Using Multiple Regression to Determine the Number of Factors to Retain in Factor Analysis

Keith Zoski
University of Toledo

Stephen Jurs
University of Toledo

A simple, analytical approach using multiple regression analysis is presented as a way to determine the number of factors to retain in a factor analysis. Two regression lines are found from the points in a scree plot and the number of retained factors is chosen at a point that maximally separates the two regression lines. Applications of the technique to data from the literature suggest that the results agree closely with solutions based on the somewhat subjective visual scree test and may be better than those from the analytical CNG method.

The number of factors to retain in a factor analysis has long been an important problem (Hakstian & Muller, 1973; Crawford, 1975; Horn & Engstrom, 1979; Hakstian, Rogers, & Cattell, 1982; Kano, 1990). This is critical because it demands a decision that affects the factor parameters and the interpretability of the factors (Lambert, Wildt, & Durand, 1990).

The most frequently used method for determining the number of factors is to select only those factors whose eigenvalues exceed 1.0 (Kaiser, 1970; Kaiser & Caffrey, 1965). Critics of this method (Gorsuch, 1983) are concerned that many times there is not a clear break among the eigenvalues at the 1.0 value and that underestimating or overestimating communalties would change the number of retained factors when the eigenvalues greater than 1.0 rule is used. Therefore, the selection or deletion of some factors may be a function of an arbitrary rule that is not sensitive to the nature or pattern of the data.

An approach that considers the relation of the eigenvalues to one another as well as their actual values is the scree test. Cattell (1966) first proposed the scree test to separate trivial from non-trivial factors. The procedure required one to plot the eigenvalues in decreasing order. The graph contained the values of the eigenvalues on the ordinate and the factors on the abscissa. A straight line could be drawn on the graph through the points associated with the smaller eigenvalues. The points near this line were judged trivial and the points above and to the left of the line were judged to be non-trivial (Cattell, 1978; Cattell & Vogelman, 1977; Cattell & Jaspers, 1967). Horn and Engstrom (1979) provided statistical support for the scree test.

Cattell and Vogelman (1977) and Cattell (1978) presented guidelines for this visual procedure. These guidelines, as summarized by Zoski and Jurs (1990), are:

1. Three sequential points form an undesirably low limit for drawing a scree.
2. The points on the part of the curve that one should consider scree should fit tightly.
3. The slope of the scree should not approach vertical. Instead, it should have an angle of 40° or less from the horizontal, that is, a slope of the tangent less than -.84.
4. In the case of multiple screes falling below 40°, the first scree on the left is the arbitrator.
5. Generally, a sharp, albeit sometimes small, break in the vertical level exists between the last point of the curve and the left-most point of the scree.

However, problems with this procedure can occur when there are multiple breaks in the eigenvalue curve, with several straight lines in the graph. It may be difficult to select as well as to justify one break over another (Gorsuch, 1983). Moreover, critics of visual approaches are concerned about researchers seeing what they want to see in the data unless they are constrained by a mechanical decision-making rule. This position is demeaning to the researchers and shifts the demand for objectivity over subjectivity to the final stages of research (decisions and conclusions) and ignores the more critical phase (research problem definition and variable selection). An analytical, programmable approach does have some appeal, if it provides results that are consistent with those obtained using the guidelines above. We propose that multiple regression techniques can be used to provide such a solution.

The Multiple Regression Approach

Gorsuch and Nelson (1981) developed an analytical

method for determining the number of factors to retain. The Cattell-Nelson-Gorsuch scree test requires one to compare the slope of the first three roots with the slope of the next three roots. Then the slope of roots 2, 3, and 4 is compared with the slope of roots 5, 6, and 7. This process continues so that all sets of three factors are compared. The number of factors is found where the difference between the slopes is greatest.

Because only three points are used to determine the slopes, the analysis is not based on as much information as is possible. Thus, we propose a somewhat different approach using multiple regression to accomplish the same thing; objective determination of the number of factors that is sensitive to the data.

