Statistical Analysis Of the Goodness of Classical Factor Analysis Regression (CFAR)

John T. Scott, Jr., and Allen Fleishman

Bulletin 759
Agricultural Experiment Station
College of Agriculture
University of Illinois at Urbana-Champaign

ABSTRACT

Classical Factor Analysis Regression is a statistical technique using factor analysis to calculate a linear function similar to ordinary least squares regression. CFAR has been recommended to replace OLS in cases where there is high multicollinearity among the explanatory variables and when there are errors in the variables as well as when there may be outliers in the data. Mathematical derivation of the distribution functions of the CFAR coefficients has so far not been done. The research reported here is a Monte Carlo study to determine the statistical goodness of CFAR compared to OLS. The results of this research show that CFAR is superior to OLS whenever there is high multicollinearity or errors in the variables. The variances of the *b* coefficients are smaller for CFAR and the biases asymptotically approach zero. Also the distributions appear to be normally distributed so statistical tests based on the normal distribution can be used.

Key Words: Factor Analysis, Principal Components, Monte Carlo, Factor Analysis Regression, Ordinary Least Squares, Multicollinearity, Errors in the Variables, Outliers.

John T. Scott, Jr., is professor of agricultural economics and Allen Fleishman is assistant in psychology at the University of Illinois at Urbana-Champaign. The authors wish to thank reviewers, particularly David A. Lins, economist, ESCS, U.S. Department of Agriculture, stationed with the Department of Agricultural Economics at the University of Illinois at Urbana-Champaign. This project was supported by the University of Illinois Research Board, the Agricultural Experiment Station, and the Department of Psychology.

Statistical Analysis of the Goodness of Classical Factor Analysis Regression (CFAR)

John T. Scott, Jr., and Allen Fleishman

Regression from factor analysis has been suggested by several authors as an alternative for ordinary least-squares (OLS) regression when the explanatory variables are subject to error or there is significant multicollinearity (Kloek and Mennes, 1960; Amemiya, 1966; Scott, 1966; Lawley, 1973).

In the case of multicollinearity when the determinant of the explanatory variables correlation matrix approaches zero, it is well known that OLS can give spurious results. The regression coefficients frequently do not correspond to either the theory or the zero-order correlation coefficients, and the variances are inconsistent. Also, when there are errors in the variables (which is normal with economic data), it has been shown (Johnston, 1963) that OLS regression coefficients are biased and that the associated variances are not only inconsistent but generally underestimate the true variances. These results follow from violation of two OLS assumptions: that the explanatory variables are independent, and that the explanatory variables are known, fixed numbers without error. For example, if we assume there are errors in all variables, then the OLS model becomes:

$$(1) \qquad (y - v) = b_1 (x_1 - u_1) + b_2 (x_2 - u_2) + \ldots + b_k (x_k - u_k) + w$$

or in matrix notation:

$$(2) \qquad (Y - V) = (X - U)B + W;$$

where Y is the $n \times 1$ vector of observed values of the dependent variable adjusted for the mean,

V is the $n \times 1$ vector of errors in Y,

X is the $n \times k$ matrix of n observations of the k explanatory variables adjusted for the mean,

U is the $n \times k$ matrix of errors in X, and

W is the $n \times 1$ vector of residuals from regression which may include specification errors as well as other errors not included in V.

Minimizing W with respect to $\hat{B}$ results in

$$(3) \qquad B = [(X - U)'(X - U)]^{-1}[(X - U)'(Y - V)], \text{ or}$$
$$\hat{B} = [(X - U)'(X - U)]^{-1}[X'Y - U'Y - X'V + U'V].$$

We can simplify the foregoing expression by making three additional assumptions not found in the classic assumptions underlying OLS: the errors in Y are uncorrelated with those in X; the errors in X are uncorrelated with Y; and the errors in Y are uncorrelated with X. These assumptions are reasonable and only moderately restrictive. Then the last three terms in equation 3 become zero and $\hat{B}$ becomes:

$$\hat{B} = [(X - U)'(X - U)]^{-1} X'Y, \text{ or} \tag{4}$$
$$\hat{B} = [X'X - 2X'U + U'U]^{-1} X'Y.$$

Assuming that the errors in X are independent of X itself (which is still another assumption), the middle term of the inverse in equation 4 drops and then equation 4 becomes:

$$\hat{B} = [X'X - U'U]^{-1} X'Y. \tag{5}$$

To estimate this modification of the OLS model, we need to know as a minimum the variance-covariance matrix of the errors in X. The problem is that this is rarely if ever known in the real world. If we make the assumption that the errors in X are uncorrelated with each other, then equation (5) becomes the ridge regression estimator:

$$\hat{B} = [X'X - \alpha I]^{-1} X'Y, \tag{6}$$

where α is another parameter which must be estimated, which is no trivial task (Marquardt, 1970; McDonald and Galarneau, 1975).

While empirical results from factor analysis regression are substantially better than those from OLS based on *a priori* expectations (Amemiya, 1966; Scott, 1966; Oehrtman, 1968; Bursch *et al.*, 1972), the statistical properties of the estimators in factor analysis regression have not been derived mathematically, nor does such a derivation appear tractable.[1] The alternative method generally acceptable for obtaining the statistical characteristics of an estimator is to perform a Monte Carlo study of the estimator. The development of such a study involving classical factor analysis regression and its results are reported here.

CLASSICAL FACTOR ANALYSIS REGRESSION

The factor analysis statistical model assumes that a large number of variables can be described adequately by a smaller number of factors:

$$Z = AF + U \tag{8}$$

[1] The senior author has worked on this problem and consulted others including R. A. Wijsman, Department of Mathematics, and Leyard Tucker and Charles Lewis, Department of Psychology, University of Illinois. All suggested the Monte Carlo approach.

where Z is the $h \times n$ matrix of n observations of all h real variables involved,

A is the $h \times m$ matrix of regression coefficients, usually referred to as factor coefficients or factor loadings, with $m < h$,

F is the $m \times n$ matrix of the n values of the m factors, and

U is the $h \times n$ matrix of the n residuals associated with the h variables.

It is assumed that $E(U) = 0$; $E(F) = 0$; $E(UU') = V$, a diagonal matrix; $E(FF') = I$; and $E(Z) = 0$; and further, that U and F are independent and have multivariate normal distributions.

A number of methods have been developed to "extract factors" or calculate the coefficient matrix to meet the foregoing statistical assumptions (Hotelling, 1933; Guttman, 1940; Lawley, 1940; Rao, 1955; Joreskog, 1962; and others).

A derivation of regression from factor analysis was developed which, for purposes of differentiation, is called "classical factor analysis regression" or "see far — CFAR" (Scott, 1970). Since CFAR is much simpler and easier to obtain than the earlier factor analysis regression derivations, it should appeal to practitioners for their research work. The results from CFAR are as good as, or better than, those from the earlier factor analysis regression methods.

Using standardized variables in ordinary least-squares regression (OLS) results in the following equation to estimate the regression coefficients:

$$\hat{B} = R_{xx}^{-1} R_{xy} \tag{9}$$

where $\hat{B}$ is the $k \times 1$ vector of regression coefficients,

R_{xx} is the $k \times k$ correlation matrix of the explanatory variables, and

R_{xy} is the $k \times 1$ vector of correlations between the dependent and the explanatory variables.

The factor analysis statistical model, equation 8, allows for errors in all variables and can be used in situations involving high multicollinearity. Factor analysis regression may also give improved results over OLS when the data set contains a number of extreme values or outliers. Thus the assumptions of factor analysis seem more appropriate for use with real economic data than do the assumptions of ordinary least squares.

Let the matrix R be the matrix of correlations among the explanatory variables augmented with the correlations between the dependent and the explanatory variable. This matrix has dimensions $k + 1$ by $k + 1$. Using matrix R, obtain the factor loading matrix A, by least

squares or maximum-likelihood (Lawley, 1940; Whittle, 1952; Rao, 1955; Joreskog, 1962). Then:

$$AA' + V = \hat{R}, \tag{10}$$

where V is a diagonal matrix and is the difference between diagonal (AA′) and I, the identity matrix; and $\hat{R}$ is the maximum-likelihood estimate of the full correlation matrix.

Then partition $\hat{R}$ into $\hat{R}_{xx}$, the k by k estimated correlation matrix of the explanatory variables, and $\hat{R}_{xy}$, the k by 1 vector of estimated correlations between the explanatory variables and the dependent variable. Use these estimated correlations in the OLS regression coefficient estimating equation to get the CFAR coefficients, $\tilde{B}$, so that:

$$\tilde{B} = \hat{R}_{xx}^{-1}\hat{R}_{xy}. \tag{11}$$

The long-run efficacy of any statistical method at least partly depends upon having knowledge of the statistical properties of the method, especially of the characteristics of the parameter estimators. We try to obtain this knowledge for the CFAR estimators in the Monte Carlo study.

