The Claremont Center for Economic Policy Studies

Working Paper Series

Department of Economics
The Claremont Graduate School
Claremont, California 91711-6165

The Claremont Colleges:

The Claremont Graduate School; Claremont McKenna College;
Harvey Mudd College; Pitzer College; Pomona College; Scripps College

The Center for Law Structures
The Lowe Institute of Political Economy

The Claremont Center for Economic Policy Studies

Working Paper Series

"The Trouble with Multicollinearity Measures"

by

Gary Smith
Pomana College

Department of Economics
The Claremont Graduate School
Claremont, California 91711-6165

The Claremont Colleges:

The Claremont Graduate School; Claremont McKenna College;
Harvey Mudd College; Pitzer College; Pomona College; Scripps College

The Center for Law Structures
The Lowe Institute of Political Economy

"The Trouble with Multicollinearity Measures"

by

Gary Smith
Pomona College

The Trouble with Multicollinearity Measures

Gary Smith
Fletcher Jones Professor of Economics
Pomona College

In the standard regression model

$$(1) \qquad \underset{t\times 1}{y} = \underset{t\times p}{X}\ \underset{p\times 1}{\beta} + \underset{t\times 1}{\varepsilon}\ ,$$

t sample observations are used to calculate the least squares estimates $\hat{\beta} = (X'X)^{-1} X'y$. The "multicollinearity problem" is a standard section in econometrics textbooks and a commonplace lament in applied work. The familiar refrain is that highly intercorrelated data cause imprecise estimates and may justify special estimation procedures. For example, John Mandel [11, p. 15] observed that "Undoubtedly, the greatest source of difficulties in using least squares is the existence of 'collinearity' in many sets of data, and most of the modifications of the ordinary least squares approach are attempts to deal with the problem of collinearity."

In the multicollinearity literature, the precision of the parameter estimates is measured by their variances or, in the case of biased estimators, by their mean squared errors. A plethora of multicollinearity measures has accumulated to explain the imprecision of the estimates ([1] and [3]), to identify the kinds of additional data most needed to improve the precision of the estimates [16], and to motivate the use of biased estimation procedures ([3], [8], [13], and [18]).

In this paper I argue that the usual multicollinearity measures may

have some descriptive value, but should not be the basis for any substantive decisions, such as the adoption of special estimation procedures. The three fundamental difficulties are the interchangeability of high multicollinearity and low variation, the ambiguous choice of parameters to scrutinize, and the arbitrary scaling of multicollinearity barometers. The placement of the regression model in a forecasting context clarifies these difficulties and suggests a potential solution.

Reparameterizations

The model (1) can be reparameterized with any nonsingular matrix A:

$$y = (XA)(A^{-1}\beta) + \varepsilon = z\gamma + \varepsilon. \tag{2}$$

A new parameterization does not change a model; it is just a different way of writing the same model. No matter what the values of X and ε, exactly the same values of y are implied by representations (1) and (2). Nor does the parameterization affect the estimates obtained by ordinary least squares: the implicit estimates $\hat{\beta} = A\hat{\gamma}$ do not depend on the choice of A. The parameterization, then, is arbitrary and should not influence substantive decisions about the adequacy of the data and the need for alternatives to ordinary least squares. Yet, virtually all multicollinearity measures are sensitive to the model's parameterization.

One common reparameterization is to center the data by subtracting the sample mean from each explanatory variable. If the first explanatory variable is a vector of ones, then the model is rearranged from

$$y = \beta_1 + \beta_2 x_2 + \beta_2 x_3 + \ldots + \beta_p x_p \tag{3}$$

to

$$y = (\beta_1 + \beta_2 \bar{x}_2 + \beta_3 \bar{x}_3 + \ldots + \beta_p \bar{x}_p) + \beta_2 (x_2 - \bar{x}_2) + \beta_2 (x_3 - \bar{x}_3) + \ldots + \beta_p (x_p - \bar{x}_p). \tag{4}$$

It clearly should not make any difference whether we estimate equation (3) or (4). Similarly, there is considerable arbitrariness in arranging polynominal expressions {15}. The equation

$$y + \beta_1 + \beta_2 x + \beta_3 x^2$$

could be rearranged as

$$y = (\beta_1+c_1\beta_2+2c_1c_2\beta_3-c_2^2\beta_3) + (\beta_2+2c_2\beta_3)(x-c_1) + \beta_3(x-c_2)^2$$

or as

$$y = (\beta_1+c_1\beta_2+c_2\beta_3) + \beta_1(x-c_1) + \beta_2(x^2-c_2)$$

for any arbitrary constants c_1 and c_2. These are three equivalent representations of the same model.

