Project 9141.103

ABARE conference
paper 91.4

35th Annual Conference of the
Australian Agricultural Economics Society
University of New England, Armidale, 11–14 February 1991

Using Partial Principal Component Analysis to Provide Better Tests of Models with Collinear Data

Bengt Hyberg and Nigel Hall

Australian Bureau of Agricultural
and Resource Economics

GPO Box 1563 Canberra 2601

Robert Abt

Associate Professor
Department of Forestry
University of Florida, 118 Newins-Ziegler Hall
Gainsville, Florida 32611

Many econometric analyses, including models of farm structure, land valuation, and private forest management, involve the use of collinear variables. In this paper a procedure termed partial principal component analysis, which decomposes only the variables involved in statistically degrading collinearity, is demonstrated. This procedure results in a well conditioned data set, preserves the identity of non-collinear variables, and increases the researcher's ability to interpret the principal components.

Analysis of gross farm revenues using partial principal component analysis indicates that the interrelationships among agricultural inputs are similar between regions, suggesting the existence of an innate structure which can be utilised in future studies.

Introduction

The design and estimation of econometric models require that the models accurately capture the salient characteristics necessary to account for variations in the dependent variable or variables. Difficulties frequently arise because the variables which embody these characteristics are often collinear. Collinearity is often seen as presenting the researcher with three unpalatable options: delete one or more of the collinear variables and use a model suffering from mis-specification bias; leave the collinear variables in the model and obtain unreliable coefficients; or use principal component analysis (PCA) and obtain results which are hard to interpret. This paper presents another option, partial PCA, which permits the development of properly specified models giving interpretable results.

Theory suggests that models of farm revenue or land value should include measures of farm size, cropped acreage, capital embodied in farm improvements, and cash expenditures. To conduct meaningful tests of hypotheses concerning influences on farm finance or valuation, statistical models are required which include variables capturing the attributes included in the theoretical models. Studies of farm finance and land valuation suggest that a number of the variables that have been used to describe landowners or their operations are collinear (Yoho 1958; Webster and Stoltenberg 1959; Muench 1964; Binkley 1980; Doll and Widdows 1982; Ervin and Ervin 1982; Boyd 1984; Hyberg 1986; Hyberg and Holthausen 1989). The interrelationships between these variables reduce the power of econometric tests of management behaviour.

Principal component analysis can be used to produce an orthogonal data set, and can be a valuable tool in the examination of collinear data. However, past abuses of the technique, combined with the difficulties associated with the interpretation of the components, have caused PCA to fall from favour. In this paper a modified version of principal component analysis, partial PCA, is introduced which transforms only the collinear variables. Using, for illustrative purposes, a model of gross revenues of Australian farms it is demonstrated that not only can partial PCA be used to identify the interrelationships between the collinear variables, but the interrelationships identified are stable between regions and over time, allowing the use of these interrelationships to impose structure on future models containing the collinear variables. Further, the interpretations of the untransformed variables are not altered, and the interpretation of the principal components is simplified.

The objective of this paper is to demonstrate that the use of partial principal component analysis (PPCA) can provide better models of farm financial performance and farm management behaviour. The difficulties associated with collinearity are discussed first, using previous studies of farm and land management as an illustration. Next, gross revenues from agricultural

operations in three distinct regions are modelled both with and without PPCA. The results of these analyses are compared and the gains available from PPCA are discussed. The implications of the study are then presented.

Data analysis with collinear variables

The effects of collinearity need to be understood if models are to be estimated correctly and appropriate statistical tests conducted. In this section the deleterious effect collinearity has on statistical estimates is discussed, a simple test is provided which researchers can employ to identify collinearity when it is suspected, and a strategy is introduced with which to address collinearity when it is detected.

Effect of collinearity on statistical estimates

Interrelationships among financial and socio-economic variables such as income, education, capital, labour, cash, and landholdings, which are often used to measure the wealth, operation size, or performance of a farm operation, reduce the power of econometric tests concerning their influence on farm management behaviour. Evidence of collinearity between these variables can be found in studies by, among others, Yoho (1958); Muench (1964); Binkley (1980); Doll and Widdows (1982); and Hyberg (1986).

