ME 310

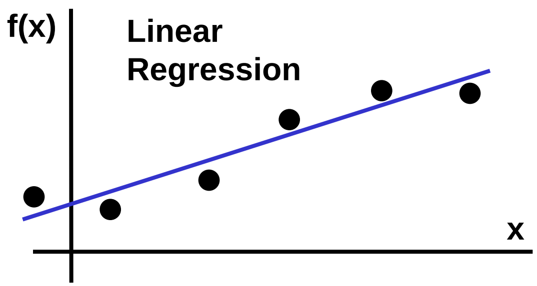
Numerical Methods

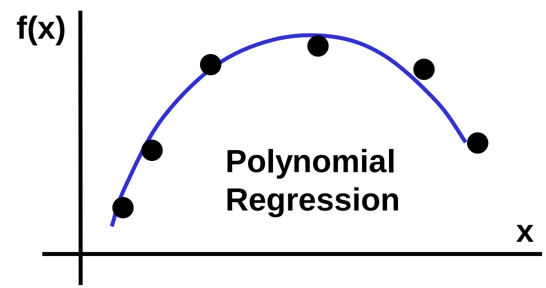
Least Squares Regression

These presentations are prepared by
Dr. Cuneyt Sert
Mechanical Engineering Department
Middle East Technical University
Ankara, Turkey
csert@metu.edu.tr

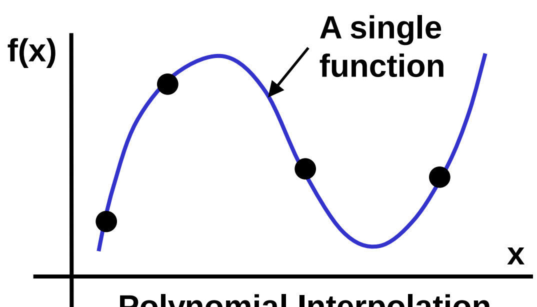
About Curve Fitting

- Curve fitting is expressing a discrete set of data points as a continuous function.
- It is frequently used in engineering. For example the emprical relations that we use in heat transfer and fluid mechanics are functions fitted to experimental data.
- **Regression:** Mainly used with experimental data, which might have significant amount of error (noise). No need to find a function that passes through all discrete points.

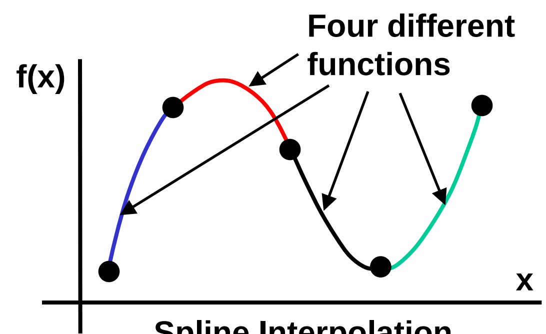




- **Interpolation:** Used if the data is known to be very precise. Find a function (or a series of functions) that passes through all discrete points.