CONCEPT AND PROCEDURE OF THE MONTE CARLO STUDY

The concept of this study was to use observations of a population with a dependent variable that is associated with observations of a set of explanatory variables, all observations assumed to be without sampling error. Then the OLS regression estimators for this set were assumed to be the parameters or expected values of the estimators. To this original population **random normal errors** were added to all variables. This new population with errors in all variables then is the observed population to be sampled for the Monte Carlo experiment. From this set of observations with measurement errors, draw a large number of random samples of various sample sizes, and estimate the regression for each sample with CFAR and OLS. Then, examine the population of coefficients obtained from these regressions by comparing the mean of each estimator with its corresponding parameter (whether or not there is a bias or how the bias behaves), and examine how closely the distribution of the estimators corresponds to the normal distribution. Desired characteristics for the CFAR coefficients would be unbiasedness, efficiency, and normality. Two additional important characteristics are compared for CFAR and OLS. These are the mean-square error for the prediction: $\Sigma(Y - \hat{Y})^2$, where Y is the original population value without

error and $\hat{Y}$ is the predicted value based on the estimation from the observed variables with error; and, the mean-square error for the regression coefficients: $\Sigma(\beta - \hat{\beta})^2$, where β is the OLS estimate from the original population without sampling error as the parameter, and $\hat{\beta}$ is the regression coefficient estimated by CFAR and OLS from the observed variables with error included. If $\Sigma(Y - \hat{Y})^2$ and $\Sigma(\beta - \hat{\beta})^2$ estimated by CFAR are less than when estimated by OLS, then this is evidence that CFAR is in some sense a better estimating procedure. These latter two criteria are usually considered more important for small sample size than are unbiasedness and normality.

MONTE CARLO PROCEDURE

Three original populations were selected, each with one dependent variable and twelve explanatory variables. Then, using four sets of associative characteristics and two variable generating procedures, 24 populations which are now called initial populations were generated having various internal characteristics.[1]

For this Monte Carlo experiment, a substantial range was generated in the associative characteristics because of the wide range of these characteristics found in empirical observations. For example, with most economic data more of the intercorrelations are positive than negative; some socioeconomic variables have high intercorrelations — as an example, prices of substitutes or economic variables over time and time series; and some socioeconomic variables occasionally have low intercorrelations, typically those from cross-section data and survey questionnaires. Also, the range in the proportion of the variance of the dependent variable explained by regression is frequently quite large. Therefore we believed that it was imperative to use different initial populations representative of a wide range of various associative characteristics.

Assuming that the observations in the initial populations were without error, we calculated the OLS regression for each of the initial 24 populations and assumed the coefficients from these regressions to be the parameters or expected values of the coefficients for each respective initial population.

A random normal error structure was added to all variables in each initial population so that we then had 24 populations with errors in all variables which became the "observable" values to be sampled. Then

[1] By associative or internal characteristics is meant the interrelationship of the variables within any one initial population.

from each of the 24 populations with errors in the variables, 100 samples each of size 16, size 64, and size 256 were drawn, with the sampling error structure potentially different with each draw, simulating drawing from an infinite population. Thus, there were 7,200 sample variance-covariance matrices drawn for this experiment. An OLS regression was run for each of the 7,200 samples, each with 12 explanatory variables.

To obtain factor analysis regression, the sample correlation matrix must be factor-analyzed and a reproduced correlation matrix calculated from the factor-loading matrix. An important consideration in factor analysis is the number of factors to be extracted from the sample correlation matrix. The factor-reproduced correlation matrix will differ, depending upon the number of factors extracted. With 12 explanatory variables, we believed a maximum of six factors should be ample to describe the underlying phenomena. Not knowing the change in characteristics of the CFAR estimators that might occur as a result of using different numbers of factors, we extracted and reproduced a correlation matrix from one factor, from two factors, etc., up to and including six factors, using the factors explaining the most cumulative variance in all cases. Thus from each sample correlation matrix there were six reproduced correlation matrices. A classical factor analysis regression equation was estimated from each of these six reproduced correlation matrices, making 43,200 CFAR equations, each with 12 explanatory variables, that were estimated for this Monte Carlo experiment.

Since factor extraction and communality estimation by least squares or maximum-likelihood is much more expensive than obtaining the principal components, the experiment included obtaining the factors by principal components as well as by a statistical routine, and calculating the regression coefficients the same way from each extraction method to compare the results between image factor-analysis extraction and principal components. There were actually 86,400 CFAR equations — half using statistical factor-analysis extraction and half using principal components.[1]

[1] L. R. Tucker, Department of Psychology, University of Illinois, suggested at the time we ran the calculations of the experiment that we factor-analyze only the explanatory variable correlation matrix rather than the augmented matrix to save computer time on such a large experiment. The estimating equation then becomes $\hat{B} = \hat{R}_{xx}^{-1} R_{xy}$ rather than $\hat{B} = \hat{R}_{xx}^{-1} \hat{R}_{xy}$. Although the difference in results is probably only marginal, we now believe that conceptually the augmented matrix should be the matrix to factor-analyze. We have no way of knowing whether a marginal improvement would have been great enough to compensate for the cost of the extra calculation.

ASSOCIATIVE CHARACTERISTICS

Tables 1 through 6 give the details of the associative or internal characteristics of each of the initial 24 populations. Table 1 shows the four initial populations with the four different associative characteristic ranges generated by the two-factor generator from the first original population (see appendix). Table 2 shows four additional initial popu-

Table 1. Characteristics of Four Initial Populations Including Regression Coefficients (1-4) Produced With the Two-Factor Generator From Original Population

Correlation range	Associative characteristics							
	Initial population 1		Initial population 2		Initial population 3		Initial population 4	
	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}
−1.0 to −.9								
−.9 to −.8								
−.8 to −.7								
−.7 to −.6								
−.6 to −.5		2						
−.5 to −.4								
−.4 to −.3	1	1		2				1
−.3 to −.2	1		1	1				1
−.2 to −.1		3	1			2	1	2
−.1 to 0	1	3	1	7	3	8	1	6
0 to .1		3	1	4	3	14	4	6
.1 to .2	1	5	1	5	6	31		5
.2 to .3	1	2		9		11	5	13
.3 to .4		6	3	11			1	8
.4 to .5		5	3	19				12
.5 to .6	2	5		8				5
.6 to .7	2	8						5
.7 to .8	3	17						2
.8 to .9		6						
.9 to 1.0								

Statistical estimator	Population regression parameters (by OLS)			
	Initial population 1	Initial population 2	Initial population 3	Initial population 4
R^2	.6806	.3474	.0647	.1195
b_1	−.0283	−.0111	−.0005	−.0277
b_2	.1071	.1068	.0593	.0284
b_3	.1129	.0850	.0526	.0549
b_4	.2457	.1575	.0904	.1959
b_5	−.1121	−.0930	−.0467	−.0450
b_6	−.0035	.0077	.0097	−.0141
b_7	.0528	.0627	.0312	.0058
b_8	.1609	.1416	.0764	.0578
b_9	−.0537	−.0443	−.0193	−.0197
b_{10}	.0881	.0896	.0510	.0222
b_{11}	.1424	.1262	.0718	.0503
b_{12}	−.0755	−.0688	−.0321	−.0225

lations with the four associative ranges generated by the four-factor generator from the first original population. Tables 3 and 4 show the corresponding eight additional initial populations generated from the second original population. Tables 5 and 6 show the corresponding eight additional initial populations generated from the third original population.

