Example

Fit a straight line to the x and y

x_i	1	2	3	4	5	6	7
y_i	0.5	2.5	2.0	4.0	3.5	6.0	5.5

Solution

$$y = a_0 + a_1 x \qquad (+ e)$$

Here

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

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

Since

$$n = 7$$

$$\sum x_i = 1 + 2 + \cdots + 7 = 28$$
$$\sum y_i = 0.5 + 2.5 + \cdots + 5.5 = 24$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{28}{7} = 4$$
$$\bar{y} = \frac{\sum y_i}{n} = \frac{24}{7} = 3.428571429$$

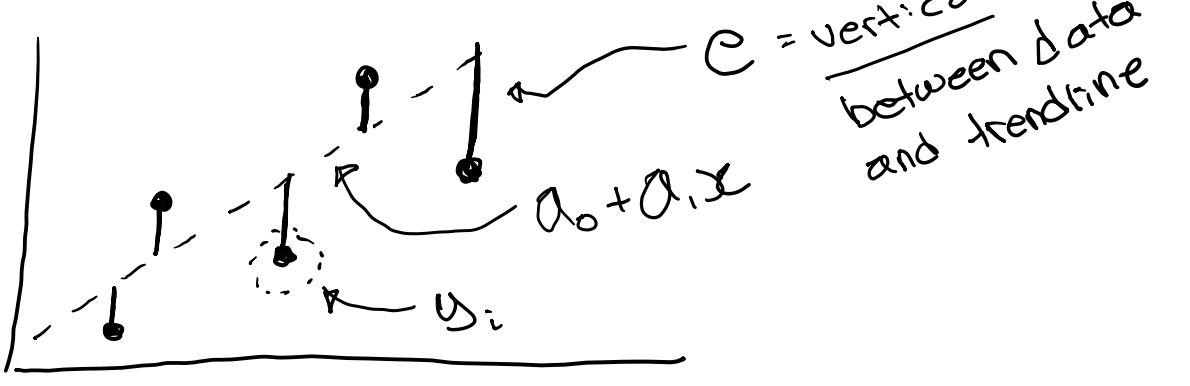
$$\sum x_i y_i = 1(0.5) + 2(2.5) + \cdots 119.5$$
$$\sum x_i^2 = 1^2 + 2^2 + \cdots + 7^2 = 140$$

Therefore

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

$$a_0 = (3.428571429) - (0.8392857)(4)$$
$$a_0 = 0.07142857$$

$$\therefore y = 0.07142857 + 0.8392857x$$



Estimate of the linear regression (error from the sampling data to the straight line):

$$S_r = \sum(y_i - a_0 - a_1 x_i)^2 \qquad which\ (= \sum e_i^2)$$

Under some conditions, the least squares regression will provide the best estimation of a_0 and a_1.
According to research found in:
Draper & Smith, 1981
Applied regression analysis

Standard error of the estimate (how spread out the data is around the best fit line):

$$s_{y|x} = \sqrt{\frac{S_r}{n-2}}$$

It quantifies the spread around the straight line.

For the data y_i, $i = 1, 2, 3, \dots, n$, define

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

Standard deviation (the quantified spread around the mean):

$$s_y = \sqrt{\frac{S_t}{n-1}}$$

Define the coefficient of determination:

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

r is called the correlation coefficient.

What does the value of r^2 represent:
$* \ 1st\ case: S_r = 0,\ r^2 = 1,$ all the data are on the straight line.
$* \ 2nd\ case: S_r = S_t,\ r^2 = 0,$ straight line fit represents no improvement (equal or worse result)

Another way to calculate r:

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

Example

Estimate the least-squares fit

x_i	1	2	3	4	5	6	7
y_i	0.5	2.5	2.0	4.0	3.5	6.0	5.5

Solution

$$\bar{y} = 3.428571429$$

$$S_t = \sum(y_i - \bar{y})^2$$
$$S_t = (0.5 - 3.428571429)^2 + (2.5 - 3.428571429)^2 + \cdots + (5.5 - 3.428571429)^2$$
$$S_t = 22.7143$$

$$a_1 = 0.8392857$$
$$a_0 = 0.07142857$$

$$S_r = \sum(y_i - a_0 - a_1 x_i)^2$$
$$S_r = \left(0.5 - 0.07142857 - 0.8392857(1)\right)^2 + \cdots + \left(5.5 - 0.07142857 - 0.8392857(7)\right)^2$$
$$S_r = 2.9911$$

$$r^2 = \frac{S_t - S_r}{S_t} = \frac{22.7143 - 2.9911}{22.7143} = 0.868$$

Then around 87% of the data can represented with a straight line – there's still some uncertainty.

