Extrapolation errors in linear regression

A.D. Franklin
Department of Physics and Astronomy
University of Maryland
College Park
MD 20742

The use of various regression techniques to fit TL-dose data (in the linear region) to a straight line was recently discussed by Rendell (1985). Variation covering a range of about 6% was found among the intercepts on the abscissa. On the other hand, the errors arising from the extrapolation itself are rather larger and may well make the differences among regression techniques unimportant, at least until much more precise data are at hand.

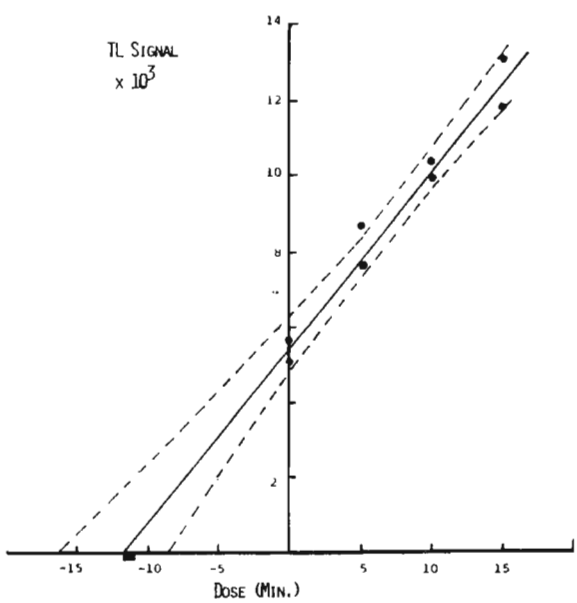

Figure 1 *Uncertainty in extrapolated Estimated Dose from TL data of Rendell (1985). Points are Rendell's data; solid line is simple linear regression line; dashed lines define 95% confidence band. The short heavy portion of the abscissa around -11 min. represents the spread of intercepts found by Rendell for various regression techniques.*

For illustration I have plotted the table of data given by Rendell in her paper in figure 1. The intersection of the confidence band with the x-axis gives the uncertainty in the intercept at the 95% confidence level. Numerically the intercept and its uncertainty are expressed by:

$$x_0 = -11.63 \begin{matrix} -3.14 \\ +4.51 \end{matrix} \quad \text{or} \quad \begin{matrix} +27\% \\ -39\% \end{matrix}$$

The statistical uncertainty is much larger than the spread in intercepts found by Rendell. Actually it is not necessary to calculate the confidence band as such. Mandel (1964) gives equations (eqn. 12.20 - 12.25, p. 281) for the confidence limits of the intercept on a specific horizontal line, in this case the x-axis.

It should be noted that where several runs are made at each dose and the means used to form the linear plot, the uncertainty in the intercept must reflect not only the scatter of the mean values about the regression line but also the uncertainty in the means. Note also the use of the 95% confidence interval for the intercept. When comparing a statistical with a systematic error, it is advisable to use a realistic confidence level for the former.

Improvement (reduction in the statistical uncertainty) can be brought about by increasing the precision of the data, the number of runs at each dose, and the number and range of doses. Because we can usually afford only a limited number of runs overall, the judicious distribution of these over the range of doses is an important question. These infuences can readily be seen using a simplified version of Mandel's eqn. 12.25. We may calculate the intervals Δx between the confidence limits for the intercept and expand it in terms containing $\hat{V}(\delta)$, the standard deviation *(? variance, Rev'r)* of the data points about the regression line. Only the leading term makes an appreciable contribution:

$$\Delta x = \frac{2\bar{y}\, t_c}{\hat{\beta}^2} \left[\hat{V}(\delta) / \sum_p n_p (x_p - \bar{x})^2\right]^{1/2}$$

The quantities t_c (the critical value of student's t) and β (the slope of the regression line) are not at our disposal, and $\sqrt{\hat{V}(\delta)}$ reflects the precision of the data, which we know we must optimize. We are left with the ratio $\bar{y} / \sqrt{\sum_p n_p (x_p - \bar{x})^2}$ to manipulate to achieve minimum error in the extrapolation, in which $\bar{x}$ and $\bar{y}$ are the mean values, the subscript p indicates the dose, and n_p is the number of values at the pth dose.

The sum $\sum_p n_p (x_p - \bar{x})^2$ can be maximized by maximizing the range of doses, and suggests putting greatest weight at the ends of the range. The average $\bar{y}$ can be minimized by making most of the runs at zero dose.

