Multicollinearity and NO-FLIRP

Thomas Johnson and T. D. Wallace*

In a widely quoted reconsideration of the problem of multicollinearity, Farrar and Glauber (1967) came very close to enunciating the principle of "no free lunch in regression" (NO-FLIRP).

Regarding solution of the multicollinearity problem, Farrar and Glauber said:

> "Economists are coming more and more to agree that the second step, correction, requires the generation of additional information." (p. 92)

> "If a model is to be retained in all its complexity, solution of the multicollinearity problem requires an augmentation of existing data to include additional informaiton." (p. 95)

> "New information must be obtained." (p. 95)

However, some doubt about NO-FLIRP was cast by the authors themselves as they recounted Kendall's (1957) procedure and then offered some procedures of their own having to do with spotting and then isolating ill-conditioning in the $X'X$ matrix.

The purpose of this paper is to argue by example that, if the notion of more information is equivalent to additional knowledge in the form of either more sample data or priors on β, then NO-FLIRP applies. Or, at the very least it is a valuable principle for synthesizing proposals for solutions to multicollinearity.

Three examples are offered. The first is a well known characterization of collinearity in a two variable model. Example two is a discussion of a relatively new method for "conditioning" $X'X$ matrices called "ridge analysis." And one final example is a brief reiteration of the Kendall procedure. Each example is related to NO-FLIRP. An application of ridge analysis is included in an appendix.

EXAMPLE ONE: Collinearity in a Two Variable Model

In a two regressor linear model with no constant term obeying

*Thomas Johnson is Associate Professor of Economics and Statistics at North Carolina State University, Raleigh, North Carolina and T. D. Wallace is Professor of Economics at Duke University, Durham, North Carolina.

the Gauss-Markov assumptions it can be shown that the variance and mean squared error of, say, $b_{y1.2}$, the o.l.s. estimator of the coefficient of one first regressor, is

(1)
$$\text{MSE}(b_{y1.2}) = \text{V}(b_{y1.2}) = \frac{\sigma^2}{\sum_{t=1}^{T} X_{1t}^2 \ (1-r_{12}^2)}$$

where σ^2 is regression variance, $\sum_{t=1}^{T} X_1^2$ is the sum of squares of the first regressor, r_{12}^2 is the squared correlation of the first and second regressors, and T is the sample size.

Collinearity in this case is related to r_{12} and as the correlation goes to one, the variance of the o.l.s. estimator explodes.

This elementary result can be related to NO-FLIRP in two immediate ways. First, for any fixed $r_{12} \neq 1$, increasing the number of observations (T) increases ΣX_1^2, thus eventually bringing the problem within any prescribed bound.[1] Hence, more information, in the form of new sample data, resolves any except perfect collinearity. Second, additional information about β_2 in the form of an exact or inexact prior can remove the collinearity problem. If, for example, $\beta_2 = 0$, the o.l.s. estimator b_{y1} has mean squared error

(2)
$$\text{MSE}\ (b_{y1} | \beta_2 = 0) = \frac{\sigma^2}{\sum_{t=1}^{T} X_1^2}$$

and any imprecision that remains can no longer be blamed on collinearity.

The two points made in this example can be generalized in rather direct ways (Toro-Vizcarrondo and Wallace, 1968) to cases of more than two independent variables.[2]

EXAMPLE TWO: Ridge Analysis

Consider the general linear model

(3)
$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim (0, \sigma^2 I)$$

1 i.e., o.l.s. estimators are consistent and asymptotically unbiased.

2 Incorporation of inexact priors in the classical framework is discussed by Theil (1963) and Theil and Goldberger (1960).

where the sample size is T, the parameter space is nominally of order m<T, and the usual definitions apply to Y, X, β and ε. Multicollinearity has to do with an "ill-conditioned" X'X matrix and the ill-conditioning can be related to its characteristic roots.

For example, Hoerl and Kennard (1970) show that the average squared distance (sum of mean squared errors) of the o.l.s. estimator from β has the following expectation.

$$(4) \qquad E(b-\beta)'(b-\beta)=\sigma^2 tr\ (X'X)^{-1}=\sigma^2 \sum_{i=1}^{m}\left(\frac{1}{\lambda_i}\right) > \frac{\sigma^2}{\lambda_m}$$

where λ_i, i=1, ..., m are the characteristic roots of X'X ordered from largest to smallest. Thus, as X'X approaches singularity, the o.l.s. estimators get worse in MSE.