Most previous investigations of landowner behaviour in farm finance, forest management, and land valuation studies have used the unmodified variables to describe the behaviour of landowners. These investigations have produced conflicting results. Some have indicated that income, landholdings or education are significant in the explanation of farm management behaviour, while other studies using different sets of independent variables have found different sets of significant variables. This inconsistency can be explained by noting that the financial and socio-economic variables used in these analyses are intrinsically collinear.

Successive analyses using variables that are collinear will result in:

- the collinear variables fluctuating between significance and non-significance as variables are entered into and dropped from the analysis; and

- the set of significant variables changing markedly as the data sets change.

These results are exactly what are observed when the empirical analyses of landowner behaviour are compared with one another. Thus, these studies have not resulted in true tests of the researchers' hypotheses.

Collinearity reduces the power of statistical tests for two reasons:

- It increases the variance associated with the estimated coefficients.

- The variables entered into the theoretical model have, in the theory that is being tested, specific properties that produce specific effects. In designing empirical tests a researcher attempts to isolate independent measurable variables having characteristics that correspond to those of the theoretical model. Unfortunately, one or more of the explanatory variables available to the researcher may have several characteristics, corresponding to a combination of several theoretical variables.

Identification of collinearity

Poorly conditioned, or collinear, data sets can be identified using the condition index statistic described by Belsley, Kuh and Welsch (1980). This statistic is available in statistical packages such as Statistical Analysis Systems (SAS) and TROLL. The values of the condition index statistics for a data set are obtained by performing a principal component decomposition of the data set and taking the ratio of the largest eigenvalue to each of the individual eigenvalues. The value of the condition index ranges from 1, which is associated with the linear combination of variables providing the most information, to infinity, for a perfectly collinear set of variables. As the value of the condition index rises the likelihood of problems with collinear data increases. Belsley, Kuh and Welsch (1980) suggest closer inspection of the data if a condition index greater than 10 occurs. A condition index greater than 30 indicates that collinearity is seriously degrading the regression.

The statistical packages also provide a decomposition of the data matrix into orthogonal vectors. This decomposition provides a means to determine which variables are contributing to the high condition index. Thus, not only can the existence of ill-conditioning be detected, but the sources of the ill-conditioning can be identified (Belsley, Kuh and Welsch 1980).

Dealing with collinearity

When statistically degrading linear dependencies are detected, the researcher must determine what is the most appropriate means of testing the hypothesis. Three methods that are frequently used are:

- dropping the collinear variables, one by one, until the matrix of independent variables is properly conditioned. This procedure will eliminate problems introduced by collinear variables, but will result in mis-specification bias if the variables removed from the regression are in fact important influences. In any case, when a hypothesis is being tested, the arbitrary elimination of a variable of interest is not an acceptable procedure.

- accepting the equation containing the collinear variables. This does not eliminate the problem; however, the diagnostic procedures developed by Belsley, Kuh and Welsch (1980) permit the identification of the weaknesses which the collinearity has introduced into the estimated equation. If the variables introducing the near-singularity into the estimation are theoretically important, this is a more appropriate procedure than dropping them. This method, while being more correct than that above, can reduce the power of tests of the hypotheses that are being conducted.

- using principal component analysis (PCA) to produce an orthogonal set of linear combinations of the variables. If an interpretation of these components can be made that corresponds to the properties of the variables which are of theoretic interest, the researcher will be able to test the hypotheses using coefficients that are not damaged by the effects of collinearity or misspecification.

The difficulty associated with PCA is the interpretation of the components. Typically, when collinearity has been addressed using principal component analysis, the decomposition has been performed on the full data set. Use of the components for hypothesis testing requires the interpretation of components consisting of linear combinations of many variables. Not only are the components difficult to interpret, but the meaning of theoretically interesting variables not involved in the collinearity is obscured by including them in the decomposition.

Given the above options, and the fact that PCA permits testing of the hypotheses while avoiding mis-specification bias, it becomes obvious that if a hypothesis requires the testing of several interrelated variables, then principal component analysis would be preferred provided that a means could be devised of preserving the identity of the non-collinear variables.

Collinear Data and Analysis of Farm Revenue

The primary goal of the following analysis is to estimate a model of farm output developed by Hall and Hyberg (1990). In this model, gross farm revenue is represented by a positive function of the size of the farm, the inputs used, and education of the farmer. It is further hypothesized that the presence of land degradation on the farm will have a negative effect on gross farm revenues. This model can be written:

(1) $\quad Q = F(S^{+}, X^{+}, ED^{+}, DG^{-})$

where Q is the sum of cash receipts and inventory changes, S is the size of the farm, X is a vector of inputs used in farm production, ED is the education of the farmer, and DG is a land degradation.

