynomial

$$a_0 n + a_1 \sum x_i + a_2 \sum x_i^2 = \sum y_i$$

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$

Polynomial regression

- Differentiating S_r equation with respect to each unknown coefficients of the polynomial

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$

$$\frac{\partial S_r}{\partial a_1} = 0$$

$$-2\sum x_i \left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right) = 0$$

$$-\sum y_i x_i + \sum a_0 x_i + \sum a_1 x_i^2 + \sum a_2 x_i^3 = 0$$

$$a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 = \sum y_i x_i$$

Polynomial regression

- Differentiating S_r equation with respect to each unknown coefficients of the polynomial

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$

$$a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 = \sum y_i x_i$$

Polynomial regression

- Differentiating S_r equation with respect to each unknown coefficients of the polynomial

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$

$$\frac{\partial S_r}{\partial a_2} = 0$$

$$-2\sum x_i^2 \left(y_i - a_o - a_1 x_i - a_2 x_i^2\right) = 0$$

$$-\sum y_i x_i^2 + \sum a_0 x_i^2 + \sum a_1 x_i^3 + \sum a_2 x_i^4 = 0$$

$$a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4 = \sum y_i x_i^2$$

Polynomial regression

- Differentiating S_r equation with respect to each unknown coefficients of the polynomial

$$a_0 n + a_1 \sum x_i + a_2 \sum x_i^2 = \sum y_i$$

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$

$$a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 = \sum y_i x_i$$

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_i \\ \sum y_i x_i^2 \end{bmatrix}$$

$$a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4 = \sum y_i x_i^2$$

Polynomial regression

For second order polynomial and ***n*** data points

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 + e$$

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$

$$a_0 n + a_1 \sum x_i + a_2 \sum x_i^2 = \sum y_i$$
$$a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 = \sum y_i x_i$$
$$a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4 = \sum y_i x_i^2$$

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_i \\ \sum y_i x_i^2 \end{bmatrix}$$

For ***mth*** order polynomial and ***n*** data points

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 + \ldots.. + a_m x_i^m + e$$

$$S_r = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_m x_i^m \right)^2$$

$$a_0 n + a_1 \sum x_i + a_2 \sum x_i^2 \cdots + a_m \sum x_i^m = \sum y_i$$
$$a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 \cdots + a_m \sum x_i^{m+1} = \sum y_i x_i$$
$$a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4 \cdots + a_m \sum x_i^{m+2} = \sum y_i x_i^2$$
$$\vdots$$
$$a_0 \sum x_i^m + a_1 \sum x_i^{m+1} + a_2 \sum x_i^{m+2} \cdots + a_m \sum x_i^{2m} = \sum y_i x_i^m$$

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^m \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{m+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \cdots & \sum x_i^{m+2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum x_i^m & \sum x_i^{m+1} & \sum x_i^{m+2} & \cdots & \sum x_i^{2m} \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{Bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_i \\ \sum y_i x_i^2 \\ \vdots \\ \sum y_i x_i^m \end{bmatrix}$$

Example: Polynomial regression ($y = a_0 + a_1x + a_2x^2$)

	x	y	x^2	x^3	xy	x^4	x^2y
1	0	2.1	0	0	0	0	0
2	1	7.7	1	1	7.7	1	7.7
3	2	13.6	4	8	27.2	16	54.4
4	3	27.2	9	27	81.6	81	244.8
5	4	40.9	16	64	163.6	256	654.4
6	5	61.1	25	125	305.5	625	1527.5
$\sum$	**15**	**152.6**	**55**	**225**	**585.6**	**979**	**2488.8**

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_i \\ \sum y_i x_i^2 \end{bmatrix}$$

$$\begin{bmatrix} 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{bmatrix} 152.6 \\ 585.6 \\ 2488.8 \end{bmatrix}$$

$$a_0 = 2.48$$
$$a_1 = 2.36$$
$$a_2 = 1.86$$

Polynomial regression: Coefficient of determination

$y_i = a_0 + a_1 x_i + a_2 x_i^2$

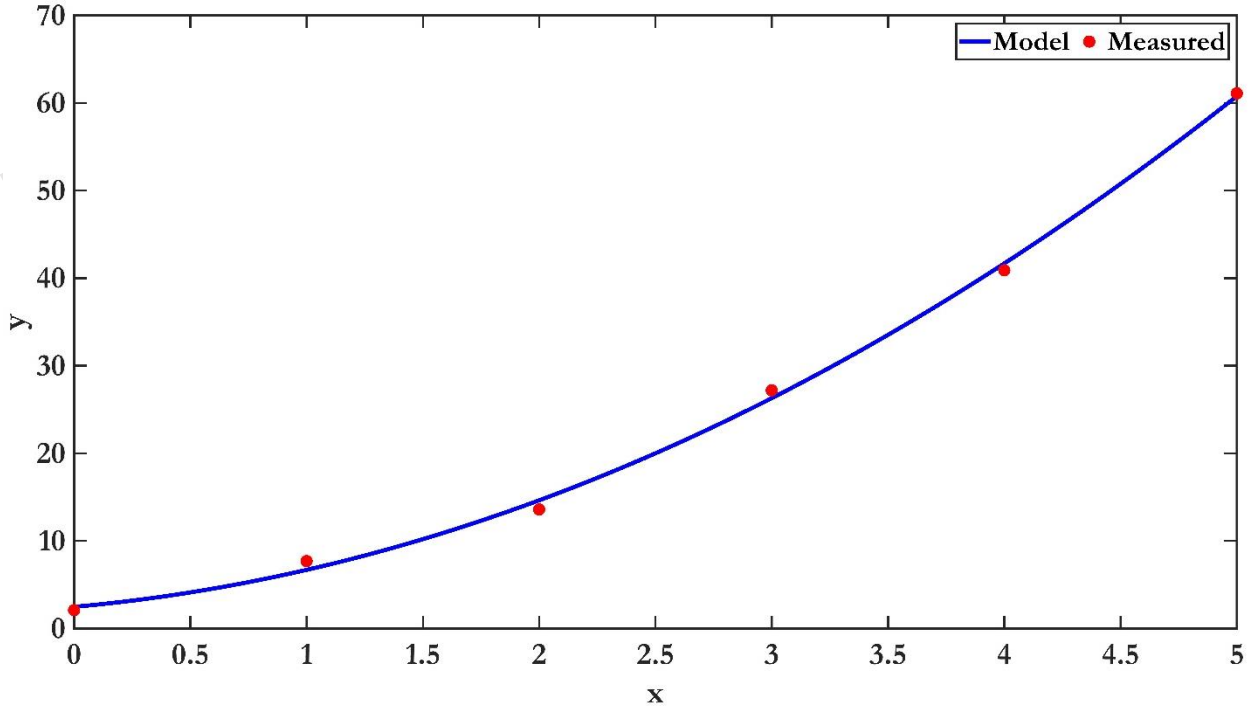
$a_0 = 2.48\,,\ a_1 = 2.36\,,\ a_2 = 1.86$

$$S_t = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \qquad S_r = \sum_{i=1}^{n} \left(y_i - y_{i,model}\right)^2$$

$$r^2 = \left[\frac{S_t - S_r}{S_t}\right] \qquad r^2 = 0.99$$

x	y	y_{model}	$(y - y_{mean})^2$	$(y - y_{model})^2$
0	2.1	2.48	544.29	0.14
1	7.7	6.7	314.35	1
2	13.6	14.64	139.95	1.08
3	27.2	26.30	3.13	0.81
4	40.9	41.68	239.32	0.61
5	61.1	60.78	1272.35	0.1

