~ Curve Fitting ~

Least Squares Regression

Chapter 17

Curve Fitting

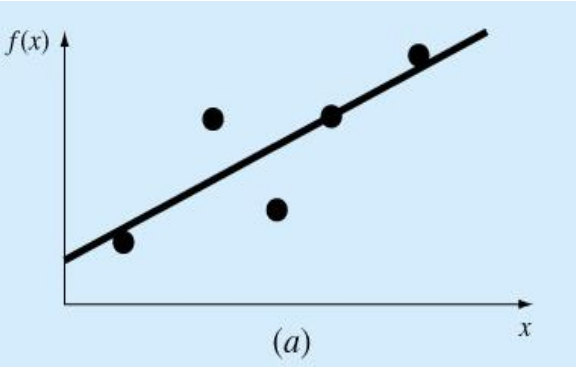
- Fit the best curve to a discrete data set and obtain estimates for other data points

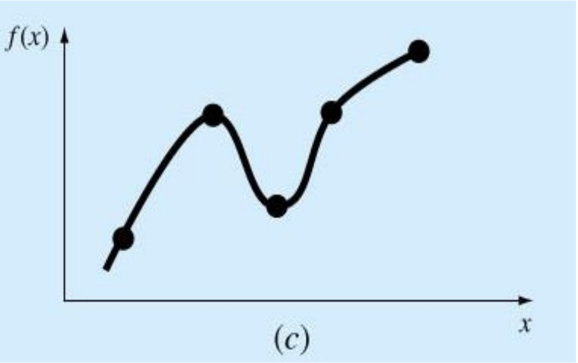
- Two general approaches:
  - *Data exhibit a significant degree of* ***scatter*** Find a single curve that represents the general trend of the data.
  - *Data is very* ***precise***. Pass a curve(s) exactly through each of the points.

- Two common applications in engineering:

Trend analysis. Predicting values of dependent variable: ***extrapolation*** beyond data points or ***interpolation*** between data points.

Hypothesis testing. Comparing existing mathematical model with measured data.





Simple Statistics

In sciences, if several measurements are made of a particular quantity, additional insight can be gained by summarizing the data in one or more well chosen statistics:

Arithmetic mean - The sum of the individual data points (y_i) divided by the number of points.

$$\bar{y} = \frac{\sum y_i}{n} \qquad i = 1, \ldots, n$$

Standard deviation – a common measure of spread for a sample

$$S_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

or *variance*

$$S_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

Coefficient of variation –

$$c.v. = \frac{S_y}{\bar{y}} 100\%$$

quantifies the spread of data (similar to relative error)

Linear Regression

- Fitting a **straight line** to a set of paired observations:
  $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

