Behavior Research Methods & Instrumentation
1983, 15 (6), 609-610

COMPREG: A BASIC program to test for differences between several linear regression lines

G. N. WIGGANS, J. S. ANDREWS, and A. SAHGAL
Newcastle General Hospital
Newcastle upon Tyne, United Kingdom

Regression analysis provides a way of identifying relationships between two variables. Occasionally, it is necessary to show that two or more regression lines differ with respect to each other: Consider the case in which we wish to assess memory as a function of age. Having divided the subjects into an arbitrary number of age groups, we may plot (the logarithm of) memory performance on the y-axis and retention interval on the x-axis. Each age group will yield a regression line, and it is desired to test for significant differences between them; in other words, we wish to compare forgetting curves for different age groups. Statistical procedures have been developed that are based on combined analysis of variance and regression analysis techniques (analysis of covariance) and that provide detailed, step-by-step assessment of differences between regression lines (Woodward, 1972). However, calculation by hand can be extremely tedious, even when only a few data points are involved. We present a program that assesses differences between a number of regression lines; the user need only supply the desired level of significance at which the tests are to be made and the various x,y coordinates.

Method. The program first tests for an overall x-y relationship by calculating a (y on x) regression line for each group of data and performing an F test over all the groups to discover if, on the whole, the lines are a good fit. If not, then it may be concluded that no meaningful relationship exists, and further analysis is unnecessary. The user may nevertheless ask for a full analysis to be performed, despite some loosely fitting curves, by typing "Y" in response to the prompt "Is a full table required?"

For each line, the slope, intercepts, mean x-y coordinate, and individual F value (for the goodness of fit to its data set) are provided. There may be any number of data groups, and if any of them do not show a statistically significant relationship, then it may not be reasonable to proceed with the analysis. If each group of data has a significant x-y relationship, then there are two possible hypotheses concerning the regression lines. For example, in Figure 1, either the lines are mutually independent and may be skewed with respect to each other, or there is a common relationship, found in both groups, and the lines could be parallel to each other. Test 1 explores the hypothesis that parallel lines with a common slope would fit the data as well as do the independent lines previously calculated. If the F test is significant, an independent relation exists for each set, and no further calculations are required. In the memory example quoted above, the subjects in different age groups forget at significantly different rates. If the F value is not significant, and the use of parallel lines introduces no significant inaccuracy, then it cannot be argued that skewed lines represent the data significantly better than do parallel lines. In our example, we would be able to conclude that the rate of forgetting is unaffected by age. This interpretation implies that although different groups achieved different initial levels of performance, the rate of forgetting was similar. (Note, in this case we are ignoring the fact that parallel forgetting curves may not provide an equivalent measure of memory.)

If parallel lines are statistically acceptable, then a third hypothesis may be tested: Test 2 explores the possibility that a single common line would fit all the data as well as would parallel lines. This line would pass through the mean center coordinate of all the data, and the slope need not be the same as that of the parallel lines. Again, an F test is used to discriminate between this hypothesis and the hypothesis that parallel lines are significantly better than is a single line. If the F value is significant, then it can be concluded that a set of parallel

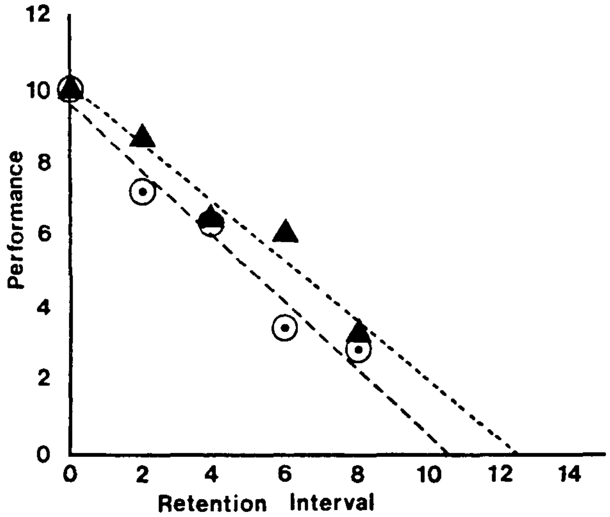

Figure 1. Key for data in graph. ⊙: data points–0,10; 2,7.3; 4,6.3; 6,3.4; 8,2.9. ▲: data points–0,10; 2,8.7; 4,6.4; 6,6.0; 8,3.3. In this example, the two lines are computed to be different, but with the same slope (–0.855).

G. N. Wiggans is now at: Department of Zoology, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom. J. S. Andrews is now at: Department of Pharmacology, Emory University, Atlanta, Georgia 30322. A. Sahgal is at: MRC Neuroendocrinology Unit, Newcastle General Hospital, Westgate Road, Newcastle upon Tyne NE4 6BE, United Kingdom.

relationships exists; otherwise, we conclude that the data can best be fitted by a single common line. If a single line is accepted (implying, in our example, that memory is unaffected by age), then a test of the significance of the relationship is made using an F test in the same way as is done for the independent lines.

Language and Program. The 4K program is written in the enhanced form of BASIC used by the BBC microcomputer. However, only standard commands and functions are used, and the program easily can be adapted for use in, for example, Apple, PET, TRS, Sinclair, and Hewlett-Packard machines. Probabilities of the computed F statistics are estimated with a subroutine provided by Ogasawara (1982) for the Apple II. In the BBC language, the numeric variables can store any positive or negative number from 2×10^{-39} to 2×10^{38}, which can include a decimal point. However, real numbers are only stored to nine-figure accuracy. This should be adequate for most purposes, although (calculated) probabilities close to the selected significance level should be carefully scrutinized.

When the program is RUN, information must be entered concerning: (1) significance level desired, (2) whether or not a full table of results is required, regardless of individual test outcomes (otherwise, the analysis is terminated at the relevant stage), and (3) the number of data sets and associated number of points. Note that the regression line(s) will pass through x and y intercepts as well as the mean center coordinate, and this information may be used to draw the line on graph paper. The slopes are always expressed as a change in the y variable for a unit increase in the x variable.

Availability. A full listing of the program, together with sample printout and comments on use, may be obtained for a fee of $5 to cover costs. The program is also available on a mini-floppy diskette for the BBC machine at a cost of $10. Please specify whether a 40- or 80-track system is in use. Money orders should be made payable to the MRC Neuroendocrinology Unit.

REFERENCES

Ogasawara, T. H. (1982). The calculation of the significance level of F, t, and r on the Apple II. *Behavior Research Methods & Instrumentation*, **14**, 492-493.

Woodward, R. H. (1972). Linear relationship between two variables. In O. L. Davies & P. L. Goldsmith (Eds.), *Statistical methods in research* (pp. 178-236). Edinburgh, Scotland: Oliver & Boyd.

(Revision accepted for publication November 28, 1983.)