$\bar{y} = 25.43$ $\quad S_t = 2513.39$ $\quad S_r = 3.74$



Linear least square model

- General least square model

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + ... + a_m z_m + e$$

- Simple linear regression in least square form

$$If \;\; z_0 = 1,\; z_1 = x_1 \Rightarrow y = a_0 + a_1 x_1 + e$$

- Multiple linear regression in least square form

$$If \;\; z_0 = 1,\; z_1 = x_1,\; z_2 = x_2, \ldots, z_m = x_m \Rightarrow y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m + e$$

- Polynomial regression in least square form

$$If \;\; z_0 = 1,\; z_1 = x,\; z_2 = x^2, ..., z_m = x^m \Rightarrow y = a_0 + a_1 x + a_2 x^2 + ... + a_m x^m + e$$

- Trigonometric regression in least square form

$$If \;\; z_0 = 1,\; z_1 = \sin(\omega t),\; z_2 = \cos(\omega t), \Rightarrow y = a_0 + a_1 \sin(\omega t) + a_2 \cos(\omega t) + e$$

Linear least square: General matrix formulation

- General least square model $y = a_0 z_0 + a_1 z_1 + a_2 z_2 + ... + a_m z_m + e$

- For n data points and m variables

$$
\begin{aligned}
y_1 &= a_0 z_{01} + a_1 z_{11} + a_2 z_{21} + \ldots + a_m z_{m1} + e_1 \\
y_2 &= a_0 z_{02} + a_1 z_{12} + a_2 z_{22} + \ldots + a_m z_{m2} + e_2 \\
&\vdots \\
y_n &= a_0 z_{0n} + a_1 z_{1n} + a_2 z_{2n} + \ldots + a_m z_{mn} + e_n
\end{aligned}
$$

$$
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} z_{01} & z_{11} & \cdots & z_{m1} \\ z_{02} & z_{12} & \cdots & z_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ z_{0n} & z_{1n} & \cdots & z_{mn} \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{Bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}
$$

$$
\{Y\} = [Z]\{A\} + \{E\} \xrightarrow{\text{Reducing sum of square of errors}} \left[[Z]^T [Z]\right]\{a\} = \left\{[Z]^T \{Y\}\right\}
$$

Linear least square: Multi-linear example

x_1	x_2	y
1	2	5.8
1.2	2.6	12.86
2	4	21.4
3	4.2	22.2
3.2	5	23
5	6	31

Multiple linear regression $y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m + e$

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{2i}x_{1i} & \sum x_{2i}^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_{1i} \\ \sum y_i x_{2i} \end{bmatrix}$$

$$\begin{bmatrix} 6 & 15.4 & 23.8 \\ 15.4 & 50.68 & 71.72 \\ 23.8 & 71.72 & 105.4 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{bmatrix} 116.26 \\ 359.23 \\ 524.88 \end{bmatrix}$$

$$a_0 = -4.77,\ a_1 = -0.91,\ a_2 = 6.68$$

General least square model

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \ldots + a_m z_m + e$$

$$Z = \begin{bmatrix} z_{01} & z_{11} & \cdots & z_{m1} \\ z_{02} & z_{12} & \cdots & z_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ z_{0n} & z_{1n} & \cdots & z_{mn} \end{bmatrix}$$

$$\left[[Z]^T [Z]\right]\{a\} = \left\{[Z]^T \{Y\}\right\}$$

$$y = a_0 + a_1 x_1 + a_2 x_2 + e$$

$$z_0 = 1,\ z_1 = x_1,\ z_2 = x_2$$

$$Z = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix}$$

$$Z = \begin{array}{c} a_0 \quad x_1 \quad x_2 \\ \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1.2 & 2.6 \\ 1 & 2 & 4 \\ 1 & 3 & 4.2 \\ 1 & 3.2 & 5 \\ 1 & 5 & 6 \end{bmatrix} \end{array}$$

$$Z^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1.2 & 2 & 3 & 3.2 & 5 \\ 2 & 2.6 & 4 & 4.2 & 5 & 6 \end{bmatrix}$$

$$Z^T Z = \begin{bmatrix} 6 & 15.4 & 23.8 \\ 15.4 & 50.68 & 71.72 \\ 23.8 & 71.72 & 105.4 \end{bmatrix} \qquad Z^T Y = \begin{bmatrix} 116.26 \\ 359.23 \\ 524.88 \end{bmatrix}$$

$$a_0 = -4.77,\ a_1 = -0.91,\ a_2 = 6.68$$

Linear least square: Coefficient of determination

$y = a_0 + a_1 x_1 + a_2 x_2$ $\quad a_0 = -4.77,\ a_1 = -0.91,\ a_2 = 6.68$

x_1	x_2	y	y_{model}	$(y - y_{mean})^2$	$(y - y_{model})^2$
1	2	5.8	7.68	184.42	3.53
1.2	2.6	12.86	11.51	42.51	1.83
2	4	21.4	20.13	4.08	1.61
3	4.2	22.2	20.56	7.95	2.7
3.2	5	23	25.72	13.1	7.39
5	6	31	30.76	135.02	0.06
		$\bar{y} = 19.38$		$S_t = 387.08$	$S_r = 17.12$

$$S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad S_r = \sum_{i=1}^{n} (y_i - y_{i,model})^2$$

$$r^2 = \left[\frac{S_t - S_r}{S_t} \right] \qquad r^2 = 0.96$$

Linear least square: Multi-linear without constant

x_1	x_2	y
1	2	5.8
1.2	2.6	12.86
2	4	21.4
3	4.2	22.2
3.2	5	23
5	6	31

Multiple linear regression

$$y = a_1 x_1 + a_2 x_2 + ... + a_m x_m + e$$

$$\begin{bmatrix} \sum x_{1i}^2 & \sum x_{2i} x_{1i} \\ \sum x_{2i} x_{1i} & \sum x_{2i}^2 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix} = \begin{bmatrix} \sum y_i x_{1i} \\ \sum y_i x_{2i} \end{bmatrix}$$