y_i : **measured value**

e : **error**

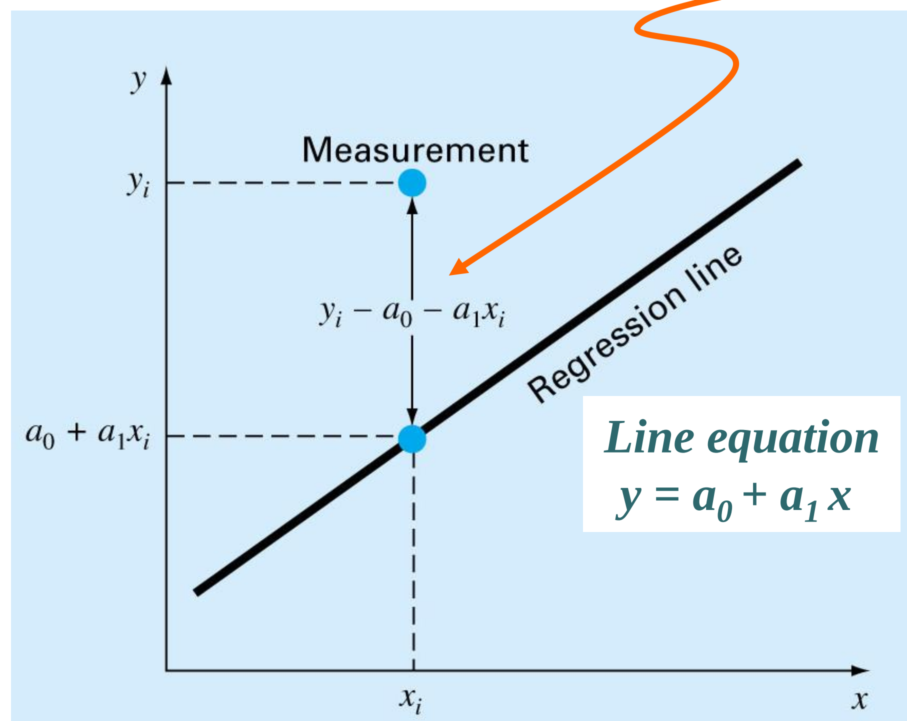
$$y_i = a_0 + a_1 x_i + e$$

$$e = y_i - a_0 - a_1 x_i$$

a_1 : *slope*

a_0 : *intercept*

e   Error



Choosing Criteria For a "Best Fit"

- **Minimize** the sum of the residual errors for all available data?

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - a_o - a_1 x_i)$$

Inadequate!

(see ➔➔➔)

- Sum of the absolute values?

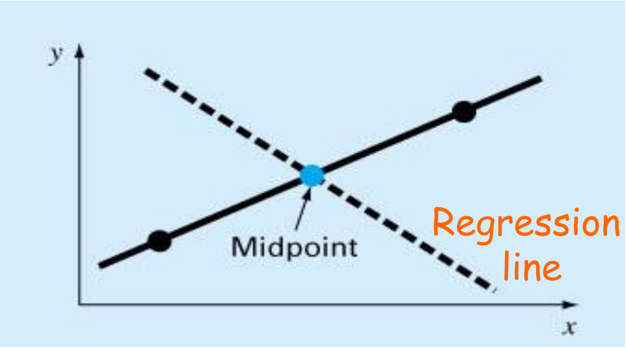
$$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - a_0 - a_1 x_i|$$

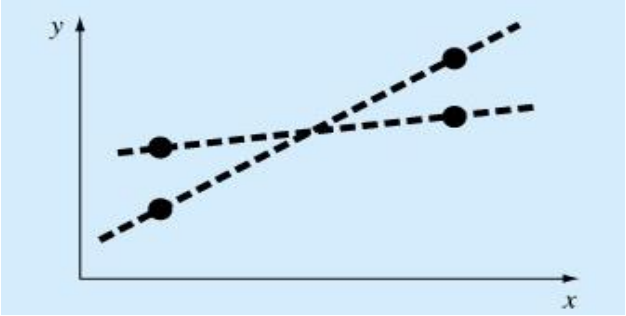
Inadequate!

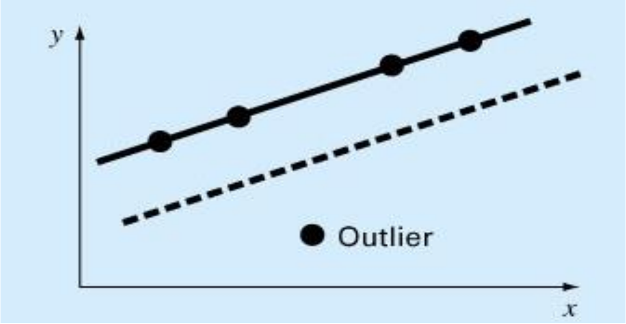
(see ➔➔➔)

- How about minimizing the distance that an individual point falls from the line?