There are also a variety of units in which most explanatory variables can be measured. Lengths can be measured in inches, feet, yards, miles, centimeters, rods, or furlongs. Weights can be measured in ounces, pounds, tons, grams, or pennyweights. Income can be measured in cents, dollars, or billions of dollars. Interest rates can be measured in hundredths, percents, or basis points. Some researchers divide each variable by its standard deviation in the particular sample at hand.

With a rescaling, the model (3) is rewritten as

$$(5) \qquad y = (\beta_1/\alpha_1)\ \alpha_1 + (\beta_2/\alpha_2)\ \alpha_2 x_2 + (\beta_3/\alpha_3)\ \alpha_3 x_3 + \ldots + (\beta_p/\alpha_p)\ \alpha_p x_p.$$

Again, it shouldn't make any difference whether the model is estimated as (3) or (5).

There are also often several plausible ways of including the explanatory variables in a model. As an economist, I'll use several economic examples. Consider, first, the consumption function, which is a key component of macroeconometric models. We might assume, as in Holmes {7}, that consumption c

depends on current income y and on permanent income y_P:

$$c = \beta_1 y + \beta_2 y_P + \varepsilon \tag{6}$$

where the difference between current and permanent income is transitory income, $y_T = y - y_P$. We could equally well write this model as

$$c = \beta_1 y_T + (\beta_1+\beta_2)\, y_P + \varepsilon, \tag{7}$$

or as

$$c = (\beta_1+\beta_2)\, y - \beta_2 y_T + \varepsilon. \tag{8}$$

Or, we could use other, less obvious, rearrangements of the explanatory variables, such as

$$c = (\beta_1-\beta_2)(.6y-.4y_P) + (2\beta_1+3\beta_2)(.2y+.2y_P) + \varepsilon. \tag{9}$$

Another specific example from economics is a model {4} in which the demand for a financial asset depends on its own rate of return and on the rate of return on at least one alternative asset. The economist can use the two rates as explanatory variables, or either rate and the yield differential between the two rates. Similarly, an interest rate and the rate of inflation are commonplace explanatory variables. The economist can, equivalently, use the rate of inflation and the real interest rate(which is the difference between the nominal interest rate and the inflation rate) or use the nominal and real interest rates.

Asset demand equations also often {4} include either the current and lagged values of wealth, lagged wealth and the change in wealth, or current wealth and the change in wealth. This arbitrary arrangement of the explanatory variables confronts every researcher who uses current and lagged data. Labor demand and supply equations might include the logarithms of either nominal wages and prices, real wages and prices, or nominal wages and real wages. Production functions could include the logarithms of either capital

and labor, labor and the capital-labor ratio, or capital and the labor-capital ratio. Consumption functions could include income and taxes, income and income minus taxes, or taxes and income minus taxes. Carl Christ {2} has argued persuasively that macroeconomic models should recognize the government budget constraint: expenditures minus taxes minus money issuance minus bond sales must equal zero. A reduced-form equation for national income should include any three of these policy variables. It doesn't matter which three, except for multicollinearity measures.

Multicollinearity measures often motivate modified estimation procedures including variable deletion, principal components, and ridge regression. But most multicollinearity measures depend on the centering, scaling, and arrangement of the explanatory variables. For some parameterizations, least squares will be retained; for other parameterizations, least squares will be abandoned. If the parameterization is chosen arbitrarily, then the estimation procedure will be, too.

High Multicollinearity or Low Variation?

In practice, simple pairwise correlations often catch researchers' attention. But statisticians scorn simple correlations, since a variable may be highly correlated with a group of variables even if it is not highly correlated with any one member of the group. Instead, there is commonly said to be a multicollinearity problem if there is a large multiple correlation coefficient R_i^2 between the i^{th} variable x_i and the remaining explanatory variables. This commonsense idea is formally advocated by Farrar and Glauber {3} and also by Marquardt {13} via his "variance inflation factor," $1/(1-R_i^2)$.

The link between R_i^2 and estimator precision is provided by the decomposition of the least squares variance

$$\text{var}(\hat{\beta}_i) = \frac{\sigma^2}{t\, s_i^2} \frac{1}{1-R_i^2} \tag{10}$$

where s_i^2 is the sample variance of the i^{th} explanatory variable. The "multicollinearity problem" is said to be that $\text{var}(\hat{\beta}_i)$ increases as R_i^2 increases, holding σ^2, t, and s_i^2 constant.