$$\begin{bmatrix} 50.68 & 71.72 \\ 71.72 & 105.4 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix} = \begin{bmatrix} 359.23 \\ 524.88 \end{bmatrix}$$

$$a_1 = 1.11, \; a_2 = 4.23$$

General least square model

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \ldots + a_m z_m + e$$

$$Z = \begin{bmatrix} z_{01} & z_{11} & \cdots & z_{m1} \\ z_{02} & z_{12} & \cdots & z_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ z_{0n} & z_{1n} & \cdots & z_{mn} \end{bmatrix}$$

$$\left[[Z]^T [Z] \right] \{a\} = \left\{ [Z]^T \{Y\} \right\}$$

$$y = a_0 x_1 + a_1 x_2 + e$$

$$z_0 = x_1, \; z_1 = x_2$$

$$Z = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{bmatrix}$$

$$Z = \begin{matrix} x_1 & x_2 \end{matrix} \begin{bmatrix} 1 & 2 \\ 1.2 & 2.6 \\ 2 & 4 \\ 3 & 4.2 \\ 3.2 & 5 \\ 5 & 6 \end{bmatrix}$$

$$Z^T = \begin{bmatrix} 1 & 1.2 & 2 & 3 & 3.2 & 5 \\ 2 & 2.6 & 4 & 4.2 & 5 & 6 \end{bmatrix}$$

$$Z^T Z = \begin{bmatrix} 50.68 & 71.72 \\ 71.72 & 105.4 \end{bmatrix} \qquad Z^T Y = \begin{bmatrix} 359.23 \\ 524.88 \end{bmatrix}$$

$$a_0 = 1.11, \; a_1 = 4.23$$

Multiple linear regression: Coefficient of determination

$$y = a_0 x_1 + a_1 x_2$$

$$a_0 = 1.11,\ a_1 = 4.23$$

x_1	x_2	y	y_{model}	$(y - y_{mean})^2$	$(y - y_{model})^2$
1	2	5.8	9.57	184.42	14.21
1.2	2.6	12.86	12.33	42.51	0.28
2	4	21.4	19.14	4.08	5.11
3	4.2	22.2	21.10	7.95	1.22
3.2	5	23	24.70	13.10	2.90
5	6	31	30.93	135.02	0.00
		$\bar{y} = 19.38$		$S_t = 387.08$	$S_r = 23.72$

$$S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad S_r = \sum_{i=1}^{n} (y_i - y_{i,model})^2$$

$$r^2 = \left[\frac{S_t - S_r}{S_t} \right] \qquad r^2 = 0.94$$

Nonlinear least-squares

- Fitting an exponential curve: $y = Ce^{Ax}$

- Least-squares requires to minimize sum of squares of residuals for n data points

$$Min\ S_r = \sum_{i=1}^{n} \left(y_i - Ce^{Ax_i} \right)^2$$

$$\frac{\partial S_r}{\partial A} = -2\sum Cx_i e^{Ax_i}\left(y_i - Ce^{Ax_i} \right) = 0 \qquad \frac{\partial S_r}{\partial C} = -2\sum e^{Ax_i}\left(y_i - Ce^{Ax_i} \right) = 0$$

$$\sum x_i y_i e^{Ax_i} - C\sum x_i e^{2Ax_i} = 0 \quad (1) \qquad \sum y_i e^{Ax_i} - C\sum e^{2Ax_i} = 0 \quad (2)$$

- Two nonlinear equations and two unknowns (A and C)

Solve using an appropriate method to determine the constant coefficients.

$$\sum x_i y_i e^{Ax_i} - C\sum x_i e^{2Ax_i} = 0$$
$$\sum y_i e^{Ax_i} - C\sum e^{2Ax_i} = 0$$

$A = 0.38,\ C = 1.61$

x	y
0	1.5
1	2.5
2	3.5
3	5
4	7.5

Nonlinear least-squares

```
clc
clear

fun = @prob;

x0 = [1 0.5]
x = fsolve(fun,x0);
```

```
function f = prob(Z)
x = [0 1 2 3 4];
y = [1.5 2.5 3.5 5 7.5];

A = Z(1); C = Z(2);

f(1) = sum(x.*y.*exp(A*x)) - C*sum(x.*exp(2*A*x));
f(2) = sum(y.*exp(A*x)) - C*sum(exp(2*A*x));
```

```
x =
  0.38      1.61
```

x	y
0	1.5
1	2.5
2	3.5
3	5
4	7.5

$$\sum x_i y_i e^{Ax_i} - C\sum x_i e^{2Ax_i} = 0$$

$$\sum y_i e^{Ax_i} - C\sum e^{2Ax_i} = 0$$

$$A = 0.38,\ C = 1.61$$

Nonlinear least-squares: Data linearization

➢Technique to fit nonlinear curves using linear least-squares.

- Step 1: $y = Ce^{Ax}$ $\xrightarrow{\text{Linearize}}$ $\ln(y) = Ax + \ln(C)$
- Step 2: Introduce the change of variables

$$Y = \ln(y) \qquad X = x \qquad a_1 = A \qquad a_0 = \ln(C)$$

- Step 3: Apply linear least squares to the linear Model

$$Y = a_0 + a_1 X$$

General least-squares model

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \ldots.. + a_m z_m + e$$

$$If \;\; z_0 = 1,\; z_1 = x_1 \Rightarrow y = a_0 + a_1 x_1 + e$$

$$Z = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \qquad Z^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix}$$

$$\boxed{\left[[Z]^T[Z]\right]\{a\} = \left\{[Z]^T\{Y\}\right\}}$$

$$\begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \end{Bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum Y_i X_i \end{bmatrix}$$

Nonlinear least-squares: Data linearization

➢Technique to fit nonlinear curves using linear least-squares.

x	y	lny
0	1.5	0.41
1	2.5	0.92
2	3.5	1.25
3	5	1.61
4	7.5	2.01

- Step 1: $y = Ce^{Ax}$ —Linearize→ $\ln(y) = Ax + \ln(C)$
- Step 2: Introduce the change of variables

$$Y = \ln(y) \qquad X = x \qquad a_1 = A \qquad a_0 = \ln(C)$$

- Step 3: Apply linear least squares to the linear Model

$$Y = a_0 + a_1 X$$

$$Z = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \qquad Z^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix}$$

$$Z = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

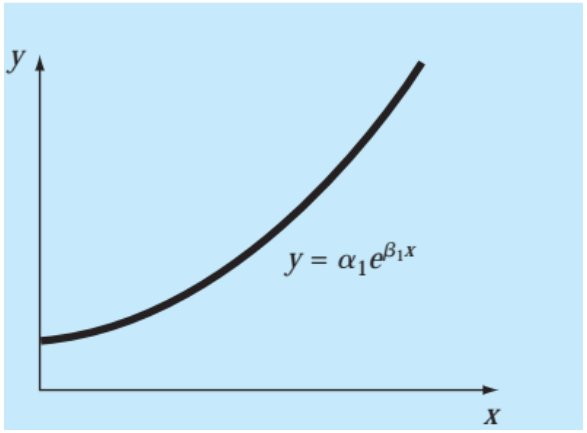
$$\begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{Bmatrix} \begin{Bmatrix} a_0 \\ a_1 \end{Bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum Y_i X_i \end{bmatrix}$$

$$\begin{bmatrix} 5 & 10 \\ 10 & 30 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \end{Bmatrix} = \begin{bmatrix} 6.2 \\ 16.29 \end{bmatrix} \qquad a_0 = 0.46,\ a_1 = 0.39$$