Table 2. Characteristics of Four Initial Populations (5-8) Produced With the Four-Factor Generator From Original Population One

Correlation range	Associative characteristics							
	Initial population 1		Initial population 2		Initial population 3		Initial population 4	
	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}
−1.0 to −.9								
−.9 to −.8								
−.8 to −.7								
−.7 to −.6								
−.6 to −.5								
−.5 to −.4	1	12						
−.4 to −.3		1						1
−.3 to −.2		5	1	3				4
−.2 to −.1		4		6		3	1	4
−.1 to 0	2	5	1	8	2	15	2	8
0 to .1	1	4	3	10	6	25	4	13
.1 to .2	1	8	1	10	4	20	2	7
.2 to .3	1	7	2	8		3	3	12
.3 to .4	2	5	3	11				7
.4 to .5		6	1	10				4
.5 to .6	2	5						5
.6 to .7	2	9						1
.7 to .8		5						
.8 to .9								
.9 to 1.0								

Statistical estimator	Population regression parameters (by OLS)			
	Initial population 1	Initial population 2	Initial population 3	Initial population 4
R^2	.6626	.316	.0513	.1123
b_1	−.0570	−.0442	−.0190	−.0158
b_2	.1346	.1230	.0608	.0381
b_3	−.0732	−.0275	−.0018	−.0681
b_4	.3206	.1944	.0999	.2408
b_5	.1251	.0966	.0408	.0403
b_6	−.05333	−.0265	−.0046	−.0336
b_7	.0337	.0422	.0216	.0033
b_8	.2080	.1766	.0846	.0641
b_9	.0184	.0062	−.0003	.0112
b_{10}	.0554	.0582	.0338	.0134
b_{11}	.1953	.1516	.0732	.0775
b_{12}	−.1626	−.1444	−.0598	−.0397

The associative characteristic sets were developed on the following criteria: Associative characteristic set 1 was to have a high R^2 and a wide range of frequency of r_{xy} and r_{xx} but with a large share of the zero-order correlations in the upper range (above 0.6); set 2 was to have a medium R^2 and zero-order correlation coefficients not as high, but still predominantly on the upper part of the range; set 3 was to have a rela-

Table 3. Characteristics of Four Initial Populations (9-12) Produced With the Two-Factor Generator From Original Population Two

Correlation range	Associative characteristics							
	Initial population 1		Initial population 2		Initial population 3		Initial population 4	
	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}
−1.0 to −.9								
−.9 to −.8								
−.8 to −.7		1						
−.7 to −.6		1						
−.6 to −.5		2						
−.5 to −.4		3		2				
−.4 to −.3		1		2				
−.3 to −.2		2		4				3
−.2 to −.1		2		2		5		5
−.1 to 0		5		7		11		9
0 to .1	2	2	2	2	3	17	3	7
.1 to .2		2	1	5	7	26	1	12
.2 to .3	1	3	2	13	2	7	2	11
.3 to .4		8	2	8			1	6
.4 to .5	2	5	3	16			2	4
.5 to .6		5	2	5			3	3
.6 to .7	2	9						4
.7 to .8	3	11						2
.8 to .9	2	4						
.9 to 1.0								

Statistical estimator	Population regression parameters (by OLS)			
	Initial population 1	Initial population 2	Initial population 3	Initial population 4
R^2	.7521	.4180	.1142	.3724
b_1	.1212	.1082	.0749	.0679
b_2	.1537	.1345	.0957	.1063
b_3	.0803	.0948	.0604	.0238
b_4	.0575	.0603	.0431	.0321
b_5	.1241	.1213	.0849	.0655
b_6	.2208	.1538	.1089	.2754
b_7	.0545	.0597	.0414	.0259
b_8	.0003	−.0023	.0006	.0059
b_9	.0242	.0317	.0176	.0017
b_{10}	.0570	.0687	.0415	.0132
b_{11}	.1601	.1276	.0926	.1484
b_{12}	.0054	.0090	.0036	−.0041

tively low R^2 and a small range of zero-order correlation coefficients; and set 4 was to have a medium to low R^2 with a wide range of zero-order correlation coefficients. Also, since we were trying to simulate socio-economic variables, we had the criterion for all sets that a major share of the correlations should be positive. These objectives are met reasonably well as shown by the frequency distribution of the data and

Table 4. Characteristics of Four Initial Populations (13-16) Produced With the Four-Factor Generator From Original Population Two

Correlation range	Associative characteristics							
	Initial population 1		Initial population 2		Initial population 3		Initial population 4	
	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}
−1.0 to −.9								
−.9 to −.8								
−.8 to −.7								
−.7 to −.6								
−.6 to −.5								
−.5 to −.4		4						
−.4 to −.3	1	4						
−.3 to −.2		2		7				
−.2 to −.1		6		6			1	8
−.1 to 0		3		6	1	19		11
0 to .1	1	5	3	7	7	30	3	15
.1 to .2	1	5	2	13	4	17	3	13
.2 to .3	1	10	4	11			2	7
.3 to .4		10	1	10			1	6
.4 to .5	4	3	2	10			2	5
.5 to .6	1	5						1
.6 to .7	1	9						
.7 to .8	2							
.8 to .9								
.9 to 1.0								

Statistical estimator	Population regression parameters (by OLS)			
	Initial population 1	Initial population 2	Initial population 3	Initial population 4
R^2	.7203	.3623	.0802	.3277
b_1	−.0106	.0131	.0035	.0913
b_2	.3159	.2277	.1345	.2866
b_3	.1937	.1871	.0910	.0610
b_4	−.0272	−.0040	.0125	−.0193
b_5	−.0791	−.0305	−.0014	−.0688
b_6	.3458	.2106	.1133	.3370
b_7	.0574	.0653	.0475	.0303
b_8	.0260	.0238	.0058	−.0060
b_9	−.0938	−.0804	−.0411	−.0395
b_{10}	.0526	.0600	.0388	.0202
b_{11}	.0537	.0653	.0614	.0694
b_{12}	.1394	.1224	.0586	.0595

the R^2's in Tables 1 through 6. The range in R^2's for set 1 is from 0.6626 to 0.8147; set 2 is from 0.3156 to 0.4682; set 3 is from 0.0512 to 0.1463; and set 4 is from 0.1123 to 0.6084.

Tables 1 through 6 also give the standardized OLS regression coefficients for each of the 24 initial populations. We assume these regression coefficients are the population parameters or expected values for each of the 24 initial populations. Tables 1 through 6 also give the values

Table 5. Characteristics of Four Initial Populations (17-20) Produced With the Two-Factor Generator From Original Population Three

Correlation range	Associative characteristics							
	Initial population 1		Initial population 2		Initial population 3		Initial population 4	
	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}
−1.0 to −.9								
−.9 to −.8								
−.8 to −.7		1						
−.7 to −.6		3						1
−.6 to −.5		1						1
−.5 to −.4		2		2				1
−.4 to −.3		2		4				3
−.3 to −.2		3		3		1		1
−.2 to −.1	1	2		2		5	1	5
−.1 to 0		2		4	1	9		3
0 to .1	1	2	1	6	3	20	1	6
.1 to .2		5	2	6	6	27	2	10
.2 to .3	1	3	1	10	2	4	1	15
.3 to .4		7	2	7			1	9
.4 to .5	2	5	5	20			3	6
.5 to .6		6	1	2			2	2
.6 to .7	4	5					1	2
.7 to .8	2	15						1
.8 to .9	1	2						
.9 to 1.0								

Statistical estimator	Population regression parameters (by OLS)			
	Initial population 1	Initial population 2	Initial population 3	Initial population 4
R^2	.8147	.4682	.1463	.6084
b_1	.1941	.1664	.1263	.2071
b_2	.1221	.1178	.0881	.1027
b_3	.1334	.1276	.0941	.1064
b_4	.0193	.0036	.0072	.1162
b_5	.1257	.1245	.0917	.0974
b_6	.1446	.1365	.1028	.1288
b_7	.0356	.0326	.0245	.0281
b_8	.1023	.0880	.0692	.1173
b_9	.2156	.1500	.1120	.3595
b_{10}	.0858	.0922	.0610	.0392
b_{11}	−.0340	−.0372	−.0246	.0049
b_{12}	.0589	.0700	.0448	.0199

of the determinant of the augmented correlation matrix as some indication of the degree of multicollinearity. The closer the determinant is to zero, the greater is the degree of multicollinearity. If the R^2 is high, then we would expect the determinant of the augmented correlation matrix to be near zero. But since the highest R^2 of any of the 24 initial populations is 0.7521, the small size of the determinants also reflects a high degree of multicollinearity among the explanatory variables.

Table 6. Characteristics of Four Initial Populations (21-24) Produced With the Four-Factor Generator From Original Population Three

Correlation range	Associative characteristics							
	Initial population 1		Initial population 2		Initial population 3		Initial population 4	
	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}	Freq. of r_{xy}	Freq. of r_{xx}
−1.0 to −.9								
−.9 to −.8								
−.8 to −.7								
−.7 to −.6		1						
−.6 to −.5								1
−.5 to −.4		1						
−.4 to −.3		4		1				
−.3 to −.2		5		5				3
−.2 to −.1		2		6		2		8
−.1 to 0	2	9	2	10	2	20	2	10
0 to .1	1	9	1	10	4	30	1	12
.1 to .2	1	2	2	8	6	15	2	13
.2 to .3	1	7	2	13			3	12
.3 to .4	1	8	2	9			2	5
.4 to .5	1	5	3	4			1	1
.5 to .6	2	7					1	1
.6 to .7	2	4						
.7 to .8	1	2						
.8 to .9								
.9 to 1.0								

Statistical estimator	Population regression parameters (by OLS)			
	Initial population 1	Initial population 2	Initial population 3	Initial population 4
R^2	.7861	.4179	.1052	.5187
b_1	.0468	.0457	.0328	.0284
b_2	.1484	.1339	.0874	.1179
b_3	.2672	.2169	.1406	.2828
b_4	.1476	.0845	.0590	.3179
b_5	−.0134	−.0040	−.0021	−.0328
b_6	.2434	.2025	.1270	.2214
b_7	.1078	.0993	.0611	.0701
b_8	−.0491	−.0343	−.0229	−.0842
b_9	.1992	.1269	.0808	.3373
b_{10}	.1687	.1635	.0923	.0782
b_{11}	−.0375	−.0280	−.0099	−.0049
b_{12}	.1529	.1471	.0831	.0829

RESULTS

Efficiency

Efficiency refers to the size of the variance of an estimator relative to the variance of another estimator or a standard estimator. The smaller the variance of an estimator, the more efficient the estimator is, and the estimator with greatest efficiency (often referred to as the efficient estimator) is the estimator with the smallest variance.