However, the terms s_i^2 and $1-R_i^2$ are interchangeable, in that offsetting changes are induced by reparameterizations. Consider the parameter β_1 with the partitioning

$$\underset{txp}{X} = \{\underset{txl}{x_1} \quad \underset{txp-1}{X_2}\}$$

and the simple transformation

$$\begin{aligned} y &= x_1\beta_1 + X_2\beta_2 + \varepsilon \\ &= (x_1 - X_2 b)\beta_1 + X_2(b\beta_1 + \beta_2) + \varepsilon \\ &= z_1\beta_1 + X_2\gamma_2 + \varepsilon. \end{aligned} \tag{11}$$

Different values of b reparameterize the model, with b = 0 corresponding to the initial parameterization (1). The parameter β_1 in (11) describes the effect on y of an increase in $x_1 - X_2 b$ holding X_2 constant; i.e., as in (1), an increase in x_1 holding X_2 constant. The interpretation, value and variance of the least squares estimate β_1 are unchanged. However, the variance of z_1 and correlation with X_2 may differ considerably. As a consequence, the explanation of the variance of $\hat{\beta}_1$ in equations such as (10) may be altered substantially. For some parameterizations, the data are orthogonal so that z_1 has a small variance but is uncorrelated with X_2. For other parameterizations, the data are almost singular so that z_1 has a large variance but is very highly correlated with X_2.*

---

*For example, if x_1 is highly correlated with X_2, then z_1 will be uncorrelated with X_2 for $b = (X_2'X_2)^{-1} X_2'x_1$. If x_1 is orthogonal to X_2, then z, will be highly correlated with X_2 for large $|b|$.

Few researchers intentionally choose parameterizations to minimize or maximize R_1^2. The point is that, if the parameterization is arbitrary, then no significance should be attached to the value of R_1^2. Some researchers may find R_1^2 helpful in describing why a parameter estimate has a high variance. But it would be entirely arbitrary to base a substantive decision on the magnitude of R_1^2.

Two Examples

To illustrate this point, consider first the consumption function's three parameterizations (6), (7), and (8). Holmes {7} used parameterization (7) and the correlation between his two explanatory variables was 0.452. If he had instead used parameterization (6) or (8), his correlation would have been 0.942 or 0.724. Milton Friedman's original hypothesis was that the correlation between y_T and $y_P$is zero. If this were so and, as in Holmes' data, the standard deviation of y_p is twice as large as the standard deviation of y_T, then the correlation coefficients would vary with parameterizations (6), (7), and (8) from 0.90 to 0.00 to 0.45. The parameterization clearly affects our multicollinearity measure.

A comparison of parameterizations (7) and (8) shows how imprecise estimates can be equivalently attributed to either high multicollinearity or low variation. Let's assume that we are interested in the parameter β_1, describing the effect on consumption of an increase in transitory income holding permanent income constant. There has been considerable debate in the economics literature about whether, as Friedman argues, this parameter is zero. Many economists have tried mightily to obtain accurate estimates, but the standard error often turns out to be disappointingly large. If we regress equation (6), we will probably complain that y and y_p are highly correlated. If we regress equation (7), we will complain that the variance of y_T is very low. These are, of course, different ways of saying the same

thing. The danger is that the person who uses equation (6) may be tempted by the multicollinearity reading to use a different estimation procedure. If there is no logical basis for choosing between parameterizations (6) and (7), then there is no logical reason to base such a decision on the multicollinearity reading.

For a second example, consider the acetylene data used by Marquardt and Snee {14}. Their quadratic model has the form

$$y = b_o + \sum_{i=1}^{3} \beta_i x_i + \sum_{1-i-j}^{3} \beta_{ij} x_i x_j + \varepsilon. \tag{12}$$

The data are highly intercorrelated. For example, the squared correlation between x_1^2 and the remaining explanatory variables is .9999996. Following the format (10), the variance of the associated parameter estimate is

$$\text{var}(\hat{\beta}_{11}) = \frac{.81258}{15(3.77\text{x}10^{10})} \quad (2.5\text{x}10^{6}) = .36 \text{ x } 10^{-5}. \tag{13}$$