The rationale for a regression approach is straightforward. It parallels the statistical work of Horn and Engstrom (1979) on Cattell's scree test using Bartlett's chi-square test (1950, 1951). The method used here provides virtually the same decision as the visual scree test but can be easily programmed. It uses a regression approach where the ordered eigenvalues are thought of as points in a scatterplot. One can then form two regression lines, one for the important factors and another for the scree or trivial factors. The decision about the number of factors to retain corresponds with the maximal differences between the two regression lines.

To use all the eigenvalues, form and compare these pairs of regression lines:

line 1	line 2
(points 1 through 3)	(points 4 through m)
line 3	line 4
(points 1 through 4)	(points 5 through m)
line 5	line 6
(points 1 through 5)	(points 6 through m)
•	•
•	•
•	•
line (m-2)	line (m-1)
(points 1, 2, ... (m-3))	(points (m-2), (m-1), m)

The slope of these regression lines will, of course, be negative and can be compared by the usual formulae (Howell, 1987, pp. 222, 239-240):

$$b = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2} \qquad [1]$$

$$t = \frac{b_1 - b_2}{s_{b_1 - b_2}} \qquad [2]$$

with

$$s_{b_1 - b_2} = \sqrt{\frac{s^2_{Y \cdot X_1}}{s^2_{X_1}(N_1 - 1)} + \frac{s^2_{Y \cdot X_2}}{s^2_{X_2}(N_2 - 1)}} \qquad [3]$$

and when homogeneity of error variances is assumed, we can pool:

$$s^2_{Y \cdot X} = \frac{(N_1 - 2)\left(s^2_{Y \cdot X_1}\right) + (N_2 - 2)\left(s^2_{Y \cdot X_2}\right)}{N_1 + N_2 - 4} \qquad [4]$$

The salient factors are those with eigenvalues in the odd numbered line of the line pair where the t-test is maximized (highest value). The even numbered line of the pair denotes the scree line. Some analysts may choose not to include the last factor. Note that neither the CNG nor the multiple regression approach would be appropriate when there are only one or two factors.

Examples

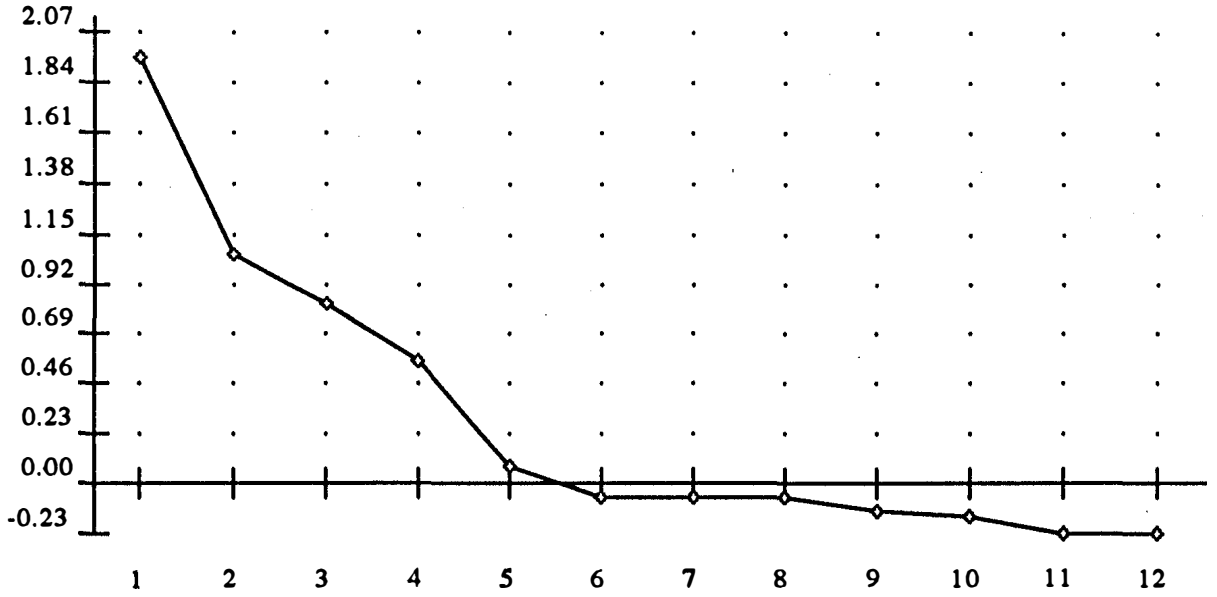
We have compared the multiple regression approach to the CNG approach using several data sets from the literature. Preliminary results indicate that the multiple regression approach usually agrees with a visual scree test and often provides a better solution than the CNG method.