We know from statistical theory (Anderson, 1958) that the OLS estimator is inefficient and that the variance is unreliable for probability estimates when there are errors in the explanatory variables.

Thus one important characteristic of the CFAR estimator to investigate is the variance of this estimator. Since there are 43,200 equations each with 12 $\tilde{b}_j$ values and it is impossible to make or report all the possible comparisons one might like to make, the variances for each estimator (calculated from each of the samples of 100) were summed and averaged over the 12 $\tilde{b}_j$ for certain Monte Carlo variables such as sample size (N = 16, 64, 256) for associative characteristics or the internal population relationships, and for each of the different numbers of factors extracted from one through six factors. We have essentially summarized the 5,184 variances related to the Monte Carlo variables.

The data relating the mean variances to the sample size and number of factors extracted are given in Table 7. The mean variance, very small when the sample size is the largest (256), remains consistently small for all factors extracted. The mean variance is still quite small for the medium sample size (64), but tends to increase as the number of factors extracted is increased. The mean variance is small even for sample size 16. The fact that the mean variance gets smaller as the sample size increases is important because it indicates that CFAR is a consistent estimator; that is, the variance asymptotically approaches a minimum as sample size increases.

Table 7. Mean Variance of the CFAR Estimators by Sample Size and Number of Factors Extracted

Factors extracted	Sample size		
	N = 16	N = 64	N = 256
1	.006176	.001803	.000764
2	.010808	.002095	.000579
3	.018060	.003286	.000960
4	.027348	.004815	.001347
5	.039844	.007020	.002084
6	.056240	.009330	.002699

Table 8. Mean Variance of the CFAR Estimators by Sample Size and Associative Characteristics

Associative characteristics	Sample size		
	N = 16	N = 64	N = 256
R_1	.023146	.004126	.001322
R_2	.026211	.004315	.001317
R_3	.028443	.005483	.001486
R_4	.627851	.004976	.001497

The mean variance increases as the number of factors extracted increases for both extraction methods. As the number of factors extracted increases, the solution approaches the OLS solution and is the same as the OLS solution when the maximum possible number of factors are extracted. This result implies that the CFAR solutions are always more efficient than the OLS solutions.

Table 8 relates the mean variances to the four selected associative characteristics and to the sample size. The four sets of associative characteristics explained earlier are designated R_1 as the populations with high R^2, R_2 as the populations with medium R^2, R_3 as the populations with a low R^2, and R_4 with a wider ranging R^2. Intercorrelations among the population variables also differ. While the associative set with the highest R^2 and the highest intercorrelations among the explanatory variables has the smallest variances for the CFAR estimators, the average variances for the other sets are also small and well behaved. The variances for the CFAR estimators drop sharply in magnitude when we go from sample size 16 to sample size 64, and again to sample size 256. This is exactly the way we would like to have the CFAR estimator behave in order to recommend it as an extremely good estimator for errors-in-the-variables regression. The variances were smallest regardless of sample size for the population characteristics which had the highest R^2 and the highest intercorrelations among the explanatory variables, also a very desirable feature.

The mean variances are related to the number of factors extracted and the associative characteristics in Table 9. These data illustrate again the increase in variance as the number of factors extracted increases. There is little difference in the variances from one associative characteristic to another. Except when only one factor is extracted, the variances are smallest for the two populations having the highest R^2 and higher intercorrelations among the explanatory variables.

Table 9. Mean Variance of the CFAR Estimator by Number of Factors and Associative Characteristics

Number of factors	Associative characteristics[a]			
	R_1	R_2	R_3	R_4
1	.003193	.003121	.003026	.002316
2	.003620	.004533	.005439	.004384
3	.005927	.007256	.008763	.007795
4	.009136	.010879	.012790	.011876
5	.014480	.015728	.017504	.017552
6	.020831	.022168	.023303	.024724

[a] In this and following tables, R_1 = High r, R_2 = Medium r, R_3 = Low r, and R_4 = Wide range r.

Normality

Normality refers to how closely the distribution of the CFAR estimator approaches the normal distribution. The method we chose to analyze this question was to calculate for each $\hat{B}_j$ the higher moments of the distribution (skewness and kurtosis) since both skewness and kurtosis of the normal distribution are zero. The Kolmogorov-Smirnov statistic, an alternative statistic, was not used because the moments are more sensitive, particularly in the tails of the distribution. The moments were calculated and averaged, again relating the mean of the moments to the Monte Carlo variables.

Skewness

Summary data for skewness are given in Table 10 with respect to sample size and the number of factors extracted. All values obtained for skewness are small. Skewness approaches zero as sample size increases and as the number of factors extracted increases. The skewness in the largest sample size is consistently small regardless of the number of factors extracted.

Skewness related to sample size and the four sets of populations with different associative characteristics is given in Table 11. While the skewness does not seem to bear a consistent relationship among the various associative characteristics for each sample size, it is clear again that the skewness approaches zero as sample size increases — the largest improvement being made as the sample size increases from 16 to 64.

Skewness related to associative characteristics and the number of factors extracted is given in Table 12. The skewness declines consistently for all associative groups as the number of factors extracted is

Table 10. Mean Skewness of the CFAR Estimator by Sample Size and Number of Factors Extracted

Number of factors	Sample size		
	N = 16	N = 64	N = 256
1	1.273076	.209734	−.013147
2	.769982	.200702	.042032
3	.597069	.121915	.056393
4	.402840	.098575	.060511
5	.276784	.076460	.043235
6	.185260	.052464	.019319

Table 11. Mean Skewness of the CFAR Estimator by Sample Size and Associative Characteristics

Associative characteristics	Sample size		
	N = 16	N = 64	N = 256
R_1	.568742	.148125	.037540
R_2	.727828	.087259	.015515
R_3	.465930	.111601	−.006986
R_4	.574174	.159582	.092826

Table 12. Mean Skewness of the CFAR Estimator by Associative Characteristics and Number of Factors

Number of factors	Associative characteristics			
	R_1	R_2	R_3	R_4
1	.280749	.419263	.440154	.819383
2	.359442	.395388	.283772	.311686
3	.304982	.315598	.194992	.218266
4	.227333	.234565	.136653	.150683
5	.191042	.185158	.056596	.095842
6	.145269	.111230	.028922	.057304

increased, except for group R_1 in going from one factor to two factors extracted. In all cases the skewness is the least for six factors extracted.

The salient points shown by this Monte Carlo experiment regarding skewness are: skewness, while generally positive, is very small in all cases; skewness approaches zero as sample size increases; skewness approaches zero as the number of factors increases; and skewness does not seem to be consistently related to the associative characteristics.

Kurtosis

Kurtosis was calculated and related to the Monte Carlo variables in the same way as variance and skewness. If the distribution is not kurtotic relative to the normal distribution, the value for kurtosis is equal to or near zero. The mean kurtosis values related to sample size and number of factors extracted are given in Table 13. All kurtosis values are small, and the kurtosis approaches zero as sample size increases. The sharpest reduction in kurtosis was made in going from sample size 16 to 64. Except for an aberration at three factors for the two larger sample sizes, the kurtosis also approached zero as the number of factors extracted increases.

Kurtosis related to sample size and associative characteristics is given in Table 14. Here again the kurtosis values for all combinations are small and approach zero as the sample size increases, with the sharpest reduction in kurtosis occurring in moving from the size 16 sample to sample size 64. The populations with high R^2 and high intercorrelations among the explanatory variables seem to have the kurtosis values closest to zero at all sample sizes.