Standard deviation (error from mean to data point)

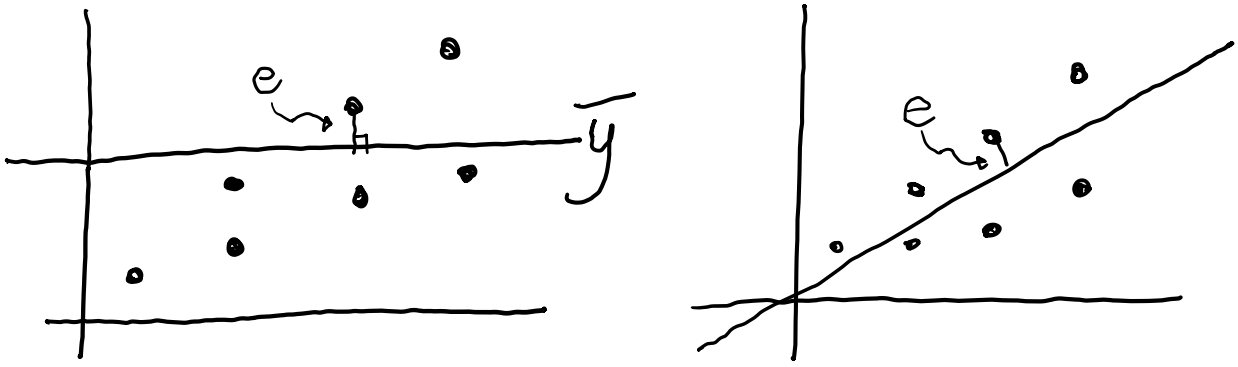
$$s_y = \sqrt{\frac{S_t}{n-1}} = \sqrt{\frac{22.7143}{7-1}} = 1.9457$$

Standard error (error from line of best fit to data point)

$$s_{y|x} = \sqrt{\frac{S_r}{n-2}} = \sqrt{\frac{2.9911}{7-1}} = 0.7735$$

$$s_y > s_{y|x}$$

Thus, straight line distribution is better than the average fit – consider the following diagram:



Linearization of Non-linear Relationships

Case 1

$$y = \alpha_1 e^{\beta_1 x}$$
$$\ln y = \ln\left(\alpha_1 + e^{\beta_1 x}\right)$$
$$\ln y = \ln \alpha_1 + \ln e^{\beta_1 x}$$
$$\ln y = \ln \alpha_1 + \beta_1 x$$

Thus,

$$a_0 = \ln \alpha_1$$
$$a_1 = \beta_1$$

Linearizing:

$$y = \ln y$$
$$x = x$$

Now:

$$y = a_0 + a_1 x$$

Thus, $\ln y$ and x are linearly related – we can get similar relationships in other cases.

Case 2

This is a typical power function:

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

Becomes:

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

Thus,

$$a_0 = \log \alpha_2$$
$$a_1 = \beta_2$$

Linearizing:

$$y = \log y$$
$$x = \log x$$

Case 3

These relationships are usually used for rates of change, in disciplines such as chemical engineering:

$$y = \alpha_3 \frac{x}{\beta_3 + x}$$

Becomes:

$$\frac{1}{y} = \frac{\beta_3 + x}{\alpha_3 x} = \frac{1}{\alpha_3} + \frac{\beta_3}{\alpha_3} \cdot \left(\frac{1}{x}\right)$$

Thus,

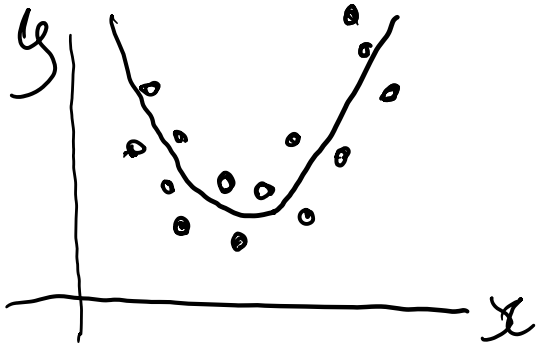
$$a_0 = 1/\alpha_3$$
$$a_1 = \beta_3/\alpha_3$$

Linearizing:

$$y = 1/y$$
$$x = 1/x$$

Polynomial Regression

Consider the following set of data:



Where the data cannot be represented by a linear line of best fit, so a second order polynomial (quadratic) line of best fit can be used.

The least-squares procedure to fit the data:

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

always exists when looking @ individual points

Define

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

The stationary conditions:

$$\frac{\delta S_r}{\delta a_0} = \sum 2\left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right) \cdot (-1) = 0$$
$$\frac{\delta S_r}{\delta a_1} = \sum 2\left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right) \cdot (-x_i) = 0$$
$$\frac{\delta S_r}{\delta a_2} = \sum 2\left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right) \cdot \left(-x_i^2\right) = 0$$

Consider:

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

$$\sum \left(a_0 x_i + a_1 x_i^2 + a_2 x_i^2 - x_i y_i\right) = 0$$
$$(\sum x_i)a_0 + \left(\sum x_i^2\right)a_1 + \left(\sum x_i^3\right)a_2 = \sum x_i y_i \quad **$$

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

Note: As long as at least two x_i are different, you can find a unique solution – they can't all be the same!

The standard error:

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

Where n is the number of data points
Where m is the degree of the polynomial