

# Empirical modelling

Note that real-world or experimental data have important purposes. Data are needed for testing a model. We often need data to estimate various model parameters. We can also use data to suggest the right model. An empirical model is a mathematical representation driven by real-world or experimental data and observation rather than underlying physical or theoretical mechanisms.

For simplicity, we consider only models in which the output variable depends on a single input variable. Suppose we have a data collection in the form  $(x_i, y_i)$ , where  $i = 1, 2, \dots, n$ . Here,  $y_i$  is a dependent variable (often called the outcome or response variable), and  $x_i$  is an independent variable (often called the predictor or input variable). We want to fit a function  $\hat{y} = f(x)$  through these data. One choice is interpolation, in which we find a function that passes through all the data points. However, interpolation does not perform well when the number of data points increases. Hence, we choose model fitting, in which we find a function that best fits all the data points. Below, we describe some model-fitting techniques.

**Model fitting using least squares:** For simplicity, we first consider the simple linear least squares. Suppose that we have data points  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . Consider the model

$$\hat{y} = a + bx$$

that best fit the data (in the sense to be defined below). The errors (residuals) are given by

$$\varepsilon_i = y_i - \hat{y}_i = y_i - a - bx_i$$

The sum of the squared errors is given by

$$S(a, b) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

To find the best fit for the data, we find the values of  $a$  and  $b$  such that  $S(a, b)$  is minimized. This can be done by taking partial derivatives with respect to  $a$  and  $b$  and setting them to zero:

$$\frac{\partial S}{\partial a} = 0 \quad \text{and} \quad \frac{\partial S}{\partial b} = 0,$$

which lead to the normal equations:

$$\begin{aligned} na + b \sum x_i &= \sum y_i \\ a \sum x_i + b \sum x_i^2 &= \sum x_i y_i \end{aligned}$$

Solving the system of normal equations, we first find the value of  $b$  as follows:

$$b = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\frac{1}{n} \sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{n} \sum x_i^2 - (\bar{x})^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Using  $b$  in the first equation, we find

$$a = \frac{\sum y_i - b \sum x_i}{n} = \bar{y} - \bar{x}b.$$

Here  $\bar{x}$  and  $\bar{y}$  denote the averages of the  $x$  and  $y$  values. We now show that  $S$  attains the minimum. For this, we calculate

$$\frac{\partial^2 S}{\partial a^2} = 2n, \quad \frac{\partial^2 S}{\partial b^2} = 2 \sum_i x_i^2, \quad \frac{\partial^2 S}{\partial a \partial b} = 2 \sum_i x_i$$

Hence,

$$\frac{\partial^2 S}{\partial a^2} \frac{\partial^2 S}{\partial b^2} - \left( \frac{\partial^2 S}{\partial a \partial b} \right)^2 = 4 \left( n \sum_i x_i^2 - \left( \sum_i x_i \right)^2 \right) = 4n \sum_i (x_i - \bar{x})^2 \geq 0$$

The above expression is equal to zero only when  $x_i = \bar{x}$  for all  $i = 1, 2, \dots, n$ , i.e., all the output data  $y_i$  are at the single input data, which is not interesting. Hence, the above expression is strictly positive and  $\frac{\partial^2 S}{\partial a^2} > 0$ . Consequently,  $S(a, b)$  indeed attains the minimum at  $a$  and  $b$  derived from the normal equations.

Here are a few observations about simple linear least squares.

i. Note that the first normal equation is

$$0 = \sum_{i=1}^n (y_i - a - bx_i) = \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n \varepsilon_i$$

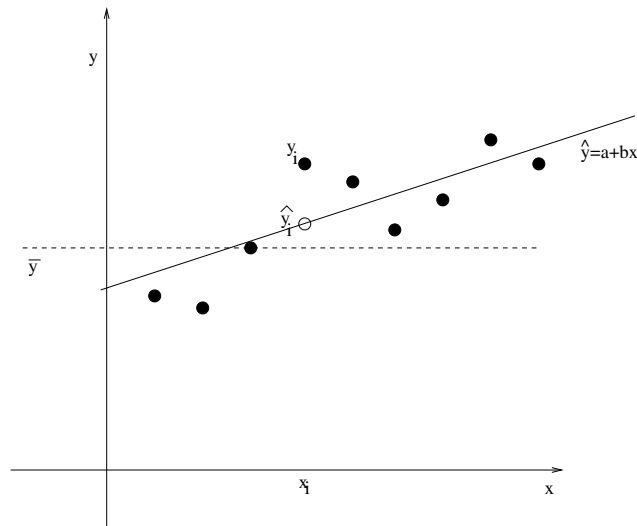
Hence,

$$\sum_{i=1}^n (y_i - \hat{y}_i) = 0 \implies \bar{y} = \bar{\hat{y}}$$

Consequently, the mean of the observed data is the same as that of the data predicted by the fitted curve.

