Applying Principal Components Regression Analysis to Time Series Demand Estimation

By Luis R. Sanint*

Abstract

Demand functions for rice in Colombia and Venezuela, estimated by means of ordinary least squares, were unsatisfactory because of problems with multicollinearity An alternative approach, principal components regression, was tried Results showed that principal components regression estimates were more consistent with theoretical expectations and were statistically more significant The cost of these gains was that the coefficients were biased However, the mean-square-error tests indicated that the reduction in variance outweighed the loss due to bias

Keywords

Principal components regression, ordinary least squares, time series demand estimation

The problem of multicollinearity occurs frequently in time series analysis A number of statistical tools have been proposed to mitigate multicollinearity, and researchers have studied their statistical properties Principal components regression (PC-OLS) is one technique proposed for use when ordinary-least-squares (OLS) parameter estimates are affected by multicollinearity But, the conceptual problems that arise from applying alternative forms of biased estimation to economic matters have not been widely discussed

In this article, I apply the principal components regression technique to time series rice demand equations for Colombia and Venezuela and enumerate the advantages and disadvantages associated with the technique

The Problem

Important changes in rice production took place in Colombia and Venezuela after 1956 Rice yields in Colombia increased from a 1956-58 average of 1 8 tons per hectare to 4 1 tons per hectare for 1976-78, while the annual rate of growth in production was 9 1 percent Rice yields in Venezuela increased only slightly less, from 1.3 tons per hectare to 3 2 tons per hectare for the 1956-78 period, but the increase and improvements in area planted allowed a 15 7 percent annual rate of growth in production for that period, the largest in Latin America (*14*) [1]

The purpose of this exercise was to estimate the impact of the lower real retail prices of rice, the higher levels of income, and the changes in the prices of other closely related foods on the level of per capita rice demand in both countries

Per capita consumption of rice in any given year, y_1, is defined as apparent disappearance divided by population Apparent disappearance is defined as production plus imports minus exports, minus changes in stocks

A doublelog OLS demand specification was selected for per capita rice consumption, y, in both countries [2] Each equation is of the form

$$y = XB + u \qquad (1)$$

where the independent variables are the logs of the price of rice, the real retail prices of seven other basic foods—corn, wheat flour, potatoes, cassava, plantains, beef, and beans—and per capital real income I assumed that rice consumption is not related to the prices of other commodities not included in the model

OLS estimation of the models revealed the following characteristic symptoms of multicollinearity on the parameters of the model, the individual contributions to the R-square value added to less than half its value, most of the t-values were low, and some of the elasticities seemed unreasonable (like a

*The author is an agricultural economist with the International Economics Division, ERS The helpful comments from David L Peacock, Fausto Medina-Lopez, R Carter Hill, Per Pinstrup Andersen, and others are gratefully acknowledged

[1] Italicized numbers in parentheses refer to items in the References at the end of this article

[2] The period of analysis covers the years 1956 77 for Colombia and 1959-77 for Venezuela, they coincide with the introduction of modern rice varieties in both countries Italicized characters are nx1 vectors, and X is a 7xn matrix, where n is the number of observations

positive own-price elasticity and an extremely high income elasticity for rice in Colombia) (table 1) Examination of the simple correlations and of the eigenvalues confirmed the presence of multicollinearity (tables 2 and 3)

Handling Multicollinearity

Perfect multicollinearity exists when a subset of the vectors x_j of the matrix of explanatory variables X are linearly dependent—that is, if there exist nonzero constants, $a_1, a_2, \ldots a_p$, so that $\Sigma a_j x_j = 0$ (*24*) Perfect multicollinearity is a problem of existence, multicollinearity is a problem of degree (*13*) The practical problem faced by researchers is severe multicollinearity Because it is a sample problem rather than a population problem, there are no definite tests Several multicollinearity measures have been suggested (*28*), the most common are the variable correlation matrix and the parameter correlation matrix Assessing the magnitude of the problem involves subjective judgment A number of alternative ways for dealing with the associated problems of multicollinearity have been proposed (*16*) A brief description of some of them follows

