Statistics in Question

SHEILA M GORE

ASSESSING METHODS— A FEEL FOR OTHER THINGS

Two examples of non-linear regression functions are given, only one of which can be transformed to become a straight line. Weighted least squares is again mentioned, and the idea of plotting interval means as a guide to the form of a regression function is emphasised in a second example.

Correlation coefficients make a useful contribution to preliminary data analysis only if supported by graphical methods. Authors are advised to report descriptively their initial impressions of complex data, giving the reader a useful background against which to see a final multivariate solution.

Non-linear relationships

(15) Transformation¹ of one or both variables sometimes induces a straight-line relationship. Find the cube root of 0.125, of 1, of 8. What is the logarithm of 1, the square root of 4, the cube of 2, and the reciprocal of 2?

—0.125^{1/3} = 0.5; 1^{1/3} = 1; 8^{1/3} = 2

—log 1 = 0; $\sqrt{4} = 2$; 2³ = 8; 2⁻¹ = $\frac{1}{2}$ = 0.5

—not all regression functions are linear

COMMENT

The list of common transformations includes taking the logarithm, or the reciprocal, or the square root. The last two are particular instances of power transformations, $w = x^c$, the original variable x being raised to the power c . Taking $c = -1$ gives the reciprocal of x , and $c = \frac{1}{2}$ gives the square root. Transformation of one or both variables (explanatory or dependent, or both) sometimes induces a straight-line relationship. For example, figure 1 shows blood lead and first-flush water lead concentrations for 949 samples.² The non-linear relationship was shown using the technique that I described in question 13.³ Samples were allocated to nine groups of approximately 100, by water lead, and the group mean blood lead concentrations were plotted (see figure 2). The steep rise in blood lead concentration as water lead concentrations increased from 0.01 to 1 $\mu\text{mol/l}$ levelled off at higher concentrations of water lead. Moore *et al* showed that the mean blood lead concentration tended to rise in proportion to the cube root of first-draw water lead concentrations so that transforming the explanatory variable in this case was sufficient to induce a straight line (see figure 3). The authors also noted from figure 1 that the variability and skewness in the distribution of blood lead concentrations were greater for high water lead concentrations, making weighted regression appropriate in addition to change of scale.

A second example of a non-linear relationship concerns the 1865 patients who survived for five years from a cohort of 3878 patients with breast cancer who were referred to the Western General Hospital, Edinburgh. The follow-up of these patients and calculation of expected mortality has been described by Langlands *et al*.⁴ Figure 4—the initial step in further

FIG 1—Scatter diagram for first-flush water.

Conversion: SI to traditional units—Blood lead: 1 $\mu\text{mol/l} \approx 0.0483 \mu\text{g}/100 \text{ ml}$.

statistical modelling by Pocock and Gore of survival corrected for age—derives from that description. Note that the ratio of observed to expected deaths among five-year survivors falls from 4.2 in the first year of subsequent follow-up to 2.6 by the 10th year after diagnosis; and thereafter, because of the smaller numbers left, the ratio varied from year to year but remained

FIG 2—Mean blood lead concentrations for nine groups at intervals of first-flush water lead.

FIG 3—Interval means plotted against the cube root of first-flush water lead concentrations. Regression function linear.

FIG 4—Five-year survivors (Western General Hospital breast cancer series). Ratio of observed to expected deaths and estimated non-linear regression function, $\beta_0 + \exp(\beta_1 - \beta_2 t)$. In three instances, 95% confidence interval for the ratio O:E is given to indicate reliability.

greater than one. Approximate 95% confidence intervals for the ratio are shown in three instances to emphasise that the ratio is estimated less reliably in later follow-up. Weighted least squares is therefore appropriate. The ratio of observed to expected deaths diminishes, not linearly, but exponentially

with length of follow-up and approaches an asymptote—a line which the curve continually approaches but never meets. The form of the regression function was therefore $\beta_0 + \exp(\beta_1 - \beta_2 t)$, the reasonableness of this form being shown in figure 4. It has the interpretation that if the asymptote β_0 is not significantly different from one, a proportion of breast cancer patients is cured. The actual results are not given here, my purpose being only to illustrate a case in which transformation could not induce linearity—taking logarithms does not help because β_0 gets in the way—and, incidentally, a further case which also needed to take account of ratios estimated with different reliabilities.