Polynomial Interpolation



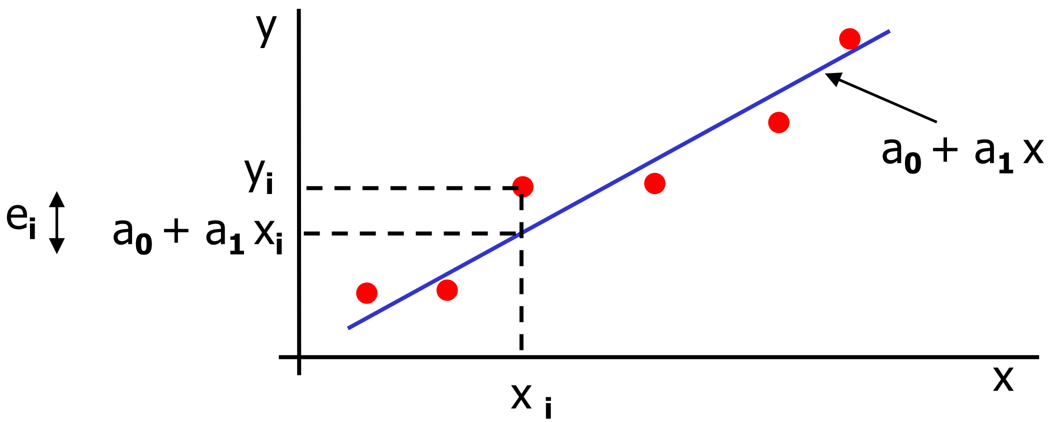
Spline Interpolation

Least Squares Regression

(Read the statistics review from the book.)

- Fitting a straight line to a set of data set (paired data points).

$$(x_1, y_1), \quad (x_2, y_2), \quad (x_3, y_3), \quad . . . , \quad (x_n, y_n)$$



a_0 : y-intercept (unknown)

a_1 : slope (unknown)

$e_i = y_i - a_0 - a_1 x_i$

Error (deviation) for the i^{th} data point

- Minimize the error (deviation) to get a best-fit line (to find a_0 and a_1). Several posibilities are:
  - Minimize the sum of individual errors.
  - Minimize the sum of absolute values of individual errors.
  - Minimize the maximum error.
  - Minimize the sum of squares of individual errors. This is the preferred strategy (Check the bookto see why others fail).

Minimizing the Square of Individual errors

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

Sum of squares of the residuals

- Determine the unknowns a_0 and a_1 by minimizing S_r.
- To do this set the derivatives of S_r wrt a_0 and a_1 to zero.

$$\frac{\partial S_r}{\partial a_0} = -2\sum_{i=1}^{n}(y_i - a_0 - a_1 x_i) = 0 \quad \rightarrow \quad n\,a_0 + \left(\sum x_i\right)a_1 = \sum y_i$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum_{i=1}^{n}[(y_i - a_0 - a_1 x_i)x_i] \quad \rightarrow \quad \left(\sum x_i\right)a_0 + \left(\sum x_i^2\right)a_1 = \sum x_i y_i$$

or

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

These are called the normal equations.

- Solve these for a_0 and a_1. The results are

$$\boxed{a_1 = \frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \qquad a_0 = \bar{y} - a_1 \bar{x}}$$

where y-bar and x-bar are the means of y and x, respectively.

Example 24:

Use least-squares regression to fit a straight line to

x	1	3	5	7	10	12	13	16	18	20
y	4	5	6	5	8	7	6	9	12	11

$$n = 10$$

$$\sum x_i = 105$$

$$\sum y_i = 73$$

$$\bar{x} = 10.5$$

$$\bar{y} = 7.3$$

$$\sum x_i^2 = 1477$$

$$\sum x_i y_i = 906$$

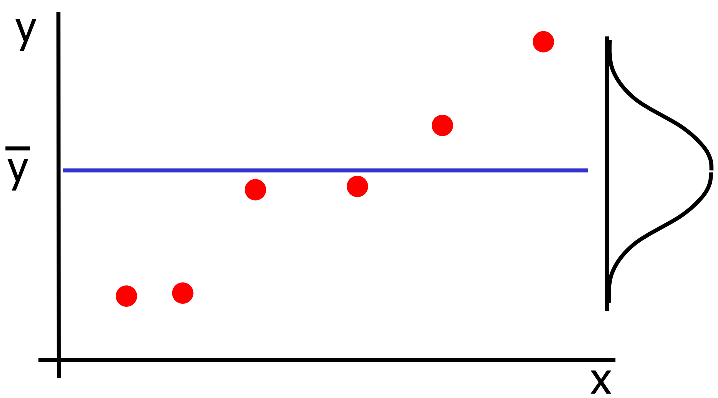
$$a_1 = \frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{10*906 - 105*73}{10*1477 - 105^2} = 0.3725$$