This does not work either! see ➔➔➔







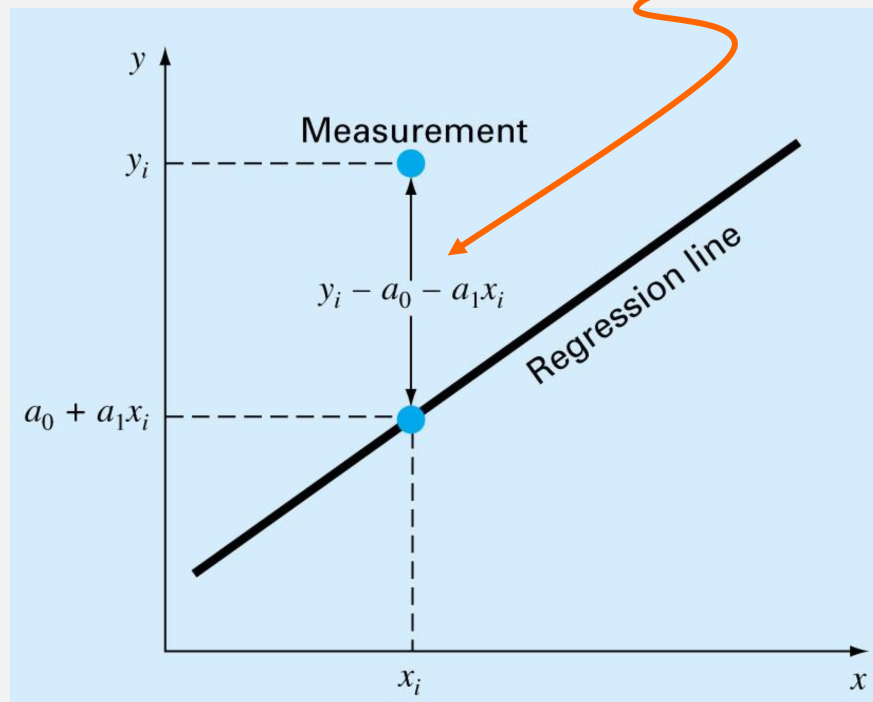
- Best strategy is to *minimize* the sum of the squares of the residuals between the *measured*-y and the y *calculated* with the linear model:

e Error

$$S_r = \sum_{i=1}^{n} e_i^2$$

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

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



- Yields a unique line for a given set of data
- Need to compute a_0 and a_1 such that S_r is minimized!

Least-Squares Fit of a Straight Line

$$\text{Minimize error}: \quad S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

$$\frac{\partial S_r}{\partial a_o} = -2\sum (y_i - a_o - a_1 x_i) = 0 \qquad \Rightarrow \qquad \sum y_i - \sum a_0 - \sum a_1 x_i = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum \left[(y_i - a_o - a_1 x_i) x_i\right] = 0 \qquad \Rightarrow \qquad \sum y_i x_i - \sum a_0 x_i - \sum a_1 x_i^2 = 0$$

Since $\sum a_0 = n a_0$

$$(1) \qquad n a_0 + \left(\sum x_i\right) a_1 = \sum y_i$$

$$(2) \qquad \left(\sum x_i\right) a_0 + \left(\sum x_i^2\right) a_1 = \sum y_i x_i$$

Normal equations which can be solved simultaneously

Least-Squares Fit of a Straight Line

$$\text{Minimize error}: \quad S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

Normal equations which can be solved simultaneously

$$\begin{aligned} n a_0 \quad + \quad \left(\sum x_i\right) a_1 &= \sum y_i \\ \left(\sum x_i\right) a_0 \quad + \quad \left(\sum x_i^2\right) a_1 &= \sum y_i x_i \end{aligned}$$

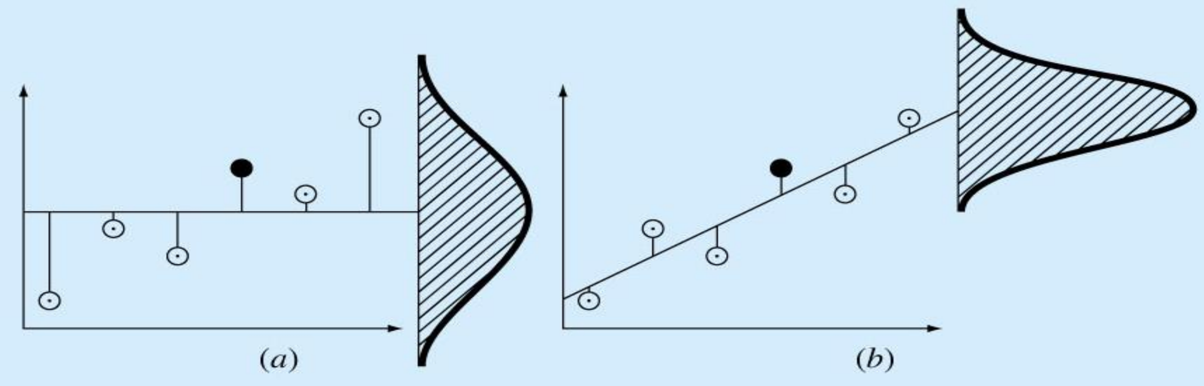
$$\begin{bmatrix} \mathrm{n} & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_i \end{bmatrix}$$

