The Use of Principal Components in Simultaneous Equations: An Empirical Application

by

Eugene Jones

Presented as a Selected Paper at
the American Agricultural Economics
Association Meetings
Iowa State University
Ames, Iowa
August 4-7, 1985

Eugene Jones is an Assistant Professor of agricultural economics at The Ohio State University, Columbus, Ohio.

Abstract

The transformation of jointly dependent and predetermined variables into principal components is shown to be an effective method for handling multicollinearity in simultaneous equations at the second stage. Estimates are obtained which are theoretically meaningful and statistically significant. Simulations from the reduced forms are realistic relative to historical values.

The Use of Principal Components in Simultaneous Equations: An Empirical Application

Introduction

Principal components (PC) are used in regression analysis for two reasons: (1) to transform a set of correlated predetermined variables (X) into orthogonal variables or vectors (P); and (2) to reduce an excess number of predetermined variables relative to the number of sample observations (N). Multicollinearity is addressed by the first method; insufficient degrees of freedom, by the second. Failure to address either problem leads to singularity or near singularity of the matrix of sums of squares and products of the predetermined variables (X'X). The use of PC provides a solution to this matrix singularity problem since PC are orthogonal variables derived as linear combinations of an original set of predetermined variables.

Several authors have illustrated the use of PC in regression analysis as a solution to multicollinearity (Pidot, 1969; Maddala, 1977; Chatterjee and Price, 1977; Mittelhammer and Baritelle, 1977). Researchers employing PC in regression analysis as a solution to insufficient degrees of freedom include (Kloek and Mennes, 1960; Amemiya, 1966; Klein, 1969). As solutions to both problems, PC have been limited to predetermined variables. This paper will illustrate the use of PC on predetermined variables as well as sets of predetermined and dependent variables (Y). Multicollinearity is the problem addressed with both uses of PC.

Specifically, a simultaneous system of the U.S. potato industry is illustrated with PC being used on the predetermined variables at

the first stage. Then, PC on predetermined and dependent variables are derived for individual equations and used at the second and third stages of estimation. Parameter estimates and the variances for the original variables are then derived. It is shown that this extension of PC to jointly dependent variables does not affect the variance properties of the estimators. Some empirical and statistical results from the estimated model are shown. Further, simulations from the reduced form of the model are shown to be realistic relative to historical values.

Model Specification

The sixteen equation system shown in table 1 was developed and subsequently estimated. The structure of the potato industry is captured by the first six equations; market equilibrium or product flow characteristics are captured by the remaining equations. The model is based on annual data for 1960-81. All endogenous variables except total potato production (QRP), shown in equation 13, are simultaneously determined.

Six areas are depicted in table 1 to be referred across the table as I, II, and III for the first six equations and IV, V and III for the remaining equations. The parameters in I account for causality among structural variables while those in II suggest how product flow variables should impact the structure in the same time period. In contrast, the parameters in IV account for the structural impact on product flow while the relationship among the flow variables is shown in V. Finally, all exogenous variables are in area III. Note that

Table 1 -- System of Equations for Analyzing the U.S. Potato Industry

Eq.	Endogenous Variables																Exogenous Variables																							
	NPP	CRP	PCM	ASR	CCA	UCA	QRP	U2F	U2P	RPF	RPP	WPP	FPR	RPT	SPR	U2O	C	MS	PCML	CT	GD	CK	PS	FPRL	MC	TR	RS	GC	TF	FF	PN	IN	WN	GR	U2PL	QRPL	ASRL
1	-1		Γ_{13}			Γ_{16}											Π_{10}	Π_{11}	Π_{12}	Π_{13}	Π_{14}															$\Pi_{1,17}$	$\Pi_{1,18}$		
2	Γ_{21}	-1		Γ_{24}													Π_{20}	Π_{21}	Π_{22}	Π_{23}		Π_{25}																	$\Pi_{2,20}$
3		Γ_{32}	-1			Γ_{36}											Π_{30}		Π_{32}																				
4		Γ_{42}		-1				Γ_{48}	Γ_{49}								Π_{40}		Π_{42}		Π_{44}															$\Pi_{4,17}$			
5	Γ_{51}				-1	Γ_{56}											Π_{50}		Π_{52}			Π_{55}																
6	Γ_{61}				Γ_{65}	-1			Γ_{69}								Π_{60}		Π_{62}																				
7							-1										Π_{70}						Π_{76}	Π_{77}		Π_{79}	$\Pi_{7,10}$										$\Pi_{7,19}$	
8								-1		$\Gamma_{8,10}$	$\Gamma_{8,11}$			$\Gamma_{8,14}$			Π_{80}											$\Pi_{8,11}$	$\Pi_{8,12}$	$\Pi_{8,13}$	$\Pi_{8,14}$	$\Pi_{8,15}$	$\Pi_{8,16}$				
9				$\Gamma_{9,4}$					-1	$\Gamma_{9,10}$	$\Gamma_{9,11}$			$\Gamma_{9,14}$			Π_{90}											$\Pi_{9,11}$	$\Pi_{9,12}$	$\Pi_{9,13}$	$\Pi_{9,14}$	$\Pi_{9,15}$	$\Pi_{9,16}$				$\Pi_{9,20}$
10										-1			1		1																								
11											-1	$\Gamma_{11,12}$					$\Pi_{11,0}$									$\Pi_{11,9}$												
12			1									-1	1																									
13							$\Gamma_{13,7}$						-1				$\Pi_{13,0}$																				$\Pi_{13,19}$	
14								.01	.01					-1																								
15															-1		$\Pi_{15,0}$							$\Pi_{15,8}$														
16						-1		1	1							1																							