Marquardt and Snee say that a variance inflation factor of 2 million "is unthinkable and unnecessary." They proceed to center and scale the data, so as to reparameterize (10) as

$$y = \left[\beta_o + \sum_{i=1}^{3} \beta_i \bar{x}_i + \sum_{1-i-j}^{3} \beta_{ij} \bar{x}_i \bar{x}_j\right] + \sum_{i=1}^{3} s_i \left[\beta_i + \beta_{ii} \bar{x}_i + \sum_{i=1}^{3} \beta_{ij} \bar{x}_j\right] z_i + \sum_{1-i-j}^{3} \beta_{ij} s_i s_j z_i z_j \tag{14}$$

where $z_i = (x_i - \bar{x}_i)/s_i$. This linear transformation has no effect on the model or the implicit estimates of any of its parameters. However, the variable $s_1^2 z_1^2$ associated with the parameter β_{11} is now less highly correlated with the remaining explanatory variables. Its correlation coefficient is reduced to .99943 and its variance inflation factor is down to 2,000. The estimate of β_{11}, however, is no more accurate,

$$(15) \qquad \text{var}(\beta_{11}) = \frac{.81258}{15(2.656\text{x}10^{7})} \; (1{,}762.58) = .36 \text{ x } 10^{-5}.$$

Comparing (15) with (13), the variance inflation factor has been reduced by a factor of 1000, but so has the variance of the associated variable. High correlation has been transformed into low variation. With parameterization (12), the estimation of β_{11} is apparently hindered by a severe multicollinearity problem. With parameterization (14), the estimation of β_{11} is apparently hampered by the small variance of the associated variable. The multicollinearity problem didn't go away. It just went into hiding, and escaped detection by a myopic multicollinearity measure.

Some General Multicollinearity Measures

The preceding section focused on interpretations of the imprecision of particular parameter estiamtes. There are also multicollinearity measures intended to give an overall assessment of the data's multicollinearity. The insightful paper by Farrar and Glauber includes the recommendation that the severity of the multicollinearity problem be measured by the determinant of the simple correlation matrix for the explanatory variables:

$$(16) \qquad |R| = |S^{-\frac{1}{2}} Z'Z \; S^{-\frac{1}{2}}|$$

where S is a diagonal matrix with the i^{th} diagonal element equal to the sample sum of squares t s_i^2 of the i^{th} explanatory variable. Since this determinant will lie 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 Z is multivariate normal, they offer as a more precise measure Bartlett's chi-square test of the null hypotheses of orthogonality in the underlying population.

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

Haitovsky {5} notes that this is a questionable null hypothesis given Farrar and Glauber's emphasis on the unimportance of the parent data. Haitovsky also argues that their test is conservative in that satisfactory estimates do not require strict orthogonality. He consequently proposes the alternative extreme of singularity as a null hypothesis, despite the acknowledged fact that this cannot be tested seriously, since an m-dimensional population will not generate data of more than m dimensions. Nevertheless, Haitovsky proposes the heuristic test statistic

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

which does give more comforting signals than Farrar and Glauber's test. For example, at the 1% level with 50 observations on two explanatory variables, Farrar and Glauber's gauge indicates a collinearity problem when the squared correlation between the two variables is greater than .13, while Haitovsky requires a squared correlation coefficient greater than .87.

Part of the problem here is the usual classical dilemma of which state should recieve the presumptive weight of being classified as the null hypothesis. It is also awkward to treat the explanatory variables as stochastic, seldom realistic to assume that they are independent draws from normal distributions, and misleading to define multicollinearity in terms of the properties of the presumed parent population. The consequences of multicollinearity are clearly due to the nature of the available sample data, rather than the source of the data. Indeed, I will argue below that the most important qualification is contrary; in a forecasting context, sample multicollinearity is more worrisome when it is not a characteristic of future data.

An even more fundamental difficulty is that reparameterizations will

cause the determinant of the correlation matrix for the explicitly displayed explanatory variables to vary arbitrarily over the interval {0,1}. With the scaler measure (16), multicollinearity seemingly can be created or dissipated at will. The explanation is, in part, the interchangeability of s_i^2 and R_i^2 and, in part, the choice of which parameters to estimate explicitly. The p parameters in β are a basis for an infinite number of linear combinations of these parameters. Any p linearly independent combinations would serve equally well as a basis. But measures of the overall multicollinearity depend on which parameters are explicitly specified, since some parameters are more accurately estimated than others and some estimated covariances are larger than others. Multicollinearity measures like (16) vary with the normalization A because the normalization determines the set of explicitly estimated parameters $\gamma = A^{-1}\beta$. The choice of a paramterization is really then a choice of which parameters to estimate explicitly and which to leave as implicit estimates.

