COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 383

MULTICOLLINEARITY AND FORECASTING

Gary Smith

December 4, 1974

MULTICOLLINEARITY AND FORECASTING*

by

Gary Smith

Much of the available economic data is so highly intercorrelated that it is incapable of distinguishing between radically different models of economic behavior. The variation of a given dependent variable can seemingly be explained equally well by a limitless variety of theoretically motivated and randomly chosen explanatory variables. And for any particular model, the confidence regions are generally so large that the point estimates and forecasts are of little interest.

One response to the inadequate informational content of the data has been shoulder shrugging, with remarks akin to Johnston's that one cannot "make bricks without straw." Some of the shruggers view the problem as annoying but intractable, while others seemingly argue that having highly correlated explanatory variables is not necessarily a problem. Raduchel, for example, has written that "collinearity should be viewed more as a condition to be recognized rather than as a problem to be corrected.... No particular virtue attaches to orthogonality nor any grievous vice to collinearity."

A more popular response is for the model builder to limit himself to a relatively small number of explanatory variables, chosen in part

---

*The research described in this paper was undertaken by grants from the National Science Foundation and from the Ford Foundation.

for their relative orthogonality. Or a more complete theoretical model is specified, and then variables are sequentially dropped when their coefficients are found to be statistically insignificant. And in between we have those who toil endlessly to find some combination of explanatory variables which will yield statistically significant and plausibly signed parameter estimates.

These practices are sometimes defended on the grounds that simpler models are easier to understand, or that it is expensive to collect data and to predict values for explanatory variables. More often, it is thought that one is informally applying a priori information by trying to get correctly signed coefficients, and formally applying statistical theory by running hypothesis tests.

Econometricians generally respond that the final estimates and statistics have little meaning and that the modeler has done little more than disguise the imprecision of his estimates. And it is often noted that incorrectly including irrelevant variables will not bias the estimated coefficients, as will incorrectly excluding variables. Thus, Johnston concludes that, "Data and degrees of freedom permitting, one should error on the side of including variables in the regression analysis rather than excluding them." In response to this, it is argued (particularly in the pretest literature) that excluding variables will reduce the variance of the estimates, and that this may more than offset the bias if a mean squared error criterion is used.

In this paper, I will argue that particularly in a forecasting context collinearity can be an extremely serious problem, which should motivate the liberal introduction of even weakly held a priori information. However, such information should not be confused with the imposition

of arbitrary or randomly selected constraints on the coefficients.

The proposed remedies for multicollinearity which focus on reducing the number of explanatory variables amount to no more than awkward ways of imposing a variety of wholly ad hoc parameter restrictions on the model. While these may be accidentally beneficial, it is difficult to rationalize their use in place of analogous a priori constraints. Similarly, it is very hard to justify the mechanical imposition of arbitrary (typically zero) restrictions on the parameters whenever these are unrejected by the data.

Instead of passively resorting to ad hoc constraints, economists should view "the collinearity problem" as an opportunity to introduce valuable a priori information of a much broader range than simple exclusion restrictions. One can informally impose exact restrictions based upon a judgmental blending of the inadequacies of the data, the faith one has in the a priori information, and the uses to which the estimates will be put. Or one can directly accomplish this with a Bayesian or quasi-Bayesian (such as mixed estimation) technique which explicitly clarifies (and exposes) one's subjective assumptions. While the informal approach myopically chooses between no information and perfect information on a parameter-by-parameter basis, a Bayesian approach permits flexible priors and multi-parameter considerations.

In Section I of this paper, I discuss the ambiguities involved in the detection and measurement of the collinearity problem. Section II is concerned with the extreme but illustrative case of exact collinearity. In Section III, I consider constrained, unconstrained and pretest estimators for imperfectly collinear explanatory variables. Particular attention is paid to out-of-sample forecasting and (for expositional purposes)

the weaknesses of a principal components approach.

The following conventional framework will be used throughout the paper. T sample observations from the assumed model

$$\underset{T\times 1}{y} = \underset{T\times p}{X}\ \underset{p\times 1}{\beta} + \underset{T\times 1}{\epsilon}$$

will be used to forecast n out-of-sample values of y which are assumed to be generated by

$$y^0 = X^0\beta + \epsilon^0$$

where

$$E\left[\begin{pmatrix}\epsilon\\ \epsilon_0\end{pmatrix}\begin{pmatrix}\epsilon\\ \epsilon_0\end{pmatrix}'\right] = \begin{bmatrix}\sigma^2 I_T & 0\\ 0 & \sigma_0^2 I_n\end{bmatrix}.$$

The forecasts will be denoted by

$$\hat{y} = X^0\hat{\beta}$$

where the parameter estimates $\hat{\beta}$ use only the in-sample data.

I. Measurement

The common sense notion of collinearity is that when two explanatory variables have been highly correlated, one cannot accurately estimate the effects of either moving separately. But it is not easy to go from this heuristic notion to a rigorous definition and an attractive scale upon which to measure the severity of the collinearity situation. The essential problem is to define a way of quantitatively tying down the harmful consequences of collinearity and to eliminate the ceteris paribus

ambiguities--what factors are being held constant as one talks about differing degrees of collinearity?

One of the more widely read approaches is Farrar and Glauber's insightful discussion, which includes the recommendation that the severity of the overall collinearity problem be measured by the determinant of the simple correlation matrix for the explanatory variables:

$$|R| = |\Sigma^{-\frac{1}{2}} X'X \Sigma^{-\frac{1}{2}}|$$

where Σ is a diagonal matrix with the i^{th} diagonal element equal to the sample sum of squares of the i^{th} explanatory variable. Since this determinant will lie down between zero (singularity) and one (orthogonality), they argue that its closeness to either of these extremes can be interpreted as a suggestive measure of how collinear the explanatory variables are. With the assumption that X is multivariate normal, they offered as a more precise measure Bartlett's chi-square test of the null hypothesis of orthogonality in the underlying population.

$$-[T - 1 - \tfrac{1}{6}(2p+5)]\text{Log}[R] \sim \chi^2_{\frac{1}{2}p(p-1)} .$$

Haitovsky correctly noted that this is a questionable null hypothesis given Farrar and Glauber's emphasis on the unimportance of the parent data. Haitovsky also argued that the test is conservative in that satisfactory estimates do not require strict orthogonality. He consequently proposed the alternative extreme of singularity as a null hypothesis, despite the acknowledged fact that this cannot be seriously tested since an m-dimensional population will not generate data of more than m dimensions. Nevertheless, Haitovsky proposed the heuristic test statistic

$$-[T - 1 - \tfrac{1}{6}(2p+5)]\text{Log}[1 - |R|] \sim \chi^2_{\frac{1}{2}p(p-1)}$$

which is at least successful in giving a more comforting signal than Farrar and Glauber's test. For example, with 50 observations on two explanatory variables, Farrar and Glauber's test would register trouble at the 1% level when the squared correlation between the two variables is greater than .13, while Haitovsky would require a squared correlation coefficient greater than .87. Part of the problem here is the usual classical dilemma of which state should receive the presumptive weight of being classified as the null hypothesis; but even more critical is the awkward struggle to define multicollinearity in terms of the properties of a presumed parent population. The consequences of multicollinearity are clearly due to the nature of the available sample data, irrespective of the source of the data. Indeed, the most important qualification to this statement is somewhat perverse: in a forecasting context, extreme sample collinearity is the most worrisome when future samples are not characterized by such collinearity. Thus, in our two-variable example, a squared correlation coefficient of, say, .99 would be less worrisome if the squared population correlation coefficient were .99 than if it were zero.

And, finally, the determinant of the simple correlation matrix is wholly arbitrary in that a linear transformation of the explanatory variables can always yield a determinant which is equal to one or arbitrarily close to zero. This (and many other important issues) can be discussed most clearly by considering in some detail the case of two explanatory variables in an equation written in deviations from means form:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon \ .$$

We have the moment matrix

$$(X_1 \quad X_2)'(X_1 \quad X_2) = \begin{bmatrix} \Sigma X_1^2 & \Sigma X_1 X_2 \\ \Sigma X_1 X_2 & \Sigma X_2^2 \end{bmatrix}$$

and covariance matrix

$$\mathrm{Cov}\begin{pmatrix} \hat{\beta}_1 \\ \beta_2 \end{pmatrix} = \sigma^2 \begin{bmatrix} \Sigma X_2^2 & -\Sigma X_1 X_2 \\ -\Sigma X_1 X_2 & \Sigma X_1^2 \end{bmatrix} \frac{1}{\Sigma X_1^2 \Sigma X_2^2 - (\Sigma X_1 X_2)^2}$$

so that

$$\mathrm{Var}(\hat{\beta}_i) = \sigma^2 \frac{1}{(\Sigma X_i^2)} \frac{1}{1 - r^2} = \frac{\sigma^2}{T} \frac{1}{\sigma_{X_i}^2} \frac{1}{1 - r^2}$$

where r is the sample correlation coefficient between X_1 and X_2. Thus, the estimates of the β_i would be more precise if the omitted factors were less important, if there was more data, if there was more variation in the explanatory variables, or if X_1 and X_2 were less correlated. The question here is which of these factors to identify as the "collinearity problem," and by what scale to assess the severity of the problem.

