10th National Convention on Statistics (NCS)
EDSA Shangri-La Hotel
October 1-2, 2007

Forecasting from an Additive Model in the Presence of Multicollinearity

by

Erniel B. Barrios and Gregorio A. Vargas III

For additional information, please contact:

Author's name	:	Erniel B. Barrios
Designation	:	Professor
Affiliation	:	School of Statistics, University of the Philippines-Diliman
Address	:	Magsaysay St., Diliman, Quezon City, Philippines
Tel. no.	:	(0632) 928-0881
E-mail	:	ernielb@yahoo.com
Co-author's name	:	Gregorio A. Vargas III
Designation	:	Instructor
Affiliation	:	School of Statistics, University of the Philippines-Diliman
Address	:	Magsaysay St., Diliman, Quezon City, Philippines
Tel. no.	:	(0632) 928-0881
E-mail	:	gregvargasiii@yahoo.com

Forecasting from an Additive Model in the Presence of Multicollinearity

by

Erniel B. Barrios and Gregorio A. Vargas III

ABSTRACT

In a regression model with time series data, whether the regressors are cointegrated or not, the structure of X'X is usually characterized by multicollinearity. An additive model is postulated and estimated via the backfitting algorithm. The model sequentially smooth the residuals by entering one variable-at-a-time into the equation until the residuals behave randomly or all the effects of the independent variables are estimated. Real and simulated data exhibits superiority of the method over ridge estimators and principal component regression in prediction. Furthermore, even with the poorest model fit, the method is capable of providing 'good' estimates of the coefficients of the variables that entered into the model early during the iterative process.

Keywords *:* multicollinearity, additive model, backfitting

I. Introduction

The problem of multicollinearity exists in linear modeling when the regressors exhibit strong pairwise and/or simultaneous correlation, causing the design matrix to become non-orthogonal or worse, ill-conditioned. Once the design matrix is ill-conditioned, the least squares estimates are seriously affected, e.g., instability of parameter estimates, reversal of expected signs of the coefficients, masking of the true behavior the linear model being explored, etc.

The problem of multicollinearity commonly exists among economic indicators that are influenced by similar policies that lead their simultaneous movement along similar directions. Whether cointegration exists or not among the predictors, simultaneous drifting away in some directions especially among time series that exhibit non-stationary behavior is common.

There are many solutions proposed in the literature to address this problem. General shrinkage estimators are intended to stabilize the variance of the least squares estimators, while some procedures attach constraints to the least squares objective function to ensure estimates that exhibit acceptable behavior. Among the class of shrinkage estimators, the variance may be effectively reduced at the expense of the sensitivity of the model to truthfully assess the effect of the predictors to the dependent variable. The variables suspected to be duplicating the effect of other 'more important' variables are also removed to ease the extent of the problem. This will however, not only distort the functional relationship that is based on some theoretical framework, but also yield bias, inconsistent and other undesirable properties of the least squares estimators.

The many useful applications of linear models may not be achieved once the multicollinearity problem exists. (Carnes and Slade, 1988) simulated apparent competition known to exist at least experimentally. They tried to delete some variables from the model to resolve the multicollinearity problem, but realized that many important features of community organization/dynamics were lost. Thus, the true measure of competition cannot be assessed from a model that missed some important indexes in the competition framework because of the 'dropped' variables.

We propose in this paper a solution to the problem of multicollinearity that benefits from the features of 'dropping' variable solution and the shrinkage estimator while keeping all the variables in the model. The additive nature of the explanatory variables is exploited in the estimation procedure.

II. Some Solutions to Multicollinearity

Dropping of variables that duplicate the role of other 'more' important variables has been proposed as a natural solution to the multicollinearity problem. However, for models that strictly adhere to some theoretical framework, this is equivalent to massive loss of information. Then principal component regression is also proposed, hoping that the linear combination of the x's as a regressor will be able to keep all the variables into the model. Depending on the structure of the relationship among the regressors, component loadings may affect parameter estimates, e.g., loadings are similar, resulting to regression coefficients that are similar for all regressors. This will cause problem in interpretation because the relative importance of predictors is being masked by the way the linear combination is formed. Then the shrinkage estimators that generally reduce the variance of the least squares estimators become attractive. The complication occurs in the choice of how the shrinkage is to be induced to achieve some 'optimality'. Oftentimes, there is no assurance that a particular choice of shrinkage would indeed result to desirable properties of the estimators.