Endogenous Variables

NPP Number of processing plants
CRP Concentration in potato processing
PCM Price cost margins
ASR Advertising-to-sales ratio
CCA Change in processing capacity
UCA Utilization of processing capacity
QRP Production of fresh potatoes
U2F Utilization of potatoes for fresh consumption
U2P Utilization of potatoes for processing
RPF Retail price of fresh potatoes
RPP Retail price of processed potatoes
WPP Wholesale price of processed potatoes
FPR Farm price of fresh potatoes
RPT Research and promotion tax
SPR Spread between retail and farm price
U2O Utilization of other potatoes

Exogenous Variables

MS Minimum efficient plant size
PCML Price cost margins lagged
CT Transportation cost
GD Geographic dispersion of potato production
CK Cost of capital
PS Price of sugar beets lagged
FPRL Farm price of fresh potatoes lagged
MC Marketing cost
TR Trend variable
RS Risk
GC Produce sales through retail grocery stores
TF Total away-from-home restaurants food sales
FF Fast-food sales as a percent of total food
PN Population
IN Income
WN Women in labor force
GR Expected growth
U2PL Utilization of potatoes for processing lagged
QRPL Production of fresh potatoes lagged
ASRL Advertising-to-sales ratio lagged

within III there are five lagged endogenous variables.

Principal Components Use at the First and Second Stages

The system outlined in table 1 can be expressed as

(1.1) $\Gamma Y + \Pi X + U = 0.$

Referring to table 1, it can be readily seen that this system satisfies the order condition, as a common characteristic among the equations is their overidentification. The system also has been verified to satisfy the rank condition for identification.

The model is based on annual data and several variables tend to be highly correlated. Such correlations create potential problems at the first and second stages of estimation. At the first stage, $Y = \theta X$, but if vectors of X are correlated, then the value of θ is questionable. An alternative would be to define an orthogonal set of vectors P such that $P = WX$, where W is a matrix of eigenvectors estimated using principal component analyses (Chatterjee and Price; Kloek and Mennes; Amemiya). As each principal component captures the maximum variance among the exogenous variables, the parameters θ generally can be estimated using some subset of the vectors of P. For example, given five exogenous variables, the first and second principal components may suffice. That is, $Y = \alpha P_1$, where P_1 is a subset matrix of P.

When a system is somewhat sparse as in table 1, similar correlation problems can arise at the second stage among the variables X and Y entering particular equations. The logic of statistical theory suggests that principal components could be equally applied

at this stage. Simultaneity in the system is maintained by defining as separate variables those parts of the components resulting from the stochastic and nonstochastic explanatory variables. Greater precision in the parameter estimates is gained by using a subset of PC at the second stage as well as at the first stage.

The use of PC at the second stage requires the orthogonal vector P_2, where $P_2 = ZW_2$, $Z = (\hat{Y},X)$, and W_2 is a new set of eigenvectors for the second stage estimation. If identities and third stage estimates are part of the system, the Y must be identified explicitly in each equation. This identity is lost if P_2 is used as in equation (1.3) below

(1.2) $Y_1 = Z\tau + U$ where $Z = (\hat{Y},X)$

(1.3) $Y_1 = P_2\delta + U$ where $\tau = W_2\delta$.

Simultaneity can be maintained by defining the variables associated with the weights applying to the endogenous variables (W_{21}) and those applying to the exogenous variables (W_{22}). Estimation at the second stage then would follow as

(1.4) $Y_1 = P_{21}\delta + P_{22}\delta + U$

where $P_{21} = \hat{Y}W_{21}$

and $P_{22} = XW_{22}$.