The most natural (and common) choice is r, the correlation between X_1 and X_2. Often, the variables are standardized

$$Y = (\beta_1 \sqrt{\Sigma X_1^2}) \left[\frac{X_1}{\sqrt{\Sigma X_1^2}} \right] + (\beta_2 \sqrt{\Sigma X_2^2}) \left[\frac{X_2}{\sqrt{\Sigma X_2^2}} \right] + \epsilon$$

$$= b_1 Z_1 + b_2 Z_2 + \epsilon$$

so that the moment matrix is a simple correlation matrix

$$(Z_1 \quad Z_2)'(Z_1 \quad Z_2) = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

with a covariance matrix given by

$$\mathrm{Cov}\begin{pmatrix} \hat{b}_1 \\ b_2 \end{pmatrix} = \sigma^2 \begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix} \frac{1}{1 - r^2} .$$

Now, given σ^2, the variances of the estimates of the b_i are directly tied to r^2, and it has been variously suggested that the extent of the collinearity problem be measured by

r : the correlation between X_1 and X_2

$-\sigma^2 \frac{r}{1 - r^2}$: the covariance between $\hat{b}_1$ and $\hat{b}_2$

$1 - r^2$: the determinant of the standardized moment matrix

$1-r$, $1+r$: the eigenvalues of the standardized moment matrix.

Consider, however, the effects of rearranging the explanatory variables. Since a linear transformation of the explanatory variables will not affect the implicit estimates of any estimable parameters, the initial arrangement of the variables is totally arbitrary and therefore should not affect our measure of the severity of the collinearity problem. To take an extreme rearrangement,

$$Y = \left(\frac{b_1 + b_2}{\sqrt{2}}\right)\left[\frac{z_1 + z_2}{\sqrt{2}}\right] + \left(\frac{b_1 - b_2}{\sqrt{2}}\right)\left[\frac{z_1 - z_2}{\sqrt{2}}\right] + \epsilon$$

$$= \gamma_1 w_1 + \gamma_2 w_2 + \epsilon \, .$$

Now the moment matrix is

$$(w_1 \quad w_2)'(w_1 \quad w_2) = \begin{bmatrix} 1+r & 0 \\ 0 & 1-r \end{bmatrix}$$

and the covariance matrix for the explicitly estimated parameters is

$$\text{Cov} \begin{matrix} \hat{\gamma}_1 \\ \gamma_2 \end{matrix} = \sigma^2 \begin{bmatrix} \frac{1}{1+r} & 0 \\ 0 & \frac{1}{1-r} \end{bmatrix}$$

so that

$$\text{Var}(\hat{\gamma}_i) = \frac{\sigma^2}{T} \frac{1}{\sigma^2_{w_i}} \, .$$

In this form of the equation, the explanatory variables are uncorrelated and the estimated parameters have no covariance, so that these two indicators (which are based on the degree of diagonality of the moment matrix) would indicate that there is no collinearity problem. If the estimates here are imprecise, it is not because the associated variables are highly correlated, but rather because they display little variation. Thus, any statement of high correlation can be equally well expressed as one of low variation. For example, one could speak of the high correlation between two interest rates or of the stability of the rated differential.

However, we surely do not want our detection of a multicollinearity problem to depend upon whether we use the two rates as explanatory variables or instead use one rate and the rate differential.

The determinant and eigenvalue tests hold up in this case, since they are unaffected by an orthonormal transformation such as was used here. It is desirable to restrict one's attention to orthonormal rearrangements since these preserve the length of the parameter vector. That is, in general, for

$$Y = Zb + \epsilon = ZH'Hb + \epsilon$$
$$= W\gamma + \epsilon$$

we have $\gamma'\gamma = (Hb)'(Hb) = b'b$.

However, orthonormal transformations do not preserve the unit variances of the explanatory variables. Thus if w_1 and w_2 had been our original variables, we would have standardized them to

$$Y = (\gamma_1\sqrt{1+r})\left[\frac{w_1}{\sqrt{1+r}}\right] + (\gamma_2\sqrt{1-r})\left[\frac{w_2}{\sqrt{1-r}}\right] + \epsilon$$

$$= \quad C_1 \quad V_1 \quad + \quad C_2 \quad V_2 \quad + \epsilon$$

with

$$(V_1 \quad V_2)'(V_1 \quad V_2) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and

$$\mathrm{Cov}\begin{pmatrix} \hat{c}_1 \\ c_2 \end{pmatrix} = \sigma^2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus, if we had started with w_1 and w_2 and calculated the determinant or the eigenvalues of the standardized moment matrix, we would have found no indication of a collinearity problem. Since neither the model nor the estimates have changed, this means that (contrary to the widely held view) the moment matrix or covariance matrix cannot alone convey the seriousness of the collinearity problem. This is because the variances can be made arbitrarily large or small simply by changing the scale of the parameters--if the variance of $\hat{\alpha}$ is $100\sigma^2$, then the variance of $\widehat{.1\alpha}$ will only be σ^2. Thus, one cannot say whether the variance of an estimated parameter is large or small without knowing the scale of the parameter and one's objectives. Specifically, a seemingly large variance may be of little concern if the parameter itself is large or uninteresting.

For a particular coefficient, one can examine the estimated parameter and a representation of its variance

$$\text{Var}(\hat{\theta}_i) = \frac{\sigma^2}{T} \frac{1}{\sigma_i^2} \frac{1}{1 - R_i^2}$$

(where σ_i^2 is the variance the associated variable and R_i^2 is the squared multiple correlation coefficient between this variable and the remaining explanatory variables) and make a subjective statement about the seriousness of the imprecision. And one can informally attribute part of this imprecision to collinearity by a calculation of how much smaller the variance of the estimate (or the confidence band) would be if, ceteris paribus R_i^2 were closer to zero.

Statements about the overall collinearity problem would seemingly require an examination of the covariance matrix for a basis for the estimable parameters and some sort of subjective loss function. It is probably

most convenient to eliminate the covariances by working with the coefficients of an orthogonal basis for the explanatory variables. In our two-variable example, we might use the principal components

$$Y = \gamma_1 w_1 + \gamma_2 w_2 + \epsilon$$

with

$$\mathrm{Cov}\begin{pmatrix} \hat{\gamma}_1 \\ \gamma_2 \end{pmatrix} = \sigma^2 \begin{bmatrix} \frac{1}{1+r} & 0 \\ 0 & \frac{1}{1-r} \end{bmatrix} = \sigma^2 \begin{bmatrix} \frac{1}{\lambda_1} & 0 \\ 0 & \frac{1}{\lambda_2} \end{bmatrix}$$

where λ_1 and λ_2 are the characteristic roots of the standardized moment matrix $Z'Z$. Now, all estimable parameters will be linear functions of γ_1 and γ_2 with variances given by linear combinations of σ^2/λ_1 and σ^2/λ_2. Consequently, a loss function for any pair of estimable parameters can be conveniently recast in terms of $\hat{\gamma}_1$ and $\hat{\gamma}_2$.

Again, such a loss function is critical to an interpretation of the collinearity problem. For example, if r is positive then a confidence interval for γ_2 is larger than one for γ_1; but the scale of γ_1 may be so much larger than that of γ_2, that one would conclude that it is the more precisely estimated parameter. This suggests pre-scaling the parameters, so that equally wide confidence intervals would be viewed as equally precise estimates, before attempting to determine in which directions the data is informative. In general, once we've subjectively specified m parameters which can be accurately estimated and $p-m$ which cannot, then we can speak loosely of the data being effectively m-dimensional and of accurately estimating no more than m linearly independent

parameters, in the sense that any larger set of independent parameters must necessarily depend upon some of the imprecise estimates. It is also true that the sum of the variances of any orthonormal transformation of the γ_i will continue to equal $\sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$.

This still leaves the question of whether the overall imprecision is serious or not. This depends upon the absolute size of the bands (and hence upon the variance of the disturbance term), and upon what values of the loss function are considered serious. Thus, r could be very close to one and yet we might conclude that there is no collinearity problem because γ_2 is very large (in part, perhaps, because the variances of the X_i are large), or because σ^2 is very small, or because it is not important that we have an accurate estimate of γ_2.