ii. With reference to the figure, we see that

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$



Thus, we find that

$$\text{Total deviation} = \text{Deviation due to the regression} + \text{Deviation due to the error}$$

Now we define total sum of squares (SST), sum of squares due to regression (SSR), and sum of squares of error/residuals (SSE), as follows:

$$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad \text{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad \text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Now we show that  $\text{SST} = \text{SSR} + \text{SSE}$ . To show this, we proceed as follows:

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y} + y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \end{aligned}$$

Now

$$\begin{aligned} \sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) &= \sum_{i=1}^n (a + bx_i - \bar{y})\varepsilon_i \\ &= 0 + b \sum_{i=1}^n \varepsilon_i x_i + 0 \\ &= b \sum_{i=1}^n (y_i - a - bx_i) x_i = 0 \end{aligned}$$

The expression in the last line is  $\frac{\partial S}{\partial b}$ , which vanishes. Hence, we arrive at

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \implies \text{SST} = \text{SSR} + \text{SSE}$$

iii. The coefficient of determination, also called R-squared, is denoted by  $R^2$ . It is defined by

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

Clearly,  $0 \leq R^2 \leq 1$ . A perfect linear relationship exists in the data when  $R^2 = 1$ , where 100% variation in  $y$  is explained by variation in  $x$ . When  $0 < R^2 < 1$ , we have a weaker linear relationship between  $x$  and  $y$ , where some but not all the variation in  $y$  can be explained by variation in  $x$ . Finally, no linear relationship between  $x$  and  $y$  when  $R^2 = 0$ . Here, none of the variation in  $y$  is explained by the variation in  $x$ .

iv. Note that

$$\begin{aligned}
R^2 &= \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\sum_{i=1}^n (a + bx_i - a - b\bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= b^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \left( \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \right)^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \left( \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \right)^2 \\
&= r^2,
\end{aligned}$$

where  $r$  is Pearson's correlation coefficient.

**Least squares power fit:** Suppose we want to fit a curve  $\hat{y} = ax^n$  thorough the points  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ . We want to minimize

$$S(a) = \sum_{i=1}^n (y_i - ax_i^n)^2$$

Taking the derivative with respect to  $a$  and setting it to zero, we find

$$\frac{\partial S}{\partial a} = 2 \sum_{i=1}^n (y_i - ax_i^n)(-x_i^n) = 0 \implies a = \frac{\sum_i x_i^n y_i}{\sum_i x_i^{2n}}$$

**Least squares logarithmic fit:** Give data points  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , we want to fit the function  $\hat{y} = a + b \log x$ . Provided all the  $x_i > 0$ , we write  $X_i = \log x_i$ . Then it becomes a linear least-squares problem over the data points  $(X_i, y_i)$ ,  $i = 1, 2, \dots, n$ . Then

$$b = \frac{n \sum (\log x_i) y_i - (\sum \log x_i) (\sum y_i)}{n \sum (\log x_i)^2 - (\sum \log x_i)^2}$$

Consequently,  $a$  is found from

$$a = \frac{\sum y_i - b \sum \log x_i}{n}$$

**Transformed least squares fit:** Sometimes we modify the data in order to use the technique of linear regression.

i. Consider the exponential model  $\hat{y} = ae^{bx}$  for the data points  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ .

Taking log in both sides, we find

$$\log \hat{y} = \log a + bx = A + bx$$

Hence, we fit a linear model for  $(x_i, \log y_i)$ ,  $i = 1, 2, \dots, n$  and we find

$$b = \frac{n \sum x_i (\log y_i) - (\sum x_i) (\sum \log y_i)}{n \sum x_i^2 - (\sum x_i)^2}$$

Then  $a$  is found from

$$\log a = A = \frac{\sum \log y_i - b \sum x_i}{n}$$

- ii. Consider the power model  $\hat{y} = ax^b$  for the data points  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ . Taking log in both sides, we find

$$\log \hat{y} = \log a + b \log x = A + b \log x$$

Hence, we fit a linear model for  $(X_i, Y_i) = (\log x_i, \log y_i)$ ,  $i = 1, 2, \dots, n$ , and we find

$$b = \frac{n \sum (\log x_i)(\log y_i) - (\sum \log x_i)(\sum \log y_i)}{n \sum (\log x_i)^2 - (\sum \log x_i)^2}$$