The literature on linear models with special focus on multicollinearity spans several decades already. However, no optimal solution has been achieved so far, so that there is still a continuing active interest on the problem.

Working on three explanatory variables, (McDonald and Galarncau, 1975), developed ridge-type estimators based on the largest eigenvalue of the design matrix. They observed that the estimators dominated the least squares estimators with respect to MSE. If instead, the ridge estimator is based on the smallest eigenvalue, mixed result has been observed. Ridge estimator does not always yield more optimal results compared to OLS. This was further followed up by (McDonald, 1980) who considered a one-parameter ridge estimator, observed that the order of ridge coefficients and some of their rates-of-change (shrinkage in variance) are unrelated to the degree of multicollinearity, but, existence and location of sign changes are explicitly related to multicollinearity. Shrinkage estimators may regulate instability of the estimates of regression coefficients, but will not always resolve the consequence of the multicollinearity problem.

The nature of shrinkage in both principal component and ridge regression is analyzed by (Butler and Denham, 2000), and compared it with partial least squares. They noted that partial least squares exhibited undesirable shrinkage properties and should not be adopted as an automatic solution to multicollinearity in regression problems. The canonical coefficients of partial least squares relative to OLS should be assessed to ensure that none are expanded substantially.

In the presence of multicollinearity, (George and Oman, 1996) proposed a multiple-shrinkage estimator which adaptively mimics the best principal components shrinkage estimator. The estimator proposed is a variant of Stein's estimator which is also guaranteed to give lower prediction mean-squared error than the least squares estimator, thus avoiding the danger of overshrinking when using principal components regression.

The varying results of different strategies in resolving the multicollinearity problem propelled the continuous interest in finding more optimal solutions to the multicollinearity problem.

III. The Additive Model

The additive model is specified through the equation $y = \alpha + \sum_{j=1}^{r} f_j(x_j) + e$. The effects of the independent variables on the dependent variables are individually assessed. This model formulation facilitates estimation since each component of the model can be addressed separately. Furthermore, the functional form of f_j need not be the same for all x's. In a more general case, the form of f_j need not be specified (nonparametric). This allows higher degree of flexibility on the postulated model, possibly compensating for the implicit assumption of additivity of the effect of the independent variables on the dependent variable. Proper construction of the function f_j can facilitate attainment of the additivity assumption in the model.

The estimation procedure called backfitting was proposed by (Hastie and Tibshirani, 1990), enables additive model-fitting using any regression-type estimation mechanism where the e are independent of the x's, $E(e) = 0$ and $var(e) = s^2$. The f_j's are arbitrary univariate functions, one for each predictor and estimated as follows:

(i) Initialize : a = ave(y_i), $f_j = f_j^0$, $j = 1,2,\ldots,r$

(ii) Cycle : $j = 1,2,..,r$

$$\hat{f}_j = S_j\left[\left(y - \sum_{k \neq j} f_k\right) \Big/ x_j\right] \tag{2}$$

(iii) Continue (ii) until the individual functions don't change where S_j denotes a smoothing of the response y against the predictor x_j.

While the algorithm started from some basic assumptions on the error term, empirical evaluation exhibited robustness of the method relative to the assumptions.

IV. A New Solution to the Multicollinearity Problem

Suppose there are p independent variables, let us decompose the linear model into two blocks of regressors as follows:

$$Y = X_1 b_1 + X_2 b_2 + e$$

The multicollinearity problem yields $tr(X'X) < p$ and $X_1^{'} X_2 \neq 0$. Following the backfiting algorithm, suppose that X_2 is ignored and least squares estimators of b_1 is obtained as

$$\begin{aligned}\hat{b}_1 &= (X'X)^{-1} X_1^{'} Y \\ &= b_1 + \left(X_1^{'} X_1\right)^{-1} X_1^{'} X_2 b_2 + e\end{aligned}$$

The residuals are computed and used in estimating b_2,

$$Y^* = Y - X_1 \hat{b}_1 = X_1 b_1 + X_2 b_2 - X_1 \hat{b}_1 = \left[I - X_1\left(X_1^{'} X_1\right)^{-1} X_1^{'}\right] X_2 b_2 + e$$

The orthogonality and additivity assumptions coincides. Additivity implies that there is no more influence of X_1 left in the residuals above. On the other hand, non-additivity implies that the residual will contain information of b_2 projected on the space of X_1. The effect of X_1 is passed on through the reparametrization of b_2 after the effect of X_1 has been accounted.