Kurtosis values related to associative characteristics and number of factors extracted are given in Table 15. Except for an aberration for group R_1 when two factors are extracted, the consistent pattern of kurtosis values indicates that they approach zero as the number of fac-

Table 13. Mean Kurtosis of the CFAR Estimator by Sample Size and Number of Factors

Number of factors	Sample size		
	N = 16	N = 64	N = 256
1	8.746820	.994613	.469777
2	5.510900	.504834	.222304
3	4.436723	.709071	.352426
4	3.701197	.425751	.070094
5	2.989154	.285380	.010045
6	2.260743	.097709	−.135707

Table 14. Mean Kurtosis of the CFAR Estimator by Sample Size and Associative Characteristics

Associative characteristics	Sample size		
	N = 16	N = 64	N = 256
R_1	3.842346	.472932	.087571
R_2	6.658900	.394044	.190024
R_3	4.925442	.469767	.109464
R_4	5.003670	.674829	.272235

tors extracted increases. Otherwise there seems to be no clear pattern of how the kurtosis values relate to the associative characteristics at corresponding factor numbers.

The findings on kurtosis can be stated as follows: all the kurtosis values are small; kurtosis values approach zero as sample size increases; and kurtosis values approach zero as the number of factors extracted increases.

Bias

The bias is the difference between the expected value or the average value over all samples of the estimator and the true values or parameter — bias = $E(\hat{B}) - B$. Because the large amount of data restricted the extent of investigation and presentation of each item, we averaged the bias over all parameters for comparison purposes. The result shown in each cell of Table 15 is the average bias calculated as follows for each subclassification of the Monte Carlo experiment:

$$(11) \qquad \sum_{j=1}^{12} \sum_{i=1}^{100} (\bar{B}_{ij} - B_j)/1200$$

where B_j is the hypothesized parameter of the j-th explanatory variable calculated by using OLS on the initial population without error. $\bar{B}_{ij}$ is the estimator for the j-th explanatory variable calculated from the i-th sample after the addition of random normal errors to all observations of all variables.

The results of calculation of the bias are given in Table 16. The bias of the OLS estimator after errors are added to the initial observations of the variables relative to the OLS values before errors are added is given in the first row of the table for comparison purposes. The results are given for both factor analysis and principal component extraction and by sample size and number of factors extracted. The OLS

Table 15. Mean Kurtosis of the CFAR Estimator by Associative Characteristics and Number of Factors

Number of factors	Associative characteristics			
	R_1	R_2	R_3	R_4
1	1.939687	3.001932	4.525533	4.147795
2	1.426016	2.013136	2.320110	2.558121
3	1.999107	1.810255	1.544697	1.976898
4	1.461554	1.619087	1.127421	1.387992
5	1.236110	1.221622	.836656	1.085051
6	.743223	.819904	.654925	.745608

bias is small for all three sample sizes, but does not appear to be consistent, going from positive to negative to positive nor does it appear to be asymptotic.

In all cases the bias for the CFAR estimators is consistent and negative, but asymptotically approaches zero as sample size increases. For the largest sample size, the bias of the CFAR estimators when four or more factors are extracted is smaller than the bias of the OLS estimator. Using principal components rather than statistical factor analysis gives comparable results. While the bias for factor analysis is equal to or smaller than the bias of principal components, the differences are sufficiently small that the extra cost of statistical factor extraction relative to the cost of principal components appears to be greater than the advantage gained.

The second part of Table 16 gives the bias by number of factors extracted and the internal population characteristics. Here the OLS estimators also seem to be inconsistent. The CFAR estimators are negatively biased and in general are smallest when there is a high intercorrelation among the variables in the population. There are several cases among these data where the principal components method results in better (smaller bias) estimates than the statistical factor extraction method. Differences in the magnitude of the bias are smaller than we had expected among the various populations. The high correlation populations have the smallest bias, but the magnitude of the bias of the CFAR estimators for the other populations is about the same.

While the most desirable outcome for the bias would be if the CFAR estimators were unbiased, the bias does appear to be well behaved; that is, the bias is negative, consistent, and asymptotically approaches zero as sample size increases. Also, the bias is smallest for the kind of population characteristics that we are most concerned about and for which this estimator was initially developed.

Table 16. Average Bias of $\hat{B}$ From the True Value of B

No. of factors extracted	N = 16		N = 64		N = 256		Overall average	
	FA	PC	FA	PC	FA	PC	FA	PC
1	−.0332	−.0317	−.0204	−.0229	−.0150	−.0186	−.0229	−.0244
2	−.0168	−.0178	−.0073	−.0102	−.0041	−.0078	−.0094	−.0119
3	−.0109	−.0113	−.0035	−.0041	−.0020	−.0017	−.0054	−.0057
4	−.0077	−.0084	−.0017	−.0023	−.0004	−.0004	−.0033	−.0037
5	−.0053	−.0070	−.0012	−.0020	−.0003	−.0004	−.0023	−.0031
6	−.0039	−.0056	−.0009	−.0016	−.0003	−.0005	−.0017	−.0028
OLS	.0012		−.0002		.0009		.0006	

Internal Characteristics

No. of factors extracted	High r		Medium r		Low r		Wide range r	
	FA	PC	FA	PC	FA	PC	FA	PC
1	−.0221	−.0229	−.0209	−.0238	−.0181	−.0020	−.0304	−.0288
2	−.0087	−.0096	−.0102	−.0113	−.0099	−.0133	−.0087	−.0134
3	−.0040	−.0030	−.0060	−.0052	−.0065	−.0088	−.0054	−.0057
4	−.0015	−.0007	−.0041	−.0034	−.0044	−.0072	−.0030	−.0034
5	−.0007	−.0005	−.0032	−.0031	−.0030	−.0058	−.0021	−.0031
6	−.0009	−.0002	−.0025	−.0026	−.0018	−.0049	−.0016	−.0026
OLS	.0002		−.0010		.0012		−.0008	

Note: FA means a statistical factor extraction; PC means extraction by principal components.

Table 17. Average Mean Square Error of $\hat{B}$ From the True Value of B

No. of factors extracted	N = 16		N = 64		N = 256		Overall average	
	FA	PC	FA	PC	FA	PC	FA	PC
1	.0120	.0100	.0070	.0085	.0059	.0080	.0083	.0089
2	.0133	.0085	.0039	.0055	.0024	.0048	.0065	.0063
3	.0194	.0099	.0040	.0044	.0015	.0030	.0083	.0058
4	.0282	.0130	.0052	.0046	.0017	.0025	.0117	.0067
5	.0406	.0176	.0074	.0057	.0024	.0027	.0168	.0087
6	.0570	.0244	.0096	.0068	.0029	.0030	.0232	.0114
OLS	.5023		.0212		.0044		.1760	

Internal Characteristics

No. of factors extracted	High r		Medium r		Low r		Wide range r	
	FA	PC	FA	PC	FA	PC	FA	PC
1	.0105	.0121	.0073	.0075	.0045	.0036	.0110	.0122
2	.0064	.0069	.0059	.0046	.0060	.0038	.0077	.0099
3	.0067	.0050	.0077	.0042	.0091	.0050	.0096	.0089
4	.0095	.0049	.0111	.0053	.0130	.0068	.0132	.0099
5	.0148	.0069	.0160	.0072	.0177	.0088	.0186	.0117
6	.0212	.0096	.0224	.0099	.0235	.0116	.0256	.0145
OLS	.1580		.1703		.1677		.2079	

Note: FA means a statistical factor extraction; PC means extraction by principal components.

Loss Function of the Estimators

Another important criterion frequently considered is the loss function of the estimators: the sum of squares of the differences between the estimator and the parameter over the number of samples. These data are given in Table 17. Mathematically the values given in this table are:

$$(12) \qquad \sum_{j=1}^{12} \sum_{i=1}^{100} (\tilde{B}_{ij} - B_j)^2/1200$$

with respect to the particular cell designations of the table. The notation is the same as for equation 11. As we look at the loss function of the estimators, we find first that it is smaller in almost all cases than the corresponding loss function for the OLS estimators. In some cases the difference is very great, being 37 times greater for the OLS estimators in one case. Moreover, the loss function for the CFAR estimators becomes smaller as sample size increases for each level of factor extraction, and the loss function increases as the number of factors extracted increases. If we realize that the CFAR estimator is the same as the OLS estimator in the limit as the number of factors extracted increases to be equal to the number of variables, we can then see the logic in the increase in the loss function as the number of factors increases. Thus the loss function for CFAR estimators should always be less than for OLS, and would have as its upper bound the value of the OLS loss function. On examining the differences between statistical factor extraction and principal component extraction, we find that at the small sample size, principal component extraction results in small loss functions for all levels of factors extracted. These results are less pronounced at the medium sample size, and statistical extraction seems better than principal components at the large sample size.

With respect to the populations with different internal characteristics, the OLS loss function again is uniformly larger than for the CFAR estimators. In most cases the principal component estimators give better results than the statistical factor procedure. Except for results from extracting only one factor, the loss function for the CFAR estimators monotonically increases with increasing number of factors for both the principal component procedure and the factor analysis procedure. Furthermore, the upper bound should again be the OLS result when all factors are extracted.