$$a_0 = 7.3 \text{ - } 0.3725*10.5 = 3.3888$$

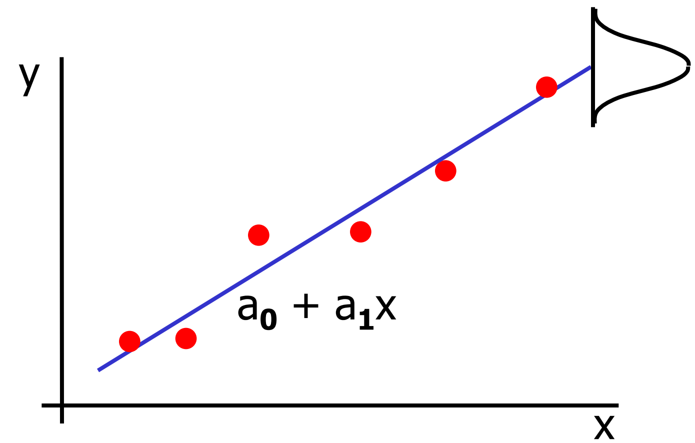
Exercise 24: It is always a good idea to plot the data points and the regression line to see how well the line represents the points. You can do this with Excel. Excel will calculate a_0 and a_1 for you.

Error of Linear Regression (How good is the best line?)

Spread of data around the mean



Spread of data around the regression line



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

$$s_y = \sqrt{\frac{S_t}{n-1}} \quad \text{std. deviation}$$

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

$$s_{y/x} = \sqrt{\frac{S_r}{n-2}} \quad \text{std. error of estimate}$$

- The improvement obtained by using a regression line instead of the mean gives a maesure of how good the regression fit is.

coefficient of determination

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

correlation coefficient

$$\mathbf{r = \frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}}$$

How to interpret the correlation coefficient?

- Two extreme cases are
  - $S_r = 0 \quad \rightarrow \quad r=1$ describes a perfect fit (straight line passing through all points).
  - $S_r = S_t \quad \rightarrow \quad r=0$ describes a case with no improvement.
- Usually an r value close to 1 represents a good fit. But be careful and always plot the data points and the regression line together to see what is going on.

Example 24 (cont'd): Calculate the correlation coefficient.

$$n = 10$$

$$\sum x_i = 105$$

$$\sum y_i = 73$$

$$\bar{x} = 10.5$$

$$\bar{y} = 7.3$$

$$\sum x_i^2 = 1477$$

$$\sum x_i y_i = 906$$

$$S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 64.1$$

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 = 12.14$$

$$r^2 = \frac{S_t - S_r}{S_t} = 0.8107$$

$$r = 0.9$$

Example 24 (cont'd): Reverse x and y. Find the linear regression line and calculate r.

$x = -5.3869 + 2.1763\, y$

$S_t = 374.5$, $S_r = 70.91$ (different than before).

$r^2 = 0.8107$, $r = 0.9$ (same as before).

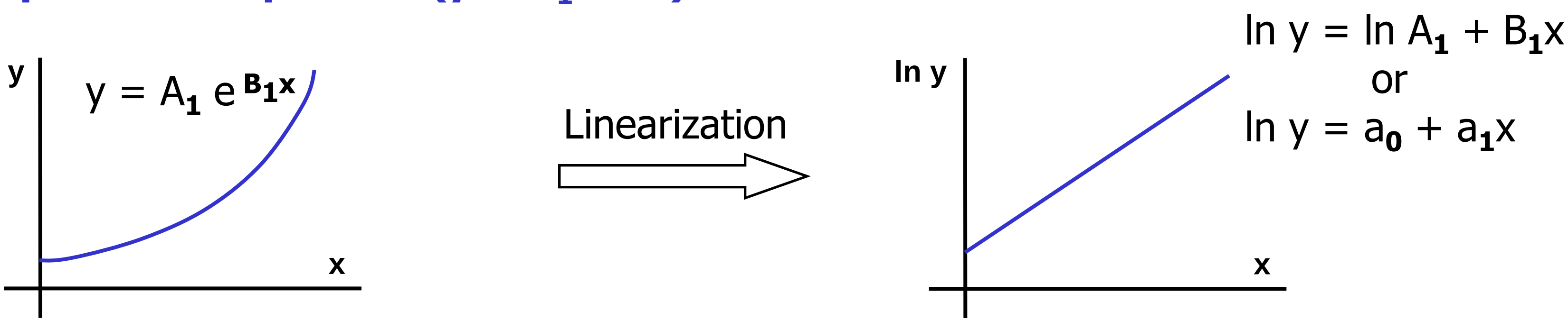
Exercise 25: When working with experimental data we usually take the variable that is controlled by us in a precise way as x. The measured or calculated quantities are y. See Midterm II of Fall 2003 for an example.

