Statistics for Nuclear Medicine
Part 4: Regression

Peter C. O'Brien, Marc A. Shampo, James S. Robertson

Section of Medical Research Statistics, Mayo Clinic and Mayo Foundation, Rochester, Minnesota

A previous paper has dealt with a hypothetical problem of measuring the increase of free thyroxine (FT_4) associated with the intravenous infusion of heparin. When a drug is found to be efficacious, we want to know what influence other factors may have upon its effect. Specifically, in this paper, we shall measure the influence of initial FT_4 level on the increase achieved by the drug.

Suppose the increase of FT_4 in 10 study participants is as listed in Table 1. The first step in analysis is to exhibit the data graphically, relating the changes to initial values by use of a scatter diagram (Fig. 1).

In order to quantify and summarize the association shown by the scatter diagram, we draw a straight line through the group of points, as illustrated in Fig. 2. How well the line fits the data is measured by the sum of the squared vertical distances of the individual points from the line. Thus the best-fitting line is the one for which this sum of squares is least, and it is called the least-squares line. (There is a formula for the calculation.)

In general terms, the least-squares line may be described by the equation:

$$y = a + bx$$

This is the linear regression equation, in which:

a = intercept, the point on the y-axis where the regression line will cross it, if extended that far (the value of y when x = 0).

b = slope, the amount of change in y per unit of increase of x. For the line shown in Figure 2, a = −0.377 and b = 0.4937; so the specific equation is:

$$y = -0.377 + 0.4937x.$$

We are also interested in measuring how closely the points cluster about the regression line. The appropriate measure, denoted by $s_{y\cdot x}$, is defined in terms of the sum of the squared vertical distances from the regression line (as shown in Fig. 2). Specifically, $s_{y\cdot x}$ is defined as the square root of the number that results from dividing the sum of squared distances by N-2. Rather than assign a special name to this statistic, statisticians usually write the symbol itself and pronounce it "s-y-dot-x." In this example, $s_{y\cdot x} = 0.2319$.

In most applications, the feature of greatest interest is the slope, which here represents the amount of post-treatment change in FT_4 that corresponds to a unit increase in initial FT_4. Although we cannot determine definitely from a sample the magnitude of the unknown true slope (usually represented by β) in the population, we can estimate it from the sample, test hypotheses about it, and establish confidence limits of the estimate as we did before in seeking a population mean.

Specifically, in our example, the true slope β is esti-

TABLE 1. DATA FOR HYPOTHETICAL EXAMPLE

Subject	FT_4 after heparin (A)	FT_4 before heparin (B)	Difference (A − B)*
1	0.60	0.75	−0.15
2	1.35	1.00	+0.35
3	1.15	1.20	−0.05
4	1.90	1.30	+0.60
5	1.75	1.50	+0.25
6	2.35	1.80	+0.55
7	2.30	2.00	+0.30
8	2.90	2.05	+0.85
9	3.05	2.35	+0.70
10	3.45	2.50	+0.95

* Positive difference values represent increases of FT_4.

Received Sept. 30, 1982; revision accepted Jan. 3, 1983.

For reprints contact: Dr. O'Brien, Mayo Clinic and Mayo Foundation, 200 First Street SW, Rochester, MN 55905.

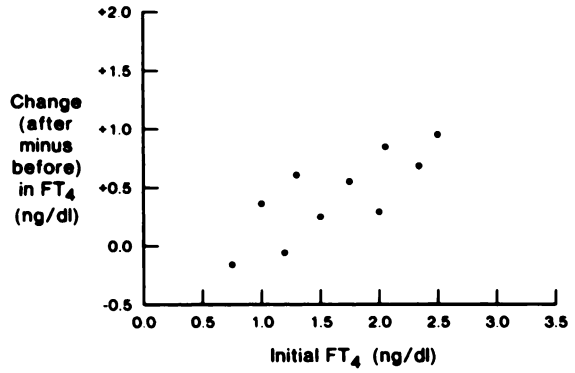

FIG. 1. Scatter diagram of initial free thyroxine level (FT_4) and change of level after administration of heparin, from Table 1.

mated by the sample slope b (0.4937). To test the hypothesis that $\beta = 0$, we begin by comparing b to its standard error (SE_b)—which depends on how closely the same points cluster about the fitted line. In the example, $SE_b = 0.1318$, so the test statistic is:

$$t = b/SE_b$$
$$= 0.4937/0.1318$$
$$= 3.746$$

From suitable tables or computing equipment, we find that, if $\beta = 0$, then the probability of obtaining a value of t as large as 3.747 is less than 0.01 ($P < 0.01$). So from our hypothetical data, we reject the hypothesis $\beta = 0$ and conclude that the drug response is affected by the initial concentration of FT_4.

