Using Simulation to Measure Bias in Principal Components Regression

Philip G. Enns
Saint Louis University

Abstract

Multicollinearity is a serious problem in regression analysis. High correlation among predictor variables can lead to unstable estimates of the regression coefficients. Principal components offers one approach to reducing the sampling variance of the coefficient estimates. However, this method produces biased estimates and the bias cannot be measured without knowledge of the true coefficients. This paper uses simulation to study the degree of bias in a model employing real economic data.

Introduction

Multiple regression continues to be one of the most widely used tools of the data analyst. The linear regression model provides one of the most simple yet powerful methods for describing the relation between variables and predicting the values of new observations. However, regression analysis is often subject to unavoidable problems of statistical estimation and practical interpretation. These problems occur with the presence of multicollinearity among the predictor variables. Severe multicollinearity gives rise to highly unstable estimates of the coefficients of the predictor variables obtained by the method of ordinary least squares (OLS). This impairs our ability to measure the marginal effect of a given predictor on the dependent variable being studied.

Various methods have been proposed for reducing the adverse effects of multicollinearity. One approach is to eliminate some of the independent variables in the ordinary least squares analysis. This can be criticized in two ways. It defeats the purpose of learning the effect of a predictor variable if that variable is eliminated from the model. It is also rather crude in that it implies a zero-one choice is necessary when selecting variables for inclusion in themodel. A more moderate procedure would attempt to use some partial information from all the potential predictor variables. [7]

The latter approach is adopted by a variety of techniques which share the property that they produce biased estimators of the regression coefficients. While OLS estimation is unbiased, high multicollinearity among the independent variables generally leads to large sampling variances for the coefficient estimates. The goal of biased regression techniques is to reduce the sampling variances sufficiently so as to compensate for the introduction of bias into the coefficient estimators. This trade-off between bias and efficiency has been the subject of extensive study by several authors in recent years[eg., 3, 5]. The problem is compounded by the fact that the degree of bias is a function of the unknown real coefficients of the underlying regression model. This is illustrated in the case of the principal components regression technique discussed below.

PRINCIPAL COMPONENTS REGRESSION

Consider the traditional form of the general linear model.

$$Y = X\beta + \varepsilon \,, \qquad (1)$$

where X is a nxk matrix of known regressors; β is a kxl vector of (unknown) regression coefficients; ε is a nxl vector of random variables, satisfying the conditions $E(\varepsilon) = 0$ and var $(\varepsilon) = E(\varepsilon\varepsilon^1) = \sigma^2 I$, which σ^2 unknown. The nxl vector Y contains observations on a dependent variable, while the n rows of X are observations on k independent variables $x_1, x_2, \ldots, x_k$. Assume that the independent variables are in standardized form, so that $\frac{1}{n}(X^1 X) = R$ is the matrix of sample correlations between the k variables.

Equation (1) expresses in matrix form the series of n observations generated by the equation

$$y_i = \beta_1 x_{i1} + \ldots + \beta_R x_{iR} + G_i \,, \qquad (2)$$

$$i = 1, 2, \ldots, n.$$

The usual interpretation of the j-th regression coefficient is that β_j measures the rate of change in y with respect to x_j when the other independent variables are held constant. This interpretation can be questioned since, in the presence of multicollinearity, it may not be meaningful to consider changing x_j and holding the other independent variables fixed [1]. Despite such questions, researchers remain attracted to regression analysis as a means of estimating the marginal response of a variable y to changes in "explanatory" variables.

The least squares estimator of the coefficient

CH1437-3/79/0153-0157$00.75

vector β is given by

$$b = (X'X)^{-1}X'Y \quad (3)$$
$$= nR^{-1}X'Y .$$

This estimator has variance-covar-iance matrix

$$\text{var}(b) = \sigma^2(X'X)^{-1}$$
$$= n\sigma^2R^{-1} \quad (4)$$

when one or more of the x variables are highly correlated, R is nearly singular, inflating the elements of the inverse and, hence, the variances associated with the ocefficient estimates.