More concretely, consider the plausible situation in which the objective of the model is to forecast the dependent variable accurately in an out-of-sample context. Using the expected value of the mean squared forecasting error as a loss function, we have

$$\text{EMSE} = E\left[\frac{(Y^0 - Y^0)'(Y^0 - Y^0)}{n}\right] = \sigma_0^2 + \frac{\sigma^2}{T}\left[\frac{h_1^2 + h_2^2 - 2rr^0 h_1 h_2}{1 - r^2}\right]$$

where $h_i = \sigma_{X_i^0}/\sigma_{X_i}$. Quite obviously, the severity of the collinearity problem depends upon a lot more than the absolute magnitude of r. Indeed it is even ambiguous whether or not an increase in the squared correlation coefficient is harmful. For example, when $h_1 = h_2 = h$,

$$\text{EMSE} = \sigma_0^2 + \frac{\sigma^2}{T} 2h^2 \left[\frac{1 - rr_0}{1 - r^2}\right].$$

Now if the out-of-sample correlation, r_0, is .9, then $[(1 - rr_0)/(1 - r^2)]$ is 1.0 if the in-sample correlation is .0 and only .55/.75 if $r = .5$. More generally, a value of r^2 close to one need not be serious if the correlation persists out-of-sample (r^0 close to r), if there is very little out-of-sample variation of the explanatory variables relative to the in-sample variation, if the in-sample variance of the disturbance term is small, or if there is a lot of sample data.

This of course does not imply that there will not be circumstances in which the high correlation between two explanatory variables is a cause for alarm. Indeed, one of the most pervasive arguments running through the remainder of this paper is that, in the presence of highly correlated explanatory variables, it is essential to liberally impose a priori information in order to make accurate forecasts in a wide variety of situations. All that is pointed out here is that there is no simple way of discerning the severity of the collinearity problem.

II. Exact Collinearity

Several general points are most easily introduced by an examination of the relatively straightforward case of perfect or exact collinearity. Consider then the situation where X is of rank $m < p$. There are $p-m$ separate linear relations among the X_i in the sample period and $p-m$ of the characteristic roots of $X'X$ are equal to zero. It is impossible to measure the full p dimensional effect of X on Y, and this is manifested in the breakdown of the ordinary least squares (OLS) regression of Y on X.

We can however find an m-dimensional basis B_1

$$\underset{Txp}{X} = \underset{Txm}{B_1}\ \underset{mxp}{C_1} \quad \text{where} \quad B_1 = \underset{(pxm)}{X\ H_1}$$

such that we can work with m explanatory variables

$$Y = B_1C_1\beta + \epsilon = \underset{Txm\ mxl}{B_1\ \gamma_1} + \epsilon \ .$$

The m elements of γ_1 can be uniquely estimated

$$\hat{\gamma}_1 = (B_1'B_1)^{-1}B_1'Y$$

but the associated m equations

$$C_1\hat{\beta} = \hat{\gamma}_1$$

cannot be solved for unique estimates of the p elements of β without the introduction of p-m linear restrictions (or pieces of <u>a priori</u> information) regarding β :

$$\underset{p-mxp\ pxl}{\alpha\quad\beta} = \underset{p-mxl}{\lambda} \ .$$

If $\begin{bmatrix} C_1 \\ \alpha \end{bmatrix}$ is nonsingular, then

$$\begin{bmatrix} C_1 \\ \alpha \end{bmatrix}\hat{\beta} = \begin{bmatrix} \hat{\gamma}_1 \\ \lambda \end{bmatrix}$$

can be solved for $\hat{\beta}$.

The out-of-sample use of the m-dimensional basis

$$X^0\beta = B_1^0\gamma = B_1^0C_1\beta = X^0H_1C_1\beta$$

will be generally correct if the exact in-sample collinearity persists

$$X^0 = B_1^0 C_1$$

or if

$$\beta = H_1 C_1 \beta \ .$$

Since $H_1 C_1$ is of rank m, this latter condition states that $C_1\beta$ is an m-dimensional basis for β --that is, it defines $p-m$ linearly independent restrictions upon β. Thus, forecasting with an m-dimensional basis

$$\hat{Y} = B_1^0 \hat{\gamma}$$

is precisely equivalent to imposing $p-m$ parameter restrictions. The accuracy of these parameter restrictions will be unimportant if the $p-m$ linear dependencies among the explanatory variables continue to hold in the forecasting period, but may be very important in more general situations.

An ancillary matter is the effect of using a different basis

$$\underset{Txm}{B_2} = \underset{Txp}{X} \ \underset{pxm}{H_2} \ .$$

For any such basis, there will be a nonsingular transformation G such that

$$B_1 = B_2 G \ .$$

Thus

$$Y = B_1\gamma_1 + \epsilon = B_2G\gamma_1 + \epsilon = B_2\gamma_2 + \epsilon$$

where $\gamma_2 = G\gamma_1$ since B_2 is of full rank m. The Gauss-Markov Theorem implies that the OLS estimates will be such that

$$\hat{\gamma}_1 = G^{-1}\hat{\gamma}_2 .$$

That is, the implicit estimates of all estimable parameters will be independent of the particular basis used for the estimation.

However, the forecasts will generally depend upon the basis used for predictions if the in-sample linear dependencies do not continue to hold out-of-sample, i.e.

$$\hat{Y}^{(2)} = B_2\hat{\gamma}_2 = B_2W^{-1}\hat{\gamma}_1 = B_1\hat{\gamma}_1 = \hat{Y}^{(1)}$$

if $B_2 = B_1W$, but only by fortuitous coincidence otherwise.

Thus, forecasting with an m-dimensional basis is equivalent to imposing $p-m$ independent parameter constraints. The particular basis chosen (or the parameter restrictions implicitly imposed) will be inconsequential as long as the $p-m$ exact in-sample linear relations among the explanatory variables continue to hold. If these evaporate, however, then the accuracy of the forecasts will depend critically upon the accuracy of the parameter restrictions, which resulted from the arbitrary and seemingly innocent choice of a basis.

For a sample example, consider two perfectly correlated explanatory variables

$$Y = \beta_1X_1 + \beta_2X_2 + \epsilon \quad \text{where} \quad X_2 = \alpha X_1$$

Using $\lambda_1 X_1 + \lambda_2 X_2$ as a basis $(\lambda_1 \neq -\alpha\lambda_2)$, we have in general

$$\beta_1 X_1 + \beta_2 X_2 = (\lambda_1 X_1 + \lambda_2 X_2)\left(\frac{\beta_1 + \alpha\beta_2}{\lambda_1 + \alpha\lambda_2}\right) + (X_2 - \alpha X_1)\left(\frac{\lambda_1\beta_2 - \lambda_2\beta_1}{\lambda_1 + \alpha\lambda_2}\right)$$

which will

$$= (\lambda_1 X_1 + \lambda_2 X_2)\left(\frac{\beta_1 + \alpha\beta_2}{\lambda_1 + \alpha\lambda_2}\right)$$

if $X_2 = \alpha X_1$ or if $\lambda_1\beta_2 = \lambda_2\beta_1$. In particular, using one of the following variety of bases is equivalent to imposing the indicated parameter restriction

λ_1	λ_2	Basis	Parameter Equivalent
1	0	X_1	$\beta_2 = 0$
0	1	X_2	$\beta_1 = 0$
1	1	$X_1 + X_2$	$\beta_1 = \beta_2$
1	-1	$X_1 - X_2$	$\beta_1 = -\beta_2$

Generalized Inverses (A Digression)

The generalized inverse is often defined as an nxm matrix A^- such that

$$X = A^- d$$

is a solution of the consistent system of equations

$$\underset{mxn}{A}\ \underset{nx1}{x} = \underset{mx1}{d}\ .$$

If the rank of A is r, this is equivalent to solving r linearly independent equations for n unknowns, where $n > r$. There are of course an infinite number of solutions ($n-r+1$ of them linearly independent if $d \neq 0$) and hence an infinite variety of generalized inverses. For this reason it is conventional to use the following four conditions to define a unique generalized inverse, A^+ :

$$AA^+A = A$$
$$A^+AA^+ = A^+$$
$$(AA^+)' = AA^+$$
$$(A^+A)' = A^+A \ .$$

Now, if X is of rank m then the normal equations

$$X'X\beta = X'Y$$

do not have a unique solution inasmuch as there are only m linearly independent equations in p $(>m)$ unknowns. Incredibly, it has been seriously proposed (by Swamy, et al.) that generalized inverses yield a solution to this problem

$$\hat{\beta} = (X'X)^-X'Y \ .$$

In fact, the problem has only been redefined, since each of the infinity of potential parameter values can indeed be written in this form. Thus, the use of a particular generalized inverse is necessarily equivalent to deleting $p-m$ linearly redundant explanatory variables and imposing $p-m$ parameter constraints. The wholly arbitrary choice of a particular generalized inverse is therefore nothing more than a disguised (but nonetheless arbitrary) selection of $p-m$ parameter restrictions.