The Volume of a Confidence Region

Willan and Watts {19} discuss a generalization of variance inflation factors. The variance inflation factor compares the variance of a parameter estimate to what that variance would be if the associated variable had the same variance, but was uncorrelated with the other explanatory variables. A natural generalization is to imagine that all of the explanatory variables have the same variances as in the actual data, but are all uncorrelated with each other. Willan and Watts propose a comparison of the volume of a confidence region for this fictitious data with the volume for the actual data.

This comparison is obviously dependent on the initial parameterization of the model, since this determines which variance will be held constant and which correlations will be assumed away. In the earlier consumption example, should the fictitious reference data be constructed by assuming that there is no correlation between y and y_P, between y_T and y_P, or between y and y_T? Table

1 shows that this arbitrary choice does make a difference.

The Willan-Watts volume ratio turns out to be simply the square root of the determinant of the correlation matrix, $|R|^{.5}$. Thus, they recommend the same calculation as Farrar and Glauber, but provide a new interpretation: $|R|^{.5}$ "tells us how much smaller the joint confidence region could have been if an efficient orthogonal design had been run instead of the actual design." (p. 409) Just as with Farrar and Glauber, the Willan-Watts multicollinearity measure is sensitive to the parameterization of the model. But this sensitivity has a new interpretation. With different parameterizations, they assume different fictitious reference data.

Characteristic Roots

Many statisticians have proposed an examination of a data set's characteristic roots (or eigenvalues) to gauge its "effective dimensionality." If the columns of A in (2) are the p orthonormal eigenvectors of X'X, then

$$A'(X'X)A = D = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix},$$

where the i^{th} diagonal element of D is the eigenvalue of X'X corresponding to the i^{th} column of A. Kendall {8}, Silvey {16}, Mansfield and Helms {12}, and Webster, et. al. {19} suggest checking whether any of the roots are close to zero, while Thisted {18} and Belsley, et. al. {1} favor an examination of the relative magnitudes. The relevance to estimate precision can be seen from

$$\sum_{i=1}^{p} \text{var}(\hat{\beta}_i) = \sigma^2 \text{Tr}(X'X)^{-1} = \sigma^2 \sum_{i=1}^{p} 1/\lambda_i.$$

For given σ, if some roots are very small, then some $\text{var}(\hat{\beta}_i)$ may be very large.

However, the individual characteristic roots are not directly related to the individual variances on β_i, but rather to those on $\gamma = A'\beta$:

$$\text{var } \hat{\gamma}_i = \sigma^2/\lambda_i.$$

I have emphasized the sensitivity of multicollinearity measures to the particular parameters that are scrutinized. Characteristic roots gauge the variances on a set of parameters, the coefficients of the eigenvectors, that are particularly difficult to interpret. These γ_i are not automatically any more or less interesting than other parameters. But they will often be unusual.

With two explanatory variables, for example, which happen to have equal lengths s^2, the eigenvectors normalize the model as

$$y = \{\sqrt{.5}\ (x_1-X_2)\}\{\sqrt{.5}\ (\beta_1-\beta_2)\} + \{\sqrt{.5}\ (x_1+X_2)\}\{\sqrt{.5}\ (\beta_1+\beta_2)\} + \varepsilon.$$

The characteristic roots are $(1+R)s^2$ and $(1-R)s^2$. If the original explanatory variables are highly positively or negatively correlated, then one of the roots will be close to zero and there will be a large variance for the estimate of either $\beta_1 - \beta_2$ or $\beta_1 + \beta_2$. But, before making a substantive decision based on the characteristic roots, there should be some rationale for focusing on the particular parameters $\beta_1 - \beta_2$ and $\beta_1 + \beta_2$.

There is also a scaling problem, in that the division of X by some scaling parameter α will multiply the parameter variances by α^2 and reduce the average value of the characteristic roots by a factor α^2. Thus, the size of the characteristic roots depends on whether we record our data in cents or dollars, inches or centimeters, millions or billions, and so on.

This scaling problem can be attacked by looking at relative magnitudes, such as the ratio of the maximum eigenvalue to the smallest. But ratios, too, have problems. Even if the ratio is large, the smallest root may still be large, too, so that all of the associated parameter variances are small. On the other hand, even if the ratio is one, all of the roots may be the same small number, with all of the parameter variances large. A root close to zero

means the researcher has identified a particular parameter estimate with a high variance. Unequal roots mean that the researcher has found two parameters with unequal variances. Neither finding is very significant.