Estimation of the above equation requires an equality restriction on δ since the parameter for P_2 applies to P_{21} and P_{22}. The identities for P_{21} must be carried as additional restrictions on the system of equations.

Note that a subset of PC is proposed for regression analysis

because regression on all PC not only would yield the same parameters as the original variables, but also would lead to very imprecise estimates (Maddala; Chatterjee and Price). While there are several methods for selecting PC, the size of the present system limits the practical selection of PC according to the size of the characteristic roots (Kloek and Mennes). Regression analysis on a subset of PC which is selected according to the size of characteristic roots leads, however, to biased parameter estimates. This biasness arises from the fact that PC are functions of the correlation matrix for the explanatory variables without regard to the dependent variables (Maddala; Mittelhammer and Baritelle). The estimated parameters, however, are functions of the dependent variables as well as the principal components. Consequently, a PC which is deleted because of its small characteristic root may have a high degree of correlation with the relevant dependent variable (Maddala).

Although biased, parameter estimates from a subset of PC with large characteristic roots have smaller variances and greater precision. This follows from the fact that the variance for any PC parameter is inversely proportional to its characteristic root. Let δ_i be the parameter estimates for the PC in equation 1 of table 1. Then, the variances for these parameters are s^2/λ_i, where s^2 is the estimated variance of the residual term, and λ_i are characteristic roots. A small λ_i would lead therefore to a very imprecise estimate for its associated parameter. This imprecise estimate would carry over to parameters for the original variables since, as shown in

equation 1.3, parameter estimates for original variables are derived from PC estimates.

Multicollinearity among a set of variables always leads to large and small characteristic roots. Dropping the PC associated with the small characteristic roots results in a gain of precision because of the relationship between the variances of the parameters for PC and the original variables. With few exceptions, there is a direct relationship between these variances. By dropping the PC associated with a small characteristic root, a sizeable reduction is realized in the estimated variance for all parameters. With this background, we now turn to the use of PC at the third-stage and the derivation of third-stage parameters and the associated variances for these parameters.

Third Stage Estimation

Collinearity among the endogenous and predetermined variables at the second stage requires, as noted, separation of the PC into their endogenous and exogenous parts. Estimation of these equations further requires equality restrictions on the parameters for the endogenous and exogenous parts of these PC. In essence, a restricted system is estimated at the second and third stages. Derivation of the parameters and variances for a restricted system with PC follows similar procedures as outlined for a restricted system without PC. Consequently, the derivation shown here follows the restricted three-stage least squares (3SLS) method as derived by Schmidt (1976, p. 243).

Note that restrictions on a system of equations can be represented as $r = RB$, where r is a Gx1 known vector, R is a GxK known

matrix of restrictions on the elements of B, and B is (Γ,Π). That is, equation (1.1) can be rewritten as

(1.5) $Y = ZB + U$ where $Z = (Y,X)$.

Schmidt (p. 243) has shown that the 3SLS restricted estimator is

(1.6) $\tilde{B} = \hat{B} + CR'(RCR')^{-1}(r-R\hat{B})$

where $\hat{B}$ is the unrestricted estimator as derived by Goldberger (p. 351). Mathematically, this unrestricted estimator is

(1.7) $\hat{B} = \{Z'X[\hat{\Sigma}^{-1}\otimes(X'X)^{-1}]X'Z\}^{-1} Z'X[\hat{\Sigma}^{-1}\otimes(X'X)^{-1}]X'Y.$

The C in equation (1.6) is the first part of equation (1.7) enclosed as $\{\ \}^{-1}$; R and r are as defined previously. Note that X and Z are diagonal matrices.

The above $\hat{B}$ is tantamount to the estimator for the PC parameters in the present system. To derive the estimator for the original variables which are embedded within the PC, the expression in (1.6) must be multiplied by the set of weights used in the transformation of the original variables to PC. As PC are standardized variables, this expression then must be divided by the vector of standard deviations corresponding to the weights or eigenvectors. The 3SLS estimator for the original variables is therefore

(1.8) $\tilde{B} = \{W[\hat{B} + CR'(RCR')^{-1}(r-R\hat{B})]\}V^{-1}.$

Note that no adjustment has been made for the intercept term, B_0. If the researcher is interested in getting an accurate estimate for B_0, then the means of the original variables must also be subtracted from the expression in 1.8. Since intercepts generally are of limited interest, this additional step has been excluded to avoid

cluttering the expression.