These conclusions can easily be made quantitative and practical for the special case in which we are confident of the linearity of the data. Then only two doses are required, zero and the largest possible within the linear range. Setting the total number of runs we can afford equal to

$$N = n_1 + n_3 \tag{2a}$$

where n_1 is the number of zero-dose and n_3 the number of maximum-dose runs (let $n_2 = 0$ for the moment), we can minimize the ratio $\bar{y} / \sqrt{\sum_p n_p(x_p - \bar{x})^2}$ with respect to n_1, holding N constant. The result is the remarkably simple prescription

$$\frac{n_1}{n_3} = \frac{y_3}{y_1} \tag{2b}$$

Equation 1 may be further manipulated, using eqns. (2a) and (2b), to yield a value for the total number of runs needed to yield a predetermined precision level in the intercept:

$$N = \frac{4k(1 + k)t_c^2}{R^2} \tag{3}$$

where R is the ratio,

$$R = (\Delta x / 2x_0) \; / \; (\sigma / y_1),$$

with σ the standard deviation of the TL value for a single run and

$$k = \frac{x_0}{x_3 - x_1} = \frac{\text{Intercept}}{\text{Maximum Laboratory Dose}}$$

We note that t_c depends upon N so eqn. 3 is solved comparing N/t_c^2 to values derived from statistical tables. Eqn. 3 is plotted in fig. 2 for several values of R for 95% confidence intervals.

To obtain the intercept with maximum efficiency, the maximum range of doses within which linearity is expected to hold is chosen and a few runs made at each end to yield a trial intercept and estimate of σ. These data are used to calculate k and R, after selection of the desired precision for the intercept. With k and R, a figure such as fig. 2, which is appropriate for a 95% confidence level, can be consulted to obtain N, the minimum total number of runs needed, and these divided between zero and maximum laboratory dose according to eqn. 2b. The value for N obtained from fig. 2 is a minimum. In practice, the number of runs at maximum dose should not be less than 3.

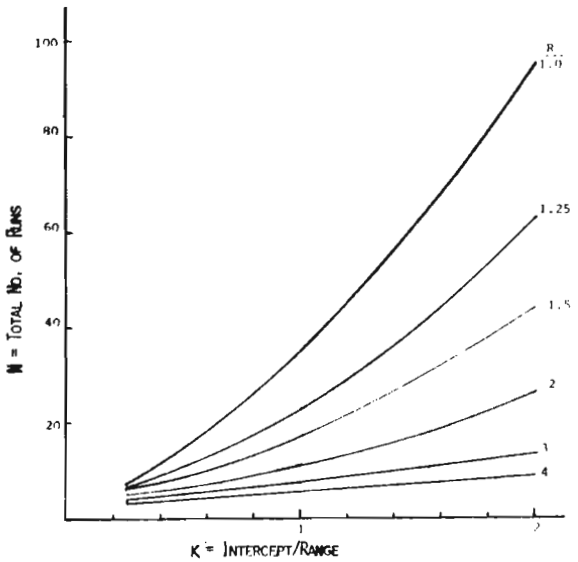

Figure 2
Total number N of runs needed to obtain a given error in the intercept. R is the ratio of ($\Delta x/2x_0$) to (σ/y_1) where Δx is the 95% confidence interval in the intercept x_0, Δ the standard deviation for replication of a single run, and y_1 the NTL.

The impact of this treatment may be illustrated again using Rendell's data, preserving the linear regression and the uncertainty in the data. Let us set $\Delta x/x_0$, the relative 95% confidence interval, at 40%, or ± 20%. The pooled standard deviation of the intensity data is 0.372 and y_1 = 5.229, leading to R = 2.8. The intercept and range are nearly equal, or k ~ 1. Consulting fig. 2, we find the minimum number of runs needed, N, to be 8, the same as the actual number. The ratio (y_3/y_1) of maximum dose/zero intensities is 2.29. If we concentrate all of the runs at zero dose and the maximum dose, and set n_1 = 6, n_3 = 2, we obtain a value for Δx = 4.9 min. compared to 7.1 min. for the uniform distribution across the 4 doses in the original data. This is a substantial reduction in error. Much of the reduction is already accomplished (Δx = 5.3 min) by dividing the runs equally between just the two doses zero and maximum.

It must be emphasized that this simple prescription can be used only when we have already proved the linearity of the data. To do that obviously requires runs in the middle of the range, as well as at the ends. Since it seems probable that the major departure from linearity arises from a quadratic term (e.g., the next term in the expansion of the exponential in a saturating curve), the most efficient test for curvature can be made using several (say $n_2 = [n_1 + n_3]/2$) runs exactly in the middle of the range (at $[x_1 + x_3]/2$). An appropriate test is to examine the F-statistic

$$F = \frac{\hat{V}(\delta)}{V_R}$$

where V_R is the replication variance of the data (the variance of the runs at each dose, pooled over all runs). If F exceeds the critical value at the confidence level chosen (e.g. 5%) for the degrees of freedom appropriate to $\hat{V}(\delta)$ (N-2) and V_R (N-1), curvature is probably present.

References

Mandel, J. (1964) The Statistical Analysis of Experimental Data, Wiley-Interscience, N.Y., Chap. 12.

Rendell, H.M. (1985) Problems with linear regression as applied to TL data, Ancient TL 3(3), 6-9.

Acknowledgements

I should like to acknowledge the support and encouragement of Professor William A. Hornyak and the Department of Physics and Astronomy of the University of Maryland. The work was carried out in conjunction with work under National Science Foundation Grant BNS 8319298.

Note A fuller report may be obtained from the author.

P.R. Reviewed by Morven Leese and accepted after revisions.