Least-Squares Regression

Chapra: Chapter-17

Curve Fitting

- Given a set of points:
  - experimental data
  - tabular data

- Fit a curve (surface) to the points so that we can easily evaluate $f(x)$ at any x of interest.

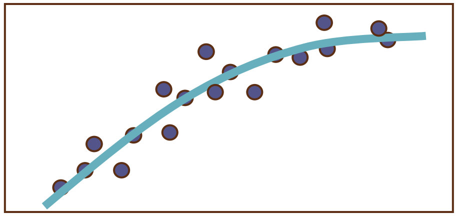
- If x within data range
  $\rightarrow$ interpolating (generally safe)

- If x outside data range
  $\rightarrow$ extrapolating (often dangerous)

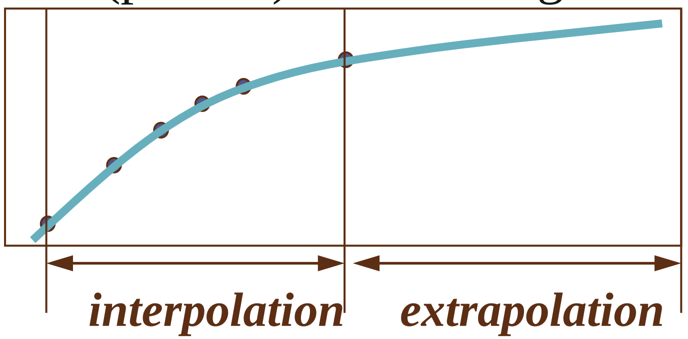
Curve Fitting

Two main methods :

1. **Least-Squares Regression**
   - Function is "best fit" to data.
   - Does not necessarily pass through points.
   - Used for scattered data (experimental)
   - Can develop models for analysis/design.



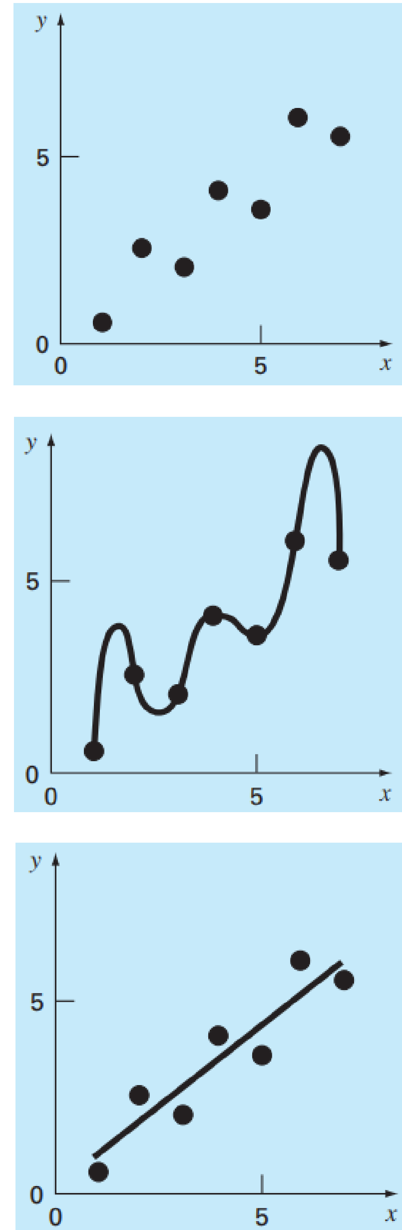
2. **Interpolation**
   - Function passes through all (or most) points.
   - Interpolates values of well-behaved (precise) data or for geometric design.