Third Stage Derivation of Variances

While Schmidt derived the estimator for a 3SLS restricted system, the author failed to show the variance properties for this estimator. Moreover, a search of the literature failed to reveal that any author had derived these variance properties. Therefore, the variance for this restricted estimator is derived below. Then it is shown how the variances for the original parameters can be derived from this restricted system with PC parameters.

The variance for $\tilde{B}$ is derived by using $\hat{B}$ as defined in (1.7) and Y as defined in (1.5). Substituting B and Y into the first part of (1.6) gives

$$(1.9)\quad \tilde{B} = \{Z'X[\hat{\Sigma}^{-1}\otimes(X'X)^{-1}]X'Z\}\ Z'X[\hat{\Sigma}^{-1}\otimes(X'X)^{-1}]\ X'(ZB + U) + CR'(RCR')^{-1}\ (r-R\hat{B}).$$

This expression reduces to

$$(1.10)\quad \tilde{B} = B + C\{Z'X[\hat{\Sigma}^{-1}\otimes(X'X)^{-1}]X'U + CR'(RCR')^{-1}(r-R\hat{B})$$

Substituting $\hat{B}$ in the last part of (1.10) yields

$$(1.11)\quad \tilde{B} = B + C\{Z'X[\hat{\Sigma}^{-1}\otimes(X'X)^{-1}]X'U + CR'(RCR')^{-1}\ r - RCZ'X[\hat{\Sigma}^{-1}\otimes(X'X)^{-1}]X'U$$

or simply

$$(1.12)\quad \tilde{B} = B + \{I - CR'(RCR')^{-1}R\}\ CZ'X[\hat{\Sigma}^{-1}\otimes(X'X)]X'U.$$

Now let $A = I - CR'(RCR')^{-1}R$. Then the variance becomes

$$(1.13)\quad E(\tilde{B} - B)\ (\tilde{B} - B)' = E\{ACZ'X[\hat{\Sigma}^{-1}\otimes(X'X)^{-1}]X'UU'X\ [\hat{\Sigma}^{-1}\otimes(X'X)^{-1}]X'ZC'A'\}.$$

once they are specified. The V can be dropped because it is unnecessary to adjust for the scale factor in deriving t-values (Chatterjee and Price, p. 171). Equation (1.17) therefore becomes

$$(1.18)\quad \text{var } \tilde{B} = W\left[Z'X\{\hat{\Sigma}^{-1} \otimes (X'X)^{-1}\}X'Z\right]^{-1} W'.$$

For two equations, this expression can be written out as

$$(1.19)\quad \text{var } (\tilde{B}) = W \begin{bmatrix} Z_1'X(X'X)^{-1} X'Z_1\sigma_{11} & Z_1'X(X'X)^{-1} X'Z_2\sigma_{12} \\ Z_2'X(X'X)^{-1} X'Z_2\sigma_{12} & Z_2X(X'X)^{-1} X'Z_2\sigma_{22} \end{bmatrix} W'$$

Now clearly the diagonal elements are variances while the off-diagonal elements are covariances. The final step requires multiplication by the weights or eigenvectors.

Empirical and Simulation Results

Multicollinear variables are known to cause serious threats to the proper specification and the effective estimation of the structural relationships underlying regression analysis (Farrar and Glauber). Because of these problems, researchers often will alter the specification of econometric models to alleviate multicollinearity and therefore improve the resulting estimates. An alternative to respecification, particularly when the model is theoretically well-specified and all the variables are important to the researcher, is the application of statistical methods to handle multicollinearity. As seen below, the use of principal components proved to be an effective method for handling this problem.

Because of space limitations, results are provided only for two of the equations in table 1. These are equations 2 estimating concentration (CRP) within the potato industry, and equation 9 estimating the utilization of potatoes for processing (UZP). The parameters and their associated statistics are shown in table 2. Simulation results are shown in figures 1 and 2.

Table 2 -- Parameters and T-Ratios

Concentration (CRP)		Parameter	T-Ratio
Intercept	(Π_{20})	64.607	223.606
NPP	(Γ_{21})	-.007	-13.684
ASR	(Γ_{24})	272.668	5.973
MS	(Π_{21})	.059	15.619
CK	(Π_{25})	.146	5.384
CT	(Π_{23})	.069	4.607
PCML	(Π_{22})	-.393	-8.751
ASRL	($\Pi_{2,20}$)	269.984	7.079