Loss Function of the Predicted Values

If we are concerned about prediction as well as the structural parameters, or if we are mainly concerned about prediction, as is frequently

Table 18. Average Mean Square Error of the Predicted Value From the True Value

No. of factors extracted	N = 16		N = 64		N = 256		Overall average	
	FA	PC	FA	PC	FA	PC	FA	PC
1	.9038	.8921	.7864	.8191	.7479	.7971	.8127	.8361
2	.8435	.8040	.6854	.7085	.6489	.6846	.7259	.7324
3	.8744	.7987	.6671	.6715	.6280	.6393	.7322	.7031
4	.9388	.8225	.6738	.6676	.6273	.6292	.7466	.7064
5	1.0322	.8609	.6874	.6750	.6315	.6307	.7837	.7222
6	1.1552	.9174	.7015	.6831	.6348	.6324	.8305	.7743
OLS	4.4568		.7689		.6424		1.9560	

Internal Characteristics

No. of factors extracted	High r		Medium r		Low r		Wide range r	
	FA	PC	FA	PC	FA	PC	FA	PC
1	.5734	.6336	.8039	.8317	1.0031	.9959	.8704	.8831
2	.3816	.4098	.7273	.7271	1.0015	.9959	.7796	.8023
3	.3347	.3376	.7271	.7016	1.0494	.9903	.7815	.7716
4	.3353	.3184	.7496	.7070	1.0933	1.0019	.8082	.7780
5	.3538	.3251	.7867	.7218	1.1460	1.0223	.8483	.7958
6	.3751	.3354	.8360	.7428	1.2140	1.0462	.8970	.8202
OLS	.8660		1.9950		2.8522		2.1110	

Note: FA means a statistical factor extraction; PC means extraction by principal components.

the case, then an important criterion is how the loss function of the predicted values compares between CFAR equations and OLS equations. The results shown in Table 18 were calculated as follows:

$$(13) \qquad \sum_{j=1}^{100} \sum_{i=1}^{12} (Y_{ij} - \hat{Y}_{ij})^2/1200$$

where Y_{ij} is the i, j-th population true value. i = 1, 2, . . ., n, where n is the sample size. j = 1, 2, . . ., 100, with 100 repeated samples.

$\hat{Y}_{ij}$ is the predicted value.

Predictions were made by the various procedures — factor analysis with six different factors, principal components with six different components, and OLS.

The values actually sampled and used to calculate the estimators and the predictions were Y_{ij}, which are $Y_{ij} + \Sigma_{ij}$, where Σ_{ij} is the random normal error added to the original population values. Thus the loss function is not the mean square error in the sample, which would be

$$(14) \qquad \sum_{j=1}^{100} \sum_{i=1}^{12} (Y'_{ij} - \hat{Y}_{ij})^2/1200,$$

which is the square of the difference between the predicted and the observed, where Y′ is the observed, but rather the loss function takes into account the difference between the predicted and the true value.

The OLS loss is largest at the small sample size. The CFAR loss is only one-fifth to one-fourth as large as the OLS loss at the small sample size, showing a considerable improvement over the OLS predictions. On the small sample size, the loss for the principal component (PC) procedure was less than the loss for the factor-analysis (FA) procedure for all six factors. For the medium- and large-size samples, the results for the PC and the FA procedures were very similar, with first one and then the other being better. For the medium-size sample CFAR was better than OLS for all factors extracted except the first and second.

The results are even more clearly in favor of CFAR when comparisons are based on differences among internal population characteristics. Here both the PC and the FA methods result in substantially smaller loss than OLS for all classifications of internal population characteristics, and for all the numbers of factors or principal components extracted. The best predictive ability for both OLS and CFAR occurred when there was high intercorrelation among the variables. The predictive ability measured by the loss function of the predicted values was from about 100 percent to as much as 300 percent better for CFAR than OLS. These results are shown in the first part of Table 18.

SUMMARY

A Monte Carlo experiment was developed to study the statistical characteristics for the beta estimators from classical factor analysis regression (CFAR), which has been proposed especially for estimating regressions when there are errors in the variables and when high multicollinearity makes ordinary least squares inappropriate or completely infeasible.

This experiment took 100 random samples of each sample size of 16, 64, and 256 from each of 24 initial populations having four different sets of associative or internal characteristics and 12 explanatory and one dependent variable. Thus there were 7,200 samples drawn. One through six factors were extracted from each and CFAR estimated from these factors, making 43,200 CFAR equations with 12 explanatory variables each. The statistical properties of the CFAR estimators were then analyzed.

It was found that the CFAR estimators behave extremely well. The

variance is small even at small sample sizes and quickly approaches a consistent minimum level as sample size is increased. For purposes of using this estimator where there is high multicollinearity, it is noteworthy that CFAR estimates for the initial populations which had the highest R^2 and the highest intercorrelations among the explanatory variables consistently had the smallest variances. The good small-sample results are important, especially for those working with time-series data for which the number of observations is often limited. The variances increase as the number of factors extracted increases, so that if this procedure were carried to the limit (where the number of factors extracted equaled the number of variables — the OLS regression case), the variances corresponding to OLS estimation would balloon up to the OLS values. Thus CFAR is clearly a very efficient estimator relative to OLS for regressions when there are errors in the variables.

The experiment also shows that the CFAR estimators are asymptotically normal both as sample size increases and as the number of factors increases. This is deduced from the behavior of the third and fourth moments (skewness and kurtosis), which are both zero in the normal distribution. Even for small samples a normal or "t" distribution could be used for probability statements about the CFAR estimators.

The CFAR estimators are consistently negatively biased, but appear to approach zero monotonically as sample size increases. Two measures often made of estimators are comparisons of the mean-square error or loss functions for the estimators themselves and of the loss function of the prediction. Here, too, CFAR shows substantial advantage over OLS, being from 100 percent in many cases to as much as 2,500 percent better than OLS.

Thus the CFAR estimator is substantially better in several respects than OLS for all applications where there is high multicollinearity or when there are errors in the variables, regardless of sample size, and CFAR is especially useful for small samples. We hypothesize that CFAR is also better when the data are plagued with outliers.

BIBLIOGRAPHY

Adelman, I., and C. T. Morris, "A Quantitative Study of Determinants of Fertility," Economic Development and Cultural Change, 14(1966), 129-157.

Amemiya, T., "On the Use of Principal Components of Independent Variables in Two-Stage Least Squares Estimation," International Economic Review, 7(1966), 283-303.

Anderson, T. W., *Introduction to Multivariate Statistical Analysis,* New York: John Wiley and Sons (1958).

———, "The Use of Factor Analysis in the Statistical Analysis of Multiple Time Series," Psychometrika, 28(1963), 1-25.

———, and H. Rubin, "Statistical Inference in Factor Analysis," Proc. Third Berkeley Symposium on Mathematics, Statistics, and Probability, 5(1956), 111-150.

Box, G. E. P., and M. E. Mueller, "A Note on the Generation of Random Normal Deviates," Annals of Mathematical Statistics, 29(1958), 210-211.

Bursch, W. G., J. T. Scott, Jr., and R. N. Van Arsdall, "Characteristics and Prospects of the Commercial Hog Feed Market in Illinois," Illinois Agricultural Experiment Station Bulletin 743 (1973).

Doll, J. P., and S. B. Chin, "A Use for Principal Components in Price Analysis," American Journal of Agricultural Economics, 52(1970), 591-593.

Flanagan, J. C., *Factor Analysis in the Study of Personality,* Stanford University Press (1935).

Guilford, J. P., "The Structure of Intellect," Psychological Bulletin, 53(1956), 267-293.

Guttman, L., "Multiple Rectangular Prediction and the Resolution into Components," Psychometrika, 5(1940), 75-79.

———, "Best Possible Systematic Estimates of Communalities," Psychometrika, 21(1950), 273-285.

Haitovsky, Y., "Multicollinearity in Regression Analysis, an Experimental Evaluation of Alternative Procedures," paper given at American Statistics Association annual meeting, New York (1969).

Hotelling, H., "Analysis of a Complex Set of Statistical Variables into Principal Components," Journal of Educational Psychology, 24(1933), 417-441.

Holzinger, K. J., and H. H. Harmon, *Factor Analysis: a Synthesis of Factorial Methods,* Chicago: University of Chicago Press (1941).

Johnston, J., *Econometric Methods,* New York: McGraw-Hill (1963).

Joreskog, K. G., "On the Statistical Treatment of Residuals in Factor Analysis," Psychometrika, 27(1962), 335-354.

Kelley, T., "Essential Traits of Mental Life," Harvard Studies in Education (1935).