- **Augmentation of Data** This is frequently mentioned as the best approach This solution is not practical in our case, as it was not possible to obtain data for a longer period

- **Restricted Least Squares (deterministic or stochastic)** This is useful when there is reliable prior information about some of the parameters involved in the multicollinearity problem This was not the case in this study

- **Variable Deletion.** This can ameliorate the degree of multicollinearity However, the fact that certain of the explanatory variables in a given model appear highly correlated should not be regarded as grounds for changing the specification of the model (*5, 9*) In our case, dropping income from the demand equations would have only worsened the situation by introducing additional errors of specification

- **Transformation of Data** This is useful when the interpretation of the structural hypothesis is not affected by the transformation or is not important to the researcher Examples of transformations are first differences, ratios of variables, and indexes

- **Ridge Regression** This was first proposed by Hoerl and Kennard It produces biased estimates, and their expected bias is greatly increased when the parameters are of opposite signs (*2*) This is the case in both equations for the expected own-price elasticity and the income elasticity

- **Factor Regression Analysis** This is based on factor analysis techniques Principal Components Regression (PC-OLS) belongs to this group In addition to mitigating the problem of multicollinearity, PC-OLS greatly reduces the influence of outliers in the data Deleting one or more components to mitigate multicollinearity implies an obvious trade-off Unless the true (and unknown) parameter vector lies in the subspace chosen for examination, the resulting estimators will be biased The trade-off is between biasedness and reduction in parameter variance

Single-equation linear models are typically estimated for one of two purposes (1) to test some theoretical or structural hypothesis, and (2) to use an equation solely as a forecasting tool In our example, the goal is to estimate the structural relationships between the dependent and the independent variables With that purpose in mind, we selected PC-OLS to mitigate the problem of multicollinearity

Principal Components Regression Analysis (PC-OLS)

Principal components are linear combinations of observed variables (the logarithms of the explanatory variables in this case) The components are orthogonal to each other The first principal component represents the largest amount of variation in the data, the second represents the second largest, and so on (*12*) It is advisable to standardize the variables to avoid scale problems The PC-OLS model of y on Z is

$$y = Zd + u \qquad (2)$$

Let Z_{nxp} be the matrix of principal components of X Thus, $Z = XC$, where C is a pxp matrix composed of characteristic vectors of $X'X$,

From the traditional OLS model, $y = XB + u$, one can find the PC-OLS of y on Z, $y = Zd + u$, where d is a pxl vector of coefficients, and u is the vector of random errors (*3, 6, 7, 8, 9, 11, 15, 17, 18*)

PC OLS represents a compromise between the criteria of unbiasedness and minimum variance T D Wallace and associates (*26*) have proposed that comparisons between restricted and unrestricted estimators be based on mean square error (MSE) The MSE criterion provides one framework for considering the problem of multicollinearity in a linear model (*29*)

McCallum (*18*) has shown that one can obtain a biased coefficient with a lower mean square error by eliminating some of the components Because the new biased PC-OLS estima-

Table 1—Doublelog OLS and PC-OLS estimation of per capita white rice demand, Colombia and Venezuela[1]

Country	Estimation method	Intercept	Price of white rice	Price of corn	Price of potatoes	Price of cassava	Price of plantains	Price of wheat flour	Price of beef	Price of beans	Per capita income	$\bar{R}^2$	SER	DW
		B_0	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9			
Colombia	OLS	-1 89 (-1 80)	0 38 (1 49)	0 01 (09)	-0 32 (-4 30)	0 31 (5 07)	-0 19 (-1 89)	0 25 (2 46)	-0 31 (-2 39)	0 44 (5 38)	2 47 (6 88)	0 97	0 0538	1 73
	PC-OLS[2]	1 71	- 69 (-31 53)	21 (1 59)	39 (-4 09)	38 (5 23)	-0 48 (-12 42)	10 (1 05)	- 14 (- 99)	52 (5 46)	93 (48 86)	92	0627	1 54
Venezuela	OLS	1 30 (53)	- 65 (-1 01)	- 74 (- 82)	26 (32)	- 46 (- 53)	1 04 (1 83)	- 42 (- 67)	05 (10)	20 (40)	85 (60)	71	1677	1 73
	PC-OLS[3]	1 92	- 59 (-8 32)	36 (1 68)	13 (- 56)	- 87 (-3 26)	1 05 (2 00)	- 65 (-2 96)	08 (2 02)	03 (25)	58 (6 86)	74	1546	1 75