Consider for example the generalized inverse

$$(X'X)^{-} = \begin{pmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

where X_1 consists of m linearly independent members of X.[1] Here

$$\hat{\beta} = (X'X)^{-}X'Y = \begin{bmatrix} (X_1'X_1)^{-1}X_1'Y \\ 0 \end{bmatrix}$$

is identical to the "estimate" obtained by directly assuming that the coefficients of X_2 equal zero.

If, instead, the previously described conventions are used to select a unique generalized inverse, then it can be shown that this will be

$$(X'X)^{+} = \sum_{i=1}^{m} \frac{1}{\lambda_i} a_i a_i'$$

where λ_i and a_i are the nonzero characteristic roots and associated orthonormal characteristic vectors of $X'X$. Using this generalized inverse to obtain estimates of all p elements of β is precisely equivalent to transforming X to its principal components and then equating to zero the coefficients (linear combinations of the elements of β)[2]

---

[1] Proof that this is a generalized inverse:

$$(X'X)(X'X)^{-}X'Y = \begin{bmatrix} X_1'Y \\ X_2'X_1(X_1'X_1)^{-1}X_1'Y \end{bmatrix} = \begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix}$$

since $X_2 = X_1H = X_1(X_1'X_1)^{-1}X_1'X_2$.

[2] In the two variable case, one would be assuming that $\beta_1 = \beta_2$ if there is a perfect positive correlation and assuming that $\beta_1 = -\beta_2$ if there is a perfect negative correlation.

of the p-m components with zero characteristic roots. This approach has been recommended by Massy as a principal components approach to perfect collinearity, and is discussed in more detail in Section III.

For now, it suffices to reiterate that if X is of rank m, then only m linearly independent coefficients can be estimated, and the estimation of all p elements of β requires p-m additional independent restrictions on the parameters. The deletion of linear combinations of variables or the use of generalized inverses to obtain complete estimates of β are no more than ad hoc techniques for imposing p-m arbitrary parameter restrictions. The only possible advantage of such techniques is that by disguising the parameter restrictions one can avoid making a decision about their reasonableness.

III. Imperfect Collinearity

Consider now the case where some of the explanatory variables are highly, though not perfectly, correlated. Remembering the ambiguities of the concept, we might say that p-m of the characteristic roots of the moment matrix are close to zero, and that while X is of full rank it has only m significant dimensions. It is consequently tempting to work with only m explanatory variables.

For generality, we can rewrite X in terms of a p-dimensional basis F

$$\underset{T\times p}{X} = XA^{-1}A = \underset{T\times p}{F}\ \underset{p\times p}{A}$$

which can be substituted

$$Y = X\beta + \epsilon = FA\beta = \epsilon + F\gamma + \epsilon$$

If Y is regressed upon all p columns of F, then the implicit estimates of the parameters associated with any such basis will be independent of the basis used for estimation, and the forecasts of Y will also be invariant to the basis used for forecasting regardless of whether or not the in-sample collinearity patterns continue. These unconstrained estimates and forecasts will be labeled here as $\hat{\gamma}^u$ and $\hat{Y}^u$:

$$\hat{\gamma}^u = (F'F)^{-1}F'Y$$

$$\hat{Y}^u = F\hat{\gamma}^u = XA^{-1}\hat{\gamma}^u = X\hat{\beta}^u \ .$$

Any such p-dimensional basis can be partitioned

$$X = XA^{-1}A = X[\underset{pxm}{H_1} \vdots \underset{pxp-m}{H_2}]A$$

$$= \underset{Txm}{[F_1} \vdots \underset{Txp-m}{F_2}] \begin{bmatrix} \underset{mxp}{A_1} \\ \underset{p-mxp}{A_2} \end{bmatrix} = F_1A_1 + F_2A_2$$

and $p-m$ dimensions put into the error term

$$Y = F_1A_1\beta + F_2A_2\beta + \varepsilon$$

$$= \underset{Txm}{F_1} \ \underset{mx1}{\gamma_1} + [\underset{Txp-m}{F_2} \ \underset{p-mx1}{\gamma_2} + \varepsilon]$$

$$= (XH_1)\gamma_1 + u \ .$$

Regressing Y on F_1 will yield estimates of the m elements of γ_1

$$\hat{\gamma}_1^c = (F_1'F_1)^{-1}F_1'Y \ .$$

However, these cannot be converted into estimates of the p elements of β without the introduction of $p-m$ independent parameter restrictions.

Using $\hat{\gamma}_1^c$ alone to forecast Y

$$\hat{Y}^c = X^0 H_1 \hat{\gamma}_1^c$$

is equivalent to imposing the $p-m$ restrictions that $A_2\beta = 0$ and thereby obtaining

$$\hat{\beta}^c = H_1 \hat{\gamma}^c$$

as a constrained estimate of β. Thus, $p-m$ parameter restrictions may be imposed upon a model by working with m linear combinations of p explanatory variables. Conversely, working with m linear combinations of p explanatory variables is precisely equivalent to imposing $p-m$ parameter restrictions and can be fruitfully evaluated on that basis.

For more generality, we can consider the imposition of nonzero parameter constraints

$$A_2\beta = \delta .$$

Now

$$Y - F_2\delta = F_1\gamma_1 + [\epsilon + F_2(\gamma_2 - \delta)]$$

and the constrained estimator is

$$\begin{aligned}\hat{\gamma}_1^c &= (F_1'F_1)^{-1}F_1'(Y - F_2\delta) \\ &= \gamma_1 + (F_1'F_1)^{-1}F_1'\epsilon + (F_1'F_1)^{-1}F_1'F_2(\gamma_2 - \delta) .\end{aligned}$$

The mean squared errors for the constrained and unconstrained estimators are displayed here as the diagonal elements in the format

$$E(\hat{\gamma}-\gamma)(\hat{\gamma}-\gamma)' = [E(\hat{\gamma}) - \gamma][E(\hat{\gamma}) - \gamma]' + E[\hat{\gamma} - E(\hat{\gamma})][\hat{\gamma} - E(\hat{\gamma})]'$$

or

$$MSE(\hat{\gamma}) = [Bias(\hat{\gamma})][Bias(\hat{\gamma})]' + Var(\hat{\gamma}) \ .$$

Thus,

$$MSE(\hat{\gamma}_1^c) = G'(\delta - \gamma_2)(\delta - \gamma_2)'G + \sigma^2(F_1'F_1)^{-1}$$

$$MSE(\hat{\gamma}_2^c) = (\delta - \gamma_2)(\delta - \gamma_2)' + 0$$

where $G = F_2'F_1(F_1'F_1)^{-1}$. In contrast, for the unconstrained OLS estimates,

$$MSE(\hat{\gamma}_1^u) = 0 + \sigma^2[(F_1'F_1)^{-1} + G'(F_2'\tilde{P}_1F_2)^{-1}G]$$

$$MSE(\hat{\gamma}_2^u) = 0 + \sigma^2[F_2'\tilde{P}_1F_2]^{-1}$$

where $\tilde{P}_1 = I - F_1(F_1'F_1)^{-1}F_1$.

Since $\tilde{P}_1$ is idempotent, $Var(\hat{\gamma}_1^u) - Var(\hat{\gamma}_1^c)$ is positive semi-definite and consequently has nonnegative diagonal elements. Thus, the imposition of exact parameter restrictions unambiguously reduces (or leaves unchanged) the variance of the estimate of each coefficient.[3] Notice, however, that the size of this variance reduction is completely independent

[3]This is true of all estimable coefficients, since

$$Var(\widehat{Z\gamma}^u) - Var(\widehat{Z\gamma}^c) = Z[Var(\hat{\gamma}^u) - Var(\hat{\gamma}^c)]Z'$$

is positive semi-definite.

of the specific values chosen for the elements of δ . In particular, deleting a variable from an equation does not reduce the variances of the estimates any more (or less) than does constraining the associated coefficient to equal some nonzero value.