Kloek, T., and L. B. M. Mennes, "Simultaneous Equations Estimation Based on Principal Components of Predetermined Variables," Econometrica, 28(1960), 45-61.

Lawley, D. N., "The Estimation of Factor Loadings by the Method of Maximum-Likelihood," Royal Society of Edinburgh Proc., A-60 (1940), 64-82.

———, and A. E. Maxwell, *Factor Analysis as a Statistical Method,* London: Butterworth (1963).

———, "Regression and Factor Analysis," Biometrika, 60(1973), 331-332.

Mangan, F. K., "A Monte Carlo Study of Linear Regression and Factor Analysis Under Multicollinearity," unpublished master's thesis, University of Washington (1970).

Marquardt, D. W., "Generalized Inverses, Ridge Regression, Biased Linear Estimating and Nonlinear Estimation," Technometrica, 12 (1970), 591-612.

McDonald, G. C., and D. I. Galarneau, "A Monte Carlo Evaluation of Some Ridge Type Estimators," Journal of American Statistical Association, 70(1975), 407-416.

Oehrtman, R. L., "A Factor Analysis of the Adjustment Problems Facing Milk Bottling Firms," paper presented at Econometric Society meeting, Chicago (1968).

Olsen, B. M., and G. Garb, "An Application of Factor Analysis to Regional Economic Growth," Journal of Regional Science, 6(1965), 51-56.

Payne, W. H., and T. G. Lewis, "Continuous Distribution Sampling: Accuracy and Speed," in *Mathematical Software*, J. R. Rice, editor, New York: Academic Press (1971), 331-345.

Pearson, K., "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, 6(1901), 559-572.

RANDU in SSP Manual System/360 Scientific Subroutine Package Version III Programmer's Manual, Program No. 360A-CM-03X, IBM Corp. (1970).

Rao, C. R., "Estimation and Tests of Significance in Factor Analysis," Psychometrika, 20(1955), 93-111.

Scott, J. T., Jr., "Factor Analysis Regression," *Econometrica*, 34(1966), 552-562.

———, "The Synthesis of Classical Regression and Factor Analysis," unpublished manuscript, Department of Agricultural Economics, University of Illinois at Urbana-Champaign (1972).

Sheth, J. N., and J. S. Armstrong, "Factor Analysis of Marketing Data: a Critical Evaluation," paper given at American Marketing Association fall conference, Department of Business Administration, University of Illinois at Urbana-Champaign (1969).

Sobolewski, J. S., and W. H. Payne, "Pseudonoise with Arbitrary Amplitude Distribution — Part I: Theory," IEEE Transactions on Computers, C-21:4(1972), 337-345.

Spearman, C., "General Intelligence, Objectively Determined and Measured," American Journal of Psychology, 15(1904), 201-292.

Thurstone, L. L., "Multiple Factor Analysis," Psychological Review, 38 (1931), 406-427.

———, "A New Rotational Method in Factor Analysis," Psychology, 3 (1938), 199-218.

———, *Multiple Factor Analysis*, Psychology Review, 38(1931), 406-427.

Tinter, G., *Econometrics*, John Wiley and Sons, New York (1952).

Tucker, L. R., R. F. Koopman, and R. L. Linn, "Evaluation of Factor Analytic Research Procedures by Means of Simulated Correlation Matrices," Psychometrika, GGNMP in IMSL manual; Library I: *International Mathematical and Statistical Library, Inc.*, 34(1969), 421-459.

Waugh, F. V., "Factor Analysis: Some Basic Principles and an Application," Agricultural Economics Research, 14(1962), 77-80.

Whittle, S. S., "Certain Generalizations in the Analysis of Variance," Biometrika, 24(1932), 471-494.

Wijsman, Robert A., "Application of a Certain Representation of the Wishart Matrix, Annals of Mathematical Statistics, 30(1959), 597-601.

APPENDIX: GENERATION OF THE ORIGINAL AND SAMPLING POPULATIONS

As this was a Monte Carlo investigation, the relevance of the results totally depended on the appropriateness of the populations used; therefore, this study varied the characteristics of the population used for the study in the same ways that characteristics have been observed to vary in many applied studies published in the literature and in other known empirical work.

The population intercorrelation matrix [P] dictates in many ways the results of any Monte Carlo experiment. The choice of an empirically derived matrix from the literature was rejected, because any single study might have had the data generated in an unique way, with an atypical error structure. The choice was therefore between an artificial but known and controllable population generation technique, and an unknown and controllable but natural data-generation technique. The simulation of P was selected. A procedure was needed which allowed errors in both the dependent and independent variables and various degrees of intercorrelation (multicollinearity) among all variables, and which paralleled the procedure that an investigator might use in selecting variables.

The methodology of Tucker, Koopman, and Linn (1969) was selected. Their procedure is based on the latent causal (or factor analytic) model. For each set of population matrices (there were two replications to be described later), it was hypothesized that the investigator had attempted to select variables to measure (to varying degrees) the underlying variables. It was assumed that the criterion was measuring all the latent variables. The degree to which variables were measuring the underlying factors was simulated by random selection of integers to sum to four. This matrix was then row-normed. By this method it was possible to get overrepresentation of loadings on certain latent variables. As indicated by the trace of the cross products, overrepresentation did not occur to a prohibitive degree. This matrix, the degree to which the variables are assumed to load on the latent variable [$\tilde{A}$], is given in Table A1. This matrix has been row-normed to be of unit length.

However, it was assumed that in practice there is a discrepancy between the *a priori* correlation of the variables with the causal variable and the actual correlation. The true loadings on the underlying constructs were generated by

$$A = \tilde{A}C_m + DX(1.0 - C^2{}_m)^{0.5} \quad (1)$$

where A = a 13 × 4 matrix of actual loadings on the underlying constructs,

Table A1. Conceptual Row-Normed Loadings on the Latent Variables (Ã)

Observed variables	Latent variables			
	I	II	III	IV
X_1	.94868	.31623	.0	.0
X_2	.0	.31623	.94868	.0
X_3	.0	.94868	.0	.31623
X_4	.0	.0	.31623	.94868
X_5	1.00000	.0	.0	.0
X_6	.0	.94868	.0	.31623
X_7	.0	.33333	.66667	.66667
X_8	.0	.0	.31623	.94868
X_9	1.00000	.0	.0	.0
X_{10}	.0	.31623	.94868	.0
X_{11}	.0	.0	1.00000	.0
X_{12}	.0	.94868	.31623	.0
Y	.50000	.50000	.50000	.50000
[Tr(A'A)]/13	.24231	.25855	.29188	.20726

C = a 4 × 4 diagonal matrix with constants for each factor representing the degree of error in specifying how well a a variable loads on a factor. Following Tucker, Koopman, and Linn, C was generated by random uniform deviates in the range of 0.70 to 0.90. The diagonal elements of C for population replication 1 were 0.83140, 0.71735, 0.74041, and 0.82244.

X = a 13 × 4 matrix of standardized normal deviates, and

D = a 13 × 13 diagonal matrix that was used to row-normalize X to unit length, where $d_{11} = (\Sigma_j X^2_{1j})^{-0.5}$.

Table A2. True Loadings for the Two-Factor Population 1

Var.	Factor 1	Factor 2
X_1	.99946	.03277
X_2	.49329	.86987
X_3	.82605	.56360
X_4	.43780	.89907
X_5	.81502	−.57943
X_6	.98863	.15035
X_7	.73175	.68157
X_8	−.19622	.98056
X_9	.96754	−.25273
X_{10}	.66333	.74833
X_{11}	.43708	.89942
X_{12}	.87765	−.47931
Y	.15204	.98837

Table A3. True Loadings for the Four-Factor Population 1

Var.	Factor 1	Factor 2	Factor 3	Factor 4
X_1	.74088	.25534	.45529	−.44262
X_2	.44165	.24887	.63778	.57987
X_3	−.00629	.89658	.37995	.22748
X_4	.15225	.39174	.24289	.87428
X_5	.69137	−.14441	−.65570	.26684
X_6	.21740	.95499	−.02229	.20058
X_7	.23414	.73280	.41418	.48646
X_8	−.16205	−.02912	−.03059	.98588
X_9	.71000	−.41126	.36330	−.44134
X_{10}	.24554	.65089	.55396	.45735
X_{11}	.63180	−.11712	.64407	.41507
X_{12}	.03940	.80663	.11615	−.57819
Y	.30116	−.12834	.19992	.92350

Number of Underlying Factors

Within each population it was desirable to vary the number of underlying latent variables; two and four variables were selected for this study. In the two-variable case, a linear sum of the first two and last two columns of matrix A was performed. This procedure insured that the two- and four-factor cases would have similar effects from the stochastic nature of X and C given in Equation 1. The loading matrix for the two- and four-factor solution was row-normalized to unity and appears for population 1 in Tables A2 and A3.

