Summary: Least squares approximation

Given a set of data points, (x_i, y_i) the simplest possible relationship we can hope to find between the data is a linear one. That is, we want to find a line $y = ax + b$ that best fits these data. Our "variables" are now the a and b, the slope and y-intercept that determine the linear relationship.

In order to find the best possible a and b, we minimize an error function. The error function we defined squares the difference between the y_i to the predicted value $ax_i + b$:

$$D(a, b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2.$$

Thus finding the best fit requires minimizing this function of two variables a and b, which is an (unconstrained) optimization problem!

Mechanics

To fit a collection of data (x_i, y_i) to a straight line that minimizes the squares of the differences between the predicted y values and the actual y_i, solve the system:

$$\left(\sum_{i=1}^{n} x_i^2\right) a + \left(\sum_{i=1}^{n} x_i\right) b = \sum_{i=1}^{n} x_i y_i \tag{1}$$

$$\left(\sum_{i=1}^{n} x_i\right) a + nb = \sum_{i=1}^{n} y_i. \tag{2}$$

where n is the number of data points.

- To fit your data (x_i, y_i) to a power law $y = cx^p$, first transform your data $(X_i, Y_i) = (\ln x_i, \ln y_i)$, then use least squares approximation. The power p will be the slope.
- To fit your data (x_i, y_i) to an exponential law $y = ce^{kx}$, first transform your data $(X_i, Y_i) = (x_i, \ln y_i)$. Then use least squares approximation. The exponential multiple k is the slope of the least squares fitting to the transformed data.

Ask Yourself

Why is it called least-squares approximation?

What is the point of drawing a straight line through data?