The principal components of the variables $x_1, x_2, \ldots, x_k$ are a set of k artificial variables $z_1, z_2, \ldots, z_k$ having the following property: the j-th principal component is a linear combination of the x's

$$Z_j = w_{1j}x_1 + w_{2j}x_2 + \ldots + w_{Rj}x_R \quad (5)$$
$$j = 1, \ldots, k.$$

such that the variance of Z_j is maximum subject to the condition that Z_j is orthogonal to Z_1, Z_2, ..., and Z_{j-1}. As is well known, the coefficients in (5) form the elements of a matrix W with columns which are the orthonormal eigenvectors of the correlation matrix R. [8] The associated eigenvectors $\lambda_1 > \lambda_2 > \ldots \lambda_R > 0$ are the variances of the respective principal components. Hence, if Λ is the diagonal matrix with diagonal element λ_j, we have the following relations

$$\Lambda = W'RW \text{ and } W'W = WW' = I \quad (6)$$

As originally proposed by [6] the regression analyst might consider transforming the original correlated independent variables into the set of orthogonal principal components: Z = XW. By regressing Y on the columns of Z using ordinary least squares, we obtain estimates of the coefficients of the principal components

$$a = (Z'Z)^{-1}Z'Y ; \quad (7)$$

it is easily shown that the OLS estimates of the original regression coefficients are reproduced by

$$b = Wa \quad (8)$$

The potential advantage of the initial regression on Z comes from deleting some of the principal components before performing the transformation (8). Consider the matrix W_p obtained by replacing the columns of W by zeros for all but the p components indexed by the set $j \; j\epsilon P$. By defining the principal components estimator

$$b^* = W_p a \quad (9)$$

certain gains over OLS may be achieved. Under the common assumption that ε is normally distributed, it can be shown that the sampling distribution of b^* is normal with variance - covariance matrix $(\sigma^2/n)W_p\Lambda^{-1}W_p$. As shown in [9] this implies that the variance of the i-th principal component regression coefficient

$$\text{var}(b_i^*) = \frac{\sigma^2}{n}\sum_{j\epsilon P}\frac{W_{ij}^2}{\lambda_j} \quad (10)$$

At the same time, the bias in b_i^* is found to be a linear function of the true regression coefficients [2]

$$\text{bias}(b_i^*) = E(b_i^*) - \beta_i$$
$$= -\sum_{j\epsilon P}W_{ij}\sum_{l=1}^{k}W_{lj}\beta_l \quad (11)$$

Equations (10) and (11) reveal the nature of the tradeoff between the reduced sampling variance and the bias of the principal components regression. Because the terms in the summation of (10) are positive, the variance of b_i^* can always be reduced by deleting more components. However, the bias in b_R^* cannot be measured without knowledge of the true regression coefficients. Therefore, it is not possible to choose between OLS and principal components regression by any meaningful criterion. Consider, for instance, the mean square error (MSE), which is the most widely used measure of estimation performance

$$MSE(b_i^*) = E[b_i^* - \beta_i]$$
$$= \text{var}(b_i^*) + [\text{bias}(b_i^*)]^2$$
$$= \frac{\sigma^2}{n}\sum_{j\epsilon P}\frac{W_{ij}^2}{\lambda_j} + [\sum_{j\epsilon P}W_{ij}\sum_{l=1}^{k}W_{lj}\beta_l]^2 \quad (12)$$

Reductions in the first term of this expression are accompanied by changes in the second term which cannot be measured without knowledge of the β_i's.

When the independent variables are highly collinear, most of the variation in X is accounted for by the first few principal components. The simpliest approach to choosing a principal components estimator is to retain these components according to some arbitrary criterion. This approach ignores the dependent variable and several authors have discussed selection procedures which attempt to measure the effect on this variable [eg. 8]. However, no such technique has been found to reliably choose the best set of components.

