

Basics of Biostatistics

Linear regression

topics

1. descriptive statistics (what data are)
2. hypothesis testing (comparing data)
3. correlation and regression analysis

Recommended readings:

- **Medical Biophysics Practices** – ed. M. Kellermayer, 3rd ed., Semmelweis Publisher: **Appendix: Biostatistics**

Further readings:

- Harvey Motulsky: Intuitive Biostatistics – A Nonmathematical Guide to Statistical Thinking, Oxford University Press
- Nature Collection: Statistics for Biologists

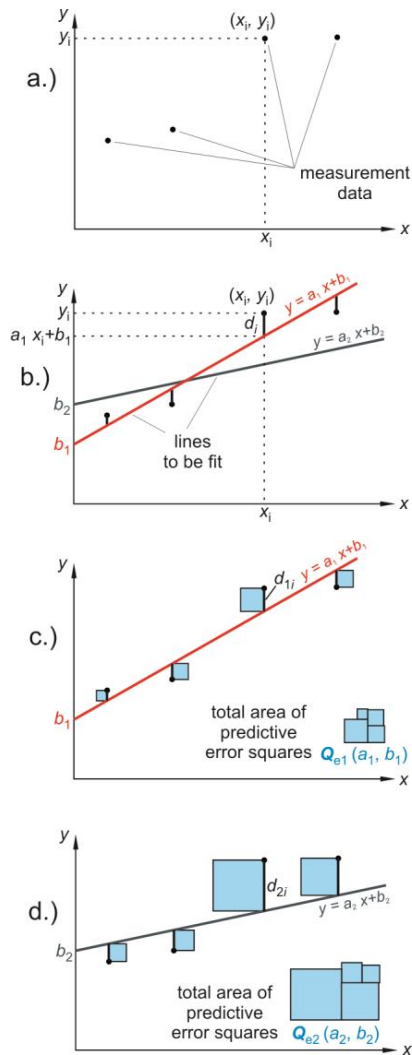


Fig. 13. Finding the line of the best fit to the datapoints.

LINEAR REGRESSION

The simplest curve, the straight line, is very easy to obtain subjectively with the aid of a paper, pencil and a ruler. However, we can hardly eliminate our doubts about the uncertainties arising from the errors of data points with the qualitative principles of drawing. Therefore, the problem of **finding the straight line of "best fit" to the data points still persists.**

The equation of a straight line is:

$$y = a \cdot x + b, \quad (15)$$

where a is the slope and b is the intercept. The y -intercept is the y value at $x = 0$, or the point of intersection of the line with the y axis. The straight line is determined explicitly by **these two parameters**. The task is to determine the actual values a^* and b^* of the parameters yielding the best fit to our data points. As a first step we will examine **what is meant by the line of best fit.**

Suppose that we have four data points (see Fig. 13a). Furthermore, let us assume that values of x_i are "exact" (without error), preset values, and only the y_i values have an error.

Let us draw an arbitrary line, determined by the parameters a_1 and b_1 ($y = a_1 x + b_1$), across the data points and calculate the vertical distance of the datapoints from this line (see Fig. 13b, vertical lines).

The distance of a selected (x_i, y_i) datapoint from the x axis is given by the y_i coordinate. At the same time the distance of the (x_i, y) line point from the x axis is obtained by substitution into the equation of the line as $(a_1 x_i + b_1)$. The difference is the **vertical distance of the point and the straight line** $d_{1i} = (y_i - (a_1 x_i + b_1))$.

This distance is positive if the point is above the line and negative if it is below. Let us calculate the squares of these distances computed for other points (predictive error squares).

Similarly to the sum of squares that we have defined in relation to the empirical standard deviation by formula (6), let us sum all the squares of distances calculated for the data points (see Fig. 13c, small blue squares) and denote the sum by Q_{e1} . Now, let us draw another line — determined by parameters a_2 and b_2 — and calculate the vertical distances (d_{2i}) of the data points from this line as well. As a result, we get a sum of squares again Q_{e2} (see Fig. 13d). Observe, that the points scatter around the line more when Q_{e2} is larger ($Q_{e2} > Q_{e1}$).