The use of backfitting in estimating a linear model in the presence of multicollinearity can yield potential benefits. Starting with the most important regressor say, x_1, its effect on the dependent variable is estimated 'optimally', ignoring the effects of all other regressors. The next regressor, say x_2, whose effect will be estimated next, will be reparametrized along the projection of x_1 first. The reparametrization will not only adjust the effect shared by both variables which has been attributed to the first one, it will also adjust the residual contribution of x_2 for appropriate size and direction. The reparametrization is expected to yield stabilizing effect on the estimates, hence, the magnitude and sign problems in OLS with the multicollinearity problem is potentially resolved.

The following interative process is proposed:

(1) Compute the correlation between the dependent variable and each of the independent variables.
(2) The dependent variable is regressed on the independent variable that exhibits the highest correlation.
(3) Compute the residuals and compute the correlation of these residuals with the remaining independent variables.
(4) Repeat Step (2), this time the correlations are based on Step (3).
(5) The iteration continues until the last independent variable.

V. Philippines Stock Exchange Index

The return of the Philippines Stock Exchange (PSE) index or change in the logarithm of the index (y_t) is regressed on change in the 91-days t-bill rate (x_{1t}), return on peso-dollar rate (x_{2t}), change in the real GDP growth (x_{3t}), and return of the S&P index 500 (x_{4t}) using the model

$$y_t = ß_1x_{1t} + ß_2x_{2t} + ß_3x_{3t} + ß_4x_{4t} + e_t$$

and quarterly data from the first quarter of 1988 until the first quarter of 2007. Overall model fit is assessed based on the root mean squared errors (RMSE). The additive model dominated other models with the lowest RMSE of 0.1204. Ridge regression with biasing parameter k=0.3 yield the second lowest RMSE, followed by principal component regression, OLS, cochranne-orcutt procedure, and backward elimination (only 1 regressor remained in the model).

In additive model-fitting, the regressor that exhibit highest correlation with the return of the PSE index is estimated first. The return of S&P index entered first (implying its importance), followed by change in peso-dollar rate, then by the change in 90-days t-bill rate, and finally by the change in GDP growth (least impact on the return of the PSE index). The three variables (S&P index, 90-days t-bill rate, and change in GDP growth) exhibited positive impact of the return in PSE index, while the peso-dollar rate yields negative impact. Volatility in the peso-dollar exchange rates can pull down the return if PSE index.

In OLS with all four regressors, cochranne-orcutt procedure, and ridge estimation, the signs of the regression coefficients are the same as in the additive model. Principal component regression expectedly estimated similar contributions among the four regressors because of the averaging effect of the component loadings. Backward elimination yields only one important variable, i.e., S&P index. The coefficient of S&P index is largest in magnitude in the additive model, although the estimates from other methods are not remarkably different. The magnitude of the peso-dollar rate (negative effect) coefficient is lowest in the additive model. The regression coefficient for the 90-days t-bill rate is lowest in additive model. The estimates from the additive model is not in the boundary only for the 'least' important predictor, the change in GDP growth where it is bounded by the cochranne-orcutt procedure above and ridge estimation below.

VI. GDP Growth Model

The additive model is also used in growth in GDP(y_t), regressed on investment growth (x_{1t}), employment growth (x_{2t}), and growth of the public expenditures on education (x_{3t}). The cobb-douglas production function is specified, using the annual data from 1981 to 2005.

Except for the cochranne-orcutt procedure which yield slightly higher RMSE, the additive model, OLS, principal component regression, ridge regression (with biasing parameter k=0.2) and back elimination (one variable is dropped) have similar predictive performance.

The additive model also yield the largest estimate for the coefficient of investment growth (the most important variable), also largest coefficient was derived for growth in employment. For the least important variable in the model, the expenditure to education, only the backward elimination's estimate is larger than that by the additive model. However, change in employment has been deleted by backward elimination. It is possible that the effect of the deleted variable has been attributed to change in education expenditures.