As a conditioning device, Hoerl and Kennard (1970) propose the estimator

$$(5) \qquad \hat{\beta}=[X'X+KI]^{-1}\ X'Y$$

where the scalar K may be chosen by considerations of the sample data.[3]

Hoerl and Kennard (1970) recognize that their estimator can be recast in a Bayesian framework of augmenting the sample data with priors on β and σ^2. However, one can use the Theil-Goldberger (1960, 1963) approach to recast ridge analysis into combining sample data with in exact restriction on β in the classical least squares framework.

E.g., consider the model in (3) above augmented by the prior

$$(6) \qquad 0=I_k\beta+U \quad \text{where} \quad U\sim(0,\gamma^2 I)$$

Combining (3) with (6) and applying the Aitken formula, the o.l.s. estimate for β is

$$(7) \qquad \hat{\beta}=\left[X'X+\frac{\sigma^2}{\gamma^2}I\right]^{-1}X'Y.$$

So the K in ridge analysis can be interpreted as the ratio of the value of the prior information to the value of the sample data, where the prior information is that β is zero apart from a random component.

3 In[2],Hoerl and Kennard (1970) choose to display the payoff function (SSE) and all b_j's as functions of K where K is varied from zero to one, with X'X normalized to the correlation matrix. In the practical problems considered, impressive stabiilty of the "ridge" estimators is achieved with only small increase in SSE.

Having K non-zero implies prior information on the variability of the elements of β either side of zero.[4]

EXAMPLE THREE: Deletion of Principal Components

The third example is the practice of deletion of principal components.

If the rank of X in equation three is m, there are m principal components which may be written

$$(8) \quad V = XG$$

Where G is mxm and its columns are characteristic vectors of X and the vectors of V are the orthogonal principal components of X[5]. The matrix V has the same informational content as X in the sense that a regression of Y on V is a full rank reparametrized regression of Y on X. Thus, if γ is the mxl vector of parameters relating V to Y,

$$(9) \quad \beta = G\gamma \text{ and}$$

$$(10) \quad b = Gc$$

where c and b are o.l.s. estimators of γ and β respectively (Massy 1965).

Some authors have suggested deletion of principal components as a solution for multicollinearity (Haitovsky 1968, Kendall 1957, Massy 1965), the heuristic motivation being that each characteristic root of X'X shows the variance of each principal component, the ordering of the roots from largest to smallest corresponding to the numbering of the principal components from first to last. Since the principal components are orthogonal, total variation in the full set is the sum of the characteristic roots of X'X. Hence, small characteristic roots indicate a correspondingly small contribution to total variation. The point to be made here is that deletion or other linear restrictions on principal components can be written as

$$(11) \quad H'\gamma = h$$

and can be recast in the form of exact priors on the original parameter

4 Given the impressive empirical results in Hoerl and Kennard (1970), priors of this type appear to be more attractive than exact zero restriction, i.e., outright deletion of some variables. However, if one were to tailor the ridge analysis approach to historical practices in economics, one might choose to condition X by $K\left(\begin{smallmatrix}00\\01\end{smallmatrix}\right)$, and thus consider the estimator $\hat{\beta} = \left[X'X + K\left(\begin{smallmatrix}00\\01\end{smallmatrix}\right)\right]^{-1} X'Y$ where the implicit inexact restrictions apply to the subset of β's which are of least interest or are under most prior "suspicion" in a particular problem.

5 See Malinvaud (1970) on the algebra of principal components.

space. Since G is orthogonal, the restrictions in (11) are equivalent to

(12) $H'G'\beta = h$.[6]

One can construct examples to show that deletion of principal components associated with least characteristic roots of $X'X$ may furnish implicit restrictions on β that are worse than restrictions associated with deletion of components associated with larger roots.[7] Basically, the central fault in the usual practice of the principal component procedure is that variability of the dependent variable enters in no way into the criteria for the evaluation of the implied priors. Conversely, through the sample estimates of β, the ridge analysis method assimilates this information. Thus, there are *a priori* grounds for arguing that *ad hoc* priors on β via ridge analysis hold more promise than indirect priors based on deleting principal components. Myoken and Uchida (1975) have proven that the ridge estimator has, in general, smaller mean square error than the principal component estimator.