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

using (1), a_0 can be expressed as $\quad a_0 = \bar{y} - a_1 \bar{x}$

Mean values

"Goodness" of our fit



The spread of data

(a) around the mean

(b) around the best-fit line

Notice the improvement in the error due to *linear regression*

- $\boldsymbol{S_r}$ = Sum of the squares of residuals around the regression line
- $\boldsymbol{S_t}$ = total sum of the squares around the mean
- $\boldsymbol{(S_t - S_r)}$ quantifies the improvement or error reduction due to describing data in terms of *a straight line* rather than as *an average value*.

r : *correlation coefficient*

$$r^2 = \frac{S_t - S_r}{S_t}$$

$$S_t = \sum (y_i - \bar{y})^2$$

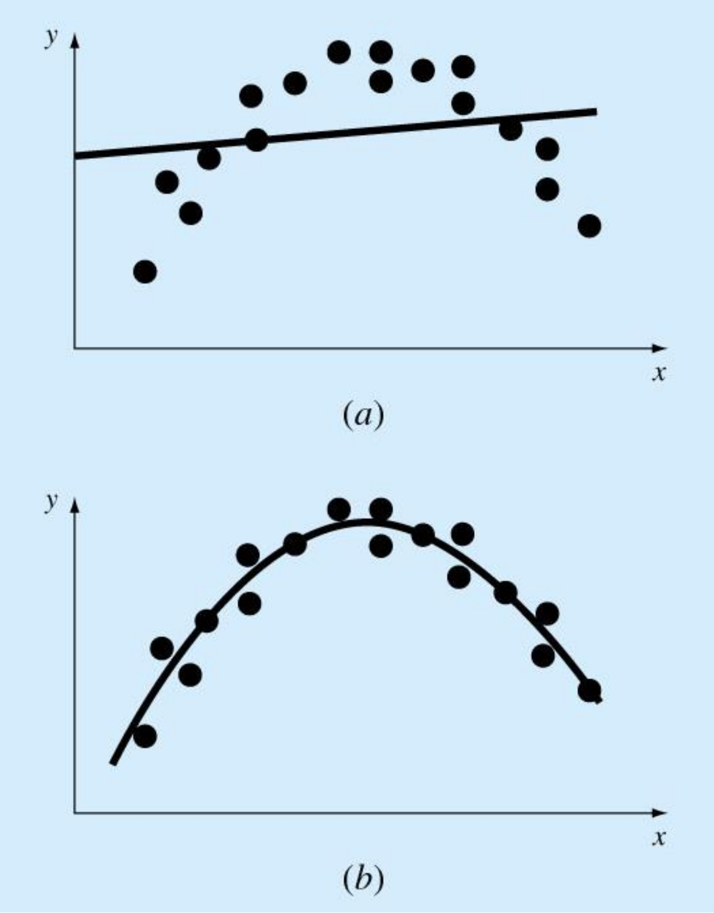
$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

- For a perfect fit $S_r = 0$ and $\boldsymbol{r = r^2 = 1}$ signifies that the line explains 100 percent of the variability of the data.
- For $\boldsymbol{r = r^2 = 0}$ ➔ $S_r = S_t$ ➔ the fit represents no improvement

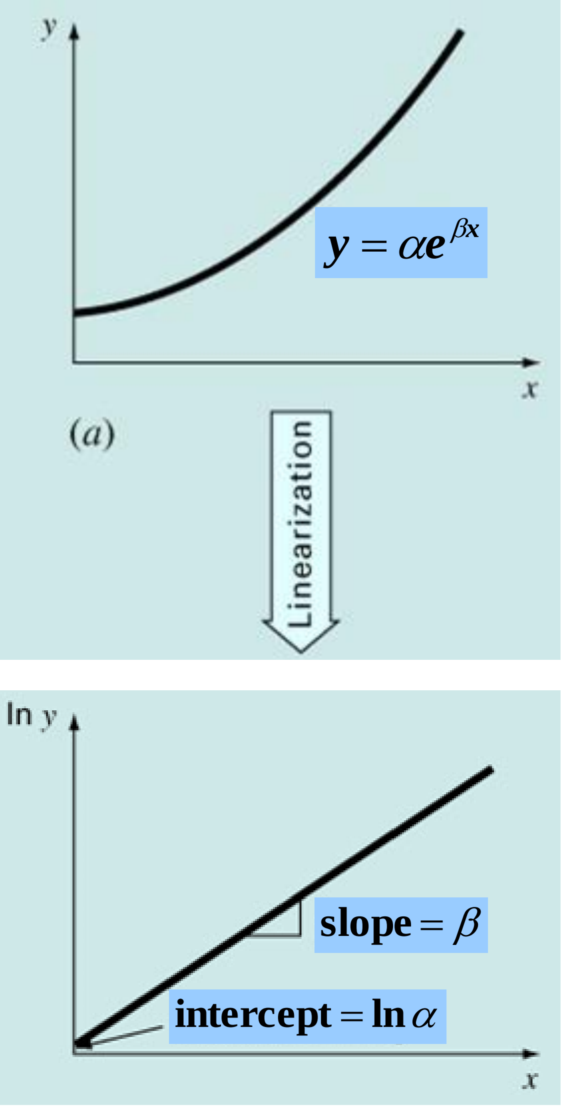
Linearization of Nonlinear Relationships