A SIMULATION STUDY OF PRINCIPAL COMPONENTS REGRESSION

Major studies have been reported recently comparing principal components estimation with other biased regression methods and with more conventional variable selection techniques such as stepwise regression. These investigations have employed simulation to generate observations for models in which the independent variables possess different degrees of multicollinearity. With the regression coefficients prespecified the researcher can measure the actual estimation bias for a given simulation. In [5] several biased regression techniques, including principal components, are studied in great detail using repeated observations to obtain average coefficient estimates. In [3] a

wide variety of techniques are explored in considerably less depth. In the latter article the authors present strong evidence of the relatively poor performance of ordinary least squares when the X matrix is ill-conditioned. They also argue that further study is needed to develop better methods of dealing with real data in a specific situation.

This last point provides the basic motivation for the study reported below. In a previous article [4] a principal components regression was performed on a set of economic data with convincing results. A set of eleven measures of final consumer demand was selected as predictors of the Federal Reserve Index of Chemical Production. Quarterly observations were recorded for the years 1967 to 1976. While the demand variables were presumed to be highly correlated, the multicollinearity was deliberately increased by employing the one-quarter lagged values for the original variables, bringing to 22 the number of independent variables in the study.

Table I displays the eigenvalues of the correlation matrix for these data. Not surprisingly, a high percentage of the total variation in X is accounted for by the first few components. The subsequent regression analysis, employing only the first two principal components, produced coefficient estimates which achieve generally much higher statistical significance than those obtained by OLS. These results are shown in Table 2.

TABLE 1

PRINCIPAL COMPONENTS FOR TEST PROBLEM

PRINCIPAL COMPONENT	PERCENT OF TOTAL VARIANCE IN X
1	65.40459%
2	18.52932
3	6.34477
4	3.92659
5	1.78232
6	1.62082
7	0.83418
8	0.48814
9	0.30245
10	0.21309
11	0.18559
12	0.10164
13	0.07773
14	0.05982
15	0.03991
16	0.03200
17	0.01718
18	0.01473
19	0.00891
20	0.00736
21	0.00514
22	0.00368

TABLE 2

FULL MODEL VERSUS TWO VARIABLE PRINCIPAL COMPONENT MODEL

INDEPENDENT VARIABLE: FRB PRODUCTION INDEX CHEMICALS & ALLIED PRODUCTS

VARIABLE	FULL MODEL COEFFICIENT	T	2 VARIABLE P.C. MODEL COEFFICIENT	T
CONSTANT	-1.66	-	-76.73	-
MV&PARTS	-0.11	-0.45	0.27	13.39
MV&PARTS*	0.19	0.73	0.27	13.55
FURNITUR	4.07	3.52	0.22	35.99
FURNITUR*	3.29	2.78	0.22	35.70
OTHDURAB	2.50	0.99	0.79	26.78
OTHDURAB*	0.38	0.13	0.79	26.19
CLOTH&SH	-2.31	-2.62	0.32	18.39
CLOTH&SH*	0.20	0.16	0.34	16.10
FOOD&BEV	0.94	2.23	0.24	38.24
FOOD&BEV*	-0.31	-0.75	0.23	32.85
GAS&OIL	-0.00	-0.01	0.25	3.41
GAS&OIL*	-1.01	-1.18	0.23	3.12
OTHNONDU	-2.51	-2.94	0.23	21.84
OTHNONDU*	-0.55	-0.56	0.22	21.94
RESINVES	0.01	0.04	0.05	7.14
RESINVES*	-0.02	-0.07	0.05	5.04
PRODDURA	0.23	0.61	0.16	37.38
PRODDURA*	-1.23	-2.82	0.12	14.89
NONRCONS	1.08	1.44	-0.31	-6.57
NONRCONS*	1.20	1.83	-0.47	-5.63
INVENTOR	0.12	1.39	0.15	3.90
INVENTOR	-0.02	-0.37	0.08	3.20

*1 QUARTER LAG.

Of course, the 2-component regression fails to reveal the extent of bias in coefficient estimates. In order to study this problem, the following procedure was adopted. The estimated coefficients from 2-component regression were used to generate simulated values of the dependent variable. The error variance was taken as the value of the error mean square from the 22-variable OLS. (This is an unbiased estimator for σ based on 40 - 22 = 18 degrees of freedom).