VII. Simulated Data

Three data sets were simulated. Thirty observations were generated for each of the three independent variables. Only the first independent variable was simulated from the normal distribution, $x_1 \sim N(10,25)$ and to generate multicollinear behavior, other variables is computed as $x_2 = 2x_1 + u(0,1)$ where u(0,1) is a random number generated from the uniform distribution, and $x_3 = \frac{1}{x_2}$. The condition number is 38,637 indeed a very strong multicollinearity among the independent variables exists. Three model fit were simulated using the same regression coefficients. For a very high model fit, the following regression model was generated: $y = 3 + 2x_1 + 1.5x_2 - 2x_3 + e$ where $e \sim N(0,1)$. For medium model fit, $y = 3 + 2x_1 + 1.5x_2 - 2x_3 + 50e$, where $e \sim N(0,1)$. And for low model fit, $y = 3 + 2x_1 + 1.5x_2 - 2x_3 + 10e$, where $e \sim \mathrm{N}(3,25)$.

Table 1 summarizes the parameter estimates from different methods as well as the root mean squared error (RMSE) assessing the predictive ability of the model. For principal component regression, only one component is used in the model since it already accounts for 80% of the total variation in the three independent variables.

The additive model estimated through the backfitting algorithm dominated the OLS, ridge estimator and the principal component regression in terms of predictive ability. Whether there is a good model fit (High) or very poor model fit (Low), the RMSE of the additive model is always the lowest.

The quality of OLS estimates worsen as the model fit deteriorates. The ridge estimator is capable of estimating the first few parameters only (even when model fit deteriorates). Regardless of model fit, principal component regression always produced parameter estimates too far from the actual values.

The additive model provides good estimates of the parameters estimated first or early during the iterative process. As the iteration progresses, the parameter estimates gets farther from the actual value. However, even with worsening model fit, the additive model can still produce estimates very close to the actual values during the early stages of iteration.

The incapability of the additive model to accurately estimate parameters later during the iterative process can be explained by the induced reparametrization in all iteration. The effect of the next variable entered into the model is first projected into the dimension of the variable previously entered into the model, resulting to 'deviating' value of the parameter in the pursuit of better fit. This also explains the superior prediction generated by the additive model.

Table 1: Parameter Estimates Using Different Estimation Methods

	Fit	Constant	X1	X2	X3	RMSE
Simulated Parameters	High	3	2	1.5	-2	
	Medium	3	2	1.5	-2	
	Low	3	2	1.5	-2	
OLS	R^2=99.95	1.8278	0.4611	2.3125[1]	-0.6177	0.69
	R^2=69.63	-55.6118	-74.9435	42.1262	67.1171	34.36
	R^2=17.80	48.5764	48.3118	-22.3013	-12.7823	57.49
Additive Model	High	1.5567	0.0011	2.5482[2]	-0.2940	0.64
	Medium	-18.8570	-0.1198	4.0318[2]	30.2326	34.11
	Low	32.5040	4.1322[2]	0.0027	-3.2947	53.90
Ridge Estimator	High	13.8539	2.0309	1.0216	-15.3726	5.48
	Medium	3.6690	2.9242	1.5033	-1.0015	39.47
	Low	36.2233	2.0450	0.8823	-10.2860	57.96
Principal Component Regression	High	53.4917	11.7780	11.7874	-8.7238	7.82
	Medium	63.3140	17.8280	17.8421	-13.2048	40.17
	Low	73.6222	9.6451	9.6528	-7.1439	56.12

[1]Significant at 5% level.
[2]First parameter estimated (highest correlation with Y).

VIII. Conclusions

Additive model can produce superior prediction or at least comparable predictive abilility to the usual solution to the problem of multicollinearity like principal component regression, ridge estimation and backward elimination.

The additive model estimated through backfitting can accurately estimate the more important variables estimated early in the iterative process. Regardless of model fit, the capability of the additive model in estimating the contributions of the more important variables holds.

References

Butler, N. and Denham, M, 2000, The Peculiar Shrinkage Properties of Partial Least Squares Regression, Journal of the Royal Statistical Society Ser. B 62(3):585-593.

Carnes, B. and Slade, N., 1988, The Use of Regression for Detecting Competition with Multicollinear Data, Ecology 69(4): 1266-1274.

George, E. and Oman, S., 1996, Multiple-Shrinkage Principal Component Regression, The Statistician 45(1): 111-124.

Hastie, T.J., Tibshirani, R.J., (1990), *Generalized Additive Model.* Great Bretain: St. Edmundsbury Press.

Mc Donald, G and Galarneau, D., 1975, A Monte Carlo Evaluation of Some Ridge-Type Estimators, Journal of the American Statistical Association 70(350): 407-416.

Mc Donald, G., 1980, Some Algebraic Properties of Ridge Coefficient, Journal of the Royal Statistical Society Ser. B 42(1): 31-34.