Linearization of Nonlinear Behavior

- Linear regression is useful to represent a linear relationship.
- If the relation is nonlinear either another technique can be used or the data can be transformed so that linear regression can still be used. The latter technique is frequently used to fit the the following nonlinear equations to a set of data.
  - Exponential equation ($y=A_1 e^{B_1 x}$)
  - Power equation ($y=A_2 x^{B_2}$)
  - Saturation-growth rate equation ($y=A_3 x / (B_3+x)$)

Linearization of Nonlinear Behavior (cont'd)

(1) Exponential Equation ($y = A_1 e^{B_1 x}$)



Example 25: Fit an exponential model to the following data set.

x	0.4	0.8	1.2	1.6	2.0	2.3
y	750	1000	1400	2000	2700	3750

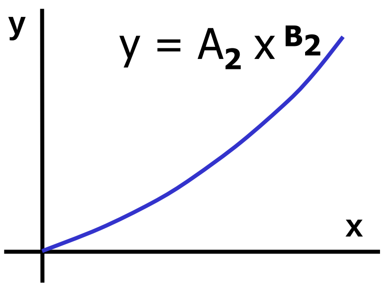
- Create the following table.

x	0.4	0.8	1.2	1.6	2.0	2.3
ln y	6.62	6.91	7.24	7.60	7.90	8.23

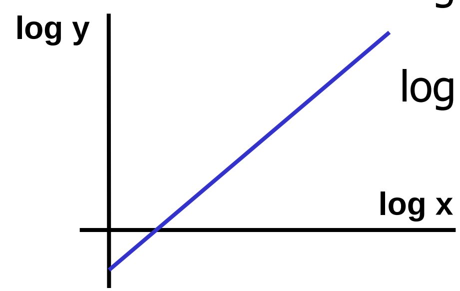
- Fit a straight line to this new data set. Be careful with the notation. You can define $z = \ln y$
- Calculate $a_0 = 6.25$ and $a_1 = 0.841$. Straight line is $\ln y = 6.25 + 0.841\,x$
- Switch back to the original equation. $A_1 = e^{a_0} = 518$, $B_1 = a_1 = 0.841$.
- Therefore the exponential equation is $y = 518\, e^{0.841x}$. Check this solution with couple of data points. For example $y(1.2) = 518\, e^{0.841\,(1.2)} = 1421$ or $y(2.3) = 518\, e^{0.841\,(2.3)} = 3584$. OK.

Linearization of Nonlinear Behavior (cont'd)

(2) Power Equation ($y = A_2\, x^{B_2}$)



Linearization



$\log y = \log A_2 + B_2 \log x$
or
$\log y = a_0 + a_1 \log x$

Example 26: Fit a power equation to the following data set.