In addition, the relative magnitudes of roots are altered by reparameterizations that involve other than an equal rescaling of all of the explanatory variables. One variable might be measured in feet and another in pounds. If we change the units of the first variable from feet to inches, the characteristic roots will be affected. The roots will also change if we rearrange the variables as in equations (6) - (9) for the consumption example.

People who use characteristic roots often standardize their variables to have zero means and unit variances. Belsley, Kuh, and Welch {1}, for example, note that the conversion from cents to dollars and other scale changes

> result in very different singular-value decompositions and condition indexes . . . Clearly the condition indexes can provide no stable information to the user of linear regression on the degree of collinearity among the X variates in such a case. It is necessary, therefore, to standardize the data matrices corresponding to equivalent model structures in a way that makes comparisons of condition indexes meaningful. A natural standardization process is to scale each column to have equal length. (p. 120)

These standardizations are intended to avoid arbitrary, yet critical, scale choices. But an equal-length scale is no less arbitrary than the alternatives. This habit merely averts one's eyes from the parameterization problem.

Consider, for example, two uncorrelated explanatory variables

$$y = x_1\beta_1 + x_2\beta_2 + \varepsilon \quad , \quad x_1'x_2 = 0.$$

The eigenvectors are A = I and the characteristic roots are $\lambda_1 = x_1'x_1$ and

$\lambda_2 = x_2'x_2$. These correspond to

$$\text{var}(\hat{\beta}_1) = \sigma^2/x_1'x_1 \text{ and } \text{var}(\hat{\beta}_2) = \sigma^2/x_2'x_2 .$$

Notice, that with unstandardized data, the roots can be very unequal even if the data are uncorrelated. A small value of λ_1 or λ_2 implies that (depending on σ^2), the variance of $\hat{\beta}_1$ or $\hat{\beta}_2$ may be large. A large ratio λ_2/λ_1 implies that β_1 and β_2 are not estimated with equal accuracy.

If we standardize these data,

$$y = (x_1/x_1'x_1)(x_1'x_1\beta_1) + (x_2/x_2'x_2)(x_2'x_2\beta_2) + \varepsilon,$$

then (if $x_1'x_2 = 0$) both roots are equal to one, indicating that the estimates of the rescaled parameters $x_1'x_1\beta_1$ and $x_2'x_2\beta_2$ are equally accurate. Unless we explicitly calculate $x_1'x_1$ and $x_2'x_2$, we won't realize that β_1 and β_2 are not equally well estimated. Standardization changes the parameters whose variances we scrutinize via characteristic roots. But there is no logical reason for focusing attention on specific parameters that are determined by the means and variances of a particular data set. And there is still the problem of linear rearrangements of the explanatory variables.

It is well known that a data set's characteristic roots are altered by nonorthonormal transformations of the explanatory variables. Consider, for example, Holmes' consumption data. Table 2 shows the roots for representations (6), (7), and (8), both with the original data and with the data scaled to have unit lengths. Notice the considerable variation in the enforced choice of parameters to estimate explicitly. None of the parameters is clearly more interesting than the others. With the variation in explicitly estimated parameters, we get considerable variation in both the absolute and relative sizes of the characteristic roots. The smallest root varies from 0.06 to 2.99. The ratio of the roots varies from 35 to 2.6.

The Parameters of Interest?

Belsley, Kuh, and Welch recognize the effects of linear rearrangements on the calculated characteristic roots. In an appendix, they admit that "there is no known relation, in general, between the condition indexes of X and those of XA" (p. 180). Their recommendation is to choose the transformation A so that γ are the "parameters of interest" (p. 179). As an example, they offer the Cobb-Douglas model $Q = AK^{\alpha}L^{\beta}$. If the researcher is interested in α and β, then the model should be parameterized as

$$\ell nQ = \ell nA + \alpha \ell nK + \beta \ell nl. \tag{17}$$

If the researcher is instead interested in $\alpha + \beta$, then the parameterization

$$\ell nQ = \ell nA + (\alpha+\beta)\ell nL + \alpha(\ell nK-\ell n\ell) \tag{18}$$

can be used. In general, they advise that

> Once a parameterization has been decided on, the data should be transformed (if need be) to conform, so that the model becomes $y = X\beta + \varepsilon$. Application of the diagnostics to X then assesses the suitability of X for estimating the specific parameters β. (p. 180).