Example 1 is based on eigenvalues taken from Cliff (1970). The eigenvalues are plotted in Figure 1. Table 1 contains the slopes of the regression lines and the t values for the multiple regression approach and the slopes and differences for the CNG approach (* indicates highest value). Note that in this case both procedures indicate that there are five factors and this agrees with a visual analysis of the plotted eigenvalues in Figure 1.

Table 1 Comparison of Multiple Regression and CNG Approaches: Example 1

	MR			CNG		
# of factors	slope 1	slope 2	t	slope 1	slope 2	difference
3	-.563	-.071	4.044	-.563	-.310	.253
4	-.441	.038	6.713	-.250	-.067	.183
5	-.426	-.032	8.448*	-.377	-.001	.376*
6	-.380	-.038	6.814	-.310	-.032	.278
7	-.323	-.042	3.752	-.067	-.043	.024
8	-.272	-.038	1.899	-.001	-.048	.047
9	-.234	-.040	0.890	-.032	-.040	.007

Figure 1 Scree Plot from Cliff (1970, p. 165, CS 600).

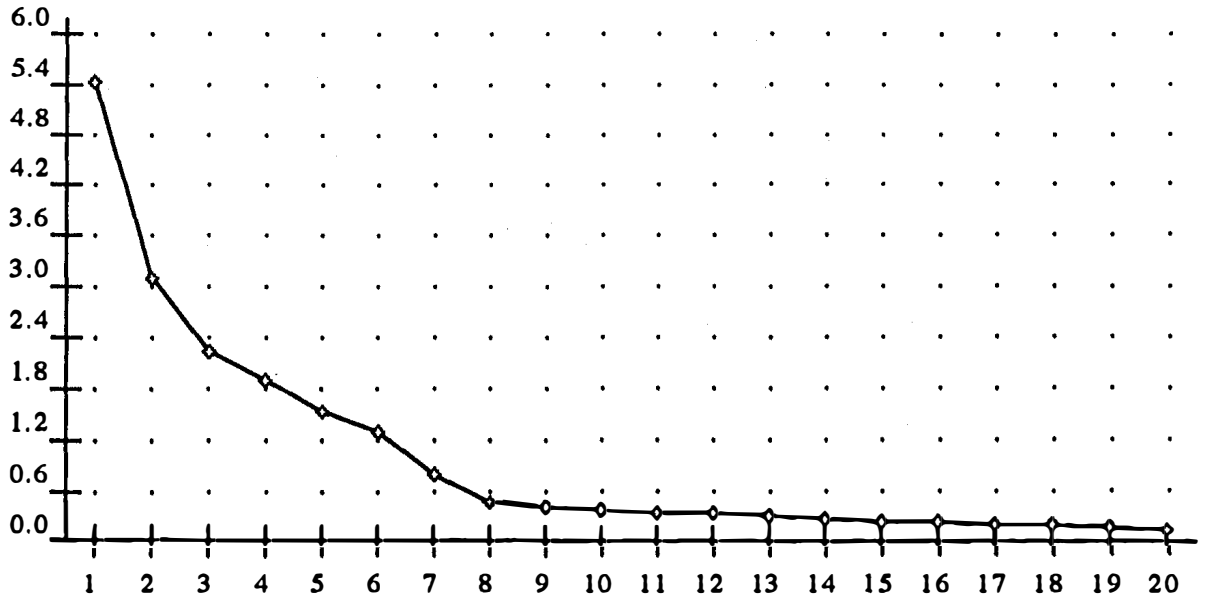


The second example is taken from Tucker, Koopman, & Linn (1969, p. 442). The plot of the eigenvalues is given in Figure 2 and the results from the multiple regression approach and the CNG approach are listed in Table 2. The data set was meant to have seven factors. The CNG approach yielded three factors and the multiple regression approach did yield the expected seven factors. Visual inspection of Figure 2 confirms that a seven factor solution is appropriate.