Summary

With three examples we argue for the principle of "no free lunch in regression" (NO-FLIRP). We can solve the problem of multicollinearity only if we pay the price to obtain more information. For any except perfect collinearity a solution may be purchased for the price of more observations on the given set of variables. For perfect multicollinearity, $|X'X| = 0$, we must provide more information about

6 It may be worth noting that, if X is of rank $p<m$, an estimable reparametrization of the less than full rank model may be made via regression of Y on the p principal components of X. In this case, the G matrix of equation (8) above is mxp and the relation $G\gamma = \beta$ furnishes the reparametrization. Partitioning G and β conformably, the reparametrization is $G_2G_1^{-1}\beta_1 = \beta_2$. Deleting principal components then imposes an additional set of restrictions on β. The two step procedure of reduction to principal components and then restricting the principal component regression by $H'\gamma = h$, is equivalent to imposing the restrictions $\left[\begin{array}{c|c} G_2G_1^{-1} & -1 \\ \hline H & G' \end{array}\right]\beta = \left[\begin{array}{c} O \\ \hline h \end{array}\right]$

7 Take the example of a two variable standardized regression with $r_{y1} = 0.1$, $r_{12} = 0.98$ and $T = 33$, where r_{y1} is the correlation of the first regressor with the dependent variable, etc. Dropping the principal component associated with the smallest characteristic root of $\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$ yields a restriction on that can be rejected as improving MSE according to the Toro-Wallace test (1968). The implicit restriction on β via dropping the first principal component is a "good" move according to the same testing procedure. For an application of ridge analysis to those data, see the Appendix.

the parameters to obtain a solution. The technique of "ridge analysis" may be used to obtain information about potential restrictions on the parameters at a relatively modest price in the form of an increase in bias of the estimates. Conversely, the procedure of deleting principal components is likely to be a poor bargain and should be avoided.

References

Haitovsky, T. "Multicollinearity in Regression Analysis," read at the Econometric Society Summer Meeting, Boulder, Colorado, 1968.

Hoerl, Arthur E. and Robert W. Kennard, "Ridge Regression: Applications to Nonorthogonal Problems," *Technometrics*, Vol. 12 (1), February, 1970, pp. 55-68.

Hoerl, Arthur E. and Robert W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, Vol. 12 (1), February, 1970, pp. 69-82.

Farrar and Glauber, "Multicollinearity in Regression Analysis: The Problem Revisited," *The Review of Economics and Statistics*, Vol. XLIX, No. 1, February, 1967.

Kendall, M. G., *A Course in Multivariate Analysis*, Charles Griff and Company, Limited, London, 1957. pp. 70-75.

Myoken, H. and Y. Uchida, "The Generalized Ridge Estimator and Improved Adjustments for Regression Parameters," paper presented at the Japanese Statistical Society, July 1975.

Malinvaud, E., *Statistical Methods of Econometrics*, Second Edition, North-Holland Publishing Company, Amsterdam London, 1970.

Massy, W. F. "Principal Components Regression in Exploratory Statistical Research," *JASA*, March, 1965.

Theil, H. and A. S. Goldberger, "On Pure and Mixed Statistical Estimation in Eocnomics," *International Economic Review*, 2. 1960.

Theil, H. "On the Use of Incomplete Prior Information in Regression Analysis," *Journal of the American Statistical Association*, June, 1963.

Toro-Vizcarrondo, Carlos and T. D. Wallace, "A Test of the Mean Square Error Criterion for Restrictions in Linear Regression," *Journal of the American Statistical Association*, June, 1968.

APPENDIX: Ridge Analysis APPLIED TO THE DATA IN NOTE SEVEN

The Appendix Table shows what happens when ridge analysis is applied to the data in note 7 of the text with K incremented by tenths from zero to one.

Appendix Table: Ridge Estimates for Two Variable Model with $r_{y1}=0.1$, $r_{y2}=0.2$, $r_{12}=0.98$

K	$b_1(K)$	$b_2(K)$	SSE
0	−2.42	2.56	.73
.1	− .34	.49	.90
.2	− .16	.30	.93
.3	− .09	.22	.98
.4	− 06	.18	.95
.5	− .04	.16	.96
.6	− .02	.14	.97
.7	− .01	.13	.98
.8	− .01	.11	.98
.9	− .002	.11	.98
1.0	.001	.10	.98

From the Appendix Table, one can see that for K between .6 and 1.0, the SSE stabilizes and estimates for β_1, β_2 stabilize at about zero and .10 respectively. Deleting the first principal component yielded estimators of β_1, β_2 of −2.5 and 2.5 while dropping the second component gave estimates of .08, .08.