[1] Numbers in parentheses are t-values, $\bar{R}^2$ is corrected R^2, SER is Standard Error of the Regression, DW is Durbin Watson statistic Critical t-value at 90-percent level is 1 34 All variables are expressed in natural logarithms
[2] One component is deleted
[3] Two components are deleted

Table 2—Eigenvalues and cumulative fractions of variance explained by each component, Colombia and Venezuela

Colombia		Venezuela	
Eigenvalue	Cumulative fraction	Eigenvalue	Cumulative fraction
3 44	0 38	4 83	0 54
1 58	56	1 84	74
1 31	70	1 24	88
1 02	82	44	93
68	89	26	96
38	95	19	98
33	97	11	99
22	99	06	99
02	1 00	02	1 00

tor has lower variance, it may lie closer to B than the OLS estimator does (*1*) The process of deleting components makes PC OLS estimates equivalent to linearly restricted OLS estimates (*17*) PC-OLS belongs to the set of Stein type estimators, as it implies giving up an unbiased estimator in favor of a reduced MSE Principal components restrictions are known to yield the maximum variance reduction of all sets of linear restrictions of equal size (*7*)

Several tests have been developed to evaluate the bias-variance tradeoff when restricted versus unrestricted OLS estimators are compared They can be classified into two groups of norms structural and predictive (*10*)

Theoretically, if one runs a regression equation using all p principal components it will yield the same transformed coefficients as the original regression The difference can be attributed to the reduction in the roundoff problem because of the decrease of the near singularity in the design matrix

The transformed coefficients for the normalized data can be obtained from the formula

$$b_i = \sum_{j=1}^{p} C_{ij} d_j \qquad (3)$$

where b_i is the OLS estimate of B_i, and C_{ij} and d_j are defined as above (*18*) The standard errors of the coefficients are

$$SE(b_i) = \left[\frac{s^2}{n} \sum_j \frac{C^2_{ij}}{E_j}\right]^{1/2} \qquad (4)$$

Where

E_j is the eigenvalue of the j_{th} component,
s is the standard error of the regression,
n is the number of observations, and
C_{ij} and b_i are defined as above (*18*)

When a subset of the components is selected, the appropriate C_{ij} and x_j are deleted from the above sum

Deletion Criteria

We considered two traditional methods in selecting a subset of the components which are used to form PC-OLS estimates the Characteristic Root Criterion (CRC) and the *t*-value criterion (TVC) (*10, 19*)

CRC deletes those principal components associated with the smallest characteristic roots of the correlation matrix of the independent variables (eigenvalues) Deletion based on small characteristic roots implies little loss in the variation of the independent variables

TVC allows the vector *y* of the dependent variables to play a role in the exclusion of the principal components Components with insignificant *t* values will be dropped Thus selec-

Table 3—Correlation matrix of the independent variables, 1956-77[1]

Commodity	Price of white rice	Price of corn	Price of potatoes	Price of cassava	Price of plantains	Price of wheat flour	Price of beef	Price of beans	Per capita income
	Dollars								
Price									
White rice	1 00	0 01	−0 29	−0 32	0 12	0 49	−0 52	0 22	−0 94
Corn	− 71	1 00	− 08	− 51	02	29	− 32	20	− 02
Potatoes	34	21	1 00	45	06	− 19	− 01	06	23
Cassava	− 72	68	− 34	1 00	14	− 38	51	− 11	34
Plantains	− 40	79	51	47	1 00	16	− 03	45	− 29
Wheat flour	59	− 71	− 22	− 55	− 67	1 00	− 56	04	− 60
Beef	− 65	51	− 29	44	11	− 32	1 00	− 15	61
Beans	31	35	09	− 01	16	− 47	− 53	1 00	− 24
Per capita income	− 92	82	− 21	80	57	− 66	54	16	1 00