The data used to test the theoretical model were gathered in the 1983-84 Australian Agricultural and Grazing Industries Survey (described in BAE 1986, pp. 45–52). Because there is no variable in the data set which accurately captures the size of a farm, the amount of native pasture (N), improved pasture (I), and cropped land (C) on a property were used as measures of farm size. In addition, the inputs were disaggregated into three categories, the labour (L), capital (K) and cash expenditures (H) employed in agricultural production. Labour was measured in weeks. The value of capital was adjusted for depreciation. Cash expenditures were adjusted by subtracting interest payments and labour costs. Because information was unavailable on the actual number of years of education or the amount of land degradation, dummy variables were used for education (ED) and land degradation (DG). A dummy value of one for education indicates the farmer completed high school. Similarly, a value of one for degradation indicates that the farmer considered the property to have a problem or potential problem of land degradation, which was the case for 37 per cent of the farms surveyed (BAE 1986, p. 21). The model estimated took the following form:

$$(2) \qquad Q = G[N, I, C, L, K, H, ED, DG]$$

The model is a linear combination of the variables, all expressed in natural log form except the dummies. The dependent variable is measured in dollars.

Separate models were estimated for three agricultural regions, the pastoral, wheat–sheep and high rainfall zones. These zones were selected because it was felt that the structure of agriculture in Australia could be represented better using regional models than by imposing a single model on farming systems facing different economic and climatic constraints. Linear regressions were estimated using the Statistical Analysis Systems's REG procedure. The results of these estimations are shown in Table 1. It can be seen that, aside from the cropland and education variables in the high rainfall zone, each of the significant variables has the expected sign.

Because collinearity was expected to be a problem, SAS's COLLIN option was used. This option provides collinearity diagnostics, including the condition indexes and the proportion of the variance of the estimate of each variable coefficient accounted for by each principal component. These diagnostics can be used to assess the conditioning of the data set. The condition indexes and variance proportions for each model are provided in Tables 2–4.

TABLE 1

Preliminary Regression Results

Zone and variable	Parameter estimate	Standard error	t for H_0: Parameter = 0	
Pastoral zone				
Intercept	12.481	0.211	59.169	***
L	0.055	0.030	1.839	*
K	–0.034	0.027	–1.255	
H	0.149	0.028	5.335	***
C	0.003	0.008	0.389	
N	0.008	0.007	1.178	
I	0.006	0.007	0.806	
DG	–0.045	0.023	–1.920	*
ED	–0.034	0.028	–1.229	
Wheat-sheep zone				
Intercept	12.761	0.052	245.108	***
L	–0.002	0.008	–0.205	
K	0.033	0.008	4.265	***
H	0.074	0.008	9.093	***
C	0.000	0.002	0.057	
N	0.002	0.001	1.691	*
I	–0.000	0.001	–0.102	
DG	–0.009	0.006	–1.413	
ED	–0.004	0.008	–0.520	
High rainfall zone				
Intercept	13.125	0.045	289.938	***
L	0.176	0.006	2.807	***
K	9.911	0.006	1.979	**
H	0.055	0.005	9.947	***
C	–0.003	0.002	–1.780	*
N	0.004	0.001	3.274	***
I	0.004	0.002	2.175	**
DG	–0.013	0.006	–2.239	**
ED	–0.016	0.008	–1.921	*

*** significant at 99 per cent level. ** significant at 95 per cent level. * significant at 90 per cent level.

An examination of the condition indexes (Tables 2–4) reveals that each model contains at least three condition indexes greater than ten, and two models contain four condition indexes greater than ten. These values indicate the existence of unacceptable linear dependences among some of the variables. The interrelationships among the collinear variables are resulting in unreliable parameter estimates for these variables, and hence introduce uncertainty into our interpretation of the model, reducing its usefulness in understanding Australian farm structure.

