In this lecture, we will look at a method for estimating parameters which is applicable in certain key situations in which the maximum likelihood principle is *not* applicable. Specifically, we will consider situations in which we are dealing with **two data sets**: a data set $(x_1, x_2, \ldots, x_n)$ which is viewed as **not random**, and a data set $(y_1, y_2, \ldots, y_n)$ which we know may be **subject to variability and noise**. We are interested in finding **correlations** between the two data sets, and in particular **estimating the parameters** which would correspond to the best **linear correlation** (linear regression) or the best **parabolic correlation** (quadratic regression). In this class, we will not go beyond these simple correlations.

The method we use to calculate the parameters for a best fit is called the **method of least squares**, which we introduce next.

1 Method of least squares for linear regression

1.1 Set up of the problem

A biologist measured the number of dividing cells in her Petri dish at three different times: $x_1 = 0$, $x_2 = 1$, and $x_3 = 2$. She found the following values at these times: $y_1 = 1$ at x_1, $y_2 = 0$ at x_2, and $y_3 = 2$ at x_3.

She knows that the phenomenon under study should follow the linear law $y = \beta x + \alpha$, and would like to find α and β which best match the data.

1.2 Least squares estimation

1.2.1 A probabilist/statistician approach

Here is the framework with which an intelligent and diligent student of Probability and Statistics would approach her problem.

She would consider the bivariate data set (x_1, y_1), (x_2, y_2), and (x_3, y_3). In her model, she would view (x_1, x_2, x_3) as nonrandom, and (y_1, y_2, y_3) as realizations of random variables Y_1, Y_2, and Y_3 such that:

$$Y_i = \alpha + \beta x_i + U_i \qquad \text{for } i = 1, 2, 3$$

The Y_i are random variables to account for the variability in the biological processes observed, as well as measurement errors. This variability is represented with the random variables U_1, U_2, and U_3, which are assumed to be independent, to have zero mean, and the same variance σ^2.

Recalling Lecture 14, one could think of using the maximum likelihood principle to provide estimates for α and β. Unfortunately, this is not possible here, because the probability mass function or probability density function of the U_i is not known. The probability and statistics student thus turns toward the method of least squares.

1.2.2 Method of least squares

The idea of the method of least squares is to find α and β which minimize the quantity

$$\underline{S(\alpha, \beta) = (y_1 - \alpha - \beta x_1)^2 + (y_2 - \alpha - \beta x_2)^2 + (y_3 - \alpha - \beta x_3)^2}$$

Graphically, the idea behind this method is clear: one finds α and β which **minimize the sum of the square of the distances between the data point y_i and the point $\alpha + \beta x_i$ on the desired line at the measuring abcissa x_i.**

Now, unless x_1, x_2, and x_3 are identical (which is not the case in our example), the graph of S is an upward facing bowl, with a unique minimum, which we find by setting the partial derivatives to 0:

$$\begin{cases} \frac{\partial S}{\partial \alpha}(\alpha, \beta) = 0 \\ \frac{\partial S}{\partial \beta}(\alpha, \beta) = 0 \end{cases} \quad \Leftrightarrow \quad \begin{cases} (y_1 - \alpha - \beta x_1) + (y_2 - \alpha - \beta x_2) + (y_3 - \alpha - \beta x_3) = 0 \\ x_1(y_1 - \alpha - \beta x_1) + x_2(y_2 - \alpha - \beta x_2) + x_3(y_3 - \alpha - \beta x_3) = 0 \end{cases}$$

This system of equations for α and β can be rewritten as

$$\begin{cases} 3\alpha + \beta(x_1 + x_2 + x_3) = y_1 + y_2 + y_3 \\ \alpha(x_1 + x_2 + x_3) + \beta(x_1^2 + x_2^2 + x_3^2) = x_1 y_1 + x_2 y_2 + x_3 y_3 \end{cases}$$