Table 2 Comparison of Multiple Regression and CNG Approaches: Example 2

	MR			CNG		
# of factors	slope 1	slope 2	t	slope 1	slope 2	difference
3	-1.595	-.084	6.346	-1.595	-.300	1.295*
4	-1.149	-.067	6.985	-.610	-.360	.250
5	-.904	-.051	7.327	-.360	-.415	.055
6	-.737	-.033	7.405	-.300	-.195	.105
7	-.651	-.023	7.665*	-.360	-.045	.315
8	-.590	-.022	7.563	-.415	-.030	.385
9	-.525	-.021	6.694	-.195	-.020	.175
10	-.465	-.021	5.507	-.045	-.025	.020
11	-.413	-.021	4.277	-.030	-.025	.005
12	-.367	-.020	3.176	-.020	-.025	.005
13	-.328	-.020	2.266	-.025	-.020	.005
14	-.295	-.019	1.555	-.025	-.020	.005
15	-.267	-.021	1.013	-.025	-.020	.005
16	-.243	-.020	.622	-.020	-.015	.005
17	-.222	-.025	.335	-.020	-.025	.005

Figure 2 Scree Plot from Tucker, Koopman and Linn (1969, p. 442, Middle 7)

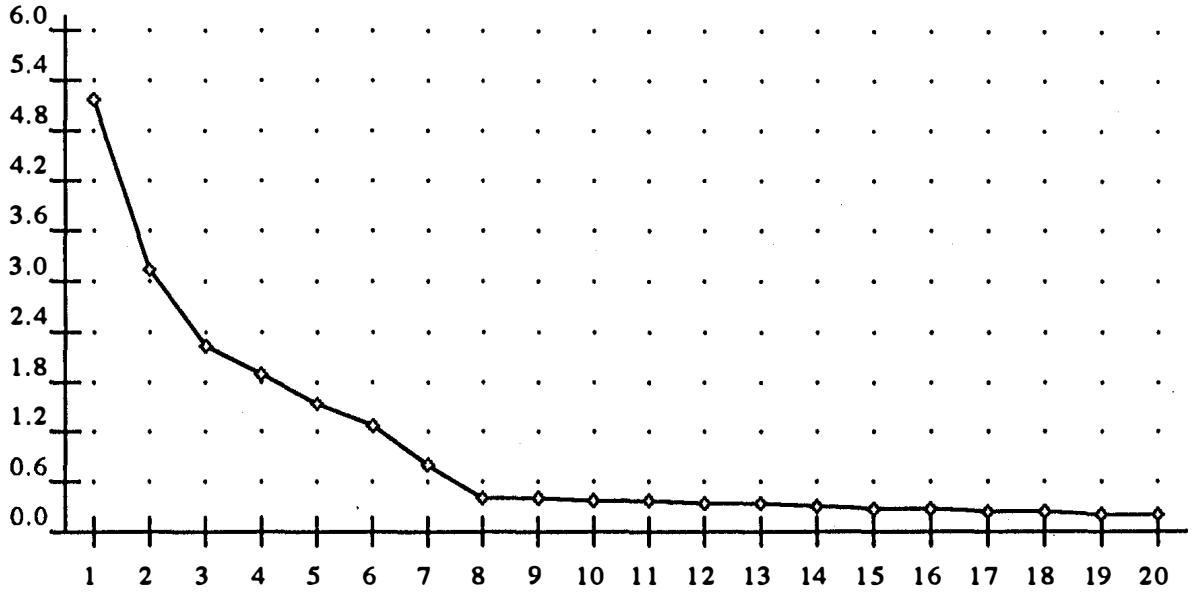


The third example was also taken from Tucker, Koopman, & Linn (1969, p. 442). This data set was intended to have seven factors and a visual inspection of the scree plot in Figure 3 suggests that there are seven factors. The analyses presented in Table 3 indicate that the CNG approach yielded only three factors and the multiple regression approach yielded eight factors. This example shows that results from the multiple regression approach may not always agree with results from a visual approach, but the technique seemed to work better than the CNG method for these data.

