# The Least Squares Regression Model

The famous German mathematician Carl Friedrich Gauss had investigated the method of least squares as early as 1794, but unfortunately he did not publish the method until 1809. In the meantime, the method was discovered and published in 1806 by the French mathematician Legendre, who quarrelled with Gauss about who had discovered the method first (Reid, 2000).

The basic idea of the method of least squares is easy to understand. It may seem unusual that when several people measure the same quantity, they usually do not obtain the same results. In fact, if the same person measures the same quantity several times, the results will vary. *What then is the best estimate for the true measurement?*

The method of least squares gives a way to find the best estimate, assuming that the errors (i.e. the differences from the true value) are *random and unbiased*. Let us consider a simple example.

**Problem: Suppose we measure a distance four times, and obtain the following results:**
**72, 69, 70 and 73 units**
**What is the best estimate of the correct measurement?**

Let us denote the estimate of the true measurement by $x$, and form the deviations (errors) from $x$, namely $x - 72$, $x - 69$, $x - 70$, and $x - 73$.

Let S be the sum of the squares of these errors, i.e.

$$S = (x - 72)^2 + (x - 69)^2 + (x - 70)^2 + (x - 73)^2.$$

We seek the value of $x$ that minimises the value of S. We can write S in the equivalent form

$$S = 4(x - 71)^2 + 10$$

We can see from this form (or we can use calculus) that the minimum value of S is 10, when $x = 71$.

So the best estimate of the true measurement is 71 units!

Note that 71 m is the *mean* or average of the original four measurements. It is always true that for $n$ measurements the minimum value of S occurs when $x$ equals the mean of the $n$ measurements. Can you prove this?

### The line of best fit

The RCS requires learners to ***estimate*** the line of best fit for a set of ordered pairs. That is not very useful, because predictions based on this model will be very vague!

The method of least squares *calculates* the line of best fit by minimising the sum of the squares of the vertical distances of the points to the line. Let's illustrate with a simple example.

**Problem: Given these measurements of the two quantities $x$ and $y$, find $y_7$:**

| | | | | | | |
|---|---|---|---|---|---|---|
| $x_1 = 2$ | $x_2 = 4$ | $x_3 = 6$ | $x_4 = 8$ | $x_5 = 10$ | $x_6 = 12$ | $x_7 = 14$ |
| $y_1 = 2$ | $y_2 = 4$ | $y_3 = 4$ | $y_4 = 5$ | $y_5 = 5$ | $y_6 = 7$ | $y_7 = ?$ |

Due to random errors in the measurements the ordered pairs $(x_i, y_i)$ do not lie on a straight line. Assume the values can be approximated by the linear function $y' = ax + b$, Let us call the deviations (errors) $d_i = y'_i - y_i = ax_i + b - y_i$ for $i = 1, 2, \ldots, 6$.

Let's solve the problem ***algebraically*** by finding the sum of the squares of the errors and minimising it:

$$S = d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$$
$$= (2a + b - 2)^2 + (4a + b - 4)^2 + (6a + b - 4)^2 + (8a + b - 5)^2 + (10a + b - 5)^2 + (12a + b - 7)^2$$
$$= 364a^2 + 84ab + 6b^2 - 436a - 54b + 183$$

We now find the partial derivative of S with respect to $a$. This means that we differentiate S with respect to $a$, and treat $b$ as if it was a constant. As with one variable, we set the derivative equal to zero. This gives

$$182a + 21b = 109 \quad \text{..... (1)}$$

We also find the partial derivative of S with respect to $b$, differentiating S with respect to $b$, and treat $a$ as if it was a constant, and set the derivative equal to zero. This gives

$$14a + 2b = 9 \quad \text{..... (2)}$$

Solving (1) and (2) we have[1] $a = \frac{29}{70}$ and $b = \frac{56}{35}$. So the line of best fit is $y = \frac{29}{70}x + \frac{56}{35}$

We can now use this *model* to find unknown information, e.g. $y_7 = y(15) = \frac{29}{70} \times 14 + \frac{56}{35} = 7{,}4$

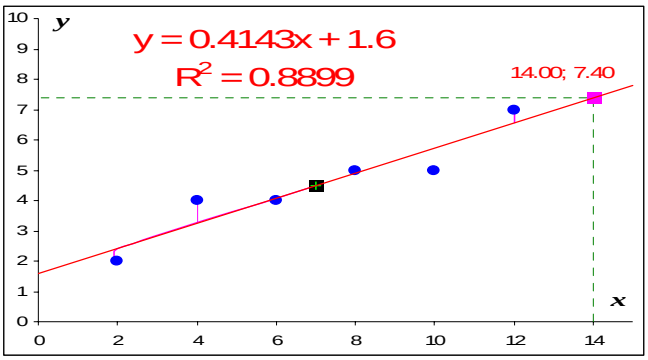
We can solve the least errors problem ***numerically*** through trial-and improvement by *systematically* varying a and b until $S = \sum (y - y')^2$ is a minimum – technology helps![2]