Linear Regression

- The simplest example of a least-squares approximation is fitting a straight line to a set of paired observations: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- The mathematical expression for the straight line is

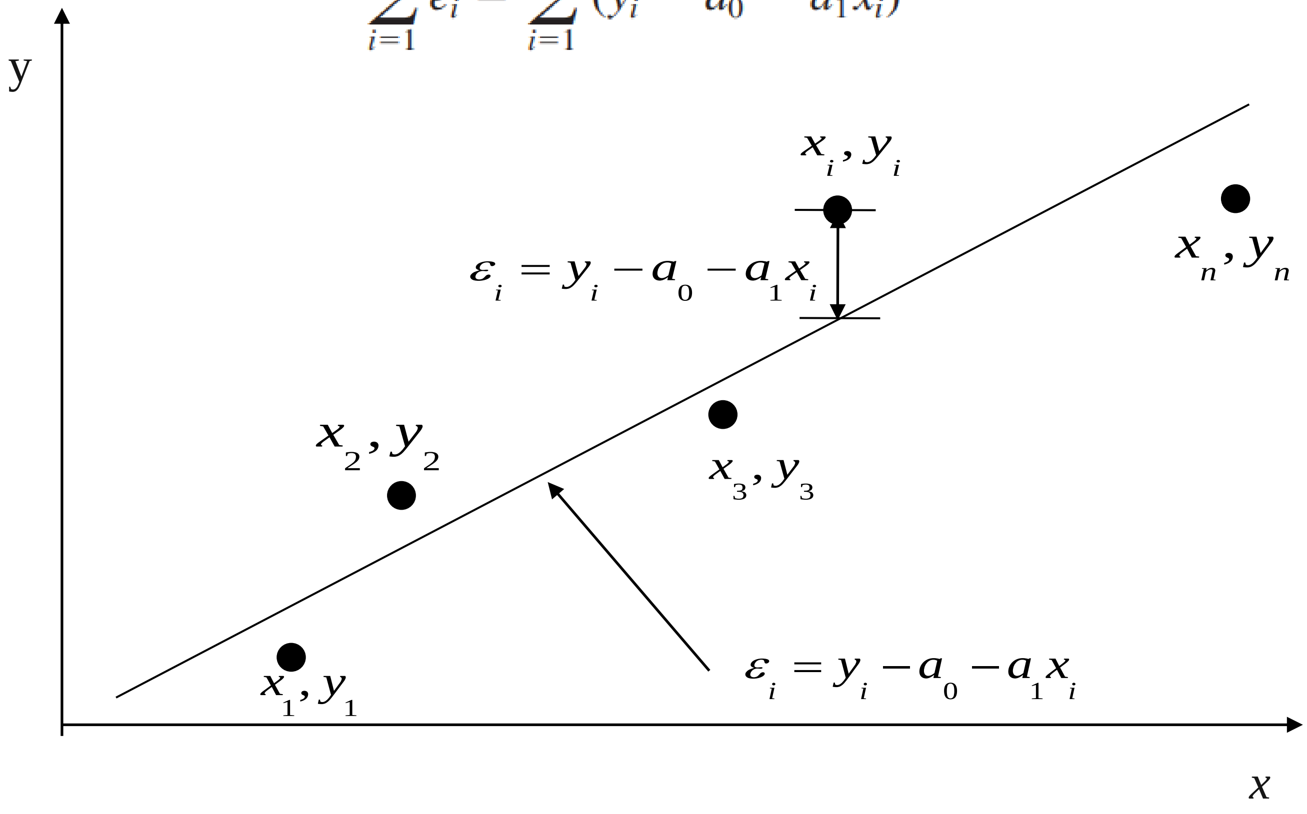
$$y = a_0 + a_1 x + e$$



Criteria for a “Best” Fit

- One strategy for fitting a “best” line through the data would be to minimize the sum of the residual errors for all the available data, as in

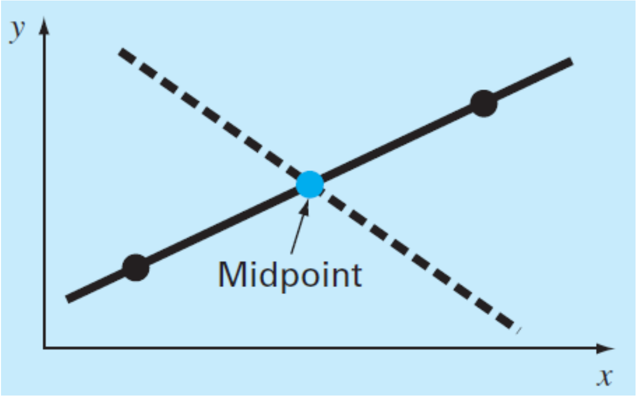
$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)$$