TABLE 2

Collinearity Diagnostics for the Pastoral Zone Model

Principal component	Eigen-value	Condition indexes	Proportions of variance								
			Intercept	*L*	*K*	*H*	*C*	*N*	*I*	*DG*	*ED*
1	6.4414	1.000	0.000	0.000	0.000	0.000	0.002	0.000	0.003	0.006	0.004
2	1.2102	2.307	0.000	0.000	0.000	0.000	0.249	0.001	0.177	0.001	0.000
3	0.6196	3.224	0.000	0.000	0.000	0.000	0.053	0.000	0.015	0.863	0.004
4	0.4965	3.602	0.000	0.000	0.000	0.000	0.407	0.000	0.632	0.084	0.004
5	0.2045	5.612	0.000	0.001	0.000	0.000	0.000	0.003	0.037	0.007	0.786
6	0.0190	18.403	0.016	**0.006**	**0.003**	**0.003**	**0.193**	0.889	0.075	0.015	0.000
7	0.0068	30.773	0.122	**0.491**	**0.002**	**0.000**	**0.016**	0.051	0.004	0.014	0.147
8	0.0012	72.525	0.704	**0.477**	**0.066**	**0.490**	**0.068**	0.055	0.008	0.009	0.000
9	0.0007	94.280	0.157	**0.024**	**0.929**	**0.507**	**0.011**	0.000	0.049	0.000	0.054

TABLE 3

Collinearity Diagnostics for the Wheat–Sheep Zone Model

Principal component	Eigen-value	Condition indexes	Proportions of variance								
			Intercept	*L*	*K*	*H*	*C*	*N*	*I*	*DG*	*ED*
1	7.7541	1.000	0.000	0.000	0.000	0.000	0.001	0.003	0.003	0.004	0.002
2	0.5384	3.795	0.000	0.000	0.000	0.000	0.000	0.000	0.016	0.871	0.007
3	0.3343	4.816	0.000	0.000	0.000	0.000	0.002	0.365	0.256	0.028	0.002
4	0.1625	6.907	0.000	0.000	0.000	0.000	0.044	0.293	0.564	0.002	0.225
5	0.1317	7.674	0.000	0.001	0.000	0.000	0.212	0.033	0.030	0.045	0.608
6	0.0706	10.481	**0.007**	**0.007**	**0.000**	**0.001**	**0.465**	0.240	0.099	0.018	0.147
7	0.0061	35.679	**0.150**	**0.652**	**0.005**	**0.000**	**0.048**	0.022	0.000	0.000	0.009
8	0.0016	69.637	**0.795**	**0.313**	**0.128**	**0.221**	**0.225**	0.008	0.011	0.008	0.000
9	0.0006	109.265	**0.047**	**0.027**	**0.865**	**0.777**	**0.002**	0.037	0.021	0.023	0.000

TABLE 4

Collinearity Diagnostics for the High Rainfall Model

Principal component	Eigenvalue	Condition indexes	Proportions of variance								
			Intercept	*L*	*K*	*H*	*C*	*N*	*I*	*DG*	*ED*
1	7.1841	1.000	0.000	0.000	0.000	0.000	0.004	0.004	0.001	0.005	0.002
2	0.6651	3.286	0.000	0.000	0.000	0.000	0.264	0.039	0.003	0.489	0.001
3	0.5715	3.545	0.000	0.000	0.000	0.000	0.258	0.090	0.001	0.488	0.001
4	0.3680	4.418	0.000	0.000	0.000	0.000	0.302	0.460	0.023	0.000	0.015
5	0.1226	7.654	0.000	0.003	0.000	0.000	0.006	0.000	0.057	0.002	0.866
6	0.0770	9.655	0.005	**0.007**	**0.001**	**0.001**	**0.064**	0.332	0.642	0.006	0.019
7	0.0079	30.078	0.091	**0.872**	**0.011**	**0.008**	**0.017**	0.039	0.020	0.000	0.077
8	0.0023	55.224	0.633	**0.107**	**0.022**	**0.537**	**0.068**	0.035	0.246	0.010	0.010
9	0.0012	76.506	0.270	**0.011**	**0.964**	**0.453**	**0.016**	0.000	0.007	0.000	0.009

The collinear variables can be determined by examining the variance proportions. They are the variables with a sizeable proportion of their variation attributable to the principal components having large condition indices (bold numbers in Tables 2–4). An examination of the variance proportions reveals that three variables, capital, labour and cash expenditures, are causing near-singularities in all three regional models; and cropland acreage is also contributing to collinearity in the model for the wheat-sheep zone. These four variables were selected for the partial principal components analysis for all three zones. The use of these four variables in the PPCA was subsequently justified by the fact that collinearity diagnostics for regressions using the components indicated no significant problems with collinearity.