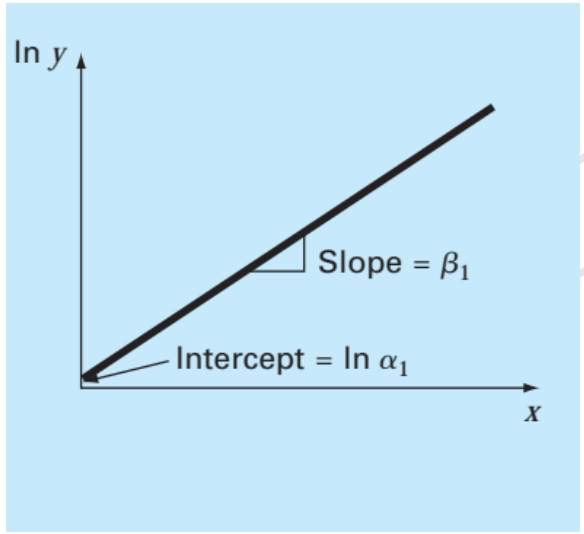
$$\left[[Z]^T [Z]\right]\{a\} = \left\{[Z]^T \{Y\}\right\}$$

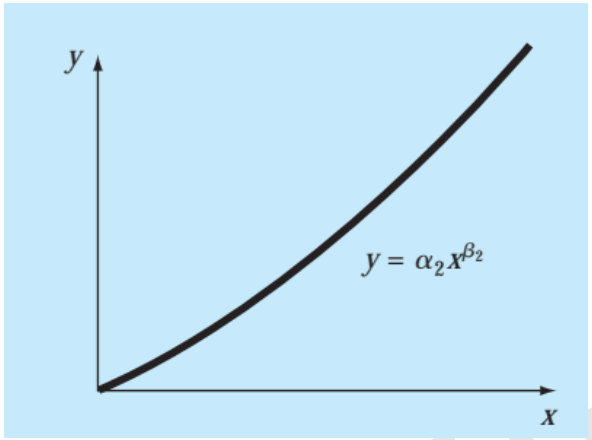
$$A = 0.39, C = 1.58$$

Data linearization

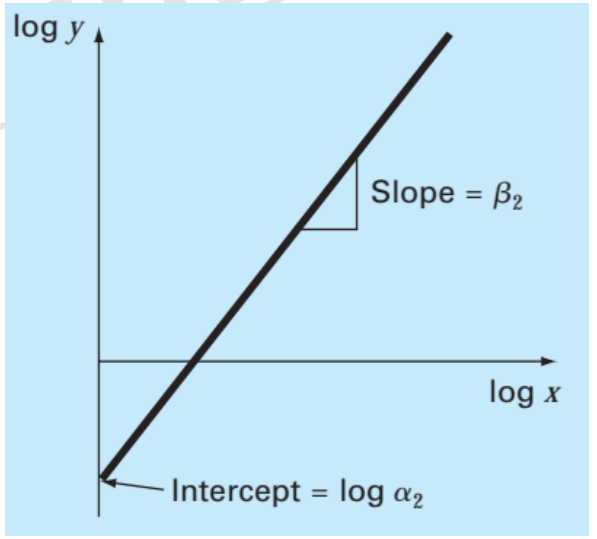


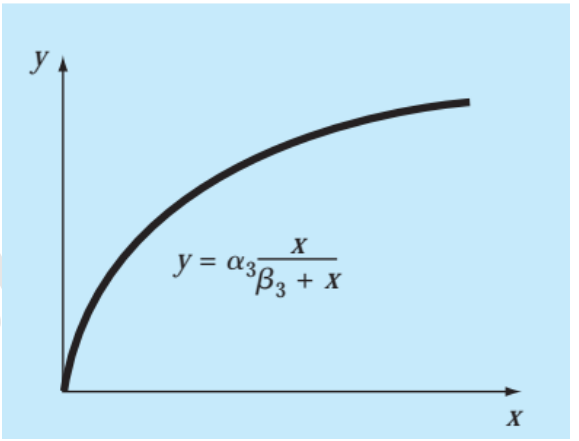
Linearization



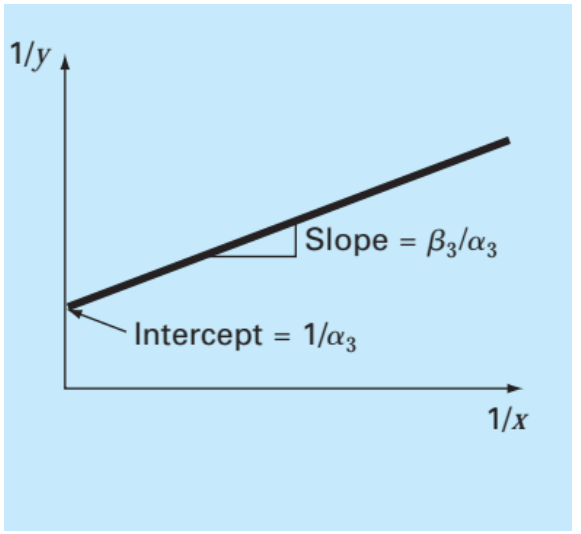


Linearization





Linearization



Data linearization

- Nonlinear models such as exponential model, power model, saturation growth model can be linearized and can be solved using linear regression.

