PRINCIPAL COMPONENTS AND THE PROBLEM OF MULTICOLLINEARITY

Bernard J. Morzuch

Multicollinearity among independent variables within a regression model is one of the most frequently encountered problems faced by the applied researcher. In a recent article in this *Journal* (Willis, *et al.*), a catalog of "remedies" for multicollinearity was presented to assist in reducing its level and associated adverse consequences.

One of these remedies—principal components—was suggested as a method of transforming a set of collinear explanatory variables into new variables that are orthogonal to each other with the first few of these transformed variables accounting for the majority of the variability in the original data set. In principal components regression, a transformed variable is determined to be important and included or unimportant and excluded in the regression model depending upon the size of the characteristic root (eigenvalue) associated with its corresponding characteristic vector (eigenvector) (Massy), the statistical significance of its regression coefficient (Mittelhammer and Baritelle), or some combination of eigenvalue size and correlation with the dependent variable (Johnson, *et al.*).

Unfortunately, this technique is widely abused and misunderstood by the applied researcher. Confusion exists with respect to (1) its relationship to other orthogonalization techniques; (2) the meaning of the orthogonalized components and the necessity of transforming the principal component estimators back to the original parameter space; (3) the implications of deleting components and the correspondence of this technique to a particular type of restricted least squares estimator; (4) the proper way to delete components and evaluate these implied restrictions; and (5) actual implementation of this estimation procedure via available computer routines.

The purpose of this note, therefore, is to place the technique of principal components in perspective and to suggest a methodology for implementing this technique correctly.

DEALING WITH MULTICOLLINEARITY

Multicollinearity is the result of a lack of selective variation among the independent variables in a regression model. It is a problem associated with passively generated data, i.e., data obtained from some outside source over which the investigator has no control or data characterized by lack of experimental design. Consequently, the problem of multicollinearity can never be cured; it can only be treated in an *ad hoc* manner.

The way to deal with multicollinearity is to introduce additional sample information to hopefully increase the selective variation among the independent variables. Such information is normally added by way of restrictions on the parameters suggested by theory. These restrictions may take the form of exact linear restrictions (Goldberger, pp. 256-8), inequality restrictions (Judge and Takayama), and stochastic restrictions (Theil and Goldberger).

In the absence of any theoretical basis for admitting restrictions on the parameters, an alternative is to place restrictions on the independent variables themselves. This can be accomplished by transforming the original variables into artificial constructs and then retaining certain of these constructs in a regression model on the basis of their contribution to variability in the original data, while eliminating—placing zero restrictions on—those constructs that contribute little or nothing to the variability in the original data. This is precisely the focus of principal components.

COMPARISON WITH OTHER ORTHOGONALIZATION TECHNIQUES

Principal components is one specific type of factor analysis. All methods of factor analysis attempt to analyze the structure of multivariate observations so as to reduce a set of data to a smaller set of latent factors. Beyond this link, any other similarity between principal components and the general body of techniques known as factor analysis is rather limited.

Principal components analysis transforms a given set of variables into a composite set of components that are orthogonal to, i.e., totally uncorrelated with, each other. No particular assumption about the underlying structure of the variables is required. In this sense, it is merely a transformation rather than the result of a fundamental model for covariance structure (Morrison, p. 259).

Factor analysis assumes that the relationships among the variables are the result of some underlying regularity in the data, i.e., each observed variable is influenced by various determinants which are common to other variables and by a component unique to itself. The common determinants in turn are smaller in number than the original variables themselves. It is looked upon as a technique for explaining the covariances among the variables and therefore as a fundamental model for covariance structure rather than merely an orthogonalization technique.

Both principal components and factor analysis can be appropriate methods for dealing with collinearity among independent variables. However, when using factor analysis in regression, the estimated coefficients on "important" factors can be interpreted only in terms of linear combinations of the original variables. Principal components on the other hand permits the coefficients on the important components to be reparameterized in terms of the original variables. This has real appeal in any economic investigation.

PRINCIPAL COMPONENTS IN REGRESSION ANALYSIS

To appreciate its use in econometric analysis, consider the model:

(1) $Y = X\beta + \epsilon$,

where y is an n × 1 vector of observations, X is an n × k matrix of observations of rank k, β is a k × 1 vector of unknown parameters, and ϵ is an n × 1 vector of $N(O,\sigma^2)$ independent and identically distributed random disturbances. Assume also a high degree of collinearity among the independent variables and little or no theory to assist in placing restrictions on the parameters. These two conditions justify using this technique.