As the data points (x_i, y_i) are always the same, Q_e is determined only by the variation of the parameters a and b . With this assignment we have defined a function with two independent variables a and b and a dependent one Q_e :

$$Q_e(a, b) = \sum_{i=1}^n [y_i - (ax_i + b)]^2. \quad (16)$$

Due to the squared distances the function $Q_e(a, b)$ is of second degree in both variables. This means that it can be represented by a quadratic surface similar to a "well" or "pit" having parabolic traces (curves of intersections of the surface with planes parallel to the coordinate planes, where $a = \text{const}$, $b = \text{const}$).

If some other conditions are fulfilled (e.g., the standard deviations of the data points are independent), then the line of the best fit is the one for which the sum of squares $Q_e(a, b)$ has an absolute minimum.

We are looking for the coordinates (a^*, b^*) for which the function $Q_e(a, b)$ has the lowest value. In other words: we have to find the coordinates (a^*, b^*) of the lowest point of the well (the vertex of the paraboloid). The corresponding fitted line is called the **regression line**. The method of finding this line is known as the **least squares method**, or **linear regression** in general.

The word regression means "return" or "reversion", expressing the inference to the connection (correlation) between the variables from the measured data. (Connection does not necessarily mean causality.)

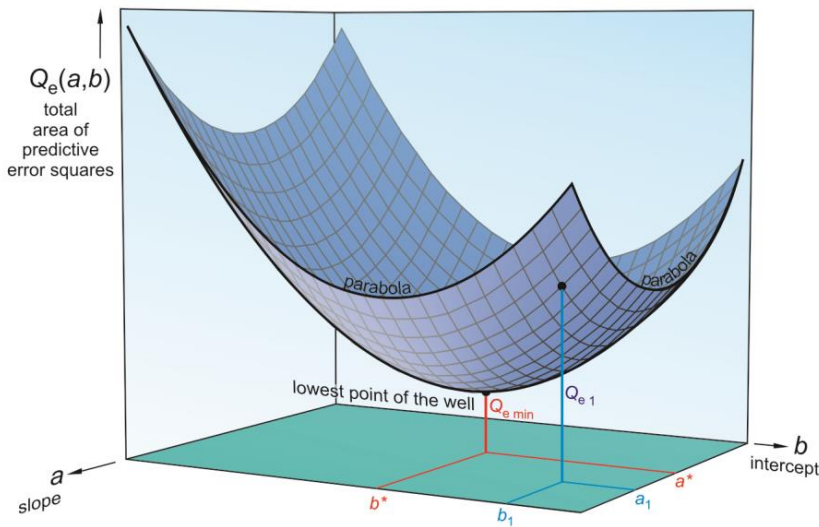


Fig. 14. The predictive error as a function of the parameters **a** and **b**. The sum of error squares has a minimum at the bottom of the well.

After finding the minimum one obtains:

$$a^* = \frac{Q_{xy}}{Q_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \text{ or } a^* = \frac{s_{xy}^2}{s_x^2}, \quad (17)$$

$$b^* = \bar{y} - a^* \bar{x} = \frac{\sum_{i=1}^n y_i}{n} - a^* \frac{\sum_{i=1}^n x_i}{n}, \quad (18)$$

where Q_{xx} and Q_{xy} corresponds to the notation introduced in (6),

— $s_{xy}^2 = Q_{xy} / (n - 1)$ is the so-called **covariance**,

— s_x^2 is the **variance** of x , and

— \bar{x} , \bar{y} are the **means**, respectively.

The line of the best fit can be calculated for any (x_i, y_i) pairs of data from the above equations, even if the points lie obviously along some curve, and not a straight line.