[1] Colombia upper triangular matrix Venezuela lower triangular matrix

Source Author's calculations

tion leads to preliminary test principal components (PTPC) (7) Fomby and Hill conclude that "when components are deleted on the basis of statistical tests, the restricted least squares formulation of PC OLS combined with the preliminary test literature make it clear that any testing procedure may not produce parameter estimates superior in mean square error (MSE) relative to OLS" (7, p 526)

In the cases treated here, the components were not highly correlated with the individual independent variables An economic interpretation of them was not feasible Keeping most of the original sample variability in the estimation process was judged as being important That implies deletion of those components with smaller eigenvalues In addition, the reduction in variance is inversely related to the eigenvalue So, it was decided to follow the CRC

CRC offers an interesting set of possibilities in the bias-precision trade-off Although the restrictions we consider are sample specific and have no economic interpretation, they may nonetheless yield useful information on the poorly planned experiment (secondary sources) which generated the data Because those components with the smallest eigenvalues are deleted, the marginal reduction in parameter variance is maximized This situation occurs because precision is directly related to the size of the eigenvalues associated with the deleted components (7)

We imposed a threshold level (95 percent) on the maximum amount of variability of the original data to be kept, that is, after those components with the smallest eigenvalues were deleted, at least 95 percent of the original variability was retained Suppose that m is the number of components that allow the 95-percent threshold level to be met The next stage is to estimate a set of PC-OLS equations deleting components one by one, starting with the component with the smallest eigenvalue, until m components are deleted From that set of equations, we chose the one with the highest corrected R-square We then tested the resulting mixed estimates against the OLS estimate for generalized MSE superiority (GMSE) based on a noncentral F distribution (27) to examine the appropriateness of the restrictions and the various trade-offs between estimator bias and variance reduction, from the structural viewpoint

Results of PC-OLS Estimation

The 95-percent threshold level imposed for the CRC was met, with up to three and four components deleted in the Colombia and Venezuela equations, respectively (table 2) The corrected R-square was highest when two components were deleted from the Venezuela equation and when one component was deleted from the Colombia equation

When an appropriate structural test was used, both PC-OLS estimates were found to be superior in the generalized MSE sense (GMSE) to the OLS estimates To test the hypothesis that a set of constrained estimators, d, is better than the unconstrained estimators, B, according to the GMSE criterion, we used the test and tables developed by Toro-Vizcarrondo and Wallace (25) The test is based on the noncentral F distribution of the statistic

$$a^* = \frac{SSE\ (d) - SSE(B)}{SSE(B)/(n - k)} \qquad (5)$$

where SSE is the sum of squared errors and $n - k$ are the degrees of freedom of the unrestricted equation The corresponding values were 8 37 and 0 37 for Colombia and Venezuela, respectively, which failed to reject the hypothesis in both cases [3]

The effects of multicollinearity on the values of the parameters and their variances were more notorious in the rice demand equation for Colombia, particularly on the own-price and income elasticities The Venezuela PC-OLS equation has a lower standard error of regression than the OLS equation, the gains in parameter precision were important (table 1)

A comparison of the results of PC-OLS and OLS leads to important conclusions about the structural nature of the equations The PC-OLS Colombia equation exhibits a negative own-price elasticity in contrast to the OLS equation It also shows a substantially smaller income elasticity Both elasticity estimates are theoretically more sound, and they are similar to results reported by others (4, 21, 22) In addition, the two variables greatly reduce parameter variance These changes are not surprising, as the correlation between them was the highest (−0 94) (table 3) Changes in parameter variance for the cross-price elasticities were relatively small, and there were no sign reversals Three commodities appear as gross substitutes of rice (corn, cassava, and beans), whereas two are complements (potatoes and plantains) and two are independent (wheat and beef)

The Venezuela equation exhibits a remarkable reduction in parameter variability for all the variables This is important, as OLS results did not allow any inferences about the individual elasticities (except for the price of plantains) because of insignificant t-values Where PC-OLS is used, seven of the nine elasticities are significant at the 90-percent level (as opposed to one of nine with OLS estimation) Corn, plantains, and beef are gross substitutes for rice, while cassava and wheat are gross complements, and potatoes and beans are independent (23)