(a) Data that is ill-suited for linear least-squares regression

(b) Indication that a parabola may be more suitable

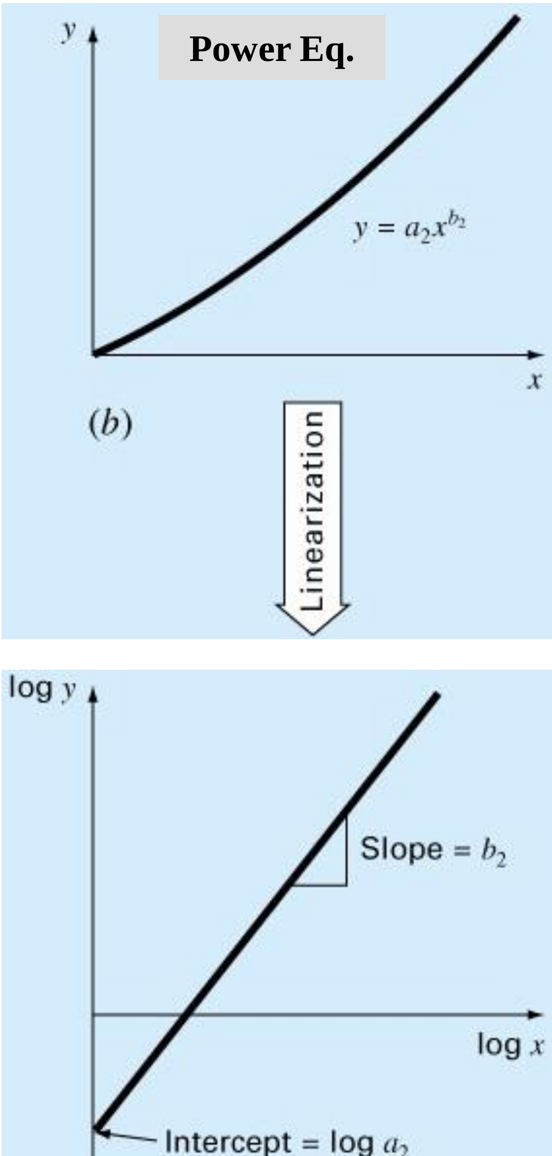


Exponential Eq.

