The basics of linear regression

In analytical chemistry, we often come across the desire to describe one variable as a function of another variable. A good example is in the preparation of standard calibration curve for instrumental analysis where we plot a series of concentrations of working standards against the corresponding instrument responses. In some situations we know the functional relationship between the two variables based on a known theoretical hypothesis, and in other situations, we may have no prior knowledge about their relationship but would like to use the observed analytical data to identify the relationship.

The simplest relationship model between two quantitative variables, x and y, is a simple linear regression, by fitting a linear equation to the observed data. The linear equation can be written as:

$$y = \alpha + \beta x \qquad (1)$$

where α (also known as the intercept) is the value of y when $x = 0$, and β, the slope (i.e. the change in y for each unit change in x, $\Delta y/\Delta x$).

An important assumption of this relationship model is: one variable is the dependent variable (y in the linear equation) whilst the other is an independent or explanatory variable (x in the regression formula).

By using this equation (1), we model y as a linear function of x in the hope that information about x will give us some information about the value of y.

Example 1. Stearic acid and digestibility of fat

A study of the % digestibility of fat from nine different levels of stearic acid proportion in the fat was conducted. Data obtained are shown in the table below, where x represents stearic acid and y, the digestibility measured in percent.

x	29.8	30.3	22.6	18.7	14.8	4.1	4.4	2.8	3.8
y	67.5	70.6	72.0	78.2	87.0	89.9	91.2	93.1	96.7

The data are scatter plotted in Figure 1.

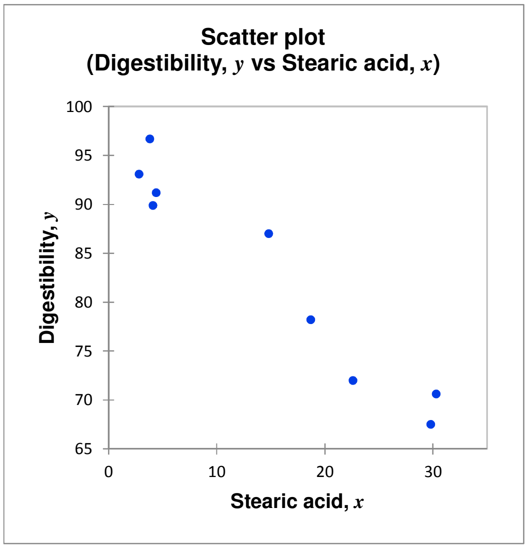

Figure 1: Scatter plot of fat digestibility for different proportions of stearic acid in the fat.

Figure 1 shows visually a trend that the stearic acid (x) and the fat digestibility (y) are having some sort of linear relationship where for each x–value of stearic acid, we have some corresponding y–value.

We may wish to draw a straight line on this scatter plot to represent this relationship but obviously this straight line will never fit all the observations perfectly. The 'best' straight line is of course, to have some of the observations above the line and some below, as shown in Figure 2.

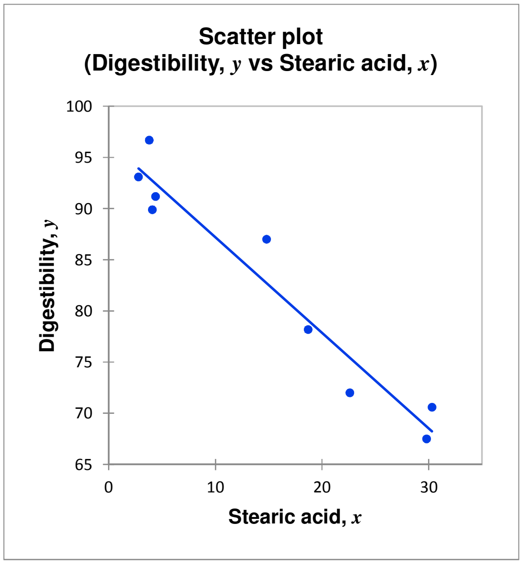

Figure 2: Fat digestibility for different proportions of stearic acid in the fat with a fitted straight line plotted

How to fit a regression line?

Fitting a regression line means identifying the "best" line, i.e., the optimal parameters to describe the observed data. This best line should be one that runs on the scatter plot with "equi-distance" from all the points.

Mathematically, let (x_i, y_i), $i = 1, \ldots, n$ denotes our n pairs of observations and assume that we somehow have "guess-estimates" of the two parameters, $\hat{\alpha}$ and $\hat{\beta}$, from a linear equation used to model the relationship between the x's and the y's.

Take notice that these $\hat{\alpha}$ and $\hat{\beta}$ indicate that the values are not necessarily the true but unknown values of α and β but *estimates.* Our model for the data is given by the line

$$y = \hat{\alpha} + \hat{\beta} x \qquad (2)$$

Hence, for any x, we can use this model to predict the corresponding y-value. In fact, we can do so for each of our original observations, $x_1, \ldots, x_n$, to find the predicted values; i.e. the y-values that the model would expect to find:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

To find out how well the model fits to the actual observed values, we can calculate the differences between the observed y-values and the predicted y-values at all x values. These differences or deviations, statistically known as the *residuals*, are defined as follows:

$$r_i = y_i - \hat{y}_i \qquad (3)$$

Indeed, the residuals measure how far away each of our actual observations (y_i's) are from the expected value given a specific straight line model. Certainly, we would like to use a model that provides small residuals because that means that the values predicted by the model are close to our observations.