Utilization of Potatoes for Processing (UZP)		Parameter	T-Ratio
Intercept	(Π_{90})	59.94	48.31
ASR	(Γ_{94})	1308.08	7.66
RPF	($\Gamma_{9,10}$)	-.20	-.26
RPP	($\Gamma_{9,11}$)	-2.35	-6.26
RPT	($\Gamma_{9,14}$)	-.81	-1.02
GC	($\Pi_{9,11}$)	18.58	2.87
TF	($\Pi_{9,12}$)	.0001	4.02
FF	($\Pi_{9,13}$)	13.97	7.50
PN	($\Pi_{9,14}$)	.09	8.95
TN	($\Pi_{9,15}$)	.0056	13.28
WN	($\Pi_{9,16}$)	26.97	3.68
ASRL	($\Pi_{9,20}$)	1159.69	8.01

These parameters were used to derive the reduced forms for the two equations and to forecast the results beyond the 1960-81 data period to 1990. The forecasts are based on changes in the exogenous variables at (1) their mean rate for the last five data periods, (2) 25 percent above mean, and (3) 25 percent below mean. Relative to historical values, these forecasts seem quite realistic.

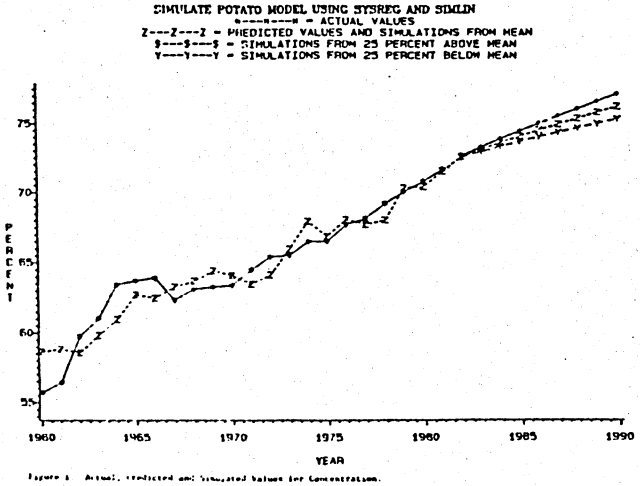

Figure 1. Actual, predicted and simulated values for Concentration.

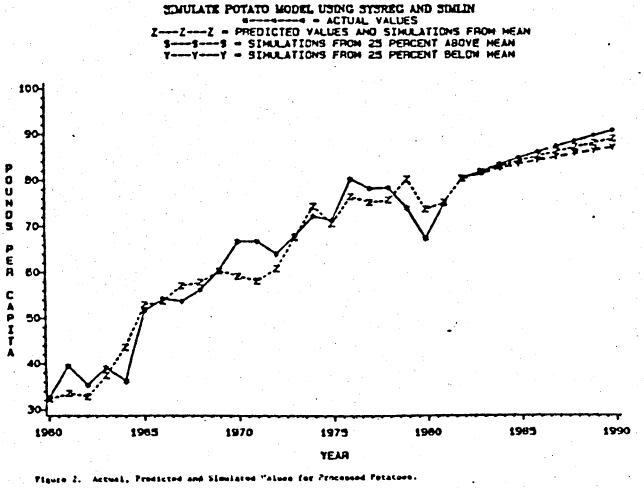

Figure 2. Actual, Predicted and Simulated Values for Processed Potatoes.

References

Amemiya, T. "On the Use of Principal Components of Independent Variables in 2SLS Estimation." International Economic Review. 7(1966): 283-303.

Chatterjee, S. and B. Price. Regression Analysis by Example. New York: John Wiley and Sons, Inc., 1977.

Farrar, D.E. and R.R. Glauber. "Multicollinearity in Regression Analysis: The Problem Revisited." Review of Economics and Statistics. 49(1967): 202-217.

Jones, E. "An Econometric Model of Structural Changes in the U.S. Potato Industry." Ph.D. dissertation, University of Florida, 1984.

Kendall, M.G. A Course in Multivariate Analysis. London: Hafner Publishing Company, 1957.

Klein, L.R. "Estimation Of Interdependent Systems in Macro-econometrics." Econometrica. 33(1969): 171-192.

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

Maddala, G.S. Econometrics. New York: McGraw-Hill, 1977.

Mittelhammer, R.C. and J.L. Baritelle. "On Two Strategies for Choosing Principal Components in Regression Analysis." American Journal of Agricultural Economics. 59(1977): 336-343.

Pidot, G.B. "A Principal Component Analysis of the Determinants of Local Government Fiscal Patterns." Review of Economics and Statistics. 51(1969): 176-188.

Schmidt, P. Econometrics. New York: Marcel Dekker, Inc., 1976.

Zellnor, A. and H. Theil. "Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations." Econometrica. 30(1962): 54-78.