Authors must try to explain or interpret complicated regression functions if they hope to persuade the reader that the relationship is meaningful clinically as well as having a mathematical basis. Common sense and the need to convince others conspire together to put an effective ban on exuberant curve-fitting.⁵ And rightly so.

Correlation coefficient

(16) The (Pearson) correlation coefficient is a measure of what sort of association?

—linear association

—correlation coefficients make a useful contribution to preliminary data analysis only if supported by graphical methods

COMMENTS

Altman⁶ carefully distinguished between regression and correlation in presenting definitive results. I shall concentrate instead on preliminary or exploratory data analysis when the distinction is blurred in practice (without serious loss), the correlation coefficient being a convenient summary statistic provided that investigators (a) always inspect the corresponding scattergrams, (b) refrain from quoting significance levels when describing this exploratory phase, and (c) include these initial impressions in a final report so that the reader can get a feel for the problem and understand why some variables but not others—either because they were too closely related or because they were uninteresting from the start—feature in a final statistical model. These remarks are illustrated with reference to the British Regional Heart Study.⁵ (a) Two hundred and fifty-three urban areas were studied. Describing their initial impression of geographic variations in cardiovascular mortality, Pocock *et al.*⁵ listed the correlation coefficient with SMR (standardised mortality ratio for cardiovascular disease) for 24 factors, each of which had a marked association with cardiovascular mortality. But they also inspected scattergrams. There are two reasons for doing this. The first is to check whether the association is reasonably a straight line and if variance is well-behaved (constant)—departures from these ideals will modify later analyses. The scattergram of water hardness plotted against SMR was shown (figure). It suggested that the effect of water hardness was non-linear, being much greater in the range from very soft to medium-hard water than thereafter. This was verified later in a multifactorial solution that took account also of other variables, such as rainfall, temperature, and socioeconomic status.

The second reason for looking at scatter plots is to ensure that an important non-linear association—U-shaped, for example—which produces an insignificant correlation coefficient has not been missed, as it will be if the exploration goes no further than scanning correlation coefficients. No such associations were reported by Pocock *et al.*, but there are plenty of examples.

Water hardness plotted against standardised mortality ratio for all men and women aged 35 to 74 with cardiovascular disease for each town. (Water hardness: 1 mmol/l = calcium carbonate equivalent 100 mg/l.)

elsewhere—a U-shaped relation between weight and total mortality; a J-shaped relation of total mortality with plasma cholesterol concentration (for which an interesting explanation has been given⁷); and a non-linear association of alcohol consumption with mortality from all causes.⁸ Therefore, do not rely only on correlation coefficients. (b) Looking at scattergrams and correlation coefficients one variable at a time against SMR gave only an outline description. This superficial view of the data may be considerably modified when several variables are taken into account together. Significance levels were not quoted for the 24 listed correlation coefficients because they would not reflect the relative importance of the different factors. Only a multivariate approach gives a solution to that conundrum. Emphasising this, Pocock *et al.*⁵ drew the reader's attention to the eight water variables among the 24 factors that showed a marked association with cardiovascular disease. Because the eight were highly dependent on each other, it was on the cards that they conveyed a single message—namely, some aspect of water quality is likely to be important. The message is made no more clear by quoting significance levels that are irrelevant when the real problem is multivariate. (c) The authors gained in the following respects by publishing their initial impressions. Firstly, the reader was acquainted with the complexity of variables—48 factors were considered as potential candidates in a multiple regression model for log SMR—and reasonably expected to see some of the 24 listed variables appearing in the final model. Because these 24 were grouped into categories—water, climatic, socioeconomic, geographical, blood group, air pollution—interest was focused on which categories stayed the course when there were many contenders instead of one, and also on how many representatives from each category were retained. Rainfall and mean temperature might offer different insights—both appeared in the final model—whereas because of their high interdependence the eight water factors might be adequately represented by one of their number, seen as a carrier for information on water quality generally. In a sense the preamble gave a forceful reminder that association is not causation. As the authors explained, total hardness was simply the water factor that they selected to show that the effect of water could not be explained away by other environmental factors such as rainfall. They could have shown this just as convincingly by using another of the eight—the concentration of water calcium, for example. Because the explanatory analysis was carefully described the reader could not fail to appreciate this point. A good general principle when reporting

results is to take the reader through the steps the investigators made so that the reader is as comfortable with them as the authors are.