The method of principal components involves the transformations:

(2) $y = XFF'\beta + \epsilon = XF\delta+\epsilon = P\delta\epsilon$,

where $F = (f_1,f_2,\ldots f_k)$ is a k × k matrix with columns f_i being characteristic vectors of $X'X$; $FF' = I$, an identity matrix of rank k;

Bernard J. Morzuch is Assistant Professor of Food and Resource Economics, University of Massachusetts—Amherst.

$F'\beta = \delta$, that is, δ is a $k \times 1$ vector and a linear combination of the matrix of eigenvectors F and the parameter vector β. Conversely, $\beta = F\delta$, that is, δ can be transformed to the original parameter space by making use of $FF' = I$. Finally, $XF = P$, that is, the $n \times k$ matrix of principal components is the product of the original data matrix and matrix of eigenvectors. More specifically, let the characteristic vectors be ordered to correspond to the relative magnitudes of the characteristic roots of $X'X$. The $n \times k$ matrix of principal components is $P = (p_1, p_2, \ldots, p_k)$ with $p_i = Xf_i$ being the ith principal component of $p_i p_i = \lambda_i$, the ith largest characteristic root of $X'X$. Johnson, *et al.* explain in greater detail the technique and the statistical properties obtained by deleting one or more of the variables p_i in equation (2).

Briefly, the method involves partitioning F in equation (2) into $[F_1 \vdots F_2]$ where F_1 is a $k \times r$ matrix of "important" eigenvectors, F_2 is a $k \times s$ matrix of "unimportant" eigenvectors, and $r + s = k$. In light of this information, equation (2) can be rewritten as:

$$(3)\quad Y = X[F_1 \vdots F_2]\delta + \epsilon.$$

Likewise, δ can be appropriately redimensioned as $[\delta_1 | \delta_2]'$ where δ_1 is an $r \times 1$ parameter vector associated with F_1, and δ_2 is an $s \times 1$ parameter vector associated with F_2. Thus

$$(4)\quad Y = XF_1\delta_1 + XF_2\delta_2 + \epsilon = P_1\delta_1 + P_2\delta_2 + \epsilon.$$

Principal component estimators are obtained by deleting the "unimportant" set of components P_2 and applying OLS to the resulting model. Since P_1 is orthogonal to P_2, the estimator d_1 of δ_1 will be unbiased, and the sample variance $\Sigma_{d_1 d_1}$ will be smaller for the retained set of components than for the entire set, i.e., Σ_{dd}.

REPARAMETERIZATION OF THE PRINCIPAL COMPONENT ESTIMATOR

Many researchers are at a loss, however, once they have performed the estimation suggested by equation (4). Their inclination is to prescribe interpretations for d_1, the estimator of δ_1. This is a difficult task because δ_1 is a coefficient vector on P_1 which itself consists of vectors that are linear combinations of the original variables. Any interpretation on d_1 is clearly deficient in that each element of d_1 cannot be associated with a particular independent variable.

However, in making use of the information suggest by equation (2), the principal component estimator can be used to generate an estimator for β, i.e., $F_1 d_1 = b^*$. Therefore, in combination with the retained set of eigenvectors, the regression coefficients d_1 can be translated back to the original parameter space.

The estimator b^* is appealing in that it amounts to a particular type of restricted least squares estimator. It is equivalent to minimizing the sum of squared residuals $(y - X\beta)'$ $(y - X\beta)$ in equation (1) subject to the exact linear restrictions $F_2'\beta = 0$ (Johnson, *et al.*). The properties of this estimator are well established (Goldberger, pp. 256-8).

DELETION OF COMPONENTS AND EVALUATION OF RESTRICTIONS

A further complication results with respect to the optimal number of components to delete. The tendency traditionally has been to delete components associated with small eigenvalues, e.g., less than one. The limitation of this approach is that components with small eigenvalues may be correlated very highly with the dependent variable. Thus, a structural norm which simultaneously considers the amount of variability accounted for by a particular component and its correlation with the dependent variable has greater appeal. A particular norm which accounts for these two measures is the F test (Fisher). Components can be sequentially deleted until a new restriction, i.e., the deletion of an additional component, causes no improvement with respect to the fit of the equation.