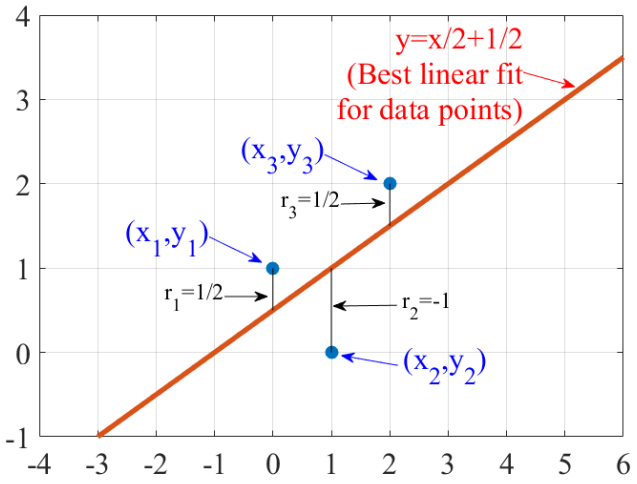

Figure 1: Data points (in blue) and best linear fit (in red), given by the equation $y = \frac{1}{2}x + \frac{1}{2}$. Also shown in the figure, in black, are the residuals $r_i = y_i - \alpha - \beta x_i$, which are introduced in Section 1.5.

$$\Leftrightarrow \begin{cases} \beta\left[(x_1+x_2+x_3)^2 - 3(x_1^2+x_2^2+x_3^2)\right] = (y_1+y_2+y_3)(x_1+x_2+x_3) - 3(x_1y_1+x_2y_2+x_3y_3) \\ 3\alpha(x_1+x_2+x_3) + 3\beta(x_1^2+x_2^2+x_3^2) = 3(x_1y_1+x_2y_2+x_3y_3) \end{cases}$$

$$\Leftrightarrow \begin{cases} \beta = \frac{(y_1+y_2+y_3)(x_1+x_2+x_3) - 3(x_1y_1+x_2y_2+x_3y_3)}{(x_1+x_2+x_3)^2 - 3(x_1^2+x_2^2+x_3^2)} \\ \alpha = \frac{y_1+y_2+y_3}{3} - \beta\frac{x_1+x_2+x_3}{3} \end{cases}$$

For our problem, we find

$$\Leftrightarrow \begin{cases} \beta = \frac{3\cdot 4 - 3\cdot 3}{3\cdot 5 - 3^2} = \frac{3}{6} = \frac{1}{2} \\ \alpha = 1 - \frac{1}{2} = \frac{1}{2} \end{cases}$$

We conclude that the best linear fit to the data is given by the line

$$y = \frac{1}{2}x + \frac{1}{2}$$

The data points and the best linear fit are shown in Figure 1. In this figure, we also show the residuals $r_i = y_i - \alpha - \beta x_i$ in black, which we will define formally in Section 1.5. The least squares method developed here minimizes the quantity

$$r_1^2 + r_2^2 + r_3^2$$

1.3 General formula

The linear regression method we just presented is obviously not limited to three data points, and can be naturally extended to n data points. The idea is to minimize the function

$$S(\alpha,\beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 \qquad \text{Linear regression} \tag{1}$$

Setting up the system for α and β in the same way as we have done, we would then find

$$\underline{\beta = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x_n}\,\overline{y_n}}{\sum_{i=1}^{n} x_i^2 - n\overline{x_n}^2}}$$

$$\underline{\alpha = \overline{y_n} - \beta\overline{x_n}}$$

with the standard definitions for $\overline{x_n}$ and $\overline{y_n}$:

$$\overline{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \overline{y_n} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

1.4 Least squares estimators

We can get a more complete understanding for the estimates for α and β just obtained by considering the estimators

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{x_n}\,\overline{Y}_n}{\sum_{i=1}^{n} x_i^2 - n\overline{x_n}^2}$$
$$\hat{\alpha} = \overline{Y}_n - \hat{\beta}\overline{x_n}$$

Using the linearity of expectation, a straightforward calculation leads to the equality

$$E[\hat{\beta}] = \beta$$

so $\hat{\beta}$ is an unbiased estimator for β.

Then, again by the linearity of expectation

$$\begin{aligned} E[\hat{\alpha}] &= E[\overline{Y}_n] - \overline{x_n}E[\hat{\beta}] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[Y_i] - \overline{x_n}\beta \\ &= \frac{1}{n}\sum_{i=1}^{n} (\beta x_i + \alpha) - \overline{x_n}\beta \\ &= \beta\overline{x_n} + \alpha - \overline{x_n}\beta = \alpha \end{aligned}$$

$\hat{\alpha}$ is an unbiased estimator for α.