Nonlinear model	Linearized model
$y = \frac{a}{x} + b$	$y = a\frac{1}{x} + b$
$y = \frac{x}{ax + b}$	$\frac{1}{y} = b\frac{1}{x} + a$
$y = a\frac{x}{b + x}$	$\frac{1}{y} = \frac{b}{a}\frac{1}{x} + \frac{1}{a}$
$y = (ax + b)^{-2}$	$y^{-1/2} = ax + b$

Linear regression using MATLAB

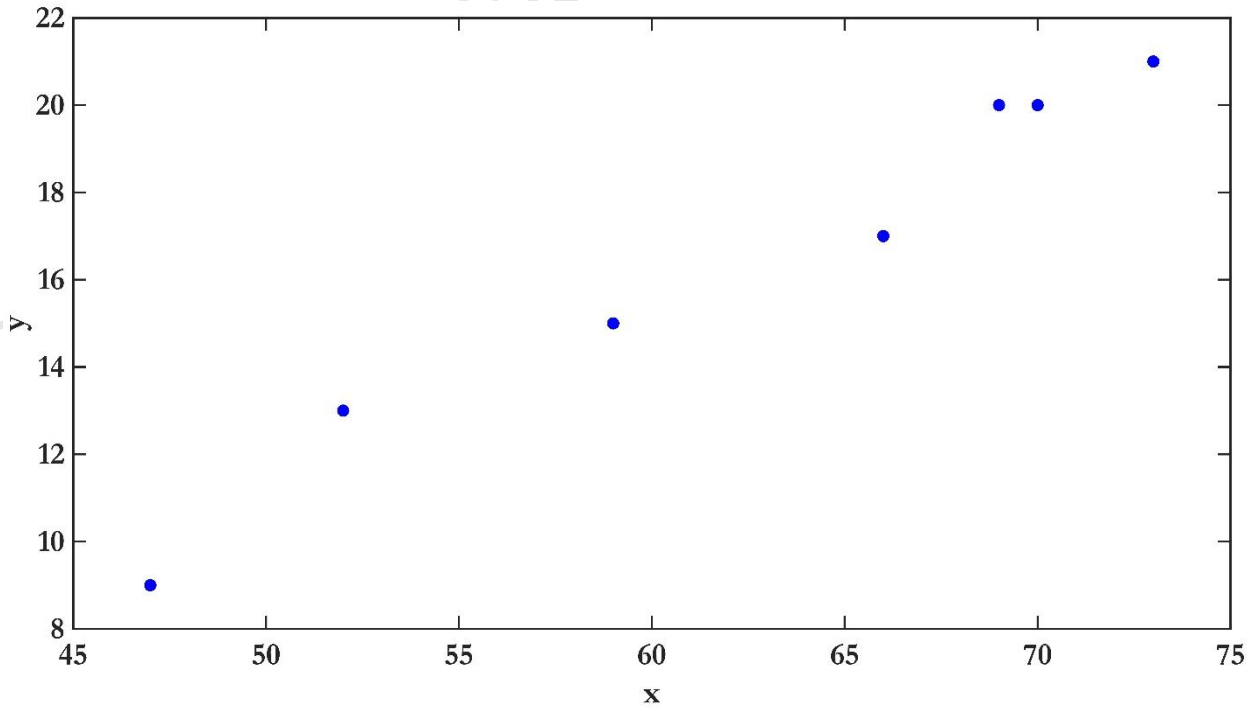
x	52	47	66	70	59	73	69
y	13	9	17	20	15	21	20

Regression model:

$$y = a_0 + a_1 x$$

[Adapted from Numerical Methods with Computer Programs in C++ by P. Ghosh]

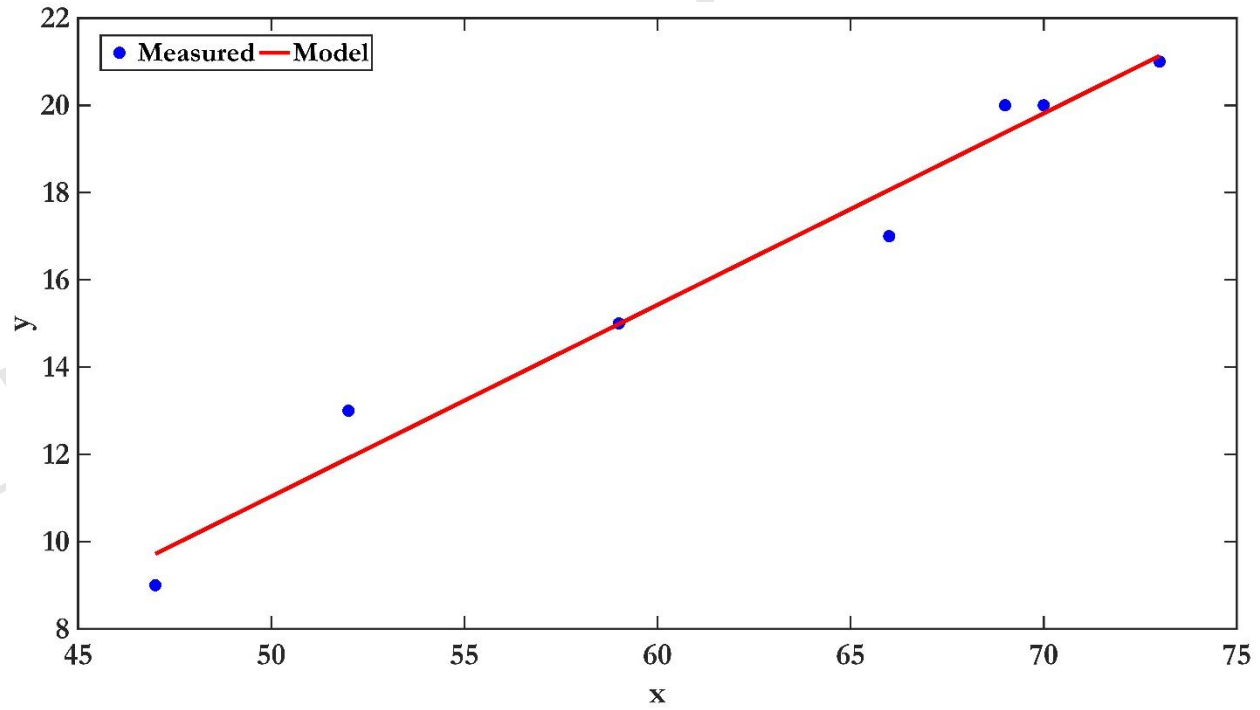
```
clc
clear
close all

x = [52;47;66;70;59;73;69];
y = [13;9;17;20;15;21;20];

plot(x,y,'b.')

xlabel ('x');
ylabel ('y');
```



Linear regression using MATLAB

```
clc
clear
close all

x = [52;47;66;70;59;73;69];
y = [13;9;17;20;15;21;20];

plot(x,y,'b.')

xlabel ('x');
ylabel ('y');

X = [ones(length(x),1) x]

[B,~,~,~,STATS] = regress(y,X);

hold on

xplot = [min(x), max(x)]
yplot = B(1) + B(2) * xplot;

plot(xplot, yplot,'r')

legend('Measured','Model')
```

Regression model: $y = a_0 + a_1 x$




```
B =
  -10.9167
    0.4390
```

```
>> STATS =
    0.9718   172.4871    0.0000    0.6520
```