Where the choice of values for δ shows up is in the extent to which the estimates are biased. If the selected δ is close to the true $A\beta_2$, then the errors in the restrictions may be offset by the reduced variances. Unfortunately this comparison is complex, ambiguous, and critically dependent upon the unknown $A\beta_2$. Looking at the difference in mean squared errors for the constrained and unconstrained estimates,

$$MSE(\hat{\gamma}_1^u) - MSE(\hat{\gamma}_1^c) = G'[MSE(\hat{\gamma}_2^u) - MSE(\hat{\gamma}_2^c)]G$$

$$MSE(\hat{\gamma}_2^u) - MSE(\hat{\gamma}_2^c) = \sigma^2[F_2'\tilde{P}_1F_2]^{-1} - (\delta - \gamma_2)(\delta - \gamma_2)' .$$

The MSE's for the constrained case are minimized by exactly correct restrictions $(\delta = \gamma_2)$, and in this happy situation the constrained estimates dominate the unconstrained estimates.

In the case of a single restriction,[4] we have the additional satisfying result that a decrease in $(\delta - \gamma_2)$ reduces (or leaves unchanged) the MSE's of each of the remaining coefficients, and that the constrained estimates are either unambiguously superior or inferior to the unconstrained

---

[4]

$$MSE(\hat{\gamma}_1^u) = \sigma^2(F_1'F_1)^{-1} + MSE(\hat{\gamma}_2^u)G'G$$

$$MSE(\hat{\gamma}_1^c) = \sigma^2(F_1'F_1)^{-1} + MSE(\hat{\gamma}_2^c)G'G$$

where $MSE(\hat{\gamma}_2^u)$ and $MSE(\hat{\gamma}_2^c)$ are scalars and $G'G$ is positive semi-definite.

estimates, depending solely upon whether or not $(\delta - \gamma_2)^2$ is smaller than the MSE of the unconstrained estimate of γ_2. Notice that again $\delta = 0$ has no special virtue, and will be thoroughly inferior to a more accurate restriction.

With more than one restriction, we have the possibility that some incorrect restrictions may offset others. Thus,

$$MSE(\hat{\gamma}_1^u) - MSE(\hat{\gamma}_1^c) = G'[MSE(\hat{\gamma}_2^u) - MSE(\hat{\gamma}_2^c)]G$$

will be semi-definite if $MSE(\hat{\gamma}_2^u) - MSE(\hat{\gamma}_2^c)$ is semi-definite. However, the diagonal elements of this latter matrix may all be nonnegative or non-positive without the matrix being semi-definite. It is consequently possible to have smaller mean squared errors for each of the elements of γ_2 and yet have larger mean squared errors for each of the elements of γ_1. In other words, if more than 1 constraint is incorrect, then the effectiveness of the constraints in reducing the mean squared errors of the remaining estimates depends upon the entire matrix, $MSE(\hat{\gamma}_2)$, and not solely upon the diagonal elements. This is true both of comparisons between alternative sets of constraints and of the choice between constrained and unconstrained estimates.

One possible avenue for simplifying the analysis is to specify a loss function with the MSE's as arguments. The comparison would then be between scalars and an estimator would not have to have smaller MSE for every coefficient in order to be deemed superior. A natural loss function arises from putting the problem in a forecasting context, with forecasting inaccuracy measured by the expected value of the mean squared error of the out-of-sample predictions

$$EMSE(\hat{Y}^0) = E[(\hat{Y}^0 - Y^0)'(\hat{Y}^0 - Y^0)/n] \ .$$

This is appropriate for a quadratic loss function and is consistent with the minimization of squared residuals in the in-sample parameter estimation. With the usual assumptions (presented earlier),

$$EMSE(\hat{Y}^0) = \sigma_0^2 + \frac{1}{n}Tr[F^0 MSE(\hat{\gamma})F^{0'}]$$

$$= \sigma_0^2 + \frac{1}{n}Tr[MSE(\hat{\gamma})F^{0'}F^0] \ .$$

This is of course no more than a format without further specification of $F^{0'}F^0$. However, none of the obvious options for specifying $F^{0'}F^0$ are very attractive: observing the actual $F^{0'}F^0$ before choosing an estimator (with potentially a different estimator of γ used for each out-of-sample forecast of Y); using $F'F$ as an estimator of $F^{0'}F^0$; or building a theoretical model as to how $F^{0'}F^0$ is generated.

Even if one of these expediencies for $F^{0'}F^0$ were adopted, there would remain the indeterminacy of $MSE(\hat{\gamma}^c)$ due to the fact that γ_2 is unknown. That is, the EMSE for the restricted estimator will depend upon the true values of γ_2 , or where we are in this p-m dimensional parameter space. One of the advantages of a linear unbiased estimator such as unconstrained OLS is that the mean squared errors of the parameter estimates (and hence EMSE) do not depend upon the true values of the parameters.

$$EMSE(\hat{Y}^u) = \sigma_0^2 + \frac{\sigma^2}{n}Tr(X'X)^{-1}(X^{0'}X^0)$$

We can consequently obtain some simple yet suggestive values for the EMSE for the unconstrained estimates by considering a few specific assumptions for $F^{0'}F^0$.

In particular, if the in-sample moment matrix is replicated out-of-sample,

$$\frac{1}{n}X^{0\prime}X^{0} = \frac{1}{T}X'X$$

Then

$$EMSE(\hat{Y}^{u}) = \sigma_0^2 + P\frac{\sigma^2}{T}$$

And one would think that normally most of this error would be due to the irreducible out-of-sample variance of the disturbance term. Thus, irrespective of how highly correlated the explanatory variables were, there would be little need for parameter restrictions if these correlations hold up in the forecasting period.

If, on the other hand, the variances of the explanatory variables are maintained while the covariances evaporate, then

$$\frac{1}{n}X^{0\prime}X^{0} = \frac{1}{T}\Sigma$$

where Σ is a diagonal matrix with the i^{th} diagonal element equal to the in-sample sum of squares of the i^{th} explanatory variable. In this case,

$$EMSE(\hat{Y}^{u}) = \sigma_0^2 + \frac{\sigma^2}{T}\sum_{i=1}^{P}\frac{1}{\lambda_i}$$

where λ_i are the characteristic roots in the in-sample correlation matrix for the explanatory variables, with $\Sigma\lambda_i = P$. If the explanatory variables are highly collinear in-sample, then some of these roots will

be close to zero and one should not be surprised by extremely large mean squared errors for out-of-sample forecasts made when this collinearity does not continue to hold. In this situation, it is clear that reasonably accurate parameter restrictions may be very useful; however a detailed comparison is too complex to really be interesting. Such a comparison is instead made below for the simple cases of a principal components approach and a two-variable example.

Pretest Estimators

An extensive literature has developed concerning the use of hypothesis tests to choose between (or to construct a weighted average of) constrained and unconstrained estimates. Some of the more frequently referenced articles are Bancroft, Bock et al., Chipman, Sclove, and Toro-Vizcarrondo and Wallace. Generally these articles take a particular loss function or risk matrix and a given set of potential parameter constraints, demonstrate that unconstrained estimates are not dominant over the entire parameter space, and explore the effects of different significance levels. Usually absent from these articles are the recognition that there are an unlimited variety of parameter restrictions available and the notion that the decision rule should depend upon the confidence one attaches to the possible constraints. The reader consequently learns that a constrained pretest estimator may be successful, but has no idea of how to optimally exploit this possibility. In this section, I will argue that the ambiguities of the pretest procedures vitiate their usefulness and that their net effect has seemingly been to give respectability to unjustifiable procedures.

To review briefly, if β is viewed as fixed, then there is no linear estimator which has a smaller[5] risk matrix $E(\hat{\beta}-\beta)(\hat{\beta}-\beta)'$ for all

possible values of β . There are, however, constrained and pretest estimators which are superior to OLS for some parts of the parameter space and inferior elsewhere. One substantial advantage of unconstrained OLS is, that being linear unbiased, the risk does not depend upon the true value of β . If β is viewed as random, then unconstrained OLS is the linear estimator which minimizes the risk matrix when the prior variances on β become infinite.

There are nonlinear Stein-James estimators which dominate least squares for the case of at least three orthonormal regressors and a scalar mean squared error loss function, $E(\hat{\beta}-\beta)'(\hat{\beta}-\beta)$. Unfortunately these results have not yet been extended to risk matrices for the general regression case. In addition, the extent of the dominance depends upon the true values of the parameters and the null hypotheses and test-levels that are utilized. As a consequence, even where operational the optimal Stein-James estimator requires a priori information.