The principal components regressions were applied to the simulated data for changing numbers of components, that is for different index sets P. The resulting coefficient estimates were compared to the "known" coefficients from the earlier study. This permitted calculation of estimated bias in each coefficient. The actual bias in the estimate of the i-th regression coefficient is computable from equation (11). As a composite measure of bias the square root of the sum of squares of the biases was used:

$$\text{Total Bias} \sqrt{\sum_{i=1}^{k} [\text{bias}(b_i^*)]^2} \qquad (13)$$

The object of the simulations was to obtain an approximation of the individual biases and the total bias after dropping the assumption of known regression coefficients. The researcher seeking to identify the best regression model will be interested in the rate at which estimation increases as more principal components are deleted. Since, in practice, the true model is not known, equation

(11) cannot be used. However, taking the model estimated using two components provides coefficient values which might be compared with "average" values generated by repeated monte carlo simulation.

RESULTS

The procedure outlined above produced two basic results, one predictable, the other unexpected. Simulations were performed with varying numbers T of repetitions, ranging from 1 to 100. As a measure of average bias for a given coefficient estimate, the simple mean was taken across all simulations:

$$\text{Mean Equation} \quad \frac{1}{T}\sum_{t=1}^{k} b^*_{it} - \beta_i = \bar{b}^*_i - \beta_i \qquad (14)$$

where b^*_{it} is the t-th principal component estimate of the i-th regression coefficient; and β_i is the assumed value of the i-th coefficient. As a composite measure of bias for the entire model, a calculation similar to that of (13) was used

$$\text{Estimated} \sqrt{\sum_{i=1}^{k} (\bar{b}^*_i - \beta_i)^2} \qquad (15)$$

As expected, the bias for the individual coefficients and for the total model, computed from equations (11) and (13), decreases as more components are added to the principal components regression. This is illustrated in Figures 1 and 2. Figure 1 depicts the bias in coefficient β_1, chosen arbitrarily from the 22 equation variables. Figure 2 shows the change in the total bias as the number of components is increased. The graphs also show changes in the estimated bias computed from (14) and (15) respectively. Not surprisingly, with T=1 these estimates increase steadily, reflecting the rise in sampling variance as more components are used in the estimation.

Unexpectedly, the effect of rising sampling variances did not disappear as the number of simulations was increased. Figures 3 and 4 depict the same measurements as Figures 1 and 2 with T=100. The pattern and magnitude of the estimated biases is quite similar to the case of one repetition. For intermediate values of T, the same patterns appeared.

This result would not seem to encourage the use of simulation in the measurement of bias in principal components regression applied to real data. However, the problem is compounded by fact that the assumed regression coefficients were obtained from the original data. An explanation of the results presented above rests on analysis which seems mathematically quite intractable.

CONCLUSION

Further study is needed to enable the investigator to assess bias in principal components regression and other alternatives to ordinary least squares. It is well established that these methods promise to be quite useful, but additional guidelines for the practical researcher need to be developed.

