Real-World Application:
The Line Of Best Fit

Given a set of p data points, $(x_1, y_1), (x_2, y_2), \ldots, (x_p, y_p)$, we would like to find the equation of the line

$$y = mx + b$$

that "best" describes this data. If the data were exactly linear, we would only need two data points to determine the line- The problem is, there is some "error"- No line will exactly go through all the data. Algebraically, we are saying that for any choice of m, b, it will probably be the case that:

$$y_i \neq mx_i + b$$

Therefore, the error taken at x_i is typically taken to be

$$(y_i - (mx_i + b))^2 = (y_i - mx_i - b)^2$$

Sum these errors together, and this is our error function:

$$E(m, b) = (y_1 - mx_1 - b)^2 + (y_2 - mx_2 - b)^2 + \ldots + (y_p - mx_p - b)^2$$

This function depends on two inputs, m, b. We can simplify this to a function of one variable by the following:

Define $\overline{x}$ to be the mean (or average) of the x's, and $\overline{y}$ to be the mean of the y's. It can be shown that the intercept will be:

$$b = \overline{y} - m\overline{x}$$

Now take $\hat{x}_i = x_i - \overline{x}$ and $\hat{y} = y_i - \overline{y}$ (this is called mean-subtracting the data). Then our new error function is:

$$E(m) = (\hat{y}_1 - m\hat{x}_1)^2 + (\hat{y}_2 - m\hat{x}_2)^2 + \ldots + (\hat{y}_p - m\hat{x}_p)^2$$

The minimum error is found by differentiating the error, setting the result to zero and solve for the slope:

$$\frac{dE}{dm} = 2(\hat{y}_1 - m\hat{x}_1)(-\hat{x}_1) + 2(\hat{y}_2 - m\hat{x}_2)(-\hat{x}_2) + \ldots + 2(\hat{y}_p - m\hat{x}_p)(-\hat{x}_p) = 0$$

Divide by 2, expand the result and isolate the slope:

$$m\left(\hat{x}_1^2 + \hat{x}_2{}^2 + \ldots \hat{x}_p^2\right) = \hat{x}_1\hat{y}_1 + \hat{x}_2\hat{y}_2 + \ldots + \hat{x}_p\hat{y}_p$$

$$m = \frac{\hat{x}_1\hat{y}_1 + \hat{x}_2\hat{y}_2 + \ldots + \hat{x}_p\hat{y}_p}{\hat{x}_1^2 + \hat{x}_2{}^2 + \ldots \hat{x}_p^2}$$

Example: Find the line of best fit through the points:

x	-1	0	1	2	3
y	-4.9	-2.0	1.4	4.3	6.8

SOLUTION:

First,

$$\overline{x} = \frac{-1+0+1+2+3}{5} = 1 \qquad \overline{y} = \frac{-4.9-2+1.4+4.3+6.8}{5} = 1.12$$

so once we find the slope m, the intercept $b = 1.12 - m(1)$.

Mean-subtract the data, then compute the slope:

x	$\hat{x}$	y	$\hat{y}$	$\hat{x}\hat{y}$	$\hat{x}^2$
-1	-2	-4.9	-6.02	12.04	4
0	-1	-2.0	-3.12	3.12	1
1	0	1.4	0.28	0	0
2	1	4.3	3.18	3.18	1
3	2	6.8	5.68	11.36	4

The sum of the column $\hat{x}\hat{y}$ is 29.7 and the sum of $\hat{x}^2$ is 10. This gives the slope and then the intercept:

$$m = \frac{29.7}{10} = 2.97 \qquad b = 1.12 - 2.97 = -1.85$$

The line of best fit is $y = 2.97x - 1.85$. We might also note that the actual error function in this case is:

$$E(m) = (-6.02+2m)^2 + (-3.12+m)^2 + (0.28-0)^2 + (3.18-m)^2 + (5.68-2m)^2$$

If we were to expand and simplify, the graph of E is a parabola.

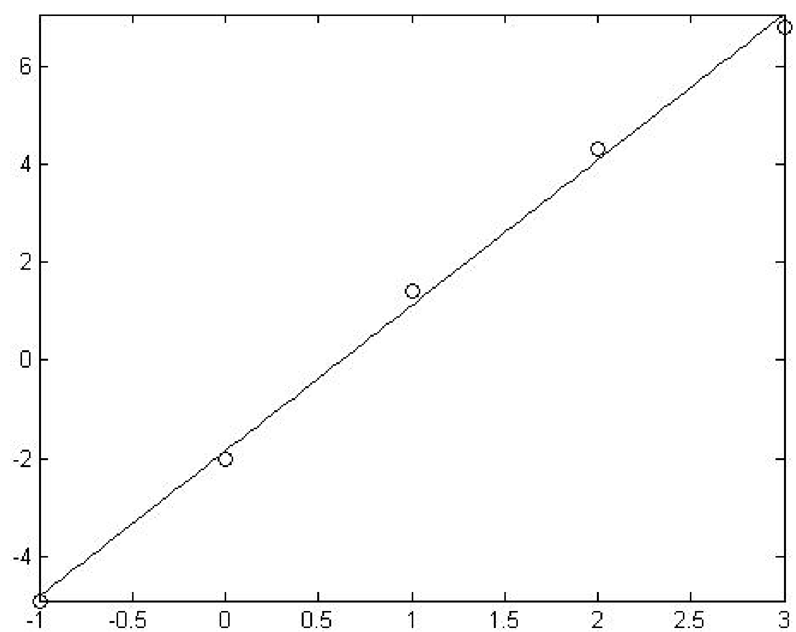

Figure 1: The data and the line of best fit.