The first flaw in their recommendation is that it is undermined by their rescaling of their data to have equal lengths. If the scaling parameters are s_K and s_L, then equation (17) is actually analyzed as

$$\ell nQ = \ell nA + (\alpha s_K)(\ell nK/s_K) + \beta s_1(\ell nL/s_L), \tag{19}$$

so that the researcher is really working with the weighted parameters αs_K and βs_L, rather than the desired parameters α and β.

The second flaw is that the characteriistic roots relate to the variances of the eigenvector coefficients, which are $\sqrt{.5}(\alpha s_K - \beta s_L)$ and $\sqrt{.5}(\alpha s_K + \beta s_L)$.

There is no one-to-one correspondence between the precision of these parameter estimates and the precision of the "parameters of interest", α and β.

The third flaw is that there are a variety of alternatives to equation (18) for estimating $\alpha + \beta$. One is

$$\ell nQ = \ell nA + (\alpha+\beta)\ \ell nK + \beta(\ell nL-\ell nK). \tag{20}$$

It is not enough to specify a random number of parameters of interest. To parameterize a model of rank p one must specify exactly p parameters of interest, no more and no less. To determine A, we need to completely specify γ. But where is that specification to come from?

In fact many of the models that Belsley, Kuh, and Welch analyze are susceptible to the transformations that I discussed at the beginning of this paper. One model (p. 163) has personal income and the change in personal income as explanatory variables. Why not instead use lagged personal income and the change in personal income? Or, lagged and current personal income? Another model (pp. 212-214) has the current and lagged values of the household net acquisition of financial assets. Why not instead use the change and either the current or lagged value? Why do it one way with one model and another way with the next? This second model also includes two interest rates. Why not either of these rates and the difference between the two rates? It is not at all easy to choose the "parameters of interest".

The Loss Function Finesse

Theoretical statisticians often begin their analyses by assuming a mean squared error loss function, such as

$$L = E(\hat{\beta}-\beta)'(\hat{\beta}-\beta) \tag{21}$$

or, perhaps, with weights B:

$$L_W = E(\hat{\beta}-\beta)'\ B(\hat{\beta}-\beta) \tag{22}$$

If they are careful, then their loss function provides the parameterization that must precede a meaningful multicollinearity measure. If the data are reparameterized, as in (2), then the loss function must be correspondingly transformed. If we begin with the loss function (21) and then center, scale, or rearrange our data via the transformation A, then our loss function becomes

$$L = E(\hat{\gamma}-\gamma)'(A'A)(\hat{\gamma}-\gamma),$$

where the new parameters are $\gamma = A^{-1}\beta$. Alternatively, if the researcher centers, scales, and arranges the data so as to work with the parameters γ, then it is tempting to assume that the implicit loss function is $E(\hat{\gamma}-\gamma)'(\hat{\gamma}-\gamma)$ $= E(\hat{\beta}-\beta)'(AA')^{-1}(\hat{\beta}-\beta)$, so that $B = (AA')^{-1}$.

The problem is that we seldom, if ever, are told the rationale for a particular loss function. Instead, the centering, scaling, and rearranging seem to be done out of force of habit and little more. Thisted {17} does advise researchers that "one chould exercise caution before one adopts the loss structure," but he offers little guidance on actually choosing a loss function.

The argument that a meaningful multicollinearity measure requires a thoughtfully chosen loss function is really just another way of arguing that a thoughtful parameterization is needed. The fundamental question is still whether or not a logical rationale can be provided for a particular parameterization (or loss function).

A Forecasting Perspective

Multicollinearity measures are sensitive to a model's parameterization. Yet the multicollinearity literature says very little about this important choice. Instead, the theoreticians simply assume a parameterization or a

loss function.* In practice, researchers generally use a convenient parameterization, selected in a very offhand manner.