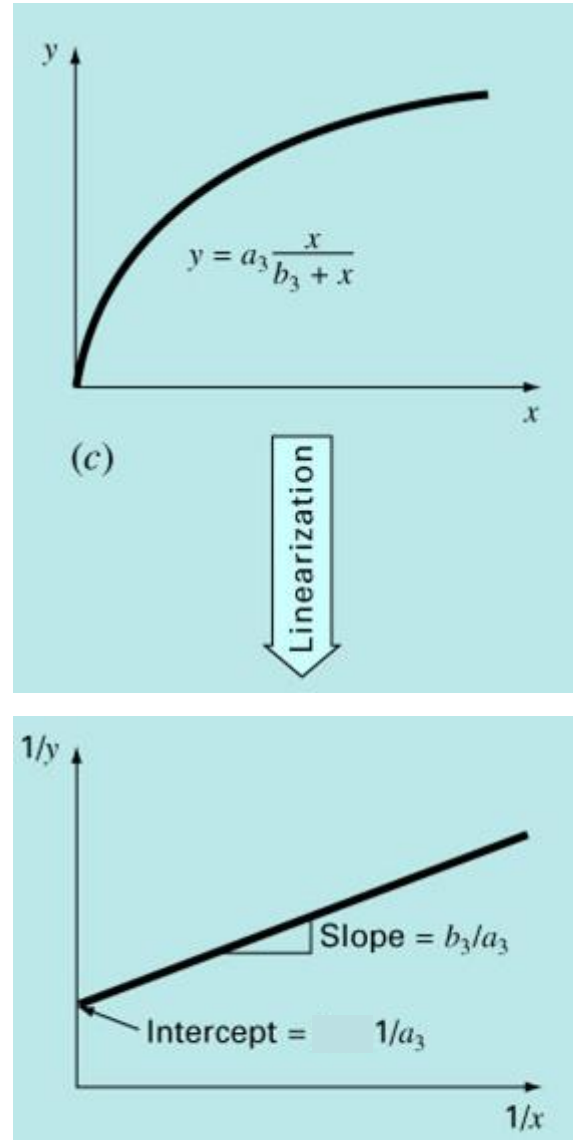


Linearization of Nonlinear Relationships

Power Eq.



Saturation growth-rate Eq.



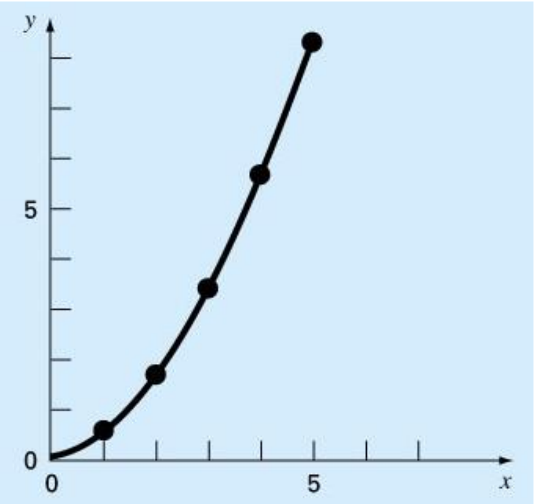
Data to be fit to the power equation:

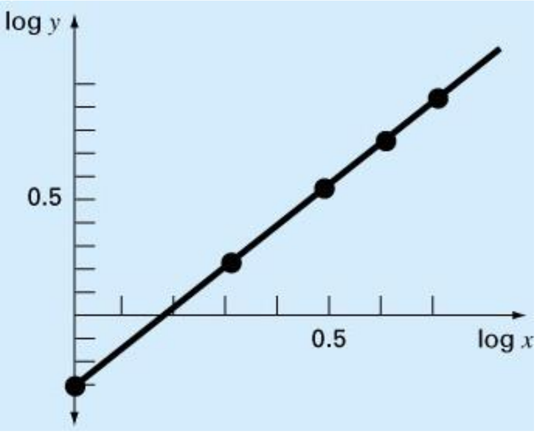
$$y = \alpha_2 x^{\beta_2}$$

$$\Rightarrow \quad \log y = \beta_2 \log x + \log \alpha_2$$

x	y
1	0.5
2	1.7
3	3.4
4	5.7
5	8.4

log x	log y
0	-0.301
0.301	0.226
0.477	0.531
0.602	0.756
0.699	0.924





Linear Regression yields the result:

log y = 1.75*log x – 0.300

After (log y)–(log x) plot is obtained

Find α_2 and β_2 using:

$\text{Slope} = \beta_2$ $\qquad$ $\text{intercept} = \log \alpha_2$

$\beta_2 = 1.75$ $\quad$ $\log \alpha_2 = -0.3 \rightarrow \alpha_2 = 0.5$

$$y = 0.5\, x^{1.75}$$

See Exercises.xls

Polynomial Regression

- Some engineering data is poorly represented by a straight line. A curve (polynomial) may be better suited to fit the data. The least squares method can be extended to fit the data to higher order polynomials.
- As an example let us consider a second order polynomial to fit the data points:

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

$$\text{Minimize error:} \quad S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$

$$\frac{\partial S_r}{\partial a_o} = -2\sum (y_i - a_o - a_1 x_i - a_2 x_i^2) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum x_i (y_i - a_o - a_1 x_i - a_2 x_i^2) = 0$$

$$\frac{\partial S_r}{\partial a_2} = -2\sum x_i^2 (y_i - a_o - a_1 x_i - a_2 x_i^2) = 0$$

$$n a_0 + \left(\sum x_i\right) a_1 + \left(\sum x_i^2\right) a_2 = \sum y_i$$

$$\left(\sum x_i\right) a_0 + \left(\sum x_i^2\right) a_1 + \left(\sum x_i^3\right) a_2 = \sum x_i y_i$$

$$\left(\sum x_i^2\right) a_0 + \left(\sum x_i^3\right) a_1 + \left(\sum x_i^4\right) a_2 = \sum x_i^2 y_i$$

Polynomial Regression

- To fit the data to an m^{th} order polynomial, we need to solve the following system of linear equations ((m+1) equations with (m+1) unknowns)

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

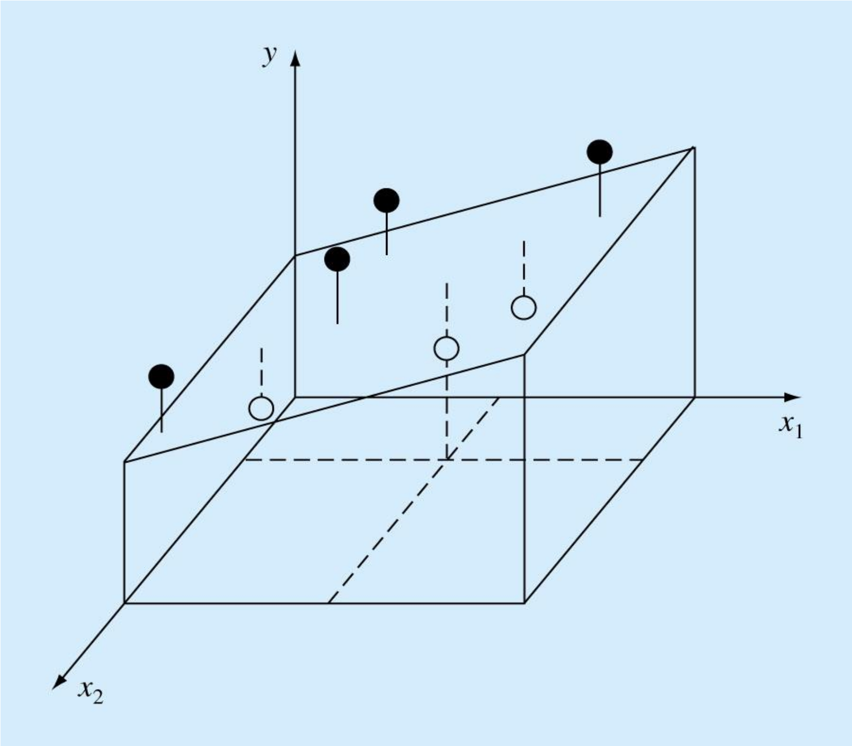
Matrix Form

Multiple Linear Regression

- A useful extension of linear regression is the case where *y* is a linear function of two or more independent variables. For example:

$$y = a_o + a_1x_1 + a_2x_2$$

- For this 2-dimensional case, the regression line becomes a plane as shown in the figure below.



Multiple Linear Regression

Example (2 - vars): Minimize error: $S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2$

$$\frac{\partial S_r}{\partial a_o} = -2\sum (y_i - a_o - a_1 x_{1i} - a_2 x_{2i}) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum x_{1i}(y_i - a_o - a_1 x_{1i} - a_2 x_{2i}) = 0$$

$$\frac{\partial S_r}{\partial a_2} = -2\sum x_{2i}(y_i - a_o - a_1 x_{1i} - a_2 x_{2i}) = 0$$

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

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

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

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

Which method would you use to solve this Linear System of Equations?

Multiple Linear Regression

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

Example 17.6

The following data is calculated from the equation $y = 5 + 4x_1 - 3x_2$

x_1	x_2	y
0	0	5
2	1	10
2.5	2	9
1	3	0
4	6	3
7	2	27

Use multiple linear regression to fit this data.

Solution:

$$\begin{bmatrix} 6 & 16.5 & 14 \\ 16.5 & 76.25 & 48 \\ 14 & 48 & 54 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 54 \\ 243.5 \\ 100 \end{bmatrix}$$

this system can be solved using Gauss Elimination.

The result is: $a_0=5$ $a_1=4$ and $a_2= -3$

$$\mathbf{y = 5 + 4x_1 - 3x_2}$$