Multiple linear regression using MATLAB

x_1	0.3	0.6	0.9	0.3	0.6	0.9	0.3	0.6	0.9
x_2	0.001	0.001	0.001	0.01	0.01	0.01	0.05	0.05	0.05
y	0.04	0.24	0.69	0.13	0.82	2.38	0.31	1.95	5.66

Use multiple linear regression to fit the following model

$$y = \alpha x_1^{\beta} x_2^{\gamma}$$

[Adapted from Applied Numerical Methods with MATLAB for Engineers and Scientists by S C Chapra]

Step 1: Linearize the model

$$\ln y = \ln \alpha + \beta \ln x_1 + \gamma \ln x_2$$

$$Y = \ln y,\ a_0 = \ln \alpha, a_1 = \beta, a_2 = \gamma$$
$$X_1 = \ln x_1,\ X_2 = \ln x_2$$

Multi-linear regression model

$$Y = a_0 + a_1 X_1 + a_2 X_2$$

```
clc
clear

x1 = [0.3;0.6;0.9;0.3;0.6;0.9;0.3;0.6;0.9];
x2 = [0.001;0.001;0.001;0.01; 0.01; 0.01;0.05;0.05;0.05];
y = [0.04;0.24;0.69;0.13;0.82;2.38;0.31;1.95;5.66];


Y = log(y);
X1 = log(x1);
X2 = log(x2);
```

Multiple linear regression using MATLAB

```
clc
clear

x1 = [0.3;0.6;0.9;0.3;0.6;0.9;0.3;0.6;0.9];
x2 = [0.001;0.001;0.001;0.01; 0.01; 0.01;0.05;0.05;0.05];
y = [0.04;0.24;0.69;0.13;0.82;2.38;0.31;1.95;5.66];

Y = log(y);
X1 = log(x1);
X2 = log(x2);
X = [ones(length(x1),1) [X1 X2]];

[B,~,~,~,STATS] = regress(Y,X);

alpha = exp(B(1));
beta = B(2);
gamma = B(3);
ymodel = alpha*x1.^beta.*x2.^gamma;
```

$$y = \alpha x_1^{\beta} x_2^{\gamma} \qquad \ln y = \ln \alpha + \beta \ln x_1 + \gamma \ln x_2$$

$$y = a_0 + a_1 X_1 + a_2 X_2$$

$$a_0 = 3.59, \alpha = 36.28$$

$$a_1 = \beta = 2.63, a_2 = \gamma = 0.53$$

y	*ymodel*
0.04	0.04
0.24	0.24
0.69	0.7
0.13	0.13
0.82	0.82
2.38	2.38
0.31	0.31
1.95	1.93
5.66	5.6

```
>> STATS =
   9.9992e-01    3.7818e+04    4.9907e-13    2.5587e-04
```

Multiple linear regression (no constant)

x_1	0.3	0.6	0.9	0.3	0.6	0.9	0.3	0.6	0.9
x_2	0.001	0.001	0.001	0.01	0.01	0.01	0.05	0.05	0.05
y	0.04	0.24	0.69	0.13	0.82	2.38	0.31	1.95	5.66

Use multiple linear regression to fit the following model

$$y = x_1^{\alpha} x_2^{\beta}$$

[Adapted from Applied Numerical Methods with MATLAB for Engineers and Scientists by S. C. Chapra

Step 1: Linearize the model

$$\ln y = \alpha \ln x_1 + \beta \ln x_2$$

$$Y = \ln y,\ a_1 = \alpha, a_2 = \beta$$
$$X_1 = \ln x_1, X_2 = \ln x_2$$

Multi-linear regression model

$$Y = a_1 X_1 + a_2 X_2$$

```
clc
clear

x1 = [0.3;0.6;0.9;0.3;0.6;0.9;0.3;0.6;0.9];
x2 = [0.001;0.001;0.001;0.01; 0.01; 0.01;0.05;0.05;0.05];
y = [0.04;0.24;0.69;0.13;0.82;2.38;0.31;1.95;5.66];


Y = log(y);
X1 = log(x1);
X2 = log(x2);
```

Multi-linear regression (no constant) using MATLAB

```
clc
clear

x1 = [0.3;0.6;0.9;0.3;0.6;0.9;0.3;0.6;0.9];
x2 = [0.001;0.001;0.001;0.01; 0.01; 0.01;0.05;0.05;0.05];
y = [0.04;0.24;0.69;0.13;0.82;2.38;0.31;1.95;5.66];

Y = log(y);
X1 = log(x1);
X2 = log(x2);
X = [X1 X2];

[B,~,~,~,STATS] = regress(Y,X);

alpha = B(1);
beta = B(2);
ymodel = x1.^alpha.*x2.^beta;
```

$$y = x_1^{\alpha} x_2^{\beta} \qquad \ln y = \alpha \ln x_1 + \beta \ln x_2$$

$$y = a_1 X_1 + a_2 X_2$$

$$a_1 = \alpha = 1.73, a_2 = \beta = -0.04$$

y	*ymodel*
0.04	0.16
0.24	0.53
0.69	1.07
0.13	0.15
0.82	0.49
2.38	0.99
0.31	0.14
1.95	0.46
5.66	0.93

```
>> STATS =
    0.4936    4.5918    0.0693    1.4001
```