Since it is rather difficult to decide subjectively how well the fitted line approximates the experimental points, the determination of the **correlation coefficient** is desirable:

$$r = \frac{Q_{xy}}{\sqrt{Q_{xx} \cdot Q_{yy}}} = \frac{s_{xy}^2}{s_x s_y}, \quad (19)$$

where the notations are the same as in formula (17).

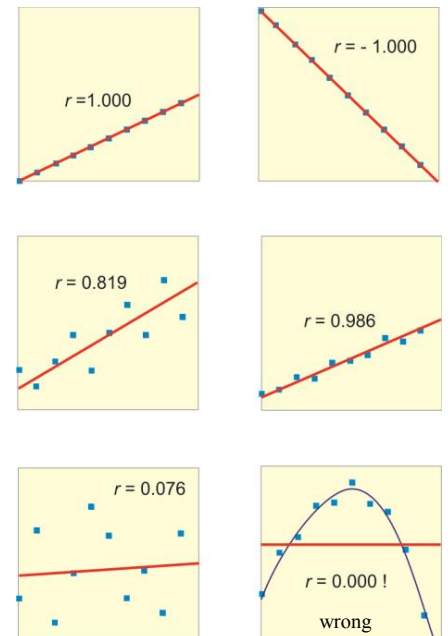


Fig. 15. Some examples for the values of the correlation coefficient.

Comment 7.

In the following table the accommodation power of the eye is listed as a function of age:

age (years)	20	25	35	45
accommodation power (dpt)	11	8.5	7	3.5

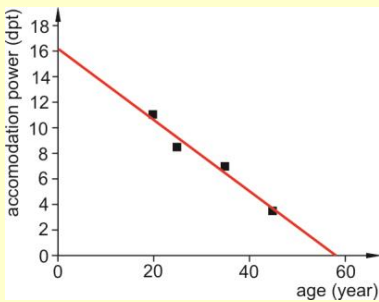


Fig. 16. Accommodation power of the eye versus age. Fitting a regression line to the data points.

After plotting the data and doing a linear regression the parameters of the fitted line are:

$$a^* = -0.28; \quad b^* = 16.2.$$

The correlation coefficient is:

$$r = -0.98.$$

Because there is no known model that would connect the two variables, parameters can be used for estimation by interpolation only. For example, we can estimate the accommodation power at the age of 40 as:

$$-0.28 \cdot 40 + 16.2 \approx 5 \text{ (dpt)}$$

This value was missing from the table. Notably, however, we know that at very young age the accommodation power is definitely lower than 16 dpt. Therefore, the obtained relationship has limited predictive power outside the sample range.

The correlation coefficient characterizes **the strength of the correlation** (connection) between the variables, and it has values between +1 and -1. Positive values of the correlation coefficient belong to a regression line with positive slope, and negative ones to a line with negative slope. If the data points are very close to the regression line, value of $|r|$ will be close to 1 (e.g., $r = 0.9860$). If the line goes through every point, then $|r| = 1$, otherwise the more $|r|$ approaches zero, the greater the deviation of the data points from the regression line (see Fig. 15 and Comment 7 for the example).

We did not discuss yet in **what situation and to what purpose** can the regression analysis be used. An important **aspect** that we need to take into account is whether there is **a model** describing a causality **relation** between the variables (x, y) with parameters of physical meaning, or the line is fitted just to represent the datapoints with the necessary accuracy. Note that even if a model is not known, causality between the variables may exist.

An example for the first case could be the relationship between the index of refraction of solutions and their concentration (see 4. [REFRACTOMETER](#)), or the relationship between the optical density of solutions and solute concentration (see 6. [LIGHT ABSORPTION](#)). In these cases the parameters of the regression line can be used to **extrapolate** values (to estimate variables outside the measured range). However, the range of validity of the formulas (physical laws) should be taken into account.

In the second example parameters can be used only for **interpolation** (estimation of variables inside the measured range) and even if the correlation coefficient was close to 1, we may not infer a causality relation (see Comment 7).