Correlation coefficients make a useful contribution to preliminary data analysis only if supported by graphical methods. Moreover, it is worth while for authors to report descriptively their initial impressions. I refer readers to an account by Altman⁶ for the sorts of error that occur when the significance of Pearson correlation coefficients is reported in definitive analysis. This test, as usually performed (under appropriate distributional assumptions), is made from the standpoint of no linear association⁹—a ridiculous standpoint when comparing, for example, two assays that purport to measure the same thing. Method comparison studies are discussed in some detail by Altman.⁶

To end with, an often repeated caution: statistical association is not, of itself, sufficient to impute causality to any observed relationship. How specific is the association—is water hardness related also to non-cardiovascular deaths?⁵ Is the relationship consistent in time or with international studies, as the negative association of completed family size with mortality from ovarian cancer is?¹⁰ Is there a possible mechanism? Is the apparent variable a proxy for some other—completed family size as a proxy for duration of breast-feeding, for example? How plausible is the alternative causality, such as proneness to ovarian cancer causing subfertility? St Leger *et al.*¹¹ who reported on factors associated with cardiac mortality referring particularly to the consumption of wine, warned against naively imputing cause. They were rewarded by a lively correspondence dealing with latitude,¹² milk,¹² sandwiches,¹³ garlic,¹⁴ and high-density lipoprotein cholesterol¹⁵—exemplary platelet activity.¹⁶

I am grateful to Dr S J Pocock and co-workers for permission to reproduce the figure under question 16, and to Dr Moore and colleagues for permission to use figures 1 and 2.

References

- Gore SM. Assessing methods—transforming the data. *Br Med J* 1981; **283**:548-50.
- Moore MR, Meredith PA, Campbell BC, Goldberg A, Pocock SJ. Contribution of lead in drinking water to blood-lead. *Lancet* 1977; **ii**:661-2.
- Gore SM. Assessing methods—recognising linearity *Br Med J* 1981; **283**: 711-3.
- Langlands AO, Pocock SJ, Kerr GR, Gore SM. Long-term survival of patients with breast cancer: a study of the curability of the disease. *Br Med J* 1979; **ii**:1247-51.
- Pocock SJ, Shaper AG, Cook DG, *et al.* British Regional Heart Study: geographic variations in cardiovascular mortality, and the role of water quality. *Br Med J* 1980; **280**:1243-9.
- Altman DG. Statistics and ethics in medical research. Analysing data. *Br Med J* 1980; **281**:1473-5.
- Rose G, Shipley MJ. Plasma lipids and mortality: a source of error. *Lancet* 1980; **ii**:523-6.
- Marmot MG, Rose G, Shipley MJ, Thomas B. Alcohol and mortality: a U-shaped curve. *Lancet* 1981; **ii**:580-3.
- Gore SM. Assessing methods—art of significance testing. *Br Med J* 1981; **283**:600-2.
- Beral V, Fraser P, Chilvers C. Does pregnancy protect against ovarian cancer? *Lancet* 1978; **ii**:1083-7.
- St Leger AS, Cochran AL, Moore F. Factors associated with cardiac mortality in developed countries with particular reference to the consumption of wine. *Lancet* 1979; **ii**:1017-20.
- Segall JJ. . . . Milk. *Lancet* 1979; **ii**:1294.
- Rao LGS. French wine and death certificates. *Lancet* 1979; **ii**:1187.
- Slater NGP. . . . Or garlic. *Lancet* 1979; **ii**:1294.
- Ricci G, Angelico F. Alcohol consumption and coronary heart-disease. *Lancet* 1979; **ii**:1404.
- Manku MS, Oka M, Horrobin DF. Alcohol consumption and coronary heart-disease. *Lancet* 1979; **ii**:1404.

Sheila M Gore, MA, is a statistician in the MRC Biostatistics Unit, Medical Research Council Centre, Hills Road, Cambridge CB2 2QH.

No reprints will be available from the author.