COMBINING REGRESSION AND FACTOR ANALYSIS FOR USE IN AGRICULTURAL ECONOMICS RESEARCH

John T. Scott, Jr.

While ordinary least squares regression has become a standard statistical technique, there are problems frequently overlooked or ignored by researchers in applying this statistical method. Two basic assumptions of the OLS regression model—(1) that the explanatory variables are independent of each other and (2) that the explanatory variables are known, fixed numbers—do not hold for most economic data, particularly time series data. This has been a consternation for econometricians, if not for the general researcher, for many years.

In the case of nonindependence of explanatory variables (multicollinearity), signs of the regression coefficients often are inconsistent with economic theory and with correlation coefficients calculated from the data. Also, variances of the estimated regression coefficients are inconsistent [6]. In practice for prediction equations, multicollinearity can usually be sufficiently reduced by either dropping one or more multicollinear variables or by indexing them and using the index as a regressor, thus circumventing the assumption regarding independence of the explanatory variables. A chi-square test for multicollinearity is available [6], and can be used as a guide to alert a researcher to the problem. Still, if a set of variables has been selected to represent a certain relationship on the basis of observation and theory, then dropping variables to reduce multicollinearity problems is a questionable practice. Yet, for economists, the multicollinearity problem persists in OLS regression analysis because there has been no good alternative method for estimating regression.

The assumption that explanatory variables are known, fixed numbers is also far from reality. In economic data, explanatory variables are just as likely to be generated by a stochastic process as the dependent variable. This problem is usually referred to as errors-in-the-variables. In the case of errors-in-the-variables, OLS regression coefficients are biased and variances are not only inconsistent, but are likely to be seriously underestimated [8]. While the two problems, multicolinearity and errors-in-the-variables, are usually treated separately both theoretically and in practice, there may in fact be confounding of the two. If the errors among the explanatory variables are not independent of each other, then they may exacerbate or even be the major cause of multicollinearity.

Modifications of the OLS regression model have been suggested in attempts to correct for errors-in-the-variables. Most of these modifications rely on *a priori* knowledge of error structure of the variables—both variances and covariances of the errors in the explanatory variables [8]. The problem with practical application of these OLS modifications is that rarely, if ever, does knowledge of the error structure exist on an *a priori* basis.

If the matrix depicting the error structure of the explanatory variables is or can be assumed to be diagonal, i.e., errors among the explanatory variables are not intercorrelated, then the ridge regression procedure might be used [17]. Also, if multicollinearity existed in this case, it would not have been the result of errors-in-the-variables. It has been

John T. Scott, Jr. is Professor of Farm Management and Production Economics, Department of Agricultural Economics, University of Illinois at Urbana, Illinois.

*This paper is a revision of a paper given at the Southern Agricultural Economics Association Annual Meeting in Mobile, Alabama February 1-4, 1976. The research reported in this paper was supported in part by the Illinois Agricultural Experiment Station.

shown in monte carlo experiments that in this situation ridge regression gives estimates with smaller mean square error than OLS estimates [18].

A number of authors in recent years have used factor analysis in their economic research [1, 2, 3, 4, 5, 7, 11, 12, 19, 25]. Factor analysis has been used in pattern analysis when there is a large variable set, to identify underlying causes in the data, to develop indexes for other use and to develop regressors to enter as regression variables in OLS analysis. These uses are fairly standard or traditional ways in which factor analysis is employed.

More recently, regression equations have been developed directly from the factor analysis model rather than using a factor index as a regressor, because assumptions of the model fit existing data better than the classic OLS regression model. Factor analysis regression has been suggested as a replacement of the OLS regression model when there are errors-in-the-variables or high multicollinearity exists among explanatory variables [7, 15, 21]. Recent monte carlo results also indicate that factor analysis regression performs much better than OLS under small sample conditions when there are errors-in-the-variables, and especially if there is multicollinearity [24].

THE FACTOR ANALYSIS MODEL

The factor analysis statistical model which has been long used in psychology can be defined as follows:

$$\mathcal{Z} = AF+U \qquad (1)$$

where

$\mathcal{Z}$ = hXn matrix of n observations of all h real variables involved
A = hXm matrix of regression coefficients, usually referred to as factor coefficients or factor loadings, m<h
F = mXn matrix of n values of the m factors
U = hXn matrix of the n residuals associated with the h variables.