Criteria for a "Best" Fit

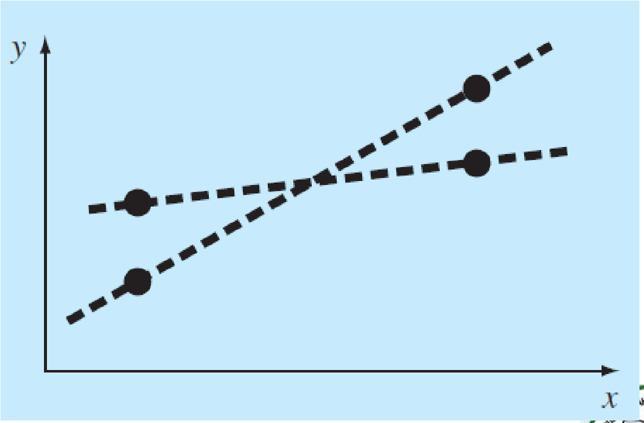
- Any straight line passing through the midpoint of the connecting line results in a minimum value equal to zero because the errors cancel.

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



- For the four points shown, any straight line falling within the dashed lines will minimize the sum of the absolute values.

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



Criteria for a "Best" Fit

Using $y=4x-4$ as the regression curve

Table. Residuals at each point for regression model $y = 4x - 4$.

x	y	$y_{predicted}$	$\varepsilon = y - y_{predicted}$
2.0	4.0	4.0	0.0
3.0	6.0	8.0	-2.0
2.0	6.0	4.0	2.0
3.0	8.0	8.0	0.0
			$\sum_{i=1}^{4} \varepsilon_i = 0$

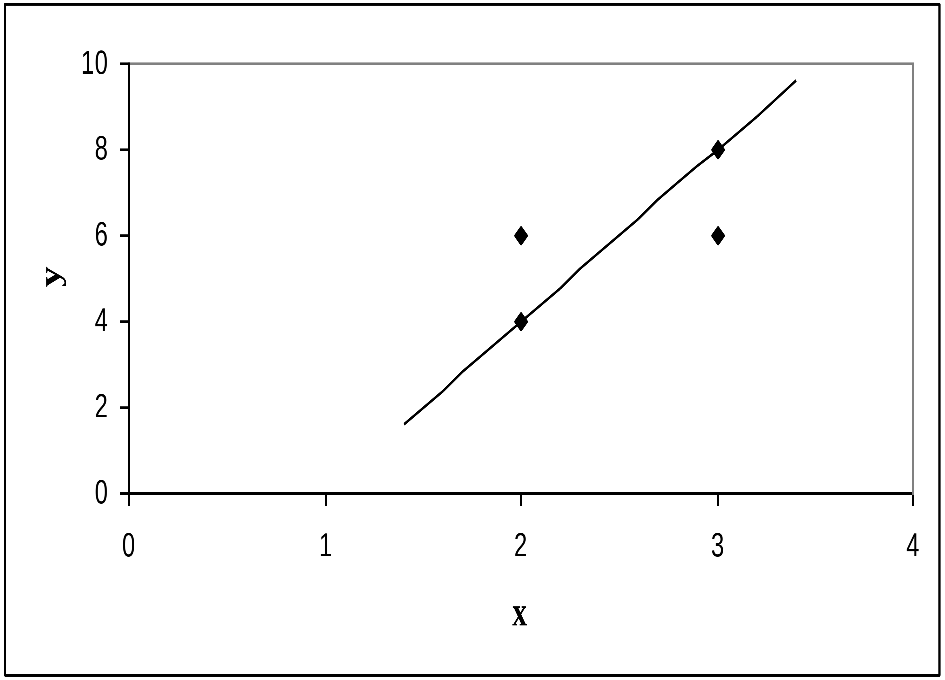


Figure. Regression curve for $y=4x-4$, y vs. x data

Criteria for a "Best" Fit

Using $y=6$ as the regression curve

Table. Residuals at each point for regression model y = 6.