[3] The critical value for the Colombia and the Venezuela equations at the 95-percent level are 8 84 and 5 99, respectively (27) In both cases, the value of a* is lower than the critical value, which indicates superiority of the restricted, over the unrestricted, estimators in the GMSE sense

There are some similarities in both PC-OLS equations The own-price elasticity and the cross-price elasticity of corn are quite similar, and rice is inelastic with respect to income in both countries However, all the other commodities differ in their relationships with rice in the two countries This could be due to the fact that the importance of carbohydrate foods in the national diets of Colombia and Venezuela is quite different (*23*)

It is not surprising to find some degree of complementarity among all these staples in the two countries All are important in their respective diets, it is common to serve two, three, or even four staples at a single meal

In the predictive sense, the total MSE of prediction for the Venezuela PC-OLS equation was lower than that of the corresponding OLS equation, that was not the case for Colombia In other words, both PC-OLS equations perform better when a test of structural form is supplied, but only the Venezuela PC-OLS equation does better when the predictive power is tested

This procedure ensured that multicollinearity was reduced and that the cost (bias) was small and was outweighed by the gains (precision) in the structural sense as indicated by the GMSE criterion An objective of the estimation process was to quantify the impact of each independent variable on the demand for rice Therefore, restricting the amount of variability deleted was desirable to increase our confidence in the estimated elasticities Thus, other PC-OLS equations may be superior in the GMSE sense to the ones selected here, but these equations would have considerably larger biases Of course, given the complexity of the trade-offs, the decision to select the best deletion criterion in a risk situation is highly subjective

Mittlehammer and Young conclude that "the researcher plagued by severe multicollinearity is unlikely to find comfort by mechanically appealing to a single estimator whose principal virtue is dominance over OLS in some sense" (*20*, p 304)

With most of the variability of the data included, the method used here results in more precise estimators which are structurally superior in the GMSE sense to the unbiased OLS estimates. This finding does not imply that the CRC criterion should be used mechanically to delete components Each problem is unique, and choosing a solution depends on the researcher's goals and preferences as well as on careful analysis of the data and knowledge of the alternatives

References

(1) Allen, David M "Mean Square Error of Prediction as a Criterion for Selecting Variables," *Technometrics*, Vol 13, No 3, 1971, pp 469-73.

(2) Brown, W G *Effect of Omitting Relevant Variables Versus Use of Ridge Regression in Economic Research* Special Report 394 Oregon State University, Oct 1973

(3) Coxe, K "Do Principal Components Solve Multicollinearity? The Longley Data Revisited " Paper presented at joint annual meeting of Biometric Society, American Statistical Association, and Institute of Mathematical Statistics, Atlanta, Ga , 1975

(4) Food and Agriculture Organization of the United Nations *Agricultural Commodity Projections, 1970-1980* Rome, 1971

(5) Fabrycy, M Z "Multicollinearity Caused by Specification Errors," *Applied Statistics*, Vol 24, No 1, 1975, pp 250-54

(6) Farebrother, R W "Principal Components Estimators and Minimum Mean Square Error Criteria in Regression Analysis," *The Review of Economics and Statistics*, Vol 54, 1972, pp 332-36.

(7) Fomby, T B , and R C Hill "Deletion Criteria for Principal Components Analysis," *American Journal of Agricultural Economics*, Vol 60, No 3, 1978, pp 524-27

(8) Greenberg, E "Minimum Variance Properties of Principal Components Regression," *Journal of the American Statistical Association*, Vol 70, 1975 pp 194-97

(9) Harvey, A C "Some Comments on Multicollinearity in Regression," *Applied Statistics*, Vol 26, No 2, 1977, pp 188 90

(10) Hill, R C , T B Fomby, and S R Johnson "Compo nent Selection Norms for Principal Components Regression," *Communications in Statistics—Theory and Methods*, Vol A6, No 4, 1977, pp 309-34