However, if we were to sum up all the r_i's of a given set of x's and y's, we would have the positive and negative residuals cancel each other and would not know the magnitude of the residuals. To overcome this problem, we use the method of *least squares*, where the residuals are squared.

When a residual is large, it means the observed data is far away from the predicted value on the 'best' regression line. The converse is true. Figure 3 shows a graphical representation of these squared residuals.

The shaded areas correspond to the square of the residuals so each observation gives rise to a square shaded area. Instead of just looking at the sum of squared residuals and trying to find a model that is as close to the observations as possible, we can try to identify a model that minimizes the sum of these squares. So, it is called the least square estimation method.

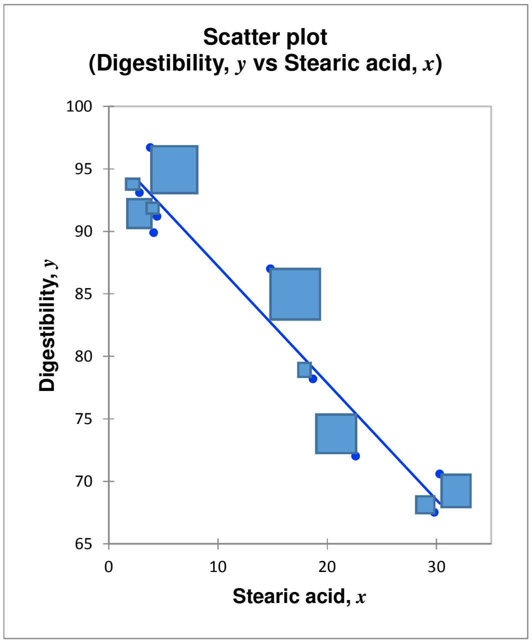

Figure 3: Squared residuals for the dataset on digestibility and stearic acid. Shaded areas represent the squared residuals for the proposed regression line.

Least squares estimation

The least squares method estimates the unknown parameters of a model by minimizing the sum of the squared deviations between the data and the model. Thus, for a linear regression model, we try to identify the parameters α and β such that

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 \tag{4}$$

becomes as small as possible. To find the maximum or minimum of a function, a standard approach is to use the functional differentiation method in calculus. We differentiate the function and identify the parameter values for which the derivative equals zero.

Let the function be

$$Q(\alpha, \beta; x, y) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 \tag{5}$$

The two partial derivatives of this function are:

$$\frac{\partial Q}{\partial \alpha} = \sum_{i=1}^{n} \frac{\partial}{\partial \alpha}\left(y_i - \alpha - \beta x_i \right)^2 = \sum_{i=1}^{n} 2\left(y_i - \alpha - \beta x_i \right)(-1)$$

$$= -2\left(\sum_{i=1}^{n} y_i - n\alpha - \beta \sum_{i=1}^{n} x_i \right) \qquad (6)$$

$$\frac{\partial Q}{\partial \beta} = \sum_{i=1}^{n} 2\left(y_i - \alpha - \beta x_i \right)(-x_i) \qquad (7)$$

To find the minima of Q, we set these two partial derivatives equal to zero, i.e.

$$\frac{\partial Q}{\partial \alpha} = 0 \;\; ; \qquad \frac{\partial Q}{\partial \beta} = 0$$

We now have to solve the two equations (6) and (7) with two unknown, α and β.

It can be shown that there is a unique minimum of equation (4), which means we can find a unique line that fits our data best. We summarize the results above as follows:

For a linear regression model, the line that best fits the data has slope (or gradient) and intercept given by:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x} \right)\left(y_i - \overline{y} \right)}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \qquad (8)$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\,\overline{x} \qquad (9)$$

Take note that by “best straight line”, we mean the one that minimizes the residual sum of squares. Also, as a consequence of equation (9), we have that the best straight line will always go through the point $(\overline{x}, \overline{y})$ since $\overline{y} = \hat{\alpha} + \hat{\beta}\,\overline{x}$.

So, the least squares estimates of the slope and intercept of our example on the digestibility study are found by inserting the data into equations (8) and (9). The details are shown in Table 2.

Table 2: Calculations for the stearic acid data

i	x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
1	29.8	67.5	15.21	-15.41	231.38	237.50	-234.42
2	30.3	70.6	15.71	-12.31	246.84	151.56	-193.42
3	22.6	72.0	8.01	-10.91	64.18	119.05	-87.41
4	18.7	78.2	4.11	-4.71	16.90	22.19	-19.37
5	14.8	87.0	0.21	4.09	0.04	16.72	0.86
6	4.1	89.9	-10.49	6.99	110.02	48.84	-73.31
7	4.4	91.2	-10.19	8.29	103.81	68.71	-84.45
8	2.8	93.1	-11.79	10.19	138.98	103.81	-120.12
9	3.8	96.7	-10.79	13.79	116.40	190.13	-148.77
Sum	131.3	746.2	0.00	0.00	1028.55	958.53	-960.40
Mean	14.589	82.911					

By calculation, we have:

$$\hat{\beta} = \frac{-960.40}{1028.55} = -0.934$$

$$\hat{\alpha} = 82.911 - (-0.9337 \times 14.589) = 96.533$$

Thus, the best regression line for the digestibility data is given by:

$$y = 96.533 - 0.934 \cdot x$$

Armed with this best line, we are able to make predictions about the digestibility percentage for stearic acid levels that we had not examined in the experiment as described in this example.

In the next blog, we shall examine a set of experimental data which do not fit into a linear regression model as discussed and shall show how we can deal with such situation.