x	y	$y_{predicted}$	$\varepsilon = y - y_{predicted}$
2.0	4.0	6.0	-2.0
3.0	6.0	6.0	0.0
2.0	6.0	6.0	0.0
3.0	8.0	6.0	2.0
			$\sum_{i=1}^{4} \varepsilon_i = 0$

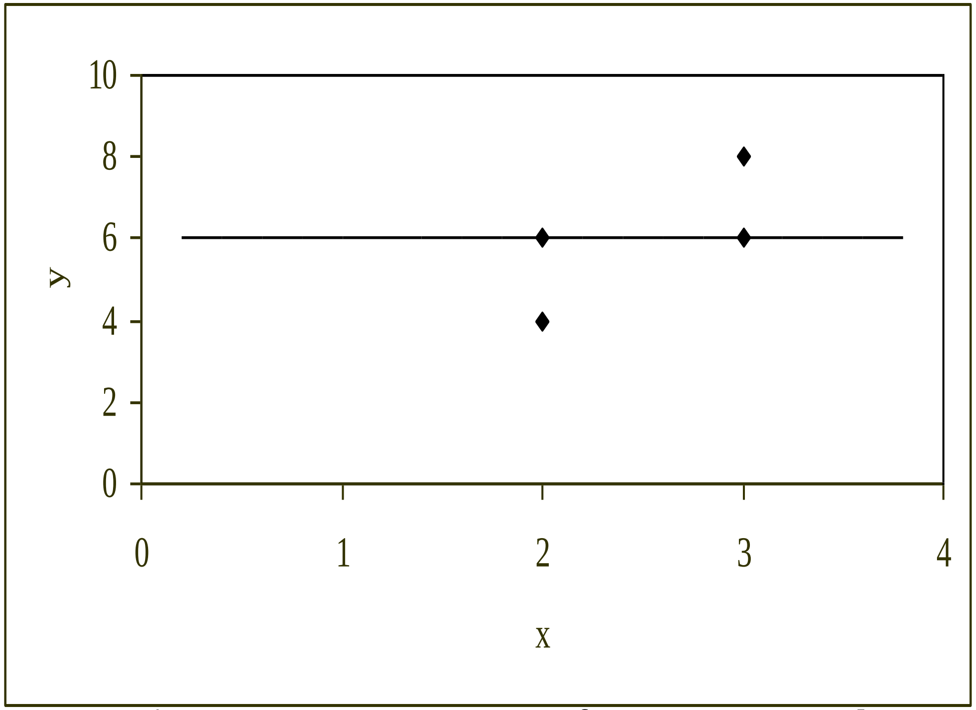


Figure. Regression curve for y=6, y vs. x data

Criteria for a "Best" Fit

- The least squares criterion minimizes the sum of the square of the residuals in the model, and also produces a unique line.

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

Least-Squares Fit of a Straight Line

- To determine values for a_0 and a_1, above equation is differentiated with respect to each coefficient:

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

$$\frac{\partial S_r}{\partial a_1} = -2 \sum [(y_i - a_0 - a_1 x_i) x_i]$$

- Setting these derivatives equal to zero will result in a minimum S_r $\quad 0 = \sum y_i - \sum a_0 - \sum a_1 x_i \quad ; \quad 0 = \sum y_i x_i - \sum a_0 x_i - \sum a_1 x_i^2$