(11) Johnson, S R , S C Reimer, and T P Rothrock "Principal Components and the Problem of Multicollinearity," *Metroeconomica*, Vol 25, 1973, pp 306-17

(12) Kim, J , and C W Mueller *Introduction to Factor Analysis What It Is and How To Do It* Paper No 13 Beverly Hills, Calif Sage Publications, 1978

(13) Kmenta, J *Elements of Econometrics* New York Macmillan Publishing Co , 1971

(14) Lazo, J "Situacion Mundial de la Productividad en Maíz, Arroz, Papa, Caña de Azúcar y Leche," *Documento Protaal No 52*, IICA, San José, Costa Rica, Mar 1980

(15) Lott, W F "The Optimal Set of Principal Component Restrictions on a Least-Squares Regression," *Communications in Statistics*, Vol 2, No 5, 1973, pp 449-64

(16) Mason, R L , R F Gunst, and J T Webster. "Regression Analysis and Problems of Multicollinearity," *Communications in Statistics*, Vol 4, No 3, 1975, pp 277-92

(17) Massy, W F "Principal Components Regression in Exploratory Statistical Research," *Journal of the*

American Statistics Association, Vol 60, No 309, 1965, pp 234-56

(18) McCallum, B T "Artificial Orthogonalization in Regression Analysis," *Review of Economics and Statistics*, Vol 52, No 1, 1970, pp 110-13

(19) Mittelhammer, R , and J Bantelle "On Two Strategies for Choosing Principal Components in Regression Analysis," *American Journal of Agricultural Economics*, Vol 59, No 2, 1977, pp 336-43

(20) Mittelhammer, R , and D L Young "Mitigating the Effects of Multicollinearity Using Exact and Stochastic Restrictions The Case of Aggregate Agricultural Production Function in Thailand Reply," *American Journal of Agricultural Economics*, Vol 63, No 2, 1981, pp 301-05

(21) Montes, G , R Candelo, and A Muñoz "La Economía del Arroz en Colombia," *Revista de Planeación y Desarrollo*, Bogotá, Vol 12, No 1, 1980, pp 73-131

(22) Ruiz-Lara, J , and L Schlesinger *Marketing Rice in Colombia* Report FG-CO-104 Bogotá, CEDE, 1965

(23) Sanint, Luis R "Analysis of the Composition of Demand for Major Carbohydrate Food Commodities in Colombia and Venezuela, 1956-77 " U S Dept Agr , Econ Res Serv , forthcoming

(24) Silvey, S D "Multicollinearity and Imprecise Estimation," *Journal of the Royal Statistics Society*, Vol B31, 1969, pp 539 52

(25) Toro-Vizcarrondo, C E , and T D Wallace "A Test of the Mean Square Error Criterion for Restrictions in Linear Regression," *Journal of the American Statistics Association*, Vol 63, 1968, pp 558-72

(26) Wallace, T D "Pretest Estimation in Regression A Survey," *American Journal of Agricultural Economics*, Vol. 59, 1977, pp 431-43

(27) _____, and C E Toro-Vizcarrondo "Tables for the Mean Square Error Test for Exact Linear Restrictions in Regression," *Journal of the American Statistics Association*, Vol 64, 1969, pp 1649-63

(28) William, A R , and D G Watts "Meaningful Multicollinearity Measures," *Technometrics*, Vol 20, No 4, 1978, pp 407-11

(29) Yancey, T A , G G Judge, and M E Bock "A Mean Square Error Test When Stochastic Restrictions Are Used In Regression," *Communications in Statistics*, Vol 3, 1974, pp 755-68

In Earlier Issues

The opposition to big business practices is frequently not upon the grounds that they are anti-competitive—rather, that the competition is too intense, too aggressive, too ruthless Economists are revising their ideas about the nature of competition Our agricultural marketing research has not given enough attention to the problem What kind of competition do we have in meat packing? In the tobacco industry? In the grocery chain systems? Is this kind of competition good for the farmers and the consumers? If not, what can and should be done about it? What are the extent and kinds of Government regulation that are needed?

Frederick V Waugh
Vol 7, No 2, April 1955, p 54