Introduction to Data Analysis and Error Handling: Part 3

The Mean as the Best "Least Squares" Fit

Suppose $x_1, x_2 \ldots x_N$ are N measurements of a quantity with a true value of X. The measurements then have errors $\varepsilon_i = x_i - X$, and the average sum of their squares is

$$S = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - X)^2.$$

If we treat the true value X as an unknown quantity, then the *Principle of Least Squares* says that the best estimate of the value X is that which minimises the sum of the squares of the errors. In other words, we minimise $S(X)$ with respect to X so that

$$0 = \frac{\partial S}{\partial X} = -\frac{2}{N}\sum_{i=1}^{N}(x_i - X) = -\frac{2}{N}\sum_{i=1}^{N}x_i + 2X,$$

from which it follows that

$$X = \frac{1}{N}\sum_{i=1}^{N}x_i = \langle x \rangle.$$

The best estimate according to the Principle of Least Squares is therefore the mean value of the data, which we know to be true from estimation theory.

Linear Regression

We often perform experiments in which we vary one parameter x and measure a second parameter y, where we expect the data to fit to a straight line

$$y = ax + b.$$

The gradient, a, and intercept, b, are then physical parameters which we want to estimate.

Suppose that we have a set of N values of x_i with corresponding measured values y_i. The value of y_i expected from the linear plot is $ax_i + b$, so the errors between theory and measurement are

$$\varepsilon_i = y_i - ax_i - b.$$

The Principle of Least Squares then says that we should minimise the sum of the squares,

$$S(a,b) = \frac{1}{N}\sum_{i=1}^{N}(y_i - ax_i - b)^2,$$

with respect to the parameters a and b. This leads us to solve the equations

$$0 = \frac{\partial S}{\partial a} = -\frac{2}{N}\sum_{i=1}^{N}x_i(y_i - ax_i - b) \quad \rightarrow \quad \langle xy \rangle = a\langle x^2 \rangle + b\langle x \rangle$$

$$0 = \frac{\partial S}{\partial b} = -\frac{2}{N}\sum_{i=1}^{N}(y_i - ax_i - b) \quad \rightarrow \quad \langle y \rangle = a\langle x \rangle + b,$$

from which it follows that

$$a = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\langle x^2 \rangle - \langle x \rangle^2}, \qquad b = \frac{\langle x^2 \rangle \langle y \rangle - \langle x \rangle \langle xy \rangle}{\langle x^2 \rangle - \langle x \rangle^2}.$$

The quantities used in the above formulae are defined by

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i, \qquad \langle y \rangle = \frac{1}{N} \sum_{i=1}^{N} y_i, \qquad \langle x^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i^2, \qquad \langle xy \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i y_i.$$

If we define the *uncorrected sample standard deviations*, s_x, s_y, and *sample covariance*, s_{xy}, by

$$s_x^2 = \langle x^2 \rangle - \langle x \rangle^2, \qquad s_y^2 = \langle y^2 \rangle - \langle y \rangle^2, \qquad s_{xy} = \langle xy \rangle - \langle x \rangle \langle y \rangle,$$

then the equation of the line may be written as

$$y - \langle y \rangle = a(x - \langle x \rangle) = \frac{s_{xy}}{s_x^2}(x - \langle x \rangle).$$

This may finally be written in the more symmetric form

$$\frac{y - \langle y \rangle}{s_y} = r_{xy} \frac{x - \langle x \rangle}{s_x} \qquad \text{where} \qquad r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

r_{xy} is known as the *sample correlation coefficient*, and is a measure of how well the data x and y are correlated.

Errors in Regression Coefficients

We have assumed that the only errors are in the measurement of the quantities y_i, with the measurement of x_i having no error. Since the equation for a may be written as

$$a = \frac{1}{N s_x^2} \sum_{i=1}^{N} (x_i - \langle x \rangle) \, y_i,$$

it follows that the error in a is related to the error in the y_i by

$$\sigma^2(a) = \frac{1}{N^2 s_x^4} \sum_{i=1}^{N} (x_i - \langle x \rangle)^2 \, \sigma^2(y) = \frac{1}{N s_x^2} \sigma^2(y).$$

Similarly, the equation for b may be written as

$$b = \frac{1}{N s_x^2} \sum_{i=1}^{N} \big(\langle x^2 \rangle - \langle x \rangle x_i \big) \, y_i,$$

from which it follows that

$$\sigma^2(b) = \frac{1}{N^2 s_x^4} \sum_{i=1}^{N} \left(\langle x^2 \rangle - \langle x \rangle x_i\right)^2 \sigma^2(y) = \frac{\langle x^2 \rangle}{N s_x^2} \sigma^2(y) = \langle x^2 \rangle \sigma^2(a).$$

The standard deviation in y may be estimated from the sum of squares,

$$\begin{aligned}
\sigma^2(y) &= \frac{1}{N} \sum_{i=1}^{N} (y_i - a x_i - b)^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} \Big[(y_i - \langle y \rangle) - a(x_i - \langle x \rangle) \Big]^2 \\
&= s_y^2 - 2 a s_{xy} + a^2 s_x^2 \\
&= s_y^2 - \frac{s_{xy}^2}{s_x^2}.
\end{aligned}$$

It follows that the error in the slope is given by

$$\sigma^2(a) = \frac{s_y^2}{N s_x^2} \left(1 - \frac{s_{xy}^2}{s_x^2 s_y^2} \right) = \frac{s_y^2}{N s_x^2} \left(1 - r_{xy}^2 \right).$$

Exercise 1: *Eight measurements of the volume of a block of iron had a mean value of* $26.52 cm^3$ *and a mean square deviation* $0.025 cm^6$. *Fifteen measurements of the volume of a block of aluminium gave the corresponding results* $8.72 cm^3$ *and* $0.058 cm^6$. *If the densities of iron and aluminium are* $7.88 gcm^{-3}$ *and* $2.70 gcm^{-3}$, *respectively, what is the best estimate of the total mass of the two blocks and the corresponding error in the measurement?*

Exercise 2: *Six measurements of the length of a wire had a mean value of* $527.3 cm$ *with mean square deviation* $0.01 cm^2$. *Twelve measurements of its diameter had a mean value of* $0.062 cm$ *with a mean square deviation of* $1.2 \times 10^{-6} cm^2$. *If the resistivity is known to be* $44.2 \times 10^{-6}\,\Omega cm$, *what is the best estimate of the resistance of the wire and the corresponding error in the measurement?*

Exercise 3: *The following table shows average life expectancy at age 10 in the United Kingdom versus year:*

Year	*Life Expectancy*
1950	61.84
1955	62.69
1960	63.00
1965	63.42
1970	63.73
1975	64.24
1980	65.14
1985	65.81
1990	66.86
1995	67.67
2000	68.89
2005	70.14
2010	70.88
2015	71.62

Find the line of best fit, $y = ax + b$, *to this data. Determine the sample correlation coefficient,* r_{xy}, *and the errors* $\sigma(a)$ *and* $\sigma(b)$ *on the regression parameters. Plot the data and the line of best fit.*