For $i = 1$ to n

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

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

$$a_1 = \frac{n \Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2}$$

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

Example

- Fit a straight line to the x and y values (1, 0.5), (2, 2.5), (3, 2.0), (4, 4.0), (5, 3.5), (6, 6.0), (7, 5.5)
- The following quantities can be computed:

$$n = 7 \qquad \sum x_i y_i = 119.5 \qquad \sum x_i^2 = 140$$

$$\sum x_i = 28 \qquad \bar{x} = \frac{28}{7} = 4$$

$$\sum y_i = 24 \qquad \bar{y} = \frac{24}{7} = 3.428571$$

$$a_1 = \frac{7(119.5) - 28(24)}{7(140) - (28)^2} = 0.8392857$$

$$a_0 = 3.428571 - 0.8392857(4) = 0.07142857$$

Example

- Computation for error analysis of the linear fit

x_i	y_i	$(y_i - \bar{y})^2$	$(y_i - a_0 - a_1 x_i)^2$
1	0.5	8.5765	0.1687
2	2.5	0.8622	0.5625
3	2.0	2.0408	0.3473
4	4.0	0.3265	0.3265
5	3.5	0.0051	0.5896
6	6.0	6.6122	0.7972
7	5.5	4.2908	0.1993
Σ	24.0	22.7143	2.9911

- Therefore, the least-squares fit is

$$y = 0.07142857 + 0.8392857x$$

Quantification of Error of Linear Regression

- To determine the spread of the points around the line is of similar magnitude along the entire range of the data

the standard deviation, $$S_y = \sqrt{\frac{S_t}{n-1}}$$

the standard error of the estimate, $$S_{y/x} = \sqrt{\frac{S_r}{n-2}}$$

The correlation coefficient of determination, $$r^2 = \frac{S_t - S_r}{S_t}$$

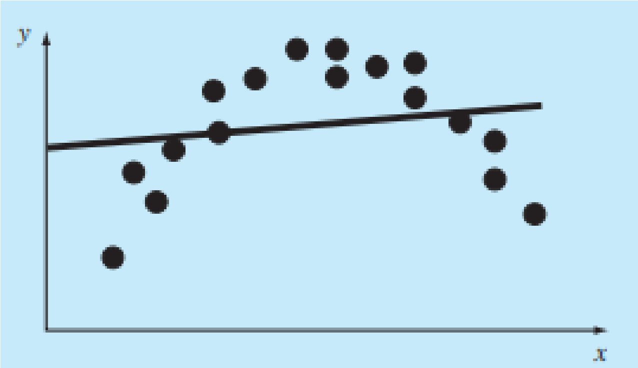
Example

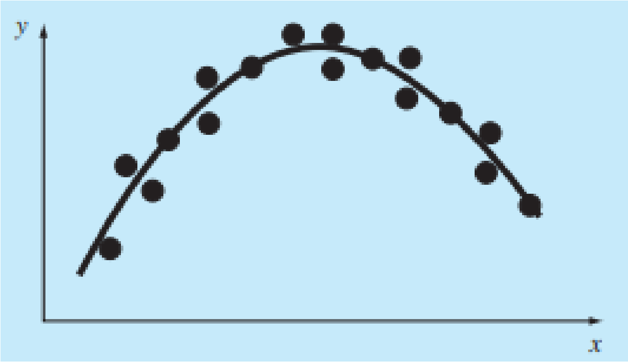
- Compute the total standard deviation, the standard error of the estimate, and the correlation coefficient for the data in the previous example.
- The standard deviation is $$s_y = \sqrt{\frac{22.7143}{7-1}} = 1.9457$$
- The standard error of the estimate is $$s_{y/x} = \sqrt{\frac{2.9911}{7-2}} = 0.7735$$
- The correlation coefficient $$r^2 = \frac{22.7143 - 2.9911}{22.7143} = 0.868$$

These results indicate that 86.8 percent of the original uncertainty has been explained by the linear model.

Polynomial Regression

- The sample data is poorly represented by a straight line
- For these cases, a curve would be better suited to fit these data.
- To fit polynomials to the data can be accomplished using *polynomial regression.*





Polynomial Regression

- Suppose we want to fit a second-order polynomial or quadratic: $y = a_0 + a_1 x + a_2 x^2 + e$
- For this case the sum of the squares of the residuals is

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

- We take its derivative with respect to each of the unknown coefficients of the polynomial

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum x_i^2 (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

Polynomial Regression

- These equations can be set equal to zero and rearranged to develop the following set of normal equations:

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

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

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

- The two-dimensional case can be easily extended to an *m*th order polynomial as $y = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x_m + e$
- The standard error is

$$s_{y/x} = \sqrt{\frac{S_r}{n - (m + 1)}}$$

Example

- Fit a second-order polynomial to the data (0, 2.1), (1, 7.7), (2, 13.6), (3, 27.2), (4, 40.2), (5, 61.1).
- From the given data,

$$
\begin{array}{lll}
m = 2 & \sum x_i = 15 & \sum x_i^4 = 979 \\
n = 6 & \sum y_i = 152.6 & \sum x_i y_i = 585.6 \\
\bar{x} = 2.5 & \sum x_i^2 = 55 & \sum x_i^2 y_i = 2488.8 \\
\bar{y} = 25.433 & \sum x_i^3 = 225 &
\end{array}
$$

- Therefore, the simultaneous linear equations are

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

Example

- Solving these equations through a technique such as Gauss elimination gives $a_0 = 2.47857$, $a_1 = 2.35929$, and $a_2 = 1.86071$. Therefore, the least-squares quadratic equation for this case is

$$y = 2.47857 + 2.35929x + 1.86071x^2$$

Example

x_i	y_i	$(y_i - \bar{y})^2$	$(y_i - a_0 - a_1x_i - a_2x_i^2)^2$
0	2.1	544.44	0.14332
1	7.7	314.47	1.00286
2	13.6	140.03	1.08158
3	27.2	3.12	0.80491
4	40.9	239.22	0.61951
5	61.1	1272.11	0.09439
Σ	152.6	2513.39	3.74657

- The standard error of the estimate based on the regression polynomial is $$s_{y/x} = \sqrt{\frac{3.74657}{6-3}} = 1.12$$

- The coefficient of determination is $$r^2 = \frac{2513.39 - 3.74657}{2513.39} = 0.99851$$

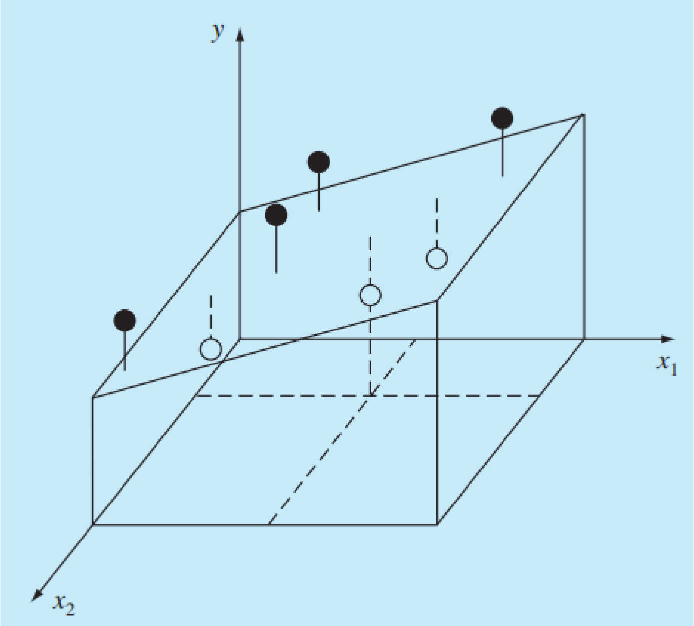
Multiple Linear Regression

- A useful extension of linear regression is the case where y is a linear function of two or more independent variables

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

- For this two-dimensional case, the regression "line" becomes a "plane".
- The "best" values of the coefficients are determined by

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



Multiple Linear Regression

- Differentiating with respect to each of the unknown coefficients

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum x_{1i}(y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum x_{2i}(y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})$$

- Setting the partial derivatives equal to zero and expressing the result in matrix form as

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

Example

- The following data were 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 these data.

- The result in matrix form

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

Example

- The following data were 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 these data.

- The result in matrix form

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

- It can be solved using a method such as Gauss elimination for $a_0 = 5$, $a_1 = 4$ and $a_2 = -3$

which is consistent with the original equation from which these data were derived.