Both F and U are assumed to be multivariate normal with zero means and uncorrelated with each other; also that $E(FF')=I$, and $E(UU')=V$, a diagonal matrix [14]. This is a simultaneous equation set which assumes all variables stochastic and errors in all variables, i.e., there is a vector of errors in U associated with each of the variables in $\mathcal{Z}$. Also, factor analysis provides for multicollinearity through the assumption that variances of all real variables can be explained by a set of factors less in number than the set of variables (m<h). As multicollinearity among the variables increases, the number of factors needed to explain the variances becomes smaller relative to the number of real variables.

COMBINING REGRESSION AND FACTOR ANALYSIS

A way to combine factor analysis with regression as reported here is, for purposes of differentiation, called "Classical Factor Analysis Regression" or CFAR. CFAR is intuitively easier to understand and easier to calculate than earlier factor analysis regression derivations.

Using standardized variables in ordinary least squares regression, the estimating equation is:

$$\hat{B} = R_{xx}^{-1} R_{xy} \qquad (2)$$

where

$\hat{B}$ = kX1 vector of regression coefficients
R_{xx} = kXk correlation matrix of the explanatory variables and
R_{xy} = kX1 vector of correlations between the dependent and explanatory variables.

This is the formulation actually used in most computer programs rather than the usual textbook formula, $\hat{B}=[X' X]^{-1} X' Y$, [8] because it reduces size and range of numbers involved and, therefore, increases both accuracy and speed of matrix inversion.

Using the same general formula, CFAR estimates of the regression are:

$$\tilde{B} = \hat{R}_{xx}^{-1} \hat{R}_{xy} \qquad (3)$$

where

$\tilde{B}$ = kX1 vector of CFAR coefficients
$\hat{R}_{xx}$ = kXk factor reproduced correlation matrix of the explanatory variables and
$\hat{R}_{xy}$ = kX1 factor reproduced correlation vector of the explanatory variables with the dependent variable.

The factor analysis reproduced correlation matrix is one of the results normally obtained by most factor analysis computer routines. If the explanatory variables are entered in the factor analysis model, equation (1), as the variables $z_1,\ldots,z_k$ and the dependent variable is z_{k+1} where k+l=h, then factor analysis is performed on the hXh correlation matrix formed from both explanatory and dependent

variables. If a maximum likelihood or least squares factor analysis routine is used to obtain the factor loading matrix, A, then:

$$AA' + V = \hat{R} \tag{4}$$

where

A = h×m factor loading matrix
V = h×h diagonal matrix of specific variances[1] and
$\hat{R}$ = h×h maximum likelihood estimate of the full correlation matrix of all the variables [9, 13, 26].

Then $\hat{R}$ can be partitioned. The upper lefthand k×k part is $\hat{R}_{xx}$ and the first k elements of the most righthand column vector of $\hat{R}$ is $\hat{R}_{xy}$, which are used in equation (3) to obtain the CFAR estimates.

The variances for $\hat{B}$ in classical regression are calculated using variance of the dependent variable from the estimated regression surface and elements in R_{xx}^{-1}. The stochastic part is variance of the dependent variable from the regression surface since the X variables in OLS are assumed to be known, fixed values.

Now, however, $\hat{R}_{xx}^{-1}$ in CFAR is also stochastic. While it has been shown that the correlation matrix of normally distributed variables follows the Wishart distribution [27], distribution of $\hat{R}$ from factor analysis, is still unknown. Therefore, there is currently no direct way to calculate a test statistic for the individual CFAR coefficients except by multiple F tests as previously suggested [21, p. 559]. Thus far, however, this test is only heuristic.

THE NUMBER OF FACTORS TO EXTRACT

When there are only errors in the variables and multicollinearity is not a problem, then the assumption is still that there are k independent explanatory variables. Therefore, there should be k factors extracted—one for each independent explanatory variable, or alternatively Lawley's X^2 test may be used to determine the optimum number of factors to extract [14].