The pretest literature generally takes note of the high variances associated with the estimates of the coefficients of highly collinear explanatory variables and argues that for the seemingly reasonable MSE criteria it may be worthwhile to impose admittedly incorrect parameter restrictions in that the biases that are introduced may be more than offset by the reduced variances. It is then explicitly demonstrated that unconstrained estimates can be bested by constrained and pretest estimates. The source of the parameter constraints is rarely discussed and much of the literature conveys the impression that a priori information is unnecessary.

---

[5] "Smaller" and "minimum" are shorthand for the difference between the risk matrices being positive semi-definite.

Even worse, the practical impact of this literature seems to have been to rationalize the common procedure of "omitting insignificant variables" by assuming that a coefficient is zero whenever that value is not rejected by the data.

One problem with this common practice is that there is rarely a predetermined single null hypothesis or nested set of hypotheses. Instead the model builder looks for hypotheses which will be accepted, and then mistakenly equates a number of single tests to a joint test or proceeds sequentially in some ad hoc fashion. If each coefficient is tested in a search for parameters for which zero cannot be rejected, then the probability of incorrectly rejecting at least one null hypothesis will be larger than the probability of rejecting each particular null hypothesis--if each test is conducted at the 5% level, then the probability of incorrectly including at least one variable will be greater than .05, though less than .05 multiplied by the number of tests. Similarly, the probability of committing at least one Type II error by incorrectly excluding a variable will also be increased. While it is clear that one should take into account the fact that more than one test will be conducted, it is not obvious in which direction the procedure should be modified; i.e., whether each test should be conducted at a higher or lower significance level. The usual modification seems to be to lower the probability of Type I error for each test, which increases the probability of incorrectly excluding a variable.

A theoretically motivated adjustment would require a resolution of the more basic question of the source of the null hypotheses that are tested. Typically, only a subset of all possible parameters are tested, namely those which are explicitly estimated. It is well-known that linear

rearrangements of the explanatory variables will have no effect on the implicit estimates of any estimable parameters nor on the variances of these estimates. Thus hypothesis tests of any estimable parameters are invariant to the way in which the explanatory variables are arranged for estimation purposes. There will always be an infinite number of combinations of coefficients for which zero cannot be rejected as a null hypothesis; and which (if any) of these hypotheses are tested typically depends solely upon the arbitrary choice of which parameters to explicitly estimate. That is, a model can always be simplified (down to one explanatory variable in fact) but how far and in what ways it is simplified by the pretest procedure depends upon a series of arbitrary choices.

Similarly (or equivalently if the intercept is introduced into the shuffling of explanatory variables), it must be asked why zero is always given the presumptive weight of being the null hypothesis. Since any point in the confidence region would be accepted if it were tested as a null hypothesis, one should wonder why the origin is to be preferred over all other unrejected hypotheses, and in particular why it should be preferable to the unconstrained point estimate. We've seen that a zero constraint has no advantage over a nonzero constraint in terms of variance reduction and consequently must be justified in terms of a smaller bias or an offset to the bias in other restrictions. But this is a comparison which necessarily requires a priori information. A major inadequacy of the usual pretest procedure is consequently that it takes no account of a priori information in choosing the constraints that are to be considered and takes no account of the confidence one has in the constraints when deciding whether or not to impose them.

A further problem with the usual pretest procedure is that it does

not accurately divulge which parameters are most in need of a priori information. Thus the fact that zero is inside a confidence band neither implies that the coefficient is close to zero nor that the band is wide. It is the variance of the estimate or the width of the band (rather than the location of the band) that most accurately reflects the informational content of the data and the susceptibility of the estimate to a priori information. More specifically, it is only when zero is outside a small confidence band that the pretestor makes a generally defensible decision, though even here firmly held a priori beliefs might have been useful. If zero is instead outside a large confidence band, then the pretestor does nothing when a restriction may be badly needed and the data is very receptive to even vague information. When zero is inside a tight band, the pretestor overrules informative data with an ad hoc restriction; and when zero is inside a wide band he resorts to adhockery when information would be welcome.

Finally, the pretestor ignores the covariances in his estimates and in his a priori information. We've seen that errors in the constraints may multiply or may cancel each other out, so that unambiguously closer restrictions may give unambiguously less accurate estimates. Thus a primary lesson of the Bayesian approach is the importance of covariances and the inadequacies of a parameter-by-parameter approach to model improvement (see for example Leamer).

Principal Components

A linear transformation of the explanatory variables will in general change the eigenvalues, eigenvectors, and principal components. For this reason, the explanatory variables are often standardized to have a mean

of zero and unit variance--i.e., here working with $X_i/\sqrt{T}\sigma_{X_i}$. Although such a transformation will not affect the unconstrained estimates or forecasts, our previous notation will now (for convenience) refer to standardized variables for both the constrained and unconstrained procedures. It should be borne in mind though that the constrained estimates will depend critically upon the essentially arbitrary choice of which linear combinations of explanatory variables to standardize.

If the rows of A are the p orthonormal eigenvectors of $X'X$, then

$$A(X'X)A' = D = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}, \quad \sum_{i=1}^{p} \lambda_i = p$$

where D is a diagonal matrix whose i^{th} diagonal element is the eigenvalue of $X'X$ corresponding to the i^{th} row of A . The principal components of X

$$F = XA'$$

are an orthogonal basis for X

$$X = FA , \quad F'F = D$$

such that the m principal components with highest associated eigenvalues (i.e., variances) minimize the total sum of squared residuals for a regression of each of the p explanatory variables on m linear combinations of the explanatory variables.

It has frequently been argued on rather vague grounds that replacing X with its principal components and then discarding $p-m$ of these

components might "alleviate the collinearity problem" or yield a "considerable parsimony in analysis." That is, as in the previous section,

$$\begin{aligned} Y &= X\beta + \epsilon = XA'A\beta + \epsilon = F\gamma + \epsilon \\ &= F_1A_1\beta + [F_2A_2\beta + \epsilon] = F_1\gamma_1 + [F_2\gamma_2 + \epsilon] \\ &= XH_1\gamma_1 + [XH_2\gamma_2 + \epsilon] \end{aligned}$$

where again

$$\underset{Txp}{F} = [\underset{Txm}{F_1} \vdots \underset{Txp-m}{F_2}] \; ; \quad \underset{pxp}{A} = \begin{bmatrix} \underset{mxp}{A_1} \\ \underset{p-mxp}{A_2} \end{bmatrix} \; ; \quad A' = A^{-1} = [\underset{pxm}{H_1} \vdots \underset{pxp-m}{H_2}]$$

Kendall has perhaps been the most influential advocate of this approach and one of the few to note its equivalence to imposing the $p-m$ parameter restrictions, $A_2\beta = 0$. Thus, he explicitly considers $H_1\hat{\gamma}_1$ to be an estimate of β , which would be appropriate if one assumed that $A_2\beta = 0$.

Although this procedure has been generally viewed by economists with some suspicion or even confusion, it offers the great pedagogical advantage of orthogonality between the omitted and retained explanatory variables $(F_1'F_2 = 0)$. The estimated coefficients of the retained variables are consequently independent of the accuracy of the implicit restrictions on the coefficients of the omitted variables. This considerably simplifies the analysis and permits a much sharper focus on some important issues.

If all p components are retained, then

$$\begin{bmatrix} \hat{\gamma}_1 \\ \gamma_2 \end{bmatrix}^u = (F'F)^{-1}F'Y = D^{-1}F'(F\gamma + \epsilon) = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} D_1^{-1}F_1'\epsilon \\ D_2^{-1}F_2'\epsilon \end{bmatrix}$$

where $D_i^{-1} = (F_i'F_i)^{-1}$ is a diagonal matrix whose diagonal elements are the inverses of the characteristic roots associated with the principal components collected as F_i .

$$\text{MSE}\begin{bmatrix} \hat{\gamma}_1 \\ \gamma_2 \end{bmatrix}^u = 0 + \sigma^2 \begin{bmatrix} D_1^{-1} & 0 \\ 0 & D_2^{-1} \end{bmatrix} .$$

Thus the MSE of the coefficient of the i^{th} component is σ^2/λ_i .

If instead it is assumed that $\gamma_2 = \delta$ ($\delta = 0$ is equivalent to dropping the $p-m$ components collected as F_2), then the resultant constrained estimates are

$$\begin{bmatrix} \hat{\gamma}_1 \\ \gamma_2 \end{bmatrix}^c = \begin{bmatrix} (F'F)^{-1}F'(Y - F_2\delta) \\ \delta \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} D_1^{-1}F_1'\epsilon \\ \delta - \gamma_2 \end{bmatrix}$$

with

$$\text{MSE}\begin{pmatrix} \hat{\gamma}_1 \\ \gamma_2 \end{pmatrix}^c = \begin{bmatrix} 0 & 0 \\ 0 & (\gamma_2 - \delta)(\gamma_2 - \delta)' \end{bmatrix} + \sigma^2 \begin{bmatrix} D_1^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

Thus, the estimates of the coefficients of the retained components are unaffected by the constraints on the remaining components, and the relevant comparison for each component reduces to the simple question of whether the coefficient can be more accurately estimated or specified.