Levels of Communality

$$R = FF' + U^2 \tag{2}$$

where F = a matrix of factor loadings, and

U^2 = a diagonal matrix containing the proportion of uncorrelated error variance of each variable (uniqueness).

Furthermore,

$$H^2 = I - U^2 \tag{3}$$

where H^2 = a diagonal matrix containing the proportion of variance each variable shares with one another (i.e., communality).

It is desirable to vary the degree of communality (or conversely, the degree of uniqueness) into four levels: high communality, $H^2 \sim U(0.70 - 0.90)$; medium communality, $H^2 \sim U(0.40 - 0.60)$; low communality, $H^2 \sim U(0.10 - 0.30)$; and wide communality, $H^2 \sim U(0.10 - 0.90)$. In order to insure comparability across communality levels, the four levels of communality were selected to be linear combinations of one another:

$$h^2 \text{ medium, } j = (h^2 \text{ high, } j) - 0.3 \text{ (observed attribute)} \tag{4}$$

$$h^2 \text{ low, } j = (h^2 \text{ high, } j) - 0.3 \text{ (observed attribute)} \tag{5}$$

$$h^2 \text{ wide, } j = 4(h^2 \text{ low, } j) - 0.3 \text{ (observed attribute)} \quad j = 1, 2, \ldots, 13 \tag{6}$$

The diagonal entries of h^2 high for population 1 were 0.84905, 0.78175, 0.88172, 0.89331, 0.80636, 0.84063, 0.70259, 0.80925, 0.77029, 0.78500, 0.82318, 0.75626, and 0.70785.

The final population intercorrelation matrix is given by

$$P = FF' + U^2 \tag{7}$$

where $F = HA$

U^2 = is given by Equation 3, and

H = a diagonal matrix containing elements given by Equation 4, 5, or 6 above.

Therefore, for each population there were two levels of factors and four levels of communalities. Thus there were eight population matrices per population replication.

The regression weights and population multiple correlation for all population 1 matrices are presented in Table A4.

Table A4. Regression Weights and R^2 for Population 1

Regression weights	Two factors				Four factors			
	High communality	Medium communality	Low communality	Wide communality	High communality	Medium communality	Low communality	Wide communality
b_1	−.0283	−.0111	−.0005	−.0277	−.0570	−.0442	−.0190	−.0158
b_2	.1071	.1068	.0593	.0284	.1346	.1230	.0608	.0381
b_3	.1129	.0850	.0526	.0549	−.0732	−.0275	.0018	−.0681
b_4	.2457	.1575	.0904	.1959	.3206	.1945	.0999	.2408
b_5	−.1121	−.0930	−.0467	−.0450	.1251	.0966	.0408	.0403
b_6	−.0035	.0077	.0097	−.0141	−.0533	−.0265	−.0046	−.0336
b_7	.0528	.0627	.0312	.0058	.0337	.0422	.0216	.0033
b_8	.1609	.1416	.0764	.0578	.2080	.1766	.0846	.0641
b_9	−.0537	−.0443	−.0193	−.0197	.0184	.0062	−.0003	.0112
b_{10}	.0881	.0896	.0510	.0222	.0554	.0582	.0338	.0134
b_{11}	.1424	.1262	.0718	.0503	.1953	.1516	.0732	.0775
b_{12}	−.0755	−.0688	−.0321	−.0225	−.1626	−.1444	−.0598	−.0397
R^2	.6806	.3474	.0647	.1195	.6626	.3156	.0513	.1123
R	.8250	.5894	.2544	.3457	.8140	.5618	.2265	.3350
Det.	$.3687 \times 10^{-6}$	.006113	.3626	.002494	$.1322 \times 10^{-4}$	.02633	.5247	.01697

Population Replication

The matrices H, C, and X which generated characteristics of P are stochastic in nature. To determine the importance of such random perturbation on the model (i.e., to determine the effect of experimenter's choice of variables), the above procedure was replicated two times, allowing A to remain the same but generating new random number matrices H, C, and X. These computations were done on an IBM 370/155 at the University of Illinois Medical Center in Chicago using double precision FORTRAN words. The random number generator used was by Lewis and Payne (1973). Test results from this generator are reported by Richardson.

The regression weights, b_i, and population multiple correlation for all population 2 matrices are given in Table A5.

Table A5. Regression Weights and R^2 for Population 2

Regression weights	Two factors				Four factors			
	High communality	Medium communality	Low communality	Wide communality	High communality	Medium communality	Low communality	Wide communality
b_1	.1212	.1082	.0749	.0679	−.0106	.0131	.0035	−.0913
b_2	.1537	.1345	.0957	.1063	.3159	.2277	.1345	.2866
b_3	.0803	.0948	.0604	.0238	.1937	.1871	.0910	.0610
b_4	.0575	.0603	.0431	.0321	−.0272	−.0040	.0125	−.0193
b_5	.1241	.1213	.0849	.0655	−.0791	−.0305	−.0014	−.0688
b_6	.2208	.1538	.1089	.2754	−.3458	.2106	.1133	.3370
b_7	.0545	.0597	.0414	.0259	.0574	.0653	.0475	.0303
b_8	.0003	−.0023	.0006	.0059	.0260	.0238	.0058	−.0060
b_9	.0242	.0317	.0176	.0017	−.0938	−.0804	−.0411	−.0395
b_{10}	.0570	.0687	.0415	.0132	.0526	.0600	.0388	.0202
b_{11}	.1601	.1276	.0926	.1484	.0537	.0653	.0614	.0695
b_{12}	.0054	.0090	.0036	.0041	.1394	.1224	.0586	.0595
R^2	.7521	.4180	.1142	.3724	.7203	.3623	.0802	.3277
R	.8672	.6465	.3379	.6102	.8487	.6019	.2832	.5724
Det.	$.7812 \times 10^{-6}$	.008115	.4070	.006693	$.3055 \times 10^{-4}$	.03676	.5864	.04599

Sample Size (N)

Most standard errors depend on the number of observations (N). It was felt that three levels should adequately span this variable and allow for possible quadratic effects. N was set at 16, 64, and 256 observations. The lower level was selected because it yields very few degrees of freedom while allowing the matrix to be nonsingular. Pilot work suggested there would be a lack of discrimination among techniques in terms of MSE if the upper limit were raised. As the variance is usually related to N^2, the intermediate level was selected on the basis of an intermediate number in a quadratic progression. To keep costs within reason, only one intermediate level was chosen.

Sampling Replications

The next step in the simulation was to generate sample intercorrelation matrices from each of the sixteen population matrices. Within each sample size and for each population, one hundred sample replications were done. A replication (given the underlying population) was done by use of Wishart (1928) matrices. Each population matrix was Choleski-decomposed into

$$P = TT' \tag{8}$$

where T = an upper triangular matrix.

Samples from P (Wijsman, 1959), were generated by forming a sum of squares matrix:

$$S = TWT' \tag{9}$$

where W = a Wishart matrix. W itself can be decomposed into

$$W = GG' \tag{10}$$

where $G = [g_{ij}]$ $g_{ij} \sim N(0,1)$ if $i > j$
G $g_{ij} \sim X_{N-i-1}$ if $i = j$
$g_{ij} = 0$ if $i < j$

Therefore,

$$S = (TG)(TG)' \tag{11}$$

For $i > j$, g_{ij} was generated by the polar variant of Box and Mueller's (1958) technique (program GGNMP of IMSL). For $i = j$, g_{ii} was generated by a variant of the Lewis-Payne generator (Payne and Lewis, 1971; and Sobolewski and Payne, 1972), given a random uniform deviate (RANDU of SSP).

A distribution of 10,000 samples of the above-normal and chi square variates was generated and was found to be distributed according to theoretical expectations. The sample unbiased covariances were obtained by

$$C = S/(N - 1) \tag{12}$$

Thus, $E(C) = P$. Each of the population variances was set arbitrarily to 1.0. A sample of means was generated by

$$\bar{x} = (N^{-0.5})Td \tag{13}$$

where d = a vector of normal deviates generated by GGNMP, as above.

Therefore $\bar{x}$ was the vector of sample means sampled from the population matrix (P) with sample size N (N = 16, 64, and 256) and population mean zero.

For each level of N there were 100 sample replications. Therefore there were 4,800 mean and covariance matrices of 12 predictors and criterion (two population replications by two levels of number of factors by four levels of communality by three levels of number of observations (N) by 100 sample replications).

Urbana, Illinois August, 1978

Publications in the Bulletin series report the results of investigations made or sponsored by the Experiment Station. The Illinois Agricultural Experiment Station provides equal opportunities in programs and employment.

3M—8-78—41701