Using partial PCA to treat collinear data sets

The use of the orthogonal rather than collinear variables increases the ability of the model to detect variables that are statistically significant. The collinear nature of capital, labour, cash expenditures and cropland, combined with the desire to understand the effect of these variables on farm revenues, leads to the use of partial principal component analysis. The next step is to determine the variables to be included in the PPCA decomposition.

Because the principal components are linear combinations of the original (log) variables, the explanatory ability of a model is not altered when the full data set or a subset of the explanatory variables are transformed using PCA. Thus, one can either decompose the entire data set or a subset of the variables without losing information. Referring again to Tables 2–4, it can be seen that although the first five components do not exhibit signs of severe collinearity, they are not totally orthogonal. One might think that some empirical gain might result from conducting a principal component decomposition on the entire data set. This has been the standard procedure in the past. However, by performing PCA on only the variables that are involved in severely degrading singularities, the variables not involved are allowed to retain their original identities. In addition, because principal components are difficult to interpret, there is a greater likelihood of obtaining a meaningful interpretation of the components if the number of variables included in the principal component analysis is limited.

There is one final consideration: principal component analysis is a form of pre-test estimation, and therefore results in statistical tests of unknown power. This difficulty can be avoided through the use of two data sets: one to perform the principal component analysis that determines the linear transformations to be used in the hypothesis testing, and another to be used to conduct the tests of the hypothesis. The linear combinations obtained from the first data set are treated as a theoretical restriction, and are imposed on the second data set.

These considerations led to the following strategy:

1. The capital, labour, cash expenditure and cropland variables, contributing significant portions of their variation to eigenvalues with an index number greater than 10, were included in the principal component decomposition. The principal component decomposition was conducted on the data set from the wheat-sheep zone, using the FACTOR procedure of SAS.

2. Because the data set was already divided into the three zones, rather than further dividing the data set it was decided to use the eigenvectors from the wheat-sheep zone to impose a structure on the relationship between the capital, labour, cash expenditures and cropland variables in the pastoral and high rainfall regions. This served to avoid introducing pre-test estimation bias for the pastoral and high rainfall zones, although pre-test bias will still affect the estimates for the wheat-sheep zone. This was done using SAS's Score procedure. An examination of the eigenvectors and factor patterns for the pastoral and high rainfall zones (Tables 5 and 6) shows that decompositions for these regions reveal similar relationships to those obtained from the wheat-sheep region, indicating that the structure imposed was appropriate.

3. The eigenvectors were examined and preliminary interpretations placed upon each principal component. These interpretations are listed in Table 7.

4. The regional models were re-estimated using the transformed variables. The parameters from these models are given in Table 8.

A Comparison of the Partial Principal Component and Untransformed Models

Although the predictive ability of a model is always of interest, the objective of this study is to isolate a set of transformed variables that describe the capital, labour, cash expenditure and cropland attributes which affect gross farm revenue in each of three regions. Because the principal components are linear combinations of the original variables, both models will have the same ability to explain gross farm revenues. For this reason, the measure of the usefulness of principal component analysis in this case will be the difference in the ability of the models to capture these attributes.

Because of the mathematical properties of PCA there is no change in the coefficients and standard errors associated with the untransformed variables. Thus, the transformation does not affect the statistical properties of these variables. An examination of the collinearity diagnostics indicates no condition index number greater than eleven, a marked improvement.

TABLE 5

Eigenvectors from Principal Component Decomposition

Zone and variable	Principal component			
	1	2	3	4(a)
Pastoral zone				
L	0.562	–0.169	0.752	0.299
K	0.573	–0.029	–0.639	0.512
H	0.582	–0.030	–0.121	–0.804
C	0.131	0.985	0.106	0.042
Wheat–sheep zone				
L	0.479	–0.543	0.672	0.153
K	0.537	–0.049	–0.565	0.625
H	0.547	–0.124	–0.316	–0.765
C	0.428	0.829	0.360	0.023
High rainfall zone				
L	0.484	–0 430	0.762	–0.006
K	0.558	–0.106	–0.419	–0.708
H	0.557	–0.125	–0.419	0.706
C	0.379	0.888	0.260	0.012

(a) The reversal of the signs in the eigenvectors for the high rainfall zone is due to the characteristics of the mathematical operations used in the principal component analysis.