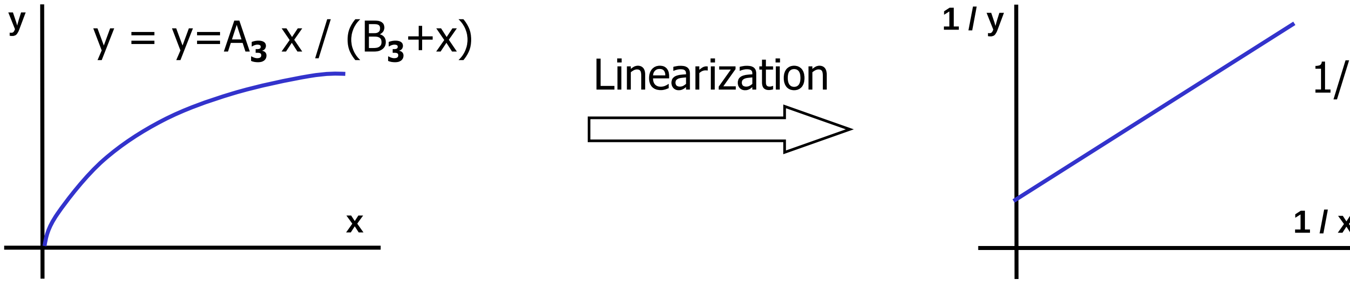
x	2.5	3.5	5	6	7.5	10	12.5	15	17.5	20
y	7	5.5	3.9	3.6	3.1	2.8	2.6	2.4	2.3	2.3

log x	0.398	0.544	0.699	0.778	0.875	1.000	1.097	1.176	1.243	1.301
log y	0.845	0.740	0.591	0.556	0.491	0.447	0.415	0.380	0.362	0.362

- Fit a straight line to this new data set. Be careful with the notation.
- Calculate $a_0 = 1.002$ and $a_1 = -0.53$. Straight line is $\log y = 1.002 - 0.53 \log x$
- Switch back to the original equation. $A_2 = 10^{a_0} = 10.05$, $B_2 = a_1 = -0.53$.
- Therefore the power equation is $y = 10.05\, x^{-0.53}$. Check this solution with couple of data points. For example $y(5) = 10.05 * 5^{-0.53} = 4.28$ or $y(15) = 10.05 * 15^{-0.53} = 2.39$. OK.

Linearization of Nonlinear Behavior (cont'd)

(3) Saturation-growth rate Equation ($y = A_3\, x \,/\, (B_3+x)$)



$1/y = 1/A_3 + B_3/A_3\, x$

or

$1/y = a_0 + a_1 x$

Example 27: Fit a saturation-growth-rate equation to the following data set.

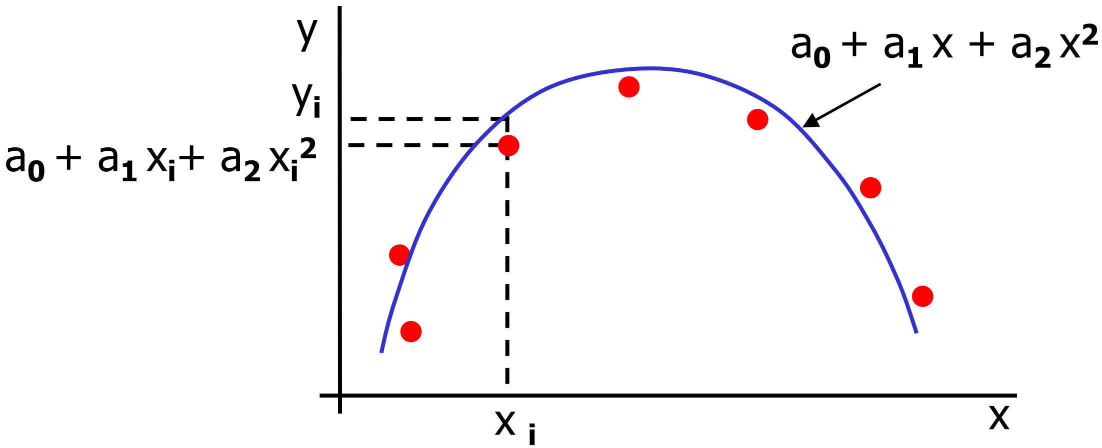
x	0.75	2	2.5	4	6	8	8.5
y	0.8	1.3	1.2	1.6	1.7	1.8	1.7