The 95% confidence interval is given by:

$$95\%\ CI = b \pm t^*_{N-2} \cdot SE_b$$

in which t^*_{N-2} is a number obtained from special tables. In this case $t^*_{N-2} = 2.306$. Upon substitution,

$$95\%\ CI = 0.4937 \pm 2.306 \cdot 0.1318$$
$$= 0.4937 \pm 0.3039$$

So the interval is from 0.1898 to 0.7976—well above zero, thus confirming our rejection of the hypothesis that the initial level and the amount of change were unrelated ($\beta = 0$). The data indicate that the initial level and the amount of change are related.

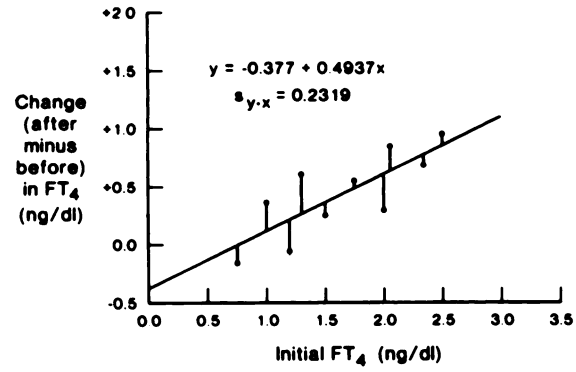

FIG. 2. Scatter diagram (Fig. 1) with regression line and lines from data points to regression line for least-squares determination.

Comment. 1. We have shown how the association between two variables may be quantified by fitting a straight line to the data. In doing so, we have considered only the simplest of situations. In practice, other factors may require attention: for example, how to modify the analysis if the scatter diagram reveals outliers or skewness or associations that are nonlinear, or how to evaluate additional variables (such as age or sex).

Two other considerations regarding the regression line should be remembered. First, in graphing the regression line the steepness of the line depends on how the axes are scaled (whether large or small units are used). Second, extension of the regression line beyond the plotted data may give rise to absurd implications.

2. You may have noticed that we have not mentioned a rather popular statistic called the correlation coefficient, usually denoted by r. Its popularity derives in part from the fact that the correlation coefficient does not depend on the units of measurement (for example, pounds or kilograms), as the slope of the regression line does. The correlation coefficient is a somewhat complicated function of the test statistic b/SE_b and the sample size. The sign of b (that is, whether the regression line runs upward or downward) determines whether r will be positive or negative; and when $b = 0$, $r = 0$. The calculation of r can be used to test the hypothesis that y is *not* related to x.

However, if y *is* related to x, r serves poorly in describing how, because it is ambiguous. Although a large

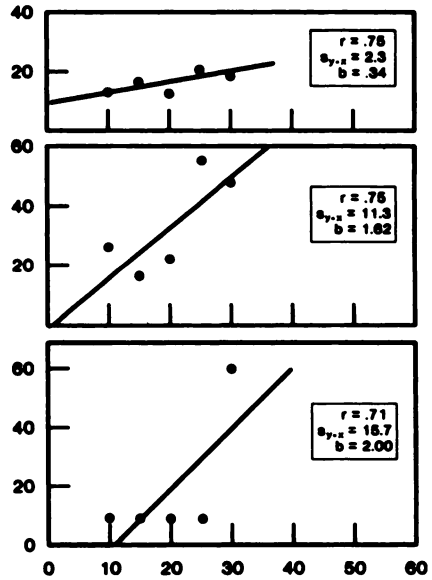

FIG. 3. Examples of similar correlation coefficients resulting from different conditions. (Top panel) Smallness of scatter about regression line. (Middle panel) Steepness of slope. (Bottom panel) Presence of outlier. (From O'Brien PC, Shampo MA. Statistics for Clinicians. 7. Regression. *Mayo Clin Proc* 56:452–454, 1981.)

value of r (within its mathematical limits of +1 and −1) suggests that the correlation is strong, faith in this simple implication may be misplaced. The value of r can be increased by increase of b and also by decrease of $s_{y \cdot x}$ (since SE_b is directly proportional to $s_{y \cdot x}$). These components are quite different: b (the slope of the regression line) indicates how large an associated change is; and $s_{y \cdot x}$ (the closeness of the data points to the regression line) indicates how consistently the change occurs. But r, as a single value, gives no indication of the relative influence of the two components in determining its value.

Notice in Fig. 3 that, although the correlation coefficient is virtually the same in each instance, the associations between y and x are much different. The high correlation coefficients are due, successively, to smallness of the scatter about the line, to steepness of the slope, and to presence of an outlier. These examples also illustrate the importance of looking at a scatter diagram whenever one does a regression analysis.

To describe the association between two variables in terms of summary statistics, it is best to use both b and $s_{y \cdot x}$.