Figure 1
Bias in 1-st
Regression Coefficient
T = 1

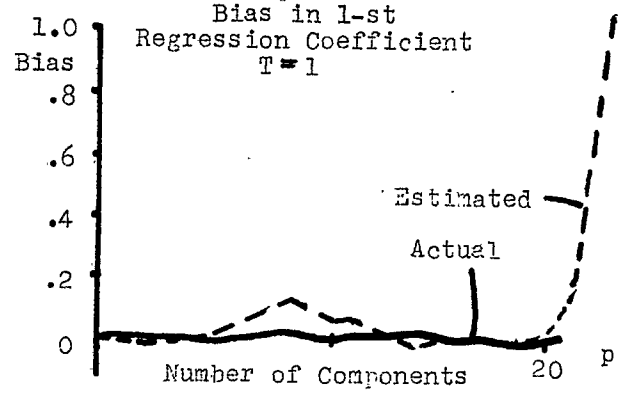


Figure 2
Total Bias
T = 1

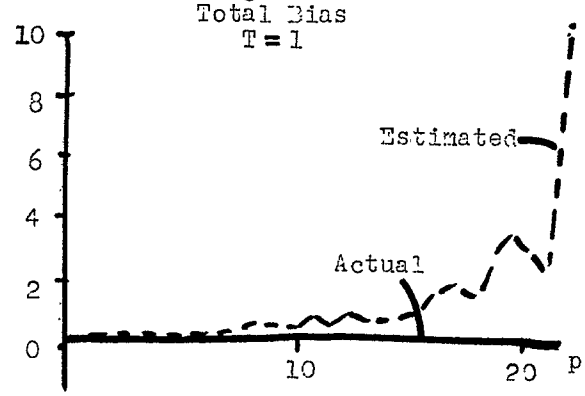


Figure 3
Bias in 1-st
Regression Coefficient
T = 100

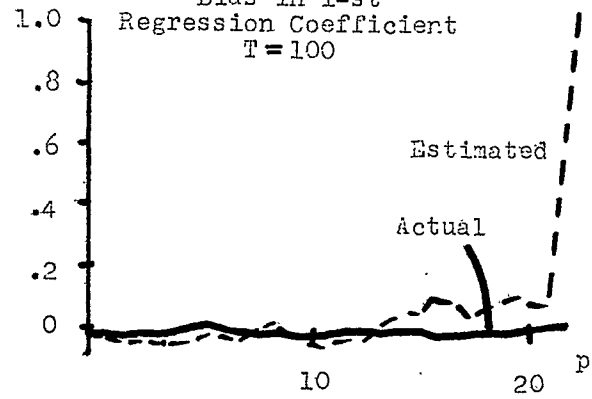
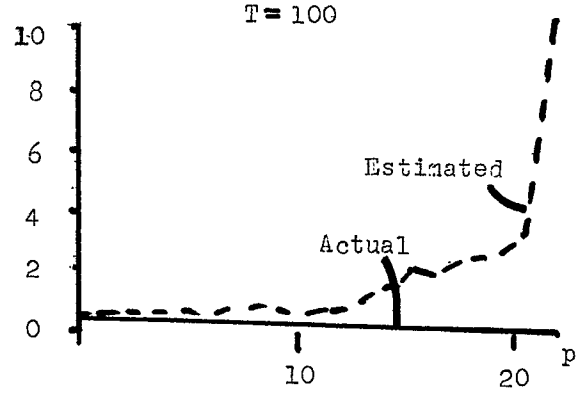


Figure 4
Total Bias
T = 100



REFERENCES

1. Allen, David M., "Comment on A Simulation of Alternatives to Ordinary Least Squares", Journal of the American Statistical Association, Vol. 72, No. 357 (March, 1977) 91-92.

2. Cheng, David C. and Iglarsh, Harvey J., "Principal Components Estimators in Regression Analysis", Review of Economics and Statistics, Vol. 58, No. 2 (May, 1976), 229-234.

3. Dempster, A.P., Schatzoff, Martin and Wermuth, Nanny, "A Simulation Study of Alternatives to Ordinary Least Squares", Journal of the American Statistical Association, Vol. 72, No. 357 (March, 1977) 77-90.

4. Enns, Philip G. and Qualls, John H., "An Approach to the Use of Principal Components Analysis in Dealing with Multicollinearity", Proceedings of Tenth Annual Meeting of American Institute for Decision Sciences (November, 1978) 281-283.

5. Gunst, Richard F. and Mason, Robert L., "Biased Estimation in Regression: An Evaluation Using Mean Squared Error", Journal of the American Statistical Association, Vol. 72, No. 359 (September, 1977) 616-628.

6. Kendall, M.G., A Course in Multivariate Analysis, London, Charles Griffin (1957).

7. Marquardt, Donald W. and Snee, Ronald D., "Ridge Regression in Practice", The American Statistician, Vol. 29, No. 1 (February, 1975). 3-19.

8. Massy. William F., "Principal Components Regression in Exploratory Statistical Research", Journal of the American Statistical Association, Vol. 60 (March, 1965) 234-256.

9. McCullum, B.T., "Artificial Orthogonalization in Regression Analysis", Review of Economics and Statistics, Vol. 52 (February, 1970) 110-113.

This research was sponsored by a grant from the Saint Louis University School of Business and Administration.