1 / x	1.333	0.5	0.4	0.25	0.1667	0.125	0.118
1 / y	1.25	0.769	0.833	0.625	0.588	0.556	0.588

- Fit a straight line to this new data set. Be careful with the notation.
- Calculate $a_0 = 0.512$ and $a_1 = 0.562$. Straight line is $1/y = 0.512 + 0.562\,(1/x)$
- Switch back to the original equation. $A_3 = 1/a_0 = 1.953$, $B_3 = a_1 A_3 = 1.097$.
- Therefore the saturation-growth rate equation is $1/y = 1.953\, x/(1.097+x)$. Check this solution with couple of data points. For example $y(2) = 1.953*2/(1.097+2) = 1.26$ OK.

Polynomial Regression (Extension of Linear Least Sqaures)

- Used to find a best-fit line for a nonlinear behavior.
- This is not nonlinear regression described at page 468 of the book. That section is omitted.



Example for a second order polynomial regression

$e_i = y_i - a_0 - a_1 x_i - a_2 x_i^2$

Error (deviation) for the i^{th} data point

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$ Sum of squares of the residuals

- Minimize this sum to get the normal equations $\frac{\partial S_r}{\partial a_0} = 0, \quad \frac{\partial S_r}{\partial a_1} = 0, \quad \frac{\partial S_r}{\partial a_2} = 0$

- Solve these equations with one of the techniques that we learned to get a_0, a_1 and a_2.

Polynomial Regression Example

- Find the least-squares parabola that fits to the following data set.

x	0	1	2	3	4	5
y	2.1	7.7	13.6	27.2	40.9	61.1

- Normal equations to find a least-squares parabola are

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

$$n = 6$$
$$\sum x_i = 15 \qquad \sum y_i = 152.6$$
$$\sum x_i^2 = 55 \qquad \sum x_i y_i = 585.6$$
$$\sum x_i^3 = 225 \qquad \sum x_i^2 y_i = 2488.6$$
$$\sum x_i^4 = 979$$

$\longrightarrow$

$$a_0 = 2.479, \quad a_1 = 2.359, \quad a_2 = 1.861$$
$$y = 2.479 + 2.359\,x + 1.861\,x^2$$
$$r^2 = \frac{S_t - S_r}{S_t} = \frac{2573.4 - 3.75}{2573.4} = 0.999$$
$$r = 0.999$$

Multiple Linear Regression

- $y = y(x_1, x_2)$
- Individual errors are $e_i = y_i - a_0 - a_1 x_{1i} - a_2 x_{2i}$
- Sum of squares of the residuals is $S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2$
- Minimize this sum to get the normal equations $\frac{\partial S_r}{\partial a_0} = 0, \quad \frac{\partial S_r}{\partial a_1} = 0, \quad \frac{\partial S_r}{\partial a_2} = 0$

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

- Solve these equations to get a_0, a_1 and a_2.

Example 28:

- Use multiple linear regression to fit

x	0	1	1	2	2	3	3	4	4
y	0	1	2	1	2	1	2	1	2
z	15	18	12.8	25.7	20.6	35	29.8	45.5	40.3

$\mathbf{n = 9}$

$\sum x_i = 20 \qquad \sum x_i y_i = 30$

$\sum x_i^2 = 60 \qquad \sum z_i = 242.7$

$\sum y_i = 12 \qquad \sum x_i z_i = 661$

$\sum y_i^2 = 20 \qquad \sum y_i z_i = 331.2$

$$\longrightarrow \begin{bmatrix} 9 & 20 & 12 \\ 20 & 60 & 30 \\ 12 & 30 & 20 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} 242.7 \\ 661 \\ 331.2 \end{Bmatrix}$$

$a_0 = 14.40, \quad a_1 = 9.03, \quad a_2 = -5.62$

$z = 14.4 + 9.03\,x - 5.62\,y$

Exercise 26: Calculate the standard error of the estimate ($s_{y/x}$) and the correlation coefficient (r).