Polynomial regression using MATLAB

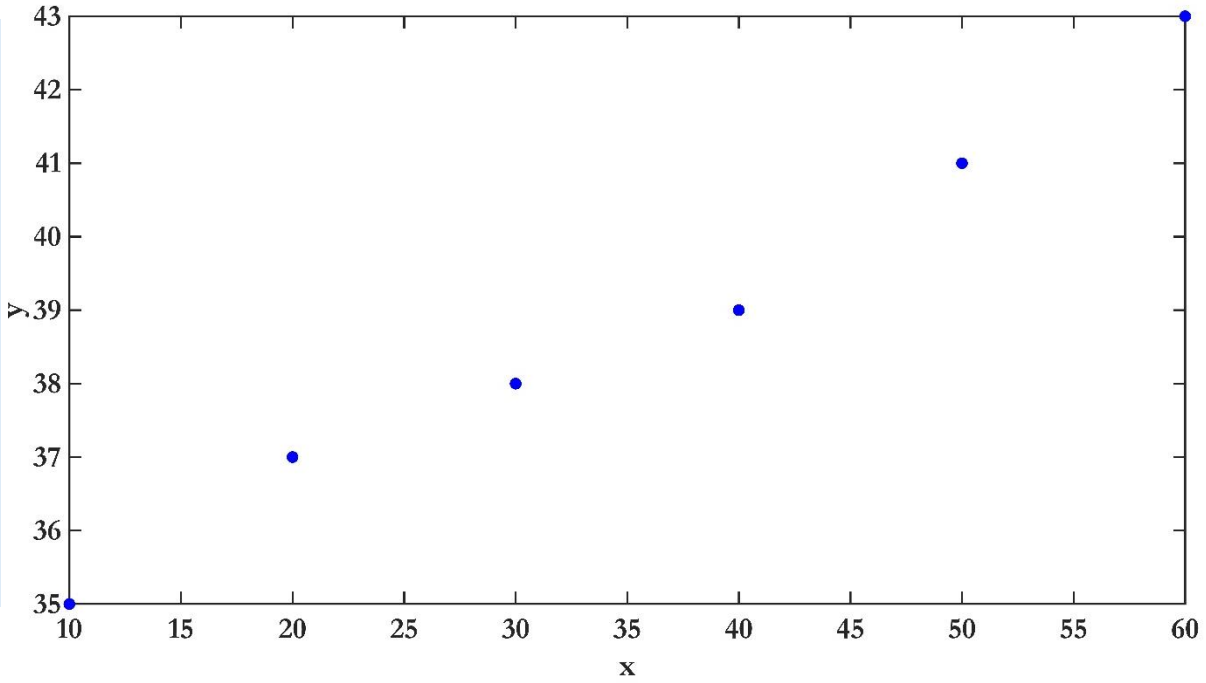
x	10	20	30	40	50	60
y	35	37	38	39	41	43

Fit a second order polynomial to this data.

$$y = a_0 + a_1 x + a_2 x^2$$

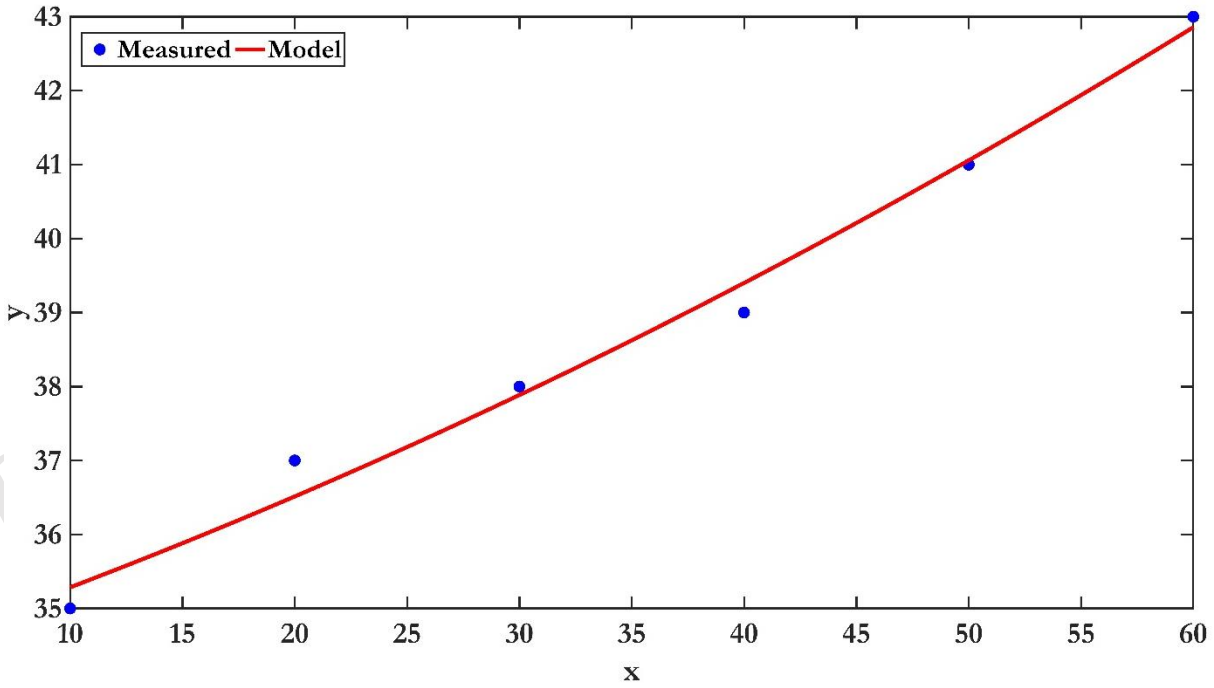
[Adapted from Numerical Methods with Computer Programs in C++ by P. Ghosh]

```
clc
clear
close all

x = [10;20;30;40;50;60];
y = [35;37;38;39;41;43];

plot( x, y,'b.')

xlabel ('x');
ylabel ('y');
```



Polynomial regression using MATLAB

```
clc
clear
close all

x = [10;20;30;40;50;60];
y = [35;37;38;39;41;43];

plot( x, y,'b.')

xlabel ('x');
ylabel ('y');

X = [ones(length(x),1) x x.^2];
[B, ~, ~, ~, STATS] = regress(y,X);

hold on
xplot = linspace(min(x),max(x),1000);
yplot = B(1) + B(2).* xplot + B(3)*xplot.^2;

plot(xplot,yplot)
legend('Measured','Model')
```




```
B =
   34.2000
    0.1014
    0.0007
```

```
>> STATS =
    0.9874  117.5972    0.0014    0.1714
```

Polynomial regression using MATLAB

```
clc
clear
close all

x = [10;20;30;40;50;60];
y = [35;37;38;39;41;43];

plot( x, y,'b.')

xlabel ('x');
ylabel ('y');

X = [ones(length(x),1) x x.^2];
[B, ~, ~, ~, STATS] = regress(y,X);

hold on
xplot = linspace(min(x),max(x),1000);
yplot = B(1) + B(2).* xplot + B(3)*xplot.^2;

plot(xplot,yplot)
legend('Measured','Model')
```

```
B =
   34.2000
    0.1014
    0.0007
```

```
clc
clear
close all

x = [10;20;30;40;50;60];
y = [35;37;38;39;41;43];
n = 2;
plot( x, y,'b.')

xlabel ('x');
ylabel ('y');

B = polyfit(x,y,n);


hold on
xplot = linspace(min(x),max(x),1000);
yplot = polyval(B,xplot);

plot(xplot,yplot)
legend('Measured','Model')
```

```
B = 0.0007   0.1014   34.2000
```