Observe that a desirable property of the least squares method is that the construction does not rely on any detailed knowledge for the random variables U_i. In fact, we can relax the assumption that the U_i all have the same variance σ^2, and the method still works as well.

If however we would like to estimate the experimental variability from the simple linear relationship obtained in linear regression, it helps to assume that the U_i all have the same variance σ^2. In that case, it is partially shown in the textbook that an unbiased estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

1.5 Residuals

If the underlying physical and biological processes are such that a linear fit is indeed a reasonable hypothesis, and if the fit was done properly, we expect to roughly have half of the points above the fitting line, and half of the points below the fitting line. We would also expect the absolute value of the difference between the data point and the corresponding point on the line to fluctuate in a random fashion.

Hence, to assess the validity of the linear regression, it can be a good idea to look at the **residuals**, defined by

$$r_i = y_i - \alpha - \beta x_i \qquad , \qquad i = 1, 2, \ldots, n \tag{2}$$

For a justified and well done linear regression, the plot of the r_i's should fluctuate about zero in a random fashion.

2 Connection with maximum likelihood estimate

If we allow ourselves to specify the distribution of the U_i, then the distribution of the Y_i is also specified, and we can view the problem of finding the optimum α and β as a maximum likelihood question.

A natural assumption is to take the U_i to be independent and normally distributed with mean 0 and variance σ^2:

$$U_i \sim N(0, \sigma^2)$$

We then have

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

Following the standard maximum likelihood approach, we can then write

$$f_{\alpha,\beta,Y_1,Y_2,\ldots,Y_n}(y_1, \ldots, y_n) = f_{\alpha,\beta,Y_1}(y_1) f_{\alpha,\beta,Y_2}(y_2) \ldots f_{\alpha,\beta,Y_n}(y_n)$$

In our case, the maximum likelihood function thus is

$$L(\alpha, \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_1-\alpha-\beta x_1)^2}{2\sigma^2}} \ldots \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_n-\alpha-\beta x_n)^2}{2\sigma^2}}$$

To simplify our calculations of the maximum, we then consider the loglikelihood function, given by

$$l(\alpha, \beta) = \ln[L(\alpha, \beta)] = -n \ln(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

The maximum of the loglikelihood function, and therefore of the likelihood function, satisfies the following system of equations:

$$\begin{cases} \frac{\partial l}{\partial \alpha} = 0 \\ \frac{\partial l}{\partial \beta} = 0 \end{cases} \quad \Leftrightarrow \quad \begin{cases} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0 \\ \sum_{i=1}^{n} x_i (y_i - \alpha - \beta x_i) = 0 \end{cases}$$

This is exactly the same system we had obtained in the least squares method, with obviously the same solutions α and β!
When the U_i are independent and normally distributed with distribution $N(0, \sigma^2)$, the maximum likelihood principle and the least squares method yield the same estimates.

3 Method of least squares for quadratic regression

For certain physical and biological processes, a linear fit may not be a satisfying scientific idea. This would for example be the case for the power P dissipated in a resistor as a function of the current I through the resistor discussed in pages 6 and 7 of Lecture 12. The data clearly shows that a quadratic fit is likely to be better suited than a linear fit. Another way to say this is to say that a linear fit would likely lead to residuals which would display a trend.

Fortunately, least squares can also be used for quadratic regression. The principle is the same: we want to minimize the sum of the squares of the distances between the data points and the corresponding points on the quadratic curve $y = \gamma x^2 + \beta x + \alpha$.

In other words, we now want to minimize

$$S(\alpha, \beta, \gamma) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i - \gamma x_i^2)^2 \qquad \text{Quadratic regression} \tag{3}$$

and the system of three equations for the three unknowns α, β, and γ is given by

$$\begin{cases} \frac{\partial S}{\partial \alpha} = 0 \\ \frac{\partial S}{\partial \beta} = 0 \\ \frac{\partial S}{\partial \gamma} = 0 \end{cases}$$

We will not give explicit formulae for the general solution to this system of equations in these notes, because it is often much more reliable to construct this system and solve it by hand than to remember the complicated formulae by heart.