We can find the line of best fit "***graphically***" by using a technology curve-fitting program[3], e.g. by using Excel's Trendline. Excel in essence *calculates* a and b using these *formulae*:

$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$, and $a = \bar{y} - b\bar{x}$

Here is a table and Trendline produced in Excel:

| DATA | | MODEL |
|---|---|---|
| $x$ | $y$ | $y' = ax+b$ |
| 2 | 2 | 2.42857 |
| 4 | 4 | 3.25714 |
| 6 | 4 | 4.08571 |
| 8 | 5 | 4.91429 |
| 10 | 5 | 5.74286 |
| 12 | 7 | 6.57143 |
| 14 | ?? | **7.4** |




Note that the point $(\bar{x}; \bar{y}) = (7; 4.5)$, i.e. the ordered pair formed by the mean of the $x$ values and the mean of the $y$ values lies on the line of best fit. This is always the case.

**Theorem:** The Least Squares Model for a set of data $(x_1, y_1)$, $(x_2, y_2)$,..., $(x_n, y_n)$ passes through the point $(x_a, y_a)$ where $x_a$ is the average of the $x_i$'s and $y_a$ is the average of the $y_i$'s.

Using this theorem enables us to simplify our method of calculating the parameters $a$ and $b$ of our line of best fit with formula $y' = ax + b$.

---

[1] I use exact values expressed as fractions to avoid the fatal error of entering rounded values into the calculator! For example, $y = 0{,}41x + 1{,}6 \Rightarrow y(15) = 7{,}34$ and $y = 0{,}414x + 1{,}6 \Rightarrow y(15) = 7{,}396$. *Postpone calculation to the final step!*

[2] An Excel workbook for this example is available online at http://www.sun.ac.za/mathed/LeastSquares.xls

[3] Curve Expert is a very good free program available at http://curveexpert.webhop.net/

Calculate the averages of the x values and of the y-values gives the point (7; 4,5). We know that this point lies on the line of best fit and therefore the ordered pair satisfies the equation:
4,5 = 7a + b
$\Rightarrow$ b = 4,5 – 7a
So we can now express the equation of the line of best fit only in terms of the gradient a:
$y' = ax + 4{,}5 - 7a$ ..... (2)

We can now use this equation to calculate the errors, then minimising the sum of the squares of the errors will give the gradient of the line of best fit.

| **DATA** | | **MODEL** | | |
|---|---|---|---|---|
| $x$ | $y$ | $y' = ax+b$ | $y - y'$ | $(y - y')^2$ |
| 2 | 2 | 4,5 – 5a | 5a – 2,5 | $25a^2 - 25a + 6{,}25$ |
| 4 | 4 | 4,5 – 3a | 3a – 0,5 | $9a^2 - 3a + 0{,}25$ |
| 6 | 4 | 4,5 – a | a – 0,5 | $a^2 - a + 0{,}25$ |
| 8 | 5 | 4,5 + a | 0,5 – a | $a^2 - a + 0{,}25$ |
| 10 | 5 | 4,5 + 3a | 0,5 – 3a | $9a^2 - 3a + 0{,}25$ |
| 12 | 7 | 4,5 + 5a | 2,5 – 5a | $25a^2 - 25a + 6{,}25$ |
| 14 | **??** | | **Sum:** | $70a^2 - 58a + 13{,}5$ |

$$a = \frac{58}{140} = \frac{29}{70}$$

In other words, for $a = \frac{29}{70}$ the sum of the squared errors is a minimum, and therefore the line with gradient $a = \frac{29}{70}$ fits the data the best. Substituting into equation (2) we there have:

$y' = \frac{29}{70}x + 1{,}6$ and

$\Rightarrow y'(14) = \frac{29}{70} \times 14 + 1{,}6 = 7{,}4$

## References:


Reid, F. (2000). The Mathematician on the Banknote: Carl Friedrich Gauss. **Parabola, 36(2)**, 2-9.

MultiChoice Africa Foundation FET Maths Educator Programme, Unit 20 [CD-ROM].



Alwyn Olivier
*Stellenbosch, June 2007*