Nonlinear regression using MATLAB

x	0.25	0.75	1.25	1.75	2.25
y	0.28	0.57	0.68	0.74	0.79

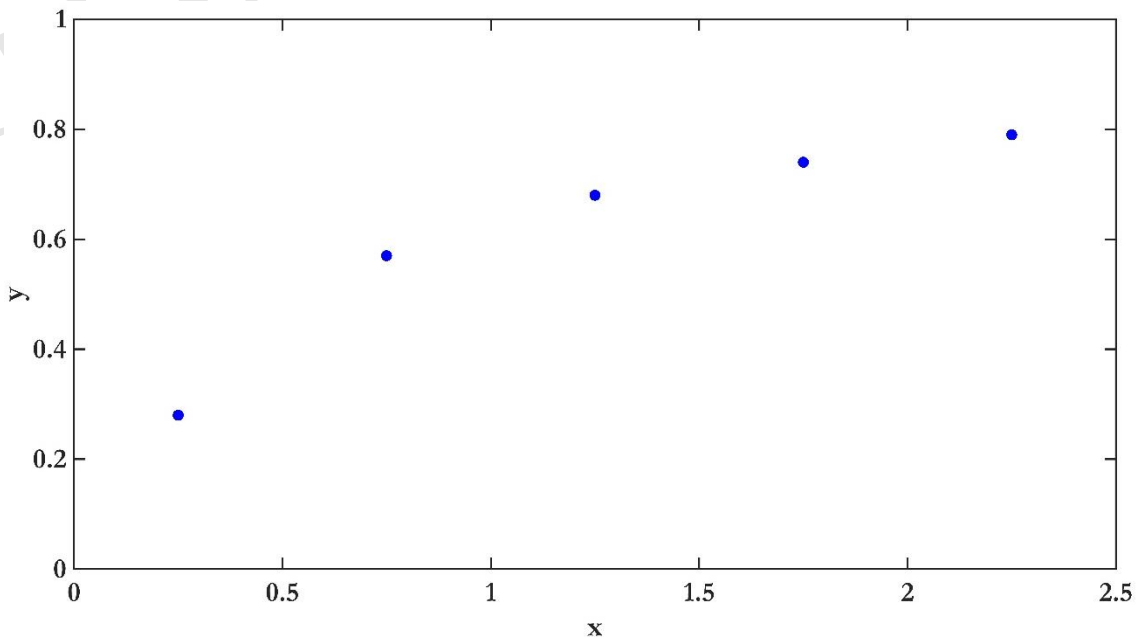
$$f\left(x;a_0,a_1\right)=a_0\left(1-e^{-a_1x}\right)$$

Initial guesses: a_0=1 ,a_1=1

```
clc
clear
close all

x = [0.25 0.75 1.25 1.75 2.25];
y = [0.28 0.57 0.68 0.74 0.79];

a = [1 1];
fun = @prob;

plot (x,y,'b.')

xlabel('x');
ylabel('y');
```

```
function f = prob(a,x)

f = a(1)*(1-exp(-a(2)*x));
```

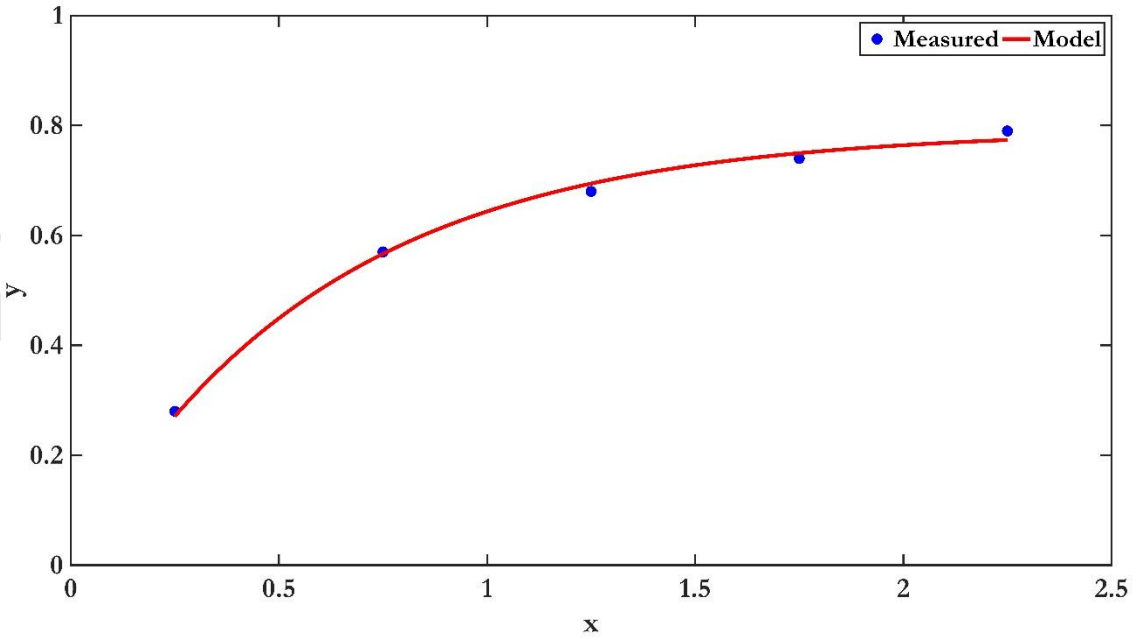
[Adapted from Numerical Methods for Engineers by S. C. Chapra & R. P. Canale]



Nonlinear regression using MATLAB

```
clc
clear
close all

x = [0.25 0.75 1.25 1.75 2.25];
y = [0.28 0.57 0.68 0.74 0.79];

a = [1 1];
fun = @prob;

plot (x,y,'b.')

xlabel('x');
ylabel('y');

BETA = nlinfit (x,y, fun, a);
hold on

xplot = linspace (min(x),max(x),1000);
yplot = fun(BETA, xplot);

plot(xplot,yplot,'r')
xlabel('x');ylabel('y');
legend('Measured','Model')
```

```
function f = prob(a,x)

f = a(1)*(1-exp(-a(2)*x));
```




```
BETA =

7.9187e-01
1.6751e+00
```

Caution !!!

- Focused on simple derivation and practical use of equations to fit data.
- Some statistical assumptions inherent in linear least squares are
  - Each x has a fixed value; it is not random and is known without error.
  - The y values are independent random variables and all have the same variance.
  - The y values for a given x must be normally distributed.
- Useful reference for understanding the aspects of regression
  - Draper, N. R., and H. Smith, *Applied Regression Analysis,* 2nd ed., Wiley, New York, 1981

References

- S. C. Chapra and R. P. Canale, Numerical methods for Engineers, Tata McGraw-Hill, 2002
- S. C. Chapra, Applied Numerical Methods with MATLAB for Engineers and Scientists, Tata McGraw-Hill, 2012
- J. H. Mathews and K. D. Fink, Numerical Methods Using MATLAB, PHI Learning, 2009
- Draper, N. R., and H. Smith, *Applied Regression Analysis,* 2nd ed., Wiley, New York, 1981

Closure

- Regression
- Linear Regression
  - Simple linear regression
  - Multiple linear regression
  - Polynomial regression
  - General linear least squares
- Non-linear Regression
- Transformations for Data Linearization
- MATLAB functions: regress, nlinfit, polyfit, polyval

Thank You !!!