On the other hand, the estimates of β will generally depend upon

all of the estimated elements of γ :

$$\hat{\beta}^u = A'\hat{\gamma}^u = \beta + A'D^{-1}AX'\epsilon = \beta + (X'X)^{-1}X'\epsilon$$

$$\hat{\beta}^c = A'\hat{\gamma}^c = \hat{\beta}^u + H_2(\delta - \hat{\gamma}_2^u)$$

$$= \beta + H_2(\delta - \gamma_2) + H_1D_1^{-1}H_1'X'\epsilon$$

with

$$MSE(\hat{\beta}^u) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} + \sigma^2 A' \begin{bmatrix} D_1^{-1} & 0 \\ 0 & D_2^{-1} \end{bmatrix} A = \sigma^2 H_1 D_1^{-1} H_1' + \sigma^2 H_2 D_2^{-1} H_2'$$

$$MSE(\hat{\beta}^c) = A' \begin{bmatrix} 0 & 0 \\ 0 & (\gamma_2-\delta)(\gamma_2-\delta)' \end{bmatrix} A + \sigma^2 A' \begin{bmatrix} D_1^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

$$= \sigma^2 H_1 D_1^{-1} H_1' + H_2(\gamma_2-\delta)(\gamma_2-\delta)'H_2'$$

Thus, for each element of β , the unconstrained estimate has a smaller bias and larger variance than the constrained estimate. The difference in MSE's

$$MSE(\hat{\beta}^u) - MSE(\hat{\beta}^c) = H_2[\sigma^2 D_2^{-1} - (\gamma_2 - \delta)(\gamma_2 - \delta)']H_2'$$

is ambiguous even if each restriction on γ_2 is more accurate than the unconstrained estimate.

For out-of-sample forecasting,

$$EMSE(\hat{Y}^u) = \sigma_0^2 + \frac{1}{n}\sigma^2 Tr[(X'X)^{-1}X^{0'}X^0] + 0$$

$$EMSE(\hat{Y}^c) = \sigma_0^2 + \frac{1}{n}\sigma^2 Tr[D_1^{-1}F_1^{0'}F_1^0] + \frac{1}{n}Tr[(\gamma_2 - \delta)(\gamma_2 - \delta)'F_2^{0'}F_2^0] \ .$$

This consists of the irreducible out-of-sample variance of the disturbance term, the effects of the errors in the estimated coefficients, and the effects of the errors in the imposed parameters constraints.

If the out-of-sample variances and covariances among the explanatory variables are identical to the in-sample ones

$$\frac{1}{n}X^{0\prime}X^{0} = \frac{1}{T}X'X$$

then

$$EMSE(\hat{Y}^{u}) = \sigma_0^2 + p\sigma^2/T$$

$$EMSE(\hat{Y}^{c}) = \sigma_0^2 + m\sigma^2/T + \frac{1}{T}Tr(\gamma_2 - \delta)(\gamma_2 - \delta)'D_2 \ .$$

Thus, if the correlation matrix for the explanatory variables is replicated out-of-sample, we are directly trading off the fewer estimated parameters against the inaccuracy of the restrictions, and these restrictions will normally have to be fairly accurate if the constrained model is to outforecast the unconstrained model.

If, on the other hand, the explanatory variables are uncorrelated out-of-sample but replicate the in-sample variances[6]

$$X^{0\prime}X^{0} = \frac{n}{T}I_P \ ,$$

then

[6]This suggests that the analysis would be considerably simplified if the initial standardization were of variables whose variances could be expected to change little.

$$EMSE(\hat{Y}^u) = \sigma_0^2 + \frac{\sigma^2}{T}\sum_{i=1}^{p}\frac{1}{\lambda_i}$$

$$EMSE(\hat{Y}^c) = \sigma_0^2 + \frac{\sigma^2}{T}\sum_{i=1}^{m}\frac{1}{\lambda_i} + \frac{1}{T}Tr(\gamma_2 - \delta)(\gamma_2 - \delta)'$$

where the m characteristic roots in the second expression are those associated with the retained components. Thus, when the in-sample multicollinearity evaporates out-of-sample, the deletion of p-m components with eigenvalues close to zero can easily be profitable even if the implicit parameter restrictions are only vaguely accurate. However, there is no limit to the potential inaccuracy of the restrictions and hence to the EMSE for the constrained forecasts. Thus, the benefits of purely ad hoc constraints are necessarily accidental. On the other hand, it does seem sensible to try to obtain parameter estimates which will provide reasonably accurate forecasts in a wide variety of situations. This implies that one should try to find reasonably accurate values for all of the coefficients in the model and not rely upon fortuitous interrelations among the explanatory variables to render some parameters unimportant.

After standardizing the parameters, an examination of the eigenvalues and eigenvectors of the moment matrix can indicate the effective dimensionality of the explanatory variables and thereby the minimum number of a priori parameter restrictions that are required for robust forecasts. The eigenvectors associated with the smallest eigenvalues will indicate those parameters for which the data provides the least information and which are consequently most receptive to a priori information. However, mechanically constraining these parameters to be zero (by routinely deleting the components associated with the smallest eigenvectors) is not

very attractive. It certainly seems preferable to consider nonzero constraints and to give some thought to what are a priori reasonable values for the parameters.[7] Unfortunately these parameters will generally be linear combinations of all p of the original β_i coefficients, and it may be quite inconvenient or even fruitless to try to imagine plausible values for the particular linear combinations which are the ripest candidates for constraints. While arbitrarily assigning values such as zeroes may occasionally be successful, it is not a dependable or defensible technique. This is particularly clear in the case where zeroes are rejected by the data in the sense of being outside a confidence region for the parameters. That is, since a large variance σ^2/λ_i does not rule out the estimated coefficient being large also, components that are of little use in explaining the explanatory variables may be very powerful in explaining the dependent variable (see Hotelling). This has led Massy and others to advocate the deletion of components which are statistically insignificant (i.e., for which zero is inside a confidence region). If, however, there is no a priori weight behind the selection of zero values, then there is no justification for mechanically selecting zeroes from the confidence region and in particular no reason for prefering zeroes to those values which maximize the likelihood function. Further, this approach permits one to impose ad hoc constraints where the data is very informative (when zero is inside a tight band) and to refrain when constraints are desperately needed (when zero is outside a large band). What is needed then is a technique in which a priori (and not ad hoc) restric-

[7] The mechanical deletion of principal components has the added discomfort that the eigenvectors and hence the implicit parameter restrictions depend upon the arbitrarily selected units for the explanatory variables.

tions are imposed on those parameters for which the data contains the least information.

Since it is often most convenient to think of a priori values for the original coefficients, it is tempting to restrict one's attention to these parameters. This would suggest examining which of the estimated parameters have the largest variances, or equivalently (as per Farrar and Glauber) which of the original explanatory variables are most highly correlated with the remaining explanatory variables as a group. Since these variances depend upon the units of the explanatory variables, it would be convenient to normalize them by choosing units for the parameters so as to equalize the precision (i.e., the subjectively specified squared error) for the proposed restrictions.

On the other hand, there may be cases where one will feel more comfortable in specifying linear combination of the β_i ; e.g., specifying $\beta_1 + \beta_2$ or $\beta_1 - \beta_2$ rather than simply β_1 or β_2 . In this situation, it would make sense to arrange the explanatory variables so that one is explicitly estimating those parameters that one would be most willing to specify a priori. By looking at the directly estimated variances (or by running the Farrar and Glauber tests) one could then see which of the potential restrictions are the most welcome in the sense of the data providing the least amount of information. However, this procedure permits only inflexible exact restrictions, ignores the covariances[8] on the priors and the estimates (and thus the possibility of errors compounding or cancelling one another out), and leaves as a vague judgmental decision

[8]The variables could be transformed so that either the sample or prior information is orthogonal, but it would be very surprising if these transformations coincided.

the weighing of the information contained in the data relative to the confidence one has in the potential constraints, unless one is willing to take the last Bayesian step and quantify the precision of one's priors by specifying a probability distribution for the parameter space. Having gone this far, that last step should be very appealing.