Table 3 Comparison of Multiple Regression and CNG Approaches: Example 3

			MR			CNG
# of factors	slope 1	slope 2	t	slope 1	slope 2	difference
3	-1.475	-.081	5.855	-1.475	-.315	1.160*
4	-1.071	-.063	6.818	-.610	-.365	.245
5	-.850	-.047	7.522	-.345	-.440	.095
6	-.702	-.029	7.944	-.315	-.210	.105
7	-.625	-.018	8.369	-.365	-.015	.350
8	-.574	-.018	8.443*	-.440	-.010	.430
9	-.513	-.018	7.401	-.210	-.010	.200
10	-.455	-.195	5.974	-.015	-.025	.010
11	-.403	-.019	4.554	-.010	-.025	.015
12	-.358	-.018	3.341	-.010	-.020	.010
13	-.320	-.178	2.365	-.025	-.015	.010
14	-.287	-.177	1.611	-.025	-.020	.005
15	-.260	-.018	1.047	-.020	-.020	.000
16	-.235	-.015	.647	-.015	-.020	.005
17	-.215	-.015	.356	-.020	-.015	.005

Figure 3 Scree Plot from Tucker, Koopman & Linn (1969, p. 442, Formal 7)



Conclusions

Multiple regression is a versatile set of techniques for which there are diverse applications. Our results indicate that multiple regression can be used successfully to determine how many factors to retain in a factor analysis. Preliminary analyses suggest that the results will usually agree with results from a visual scree test and the results often are better than those from alternative analytic techniques such as the CNG method. Further use of the multiple regression method will identify the strengths and limitations of this approach.

References

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, 3,* 77-85.

Bartlett, M. S. (1951). A further note on tests of significance in factor analysis. *British Journal of Psychology, 4,* 1-2.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245-276.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavior and life sciences.* New York: Plenum.

Cattell, R. B. & Jaspers, J. (1967). A general plasmode No. 30-10-5-2 for factor analytic exercises and research. *Multivariate Behavioral Research Monographs, 2,* 1-212.

Cattell, R. B. & Vogelman, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research, 12,* 289-325.

Cliff, N. (1970). The relation between sample and population characteristic vectors. *Psychometrika, 35,* 163-178.

Crawford, C. B. (1975). Determining the number of interpretable factors. *Psychological Bulletin, 82,* 226-237.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Gorsuch, R. L. & Nelson, J. (1981). *CNG scree test: An objective procedure for determining the number of factors.* Paper presented at the annual meeting of the Society for Multivariate Experimental Psychology.

Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number-of-factors rules with simulated data. *Multivariate Behavioral Research, 17,* 193-219.

Hakstian, A. R. & Muller, V. J. (1973). *Users manual to accompany the Alberta General Factor Analysis Program,* Edmonton, Alberta: Division of Educational Research Services, University of Alberta.

Horn, J. L. & Engstrom, R. (1973). On the subjective character of the empirical base of the structure-of-intellect model. *Psychological Bulletin, 80,* 33-43.

Howell, D. C. (1987). *Statistical methods for psychology.* Boston: Duxbury Press.

Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika, 35,* 401-415.

Kaiser, H. F. & Caffrey, J. (1965). Alpha factor analysis. *Psychometrika, 30,* 1-14.

Kano, Y. (1990). Noniterative estimation and the choice of the number of factors in exploratory factor analysis. *Psychometrika, 55,* 277-291.

Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1990). Assessing sampling variation relative to the number of factors criteria. *Educational and Psychological Measurement, 50,* 33-48.

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969) Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika, 34,* 421-459.

Zoski, K. & Jurs, S. (1990). Priority determination in surveys: An application of the scree test. *Evaluation Review, 14,* 214-219.