When the additional problem of multicollinearity is involved, the real explanatory vector space is reduced and the number of factors to extract will be equal to the number of independent explanatory vectors rather than the number of explanatory variables. This will be equal to the number of subsets of multicollinear variables plus the number of independent variables. For example, in the following correlation matrix, (5) of six explanatory variables, X_1 and X_2 are one subset of multicollinear variables, while each is relatively independent of all others: X_3, X_4 and X_5 form another such subset, and X_6 is relatively independent of all of the other five variables. Here the recommendation would be to extract three factors—one for each of the two multicollinear subsets and one for the variable which is relatively independent of all others. Decision on the number of factors to extract may not always be this clear cut.

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1.00					
X_2	.92	1.00				
X_3	.33	.21	1.00			
X_4	.26	.17	.93	1.00		
X_5	.15	.27	.87	.89	1.00	
X_6	.23	.12	.22	.09	.14	1.00

(5)

Then one can resort to the Lawley criterion suggested earlier.

The concept underlying these recommendations flows from the assumptions and results obtained from factor analysis. If the factor loading matrix A shown in the factor model (equation 1) is h×h, which means that the number of factors extracted is h, then AA' in equation (4) will reproduce the original empirical correlation matrix exactly, implying that there are no errors in the variables. If m is the number of factors extracted and m=h—1, then there is allowance for errors in the variables. The factor analysis model is, then, a true stochastic model rather than the mathematical principal components model which has no error component in the model:

$$\not{Z} = AP \tag{6}$$

where

$\not{Z}$ = h×1 vector of real variables
P = h×1 vector of principal components and
A = coefficient matrix, but is now h×h.

As m is further reduced, the inference is that multicollinearity exists among the explanatory variables and explanatory vector space is less than explanatory variable space. Therefore, a researcher may use the heuristic recommendations suggested in this section; or he may prefer to use the Lawley test [14], which is a statistical test to determine the number of factors to extract.

[1] In equation (4), V is also I-diagonal (AA'), where I is the unity matrix.

MONTE CARLO RESULTS

A monte carlo study of the characteristics of CFAR estimators relative to those of ordinary least squares was recently conducted. The approach used was that a dependent variable was to be explained by 12 explanatory variables when there were errors in all variables, with different sample size, and different degrees of statistical dependence among explanatory variables [24].

In this study, the ordinary least squares estimates were taken as the parameters before adding random errors to the variable set. Then, random normal errors were added repeatedly to all variables. Each time the regression coefficients were estimated by both CFAR and OLS. These parameter estimates with errors-in-the-variables were then compared with parameters estimated before adding the errors.

Normality of the parameter estimates from CFAR was inferred by examining skewness and kurtosis, both of which approached zero as sample size increased. This is what we would expect if the CFAR estimators were normally distributed.

CFAR estimators appear to have a small negative bias, but the bias assymptotically approaches zero as sample size increases.

Mean square errors for both the estimators themselves and for the prediction of Y were examined for both OLS and CFAR. These results show that CFAR has a very substantial advantage over OLS. Depending upon sample size, the CFAR mean square error for the estimators ranges from eight to as much as 50 times smaller than for OLS with greatest advantage for CFAR at small sample sizes. This result is most important to economists since we generally deal with relatively small samples. The mean square error for the prediction also is smaller for CFAR estimators, again with greatest advantage at small sample sizes where the MSE from CFAR predictions is about one-fourth to one-fifth as large as the MSE for predictions from OLS.

The other variable in the monte carlo study mentioned earlier was statistical dependence or level of multicollinearity among explanatory variables. Results here are also important to economists: as multicollinearity among explanatory variables increased, advantages are greater for CFAR relative to OLS based on the MSEs of both estimators and predictions.

Conclusions drawn are that CFAR is an excellent alternative to OLS, especially for economists, because of identified problems where CFAR shows the greatest advantage. These problems are errors-in-the-variables, multicollinearity and small sample size.

REFERENCES

[1] Adelman, Irma and G. Dalton. "A Factor Analysis of Modernization in Village India," *Economics Journal*, 81(323):563-379, September 1971.

[2] Adelman, Irma and Cynthia Taft Morris. "A Quantitative Study of Determinants of Fertility," *Economic Development and Cultural Change*, 14, pp. 129-157, January 1966.