A Two-Variable Example

Consider the model

$$\text{in-sample:} \quad Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

$$\text{out-of-sample:} \quad Y^0 = \beta_1 X_1^0 + \beta_2 X_2^0 + \epsilon$$

where

$$\begin{pmatrix} X_1' \\ X_2' \end{pmatrix} (X_1 \quad X_2) = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} ; \quad \begin{pmatrix} X_1^{0'} \\ X_2^{0'} \end{pmatrix} (X_1^0 \quad X_2^0) = \begin{pmatrix} 1 & r_0 \\ r_0 & 1 \end{pmatrix}$$

and

$$E \begin{pmatrix} \epsilon \\ \epsilon^0 \end{pmatrix} \begin{pmatrix} \epsilon \\ \epsilon^0 \end{pmatrix}' = \begin{bmatrix} \sigma^2 I & 0 \\ 0 & \sigma_0^2 I \end{bmatrix} .$$

The eigenvalues and orthonormal eigenvectors are

$$D = \begin{pmatrix} 1 + r & 0 \\ 0 & 1 - r \end{pmatrix} \qquad A = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} .$$

The principal components approach would rearrange the model as

$$Y = \left(\frac{X_1 + X_2}{\sqrt{2}}\right)\left(\frac{\beta_1 + \beta_2}{\sqrt{2}}\right) + \left(\frac{X_1 - X_2}{\sqrt{2}}\right)\left(\frac{\beta_1 - \beta_2}{\sqrt{2}}\right) + \varepsilon$$

and then consider dropping one of the components, which would be equivalent to assuming that $\beta_1 + \beta_2 = 0$ or that $\beta_1 - \beta_2 = 0$. Table I displays the consequences of retaining both components, assuming that $\beta_1 + \beta_2 = \sqrt{2}\ \delta_2$, assuming that $\beta_1 - \beta_2 = \sqrt{2}\ \delta_1$, or assuming that $\beta_2 = b_2$.

The EMSE for the unconstrained method becomes very large as r approaches ± 1, unless r_0 is close to r. This suggests that there is a great need for parameter restrictions when the explanatory variables are highly correlated, unless one is willing to rely on that collinearity persisting. The effectiveness of the parameter restrictions that are imposed will depend upon how accurate they are, how difficult the constrained parameters are to estimate, and how important it is in the out-of-sample period to have accurate values for the constrained parameters. Consider, for instance, the deletion of a principal component when r_0 is not close to r. If r is close to 1, then deleting the component associated with $\lambda_2 = 1-r$ can be extremely beneficial while deleting the component associated with $\lambda_1 = 1+r$ will be of little help; for r close to -1, the situation is reversed. The usefulness of the first deletion will however require that β_1 be fairly close to β_2, while the value of the second depends upon β_1 being close to $-\beta_2$. It is clear that an a priori consideration of non-zero restrictions can greatly improve the potential usefulness of the restrictions.

In the two-variable case, the only effect on EMSE of imposing $\beta_1 = b_1$ rather than $\beta_2 = b_2$ is to replace $(\beta_2 - b_2)^2$ with $(\beta_1 - b_1)^2$.

TABLE I

Estimation Equation	Effective Constraint	EMSE ($\hat{Y}$)		
		In General	When $r_0 = r$	When $r_0 = 0$
$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$	None	$\sigma_0^2 + \frac{2\sigma^2}{T}\frac{1 - rr_0}{1 - rr}$	$\sigma_0^2 + \frac{2\sigma^2}{T}$	$\sigma_0^2 + \frac{2\sigma^2}{T}\frac{1}{1-r^2}$
$Y - F_2\delta_2 = F_1\gamma_1 + U_1$ (F_2 = principal component associated with $\lambda_2 = 1-r$)	$\frac{\beta_1 - \beta_2}{\sqrt{2}} = \delta_2$	$\sigma_0^2 + \frac{\sigma^2}{T}\frac{1+r_0}{1+r} + \frac{1-r_0}{T}\left[\frac{\beta_1-\beta_2}{\sqrt{2}} - \delta_2\right]^2$	$\sigma^2 + \frac{\sigma^2}{T} + \frac{1+r}{T}\left[\frac{\beta_1-\beta_2}{\sqrt{2}} - \delta_2\right]^2$	$\sigma^2 + \frac{\sigma^2}{T}\frac{1}{1+r} + \frac{1}{T}\left[\frac{\beta_1-\beta_2}{\sqrt{2}} - \delta_2\right]^2$
$Y - F_1\delta_1 = F_2\gamma_2 + U_2$ (F_1 = principal component associated with $\lambda_1 = 1+r$)	$\frac{\beta_1 + \beta_2}{\sqrt{2}} = \delta_1$	$\sigma_0^2 + \frac{\sigma^2}{T}\frac{1-r_0}{1-r} + \frac{1+r_0}{T}\left[\frac{\beta_1+\beta_2}{\sqrt{2}} - \delta_1\right]^2$	$\sigma_0^2 + \frac{\sigma^2}{T} + \frac{1+r}{T}\left[\frac{\beta_1+\beta_2}{\sqrt{2}} - \delta_1\right]^2$	$\sigma_0^2 + \frac{\sigma^2}{T}\frac{1}{1-r} + \frac{1}{T}\left[\frac{\beta_1+\beta_2}{\sqrt{2}} - \delta_1\right]^2$
$Y - X_2b_2 = X_1\beta_1 + U_3$	$\beta_2 = b_2$	$\sigma_0^2 + \frac{\sigma^2}{T} + \frac{1 + r^2 - 2rr_0}{T}(\beta_2 - b_2)^2$	$\sigma_0^2 + \frac{\sigma^2}{T} + \frac{1-r^2}{T}(\beta_2 - b_2)^2$	$\sigma_0^2 + \frac{\sigma^2}{T} + \frac{1+r^2}{T}(\beta_2 - b_2)^2$

r = in-sample correlation between X_1 and X_2

r_0 = out-of-sample correlation between X_1 and X_2

Thus, the choice between constraining β_1 or β_2 depends only upon the accuracy of the constraint. Although it is camoflauged by the use of standardized variables, this comparison is influenced by the variances of the underlying explanatory variables. That is, if we represent the unstandardized (deviations from mean) explanatory variable by Z_i and the associated coefficient by b_i, then

$$\beta_i X_i = \sqrt{T}\sigma_{Z_i} b_i \left[\frac{Z_i}{\sqrt{T}\sigma_{Z_i}} \right]$$

and, for given accuracy in specifying b_i, the larger the variance of Z_i the greater will be the level of inaccuracy in specifying β_i. This is but a reflection of the fact that the greater the variance of an explanatory variable, the more accurate a restriction on the associated coefficient will have to be in order to be helpful.

REFERENCES

Bancroft, T. A. "On Biases in Estimation Due to the Use of Preliminary Tests of Significance," Annals of Mathematical Statistics, 15 (1944), 190-204.

_____. "Analysis and Inference for Incompletely Specified Models Involving the Use of Preliminary Tests of Significance," Biometrics, 20 (1964), 427-442.

Bock, M.; T. Yancey; and G. Judge. "The Statistical Consequences of Preliminary Test Estimators in Regression Analysis," Journal of the American Statistical Association, 68 (1973), 109-116.

Chipman, J. S. "On Least Squares with Insufficient Observations," Journal of the American Statistical Association, 59 (1964), 1078-1111.

Farrar, D. and R. Glauber. "Multicollinearity in Regression Analysis: The Problem Revisited," Review of Economics and Statistics, 49 (1967), 92-107.

Haitovsky, Y. "Multicollinearity in Regression Analysis: Comment," Review of Economics and Statistics, 51 (1969), 486-489.

Hotelling, H. "The Relation of the Newer Multivariate Statistical Methods to Factor Analysis," British Journal of Statistical Psychology, 10 (1957), 69-79.

Johnston, J. Econometric Methods, second edition. New York: McGraw-Hill, 1972.

Kendall, M. A Course in Multivariate Statistical Analysis, third edition. New York: Hafner, 1965.

Leamer, E. "Multicollinearity: A Bayesian Interpretation," Review of Economics and Statistics, 55 (1973), 371-380.

Massy, W. "Principal Components in Exploratory Statistical Research," Journal of the American Statistical Association, 60 (1965), 234-256.

Raduchel, W. "Multicollinearity Once Again," Harvard Institute of Economic Research, Discussion Paper No. 205 (1971).

Sclove, S. L. "Improved Estimators for Coefficients in Linear Regression," Journal of the American Statistical Association, 63 (1968), 596-606.

Swamy, P. A. V. D.; G. S. Maddala; and J. Holmes. "Use of Undersized Samples in the Estimation of Simultaneous Equation Systems," Discussion Paper No. 45, Department of Economics, SUNY at Buffalo (1969).

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, 63 (1968), 558-572.