TABLE 6

Factor Patterns

Zone and variable	Factor			
	1	2	3	4
Pastoral zone				
L	0.900	–0.168	0.379	0.132
K	0.919	–0.029	–0.321	0.226
H	0.933	–0.030	–0.061	–0.354
C	0.210	0.976	0.054	0.019
Wheat–sheep zone				
L	0.821	–0.427	0.376	0.057
K	0.919	–0.038	–0.316	0.234
H	0.937	–0.097	–0.177	–0.286
C	0.732	0.651	0.201	0.008
High rainfall zone				
L	0.782	–0.374	0.498	–0.003
K	0.903	–0.092	–0.274	–0.318
H	0.902	–0.108	–0.274	0.317
C	0.613	0.771	0.170	0.005

TABLE 7

Interpretations Adopted for the Principal Components

Component 1	Captures the general magnitude of a farm operation
Component 2	Primarily a measure of the importance of cropping
Component 3	Primarily a measure of the labour/capital ratio
Component 4	Primarily a measure of the ratio of fixed capital to variable costs(a)

(a) See note to Table 5.

TABLE 8

Principal Component Regression: Parameter Estimates

Zone and variable	Parameter estimate	Standard error	t for H_0: Parameter = 0	
Pastoral zone				
Intercept	13.939	0.075	186.258	***
Component 1	0.387	0.042	9.191	***
Component 2	-0.017	0.076	-0.224	
Component 3	0.158	0.122	1.288	
Component 4	-0.504	0.137	-3.689	***
N	0.008	0.007	1.178	
I	0.006	0.007	0.806	
DG	-0.045	0.023	-1.920	*
ED	-0.034	0.028	-1.229	
Wheat–sheep zone				
Intercept	13.933	0.012	1172.564	***
Component 1	0.553	0.022	24.929	***
Component 2	-0.095	0.047	-02.005	**
Component 3	-0.409	0.064	-6.349	***
Component 4	-0.344	0.098	-3.498	***
N	0.002	0.001	1.691	*
I	-0.000	0.001	-0.102	
DG	-0.009	0.006	-1.413	
ED	-0.004	0.008	-0.520	
High rainfall zone				
Intercept	13.873	0.012	1116.517	***
Component 1	0.406	0.025	16.0.93	***
Component 2	-0.186	0.036	-5.121	***
Component 3	-0.209	0.049	-4.287	***
Component 4	0.320	0.069	4.657	***
N	0.004	0.001	3.274	***
I	0.004	0.002	2.175	**
DG	-0.013	0.006	-2.239	**
ED	-0.016	0.008	-1.921	*

*** significant at 99 per cent level. ** significant at 95 per cent level. * significant at 90 per cent level.

Pastoral zone

In both models (see Tables 1 and 8), for the pastoral zone the intercept is positive and highly significant, while degradation has a significant (90 per cent) negative effect on gross farm revenue. The difference between the two is that the model using the transformed variables has two components entering at the 99 per cent level, while in the original model only cash expenditures are highly significant. In the model using principal components, the first component is positive, indicating that increasing the amount of inputs has a positive effect on gross farm revenues. The fourth component is negative, suggesting that capital has a negative effect on gross revenues, while cash expenditures have a positive influence. Thus, in modeling gross farm revenues in the pastoral zone, the use of principal component analysis isolates the effects of the inputs, while preserving the interpretation of the degradation variable.

Wheat–sheep zone

In both models the intercept is significant at the 99 per cent level, while the amount of native pasture is positive and significant (90 per cent). The degrading effects of collinearity can be seen from the fact that, while cash expenditures and capital are positive and highly significant in the untransformed model, all four components are significant in the transformed model and three of them are highly significant. It appears that the collinearity between the input variables in the original model resulted in statistical degradation which masked the contribution of variables other than cash expenditures and farm size.

Except for the first component, the signs on the transformed variables are negative. This suggests that the magnitudes of both total inputs and cash expenditures have a positive effect on gross revenues, while the effects of labour and cropland on gross farm revenue are negative once their contribution to the first component has been taken into account.