[3] Amemiya, T. "On the Use of Principal Components of Independent Variables in Two-Stage Least Squares Estimation," *International Economic Review*, 7-3, pp. 283-303, September 1966.

[4] Bursch, William G., John T. Scott, Jr. and Roy N. Van Arsdall. "Characteristics and Prospects of the Commercial Hog Feed Market in Illinois," Bulletin 743, College of Agricultural, University of Illinois, Urbana, Illinois, 1972.

[5] Doll, John P. and Sean B. Chin. "A Use for Principal Components in Price Analysis," *American Journal of Agricultural Economics*, 52-4, pp. 591-593, November 1970.

[6] Farrar, Donald E. and Robert R. Glauber. "Multicollinearity in Regression Analysis: The Problem Revisited," *The Review of Economics and Statistics*, 49, pp. 92-107, 1967.

[7] Holdren, Bob R. "Some Applications of Maximum Likelihood Factor Analysis in Econometrics," paper presented at the Econometric Society Meeting, Chicago, December 1968.

[8] Johnston, J. *Econometric Methods*, New York, 1963.

[9] Joreskog, K. G. "On the Statistical Treatment of Residuals in Factor Analysis," *Psychometrika*, 27, pp. 335-354, 1962.

[10] King, Benjamin F. "Comment on 'Factor Analysis and Regression,' " *Econometrica*, 37(3):538-540, July 1969.

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

[12] Ladd, George W. and R. L. Oehrtman. "Factor Analysis of the Market Studies of the Fluid-Milk Bottling Industry in the North Central Region," Iowa State University Experiment Station, Ames, Iowa, Research Bulletin 573:541-575, October 1971.

[13] Lawley, D. N. "The Estimation of Factor Loadings by the Method of Maximum-Likelihood," *Royal Society of Edinburgh Proc.*, A-60, pp. 64-82, 1940.
[14] Lawley, D. N. and A. E. Maxwell. "Factor Analysis as a Statistical Method," London, 1963.
[15] Lawley, D. N. "Regression and Factor Analysis," *Biometrika*, 60-2, pp. 331-332, 1973.
[16] Mangan, Frederick K. "A Monte Carlo Study of Linear Regression and Factor Analysis Under Multicollinearity," Unpublished Masters Thesis, Seattle, Washington, University of Washington, 1970.
[17] Marquardt, D. W. "Generalized Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation," *Technometrica*, Volume 12, pp. 591-612, 1970.
[18] McDonald, G. C. and D. I. Galarneau. "A Monte Carlo Evaluation of Some Ridge Type Estimators," *Journal of the American Statistical Association*, 70, pp. 407-416, 1975.
[19] Olsen, Bernard M. and Gerald Garb. "An Application of Factor Analysis to Regional Economic Growth," *Journal of Regional Science*, 6, pp. 51-56, 1965.
[20] Pruzek, Robert M., John R. Sherry and Mary E. Huba. "Improving Least Squares Regression Estimators Through Matrix Factoring," paper given at the Annual Spring Meeting of the Psychometric Society, Iowa City, Iowa, April 1975.
[21] Scott, John T., Jr. "Factor Analysis Regression," *Econometrica*, 34, pp. 552-562, 1966.
[22] Scott, John T., Jr. "Factor Analysis Regression Revisited," *Econometrica*, 37(4):719, October 1969.
[23] Scott, John T., Jr. "The Synthesis of Classical Regression and Factor Analysis," Urbana, Illinois, Department of Agricultural Economics paper AE-4393, April 1970.
[24] Scott, John T., Jr. and Allen Fleishman. "Classical Factor Analysis Regression (CFAR) and its Statistical Properties," paper at the Western Economics Association Meeting in San Francisco, June 27, 1976.
[25] Shetty, W. S. "A Factor Analysis of Use of Fertilizers by Farmers," *Indian Journal of Agricultural Economics*, 24(1):50-61, March 1969.
[26] Whittle, P. "On Principal Components and Least-Squares Methods of Factor Analysis," *Skandinavisk Aktuarietidskrift*, 35, pp. 223-239, 1952.
[27] Wilks, S. S. "Certain Generalizations in the Analysis of Variance," *Biometrika*, 24, pp. 471-494, November 1932.