Then  $a$  is found from

$$\log a = A = \frac{\sum \log y_i - b \sum \log x_i}{n}$$

**Note:** Instead of the transformation, if we had used least squares on the original formulation, then our problem is to minimize

$$S(a, b) = \sum_i (y_i - ae^{bx_i})^2$$

Now taking derivative with respect to  $a$  and  $b$ , we find

$$\begin{aligned} \sum_i (y_i - ae^{bx_i})e^{bx_i} &= 0 \\ \sum_i (y_i - ae^{bx_i})x_i e^{bx_i} &= 0 \end{aligned}$$

Since the above system is nonlinear in  $a$  and  $b$ , we need to use a nonlinear solver. Also, the values of  $a$  and  $b$  obtained from the nonlinear solver in general are not equal to those obtained from the transformed least squares equations.

**Curve fitting using the Chebyshev approximation criterion:** This is also known as the minimax or uniform norm criterion that seeks an approximation function that minimizes the maximum absolute deviation (the worst-case error). Given  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , we need to fit  $\hat{y} = f(x)$  such that the maximum absolute value of  $y_i - f(x_i)$  is minimized. Let

$$r_i = y_i - f(x_i) \quad \text{for } i = 1, 2, \dots, n \quad \text{and} \quad r = \max_{1 \leq i \leq n} |r_i|.$$

Our problem is to minimize  $r$  subject to  $|r_i| \leq r$  for  $i = 1, 2, \dots, n$ , i.e.

Minimize  $r$

subject to

$$r - r_i \geq 0 \quad \text{and} \quad r + r_i \geq 0 \quad \text{for } i = 1, 2, \dots, n.$$

Thus, it becomes a linear programming problem. The example given below is solved graphically.

**Example:** We use the Chebyshev criterion to fit the model  $\hat{y} = cx$  through  $(1, 2), (2, 5), (3, 8)$ .

Here, the problem becomes minimize  $r$  subject to

$$r - (2 - c) \geq 0, \quad (1)$$

$$r + (2 - c) \geq 0, \quad (2)$$

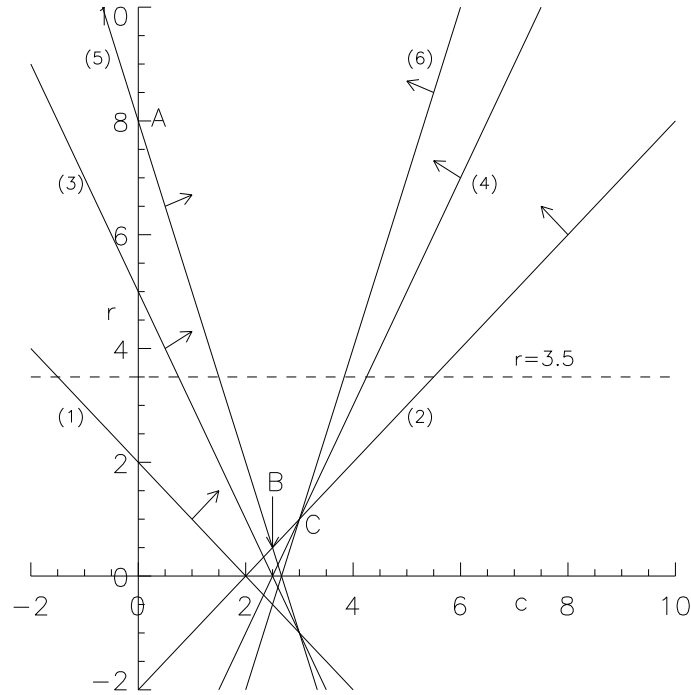
$$r - (5 - 2c) \geq 0, \quad (3)$$

$$r + (5 - 2c) \geq 0, \quad (4)$$

$$r - (8 - 3c) \geq 0, \quad (5)$$

$$r + (8 - 3c) \geq 0 \quad (6)$$

The straight lines corresponding to these inequalities (constraints) are shown in the figure. The feasible solution (i.e., the set of values for the decision variables that satisfies all the problem's constraints) is bounded by the line segment BC and the straight lines corresponding to (5) and (6) above B and C, respectively.



Note that the line  $r = 3.5$  is parallel to the  $c$ -axis and a part of it lies in a feasible solution. Since we want to minimise  $r$ , we can move below  $r = 3.5$  while remaining within the feasible solution. Clearly, this line of constant  $r$  exits the feasible solution at  $B$ . Thus, the minimum value is attained at  $B$ , which is the intersection of the lines corresponding to (2) and (5). The coordinate of  $B$  is  $(c = \frac{5}{2}, r = \frac{1}{2})$ . Hence, the fitted line is  $\hat{y} = \frac{5}{2}x$ .