By multiplying the coefficient associated with each component by the appropriate element of each eigenvector, one can make a closer examination of the effect of a unit increase in the original variable. When this is done for farm capital it is seen that, after the effects of the first component are taken into account, capital has a marginally positive effect on gross farm revenues in the wheat–sheep zone (Table 9).

High rainfall zone

All variables enter significantly in both the original and transformed equations. Here again the intercept is positive and highly significant. In both models, native and improved pasture are positive variables (at the 99 per cent and 95 per cent levels respectively), while degradation has a negative effect on gross farm revenue which is significant at the 95 per cent level. The variable for education is significant at the 90 per cent level. (In the other two zones it has the same negative sign but is not significant.) Education is known to be correlated negatively with

TABLE 9

Net Effects of Financial Inputs on Gross Farm Revenues(a)

Zone	Capital	Labour	Cash	Cropland
Pastoral	–		+	
Wheat–sheep	+	–	+	–
High rainfall	+	+	–	–

(a) Estimated from the coefficients of the principal components in Table 8.

age and positively with off-farm work. It is suggested that younger operators working off farm are more likely to be in a property development phase of their life cycle and so to be producing less output per unit of inputs than more established operators specialising in farming.

In the untransformed model, labour and cash expenditures enter as positive, highly significant variables, while capital is positive and significant (95 per cent) and cropland is negative and significant (90 per cent). In the model using principal components, all four components are highly significant, with components one and four positive, and components two and three negative. After taking into account the effects of the first component, the effect of capital and labour are positive, although that of labour is small. The effects of cash expenditures and cropland are negative. Thus, the use of principal component analysis identifies a size effect associated with input use and isolates the negative effect of cash expenditures on gross revenues.

Differences in regional farm structure

The use of partial PCA permits a clearer understanding of the farm structure of the three regions. While the intercept and first component are positive and highly significant in all regions, the signs and significances of the other variables change from region to region.

The variables measuring size, the first component, native pasture, and improved pasture generally behave as expected. The first component is positive and highly significant in each of the three models, while the amount of pasture has a generally positive effect. Improved pasture in the wheat–sheep zone is the only variable measuring size that does not have a positive coefficient, and even in that case the coefficient is zero rather than negative. Native pasture is highly significant in the high rainfall zone and significant (90 per cent) in the wheat–sheep zone, while improved pasture is significant in the high rainfall zone. The amount of improved pasture land provides little information concerning gross farm revenues in the pastoral zone, which is not surprising given that most land in the pastoral zone is native pasture and the first component provides information on the size of an operation. On the other hand the amounts of

pasture land provide considerable information concerning gross farm revenues in the high rainfall zone.

The land degradation variable also behaves as expected. It has an apparent negative effect on gross farm revenues, significant in the pastoral (90 per cent) and high rainfall zones (95 per cent). The larger coefficient on land degradation in the pastoral zone suggests that land there is more fragile than land in other areas. However, the effects of land degradation on gross farm revenues in the pastoral zone are more variable. These results are indicated by a land degradation coefficient more than three times greater, and a standard error nearly four times greater than for the other zones.

The directions of the effects of the transformed variables are presented in Table 9. These results were derived using only the significant components. It is difficult to find a distinct common pattern in these results, again highlighting the differences in farm structure between zones.

Cropland does not enter the pastoral zone model, which is not surprising given the absence of cultivation. Cropland is a negative regressor of gross farm revenues in the wheat-sheep and high rainfall zones; this could be more a consequence of the profitability of wheat relative to sheep in 1983-84 than of any other factor.

Cash expenditures have an unambiguously positive effect for the pastoral and wheat-sheep zones, and an equally unambiguously negative effect for the high rainfall zone.

Capital has a strong negative effect in the pastoral zone, and weakly positive effects in the wheat-sheep and high rainfall zones. Labour has a weakly negative effect in the wheat-sheep zone and a weakly positive effect in the high rainfall zone.

The reason for the differences between the zones is not clear. A number of reasons including weather and market conditions can be hypothesized, but a comprehensive economic model which explains these results does not yet exist. However, a clearer understanding of the different effects of economic inputs for farms in different regions is now available to help in the construction of such a model.