One faces the choice regarding whether or not the restrictions are consistent with the sample data, i.e., whether the restrictions are true or involve some degree of inconsistency. Indeed, an investigator never knows the truth or falsity of the implied restrictions. Hence they can be analyzed by way of alternative norms which account for each of these possibilities. Testing the truth of the restrictions can be analyzed by way of the classical (central) F test (with centrality parameter zero). Since restricted least squares estimators may be baised, it is likewise appropriate to test whether the imposed restrictions result in a reduction of some measure of mean square error (MSE). A very workable criterion to test for MSE improvement if the weak MSE criterion suggested by Toto-Vizcarrondo and Wallace. The test statistic used is the non-central F (with non-centrality parameter one-half). Quite a bit of misunderstanding revolves around its implementation. It is used in precisely the same manner as the central F, the difference being that critical values for given degrees of freedom are different than the central F due to the noncentrality parameter (see Wallace and Toro-Vizcarrondo). Thus, depending upon the structural norm employed, i.e., depending upon the intent of the researcher regarding the truth or falsity of the restrictions, evaluation of sets of restrictions may lead to different results.

LIMITATIONS INVOLVED WITH SEQUENTIAL DELETION

The above procedures suggest a much more rigorous manner of selecting and judging the appropriateness of restrictions than using an arbitrary norm such as size of eigenvalues. However, statistical properties of the ensuing estimators in the suggested framework must be viewed with caution. As the same set of sample data is used to sequentially test an additional restriction, i.e. the deletion of another component, the resulting estimators are preliminary test estimators. The tests suggested above to judge the restrictions thus are merely rough guides as to the consistency of the restrictions with the sample data (Wallace).

COMPUTER ROUTINES

Computer routines exist which conveniently perform all of the manipulations required for principal components analysis. Especially useful regression routines are available within the SHAZAM computer package put out by Rice University and Biomedical Computer Programs (BMDP) put out by the University of California Press. The attractiveness of each of these programs is that estimators on the deleted component regression models are translated back to the original parameter space. All pertinent summary information is supplied so that the restrictions can be evaluated via a central or non-central F test.

REFERENCES

Dixon, W. J. ed. *Biomedical Computer Programs (BMDP).* Berkeley: University of California Press, 1975.

Fisher, Franklin M. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions: An Expository Note." *Econometrica.* 38(1970):361-6.

Goldberger, Arthur S. *Econometric Theory.* New York: John Wiley & Sons, Inc. 1964.

Johnson, S. R., Steven C. Reimer, and Thomas P. Rothrock. "Principal Components and the Problem of Multicollinearity." *Metroeconomica.* 25(1973):306-17.

Judge, G. G. and T. Takayama. "Inequality Restrictions in Regression Analysis." *J. Amer. Statist. Assoc.* 61(1966):166-81.

Massy, William. "Principal Components Regression in Exploratory Statistical Research." *J. Amer. Statist. Assoc.* 60(1965):234-56.

Mittelhammer, Ron C. and John L. Baritelle. "On Two Strategies for Choosing Principal Components in Regression Analysis." *Amer. J. Agr. Econ.* 59(1977):336-43.

Morrison, Donald F. *Multivariate Statistical Methods.* New York: McGraw-Hill Book Co. 1967.

Theil, H. and A. S. Goldberger. "On Pure and Mixed Statistical Estimation in Economics." *Intern. Econ. Rev.* 2(1961):65-78.

Toro-Vizcarrondo, Carlos, and T. D. Wallace. "A Test of the Mean Square Error Criterion for Restrictions in Linear Regression." *J. Amer. Statist. Assoc.* 63(1968):558-72.

Wallace, T. D. "Pretest Estimation in Regression: A Survey." *Amer. J. Agr. Econ.* 59(1977):431-43.

Wallace, T. D. and C. E. Toro-Vizcarrondo. "Tables for the Mean Square Error Test for Exact Linear Restrictions in Regression." *J. Amer. Statist. Assoc.* 64(1969):1644-93.

White, Kenneth J. *SHAZAM (Computer Manual).* Version 2.3. Department of Economics, Rice University. August 1978.

Willis, Cleve E. and Robert D. Perlack, *et al.* "Multicollinearity: Effects, Symptoms, and Remedies." *Journal of the Northeastern Agricultural Economics Council.* 7(1978):55-61.