Discussion

The results of both sets of regressions, taken together, indicate that principal component analysis can be used to isolate characteristics of significance when there are several collinear variables providing information. This facility can be used to gain a better understanding of innate economic relationships, provide more powerful tests of hypotheses involving the socio-economic characteristics of landowners, and impose structure on collinear data sets.

The usefulness of this technique is not dependent on the researcher's ability to place meaningful interpretations on the principal components, because the knowledge that innate relationships exist permits the utilisation of these relationships to forecast economic performance and behaviour, provided they are constant over time. However, the use of partial PCA increases the researchers ability to place a meaningful interpretation on a set of components, expanding the understanding of economic phenomena.

In the analysis conducted in this paper, models with transformed and untransformed variables had similar abilities to predict gross farm revenues. The primary differences between the two models were:

- The models using principal components permitted identification of a size component of the collinear variables which, when isolated, permitted the examination of additional effects of these variables. For the untransformed models it was not possible to separate the contribution of the financial variables via farm size from their direct contributions.

- The statistical significance of the transformed financial variables was greater than those in the untransformed model.

While the eigenvectors and factor patterns from the sets of data presented here are similar between zones, the study should be repeated by other researchers to ensure that the results obtained are not due to chance. If the results of such future studies are consistent with those found in this study, there would be good reason to accept the associations between the financial variables as innate relationships.

Conclusions

Past models of landowner behaviour have either used collinear data and obtained inferior statistical estimates, or have omitted some variables and been mis-specified. In this paper a method which can be used to resolve the problem of collinear data has been demonstrated.

This result is important for two reasons. First, researchers are interested in both interpreting previous results and finding the correct perspective to use in describing economic phenomena. To do this, variables measuring the attributes hypothesised to affect farm management behaviour must be used if the statistical tests are to be valid. Second, researchers are often asked to provide guidance to administrators and legislators on the means that will best achieve

policy goals. Both of these tasks are facilitated by the insights provided by the use of accurate, well specified models.

Farm management behaviour is a complex phenomenon. The variables which are available for the analysis of landowner farm management, such as capital and labour inputs, farm values, farm revenue and landholdings, are often highly collinear. The use of principal component analysis to modify the available social and economic variables allows a more accurate representation of the landowner characteristics influencing farm management decisions by eliminating the collinearity.

In conclusion, farm management models require more sophisticated statistical treatment than they have received in the past. Collinearity and mis-specification problems have resulted in models that have generated conflicting results. In order to model the farm management problem more correctly, researchers must develop accurate measures of the attributes that influence landowner behaviour. Principal component analysis provides the researchers with a tool to do just this.

References

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley and Sons, New York.

Binkley, C.S. (1981), *Timber Supply from Private Nonindustrial Forests*, School of Forestry and Environmental Studies Bulletin 92, Yale University, Connecticut.

Boyd, R. (1984), 'Government support of nonindustrial production: the case of private forests', *Southern Journal of Economics* **51**, 89–107.

BAE (1986), *Farm Surveys Report*, Bureau of Agricultural Economics, AGPS, Canberra.

Doll, J.P. and Widdows. R. (1982), *The Value of Agricultural Land in the United States: Some Thoughts and Conclusions*, ERS Staff Report No. AGES820323, Washington DC.

Ervin, C.A. and Ervin, D.E. (1982), 'Factors affecting the use of soil conservation practices: hypotheses, evidence, and policy implication', *Land Economics* **58**(3), 277–92.

Hall, N. and Hyberg, B.T. (1990), Effects of land degradation on farm output. Unpublished manuscript, ABARE.

Hyberg, B.T. (1986), The development and empirical testing of a utility maximization model for NIPF landowner forest management. PhD dissertation, North Carolina State University, Raleigh, NC.

Hyberg, B.T. and Holthausen, D. (1989), 'The behaviour of nonindustrial forest landowners', *Canadian Journal of Forest Research* **19**, 1014–23.

Muench, J. (1964), *Private Forests and Public Programs in North Carolina*, The American Forestry Association, Raleigh, NC.

Webster, H.H. and Stoltenberg, C.H. (1959), 'What ownership characteristics are useful in predicting response to forestry programs', *Land Economics* 35, 292–95.

Yoho, J.G. and James, L.M. (1958), 'Influence of some public assistance programs on forest landowners in North Michigan', *Land Economics* **34**, 363–64.