A reasonable alternative is to specify explicitly the situations in which the model will be used. Models are typically intended for forecasting, either actual predictions using future values of the explanatory variables or hypothetical calculations of the consequences of selected changes in one or more of the variables. This latter category includes historical review, identification and assessment of the importance of certain explanatory variables, and policy analysis of the effects of changes in some of the variables. If the model's purpose can be interpreted in terms of forecasting accuracy, then we can obtain a natural scalar measure of estimation precision.**

It is important to recognize that "forecasting" is interpreted very broadly here. If the researcher is interested in the effects of ceteris paribus changes in one of the explanatory variables, say x_1, then the accuracy of the parameter estimate $\hat{\beta}_1$ is paramount. The relevant forecasting situation would be described by assigning a variety of values to x_1 with fixed values for the other variables. In general, we want to specify values $\underset{\sim}{X}$ which pose the questions that are to be asked of the model. We then want

---

*It's the old joke about the physicist, chemist, and mathmatician locked in a room. The physicist builds a lever, the chemist concocts an explosive, and the mathematician assumes a key.

**Willan and Watts {19} also discuss the use of forecasting objectives to gauge multicollinearity. But they compare the actual prediction variances to what the prediction variances would be if the data had the same variances but were uncorrelated. As noted earlier, such comparisons depend on the initial parameterization of the model.

to forecast n out-of-sample values generated by

$$\underset{nx1}{\tilde{y}} = \underset{nxp}{\tilde{X}} \underset{px1}{\beta} + \underset{tx1}{\tilde{\varepsilon}} \tag{23}$$

where

$$E\begin{pmatrix}\varepsilon \\ \tilde{\varepsilon}\end{pmatrix}\begin{pmatrix}\varepsilon \\ \tilde{\varepsilon}\end{pmatrix}' = \begin{bmatrix}\sigma^2 I_t & o \\ o & \tilde{\sigma}^2 I_n\end{bmatrix}. \tag{24}$$

The mean squared forecasting error is

$$\begin{aligned} MSFE &= E(\hat{\tilde{y}} - \tilde{y})'(\hat{\tilde{y}} - \tilde{y}) \\ &= n\tilde{\sigma}^2 + E(\hat{\beta} - \beta)'\tilde{X}'\tilde{X}(\hat{\beta} - \beta). \end{aligned} \tag{25}$$

The explicit forecasting objectives justify the loss function (22), with $B = \tilde{X}'\tilde{X}$. For least squares estimates,

$$MSFE = n\tilde{\sigma}^2 + \sigma^2 \text{Trace}(\tilde{X}'\tilde{X}(X'X)^{-1}). \tag{26}$$

The multicollinearity measure, trace $(X'X)^{-1} = \sum_{i=1}^{p} (1/\lambda_i)$, is relevant to the special case $\tilde{X}'\tilde{X} = I$. More complicated assumptions are required to salvage the relevance of $|R|$. In general, the mean squared forecasting error is a specific weighted average of parameter mean squared errors, covariances as well as variances, with weights determined by the characteristics of the out-of-sample data.

If a reparameterization A is selected so that $\tilde{Z}'\tilde{Z} = A'(\tilde{X}'\tilde{X})A = nI$, then

$$\begin{aligned} MSFE &= n\sigma^2 + n\sigma^2 \text{Trace}(Z'Z)^{-1} \\ &= n\tilde{\sigma}^2 + n\sigma^2 \sum_{i=1}^{p} \text{var}(\hat{\gamma}_i) \\ &= n\tilde{\sigma}^2 + n\sigma^2 \sum_{i=1}^{p} 1/\mu_i \end{aligned} \tag{27}$$

where the μ_i are eigenvalues of $Z'Z$ when $AA' = n(\tilde{X}'\tilde{X})^{-1}$. The sum of the

variances of the $\hat{\gamma}_i$ (and of the inverse roots for this parameterization) directly measures forecasting accuracy. While the specific elements of A are not unique, it can easily be shown that the requirements that A be nonsingular and that $AA' = n(\tilde{X}'\tilde{X})^{-1}$ imply that the roots of Z'Z are unique.* Although imperfectly related to MSFE, the alternative measures of the smallest root of Z'Z, the ratio of the largest root to the smallest, and the determinant $|Z'Z| = \prod_{i=1}^{p} \mu_i$ are all at least fixed by the suggested parameterization. This is not true of the correlations among the z_i and of the determinant of the correlation matrix for Z, in that there is some continuing interchangeability of s_i^2 and $1 - R_i^2$ for Z which precludes a substantive interpretation of $1 - R_i^2$ and $|R|$.

Scaling a Multicollinearity Barometer

A persistent issue for proposed multicollinearity measures is the scale calibration. For what values of the characteristic roots or correlation coefficients are the data to be considered "ill-conditioned" and estimates "degraded"? Correlation coefficients have a finite range but the interchangeability of $1 - R_i^2$ and s_i^2 prevents a meaningful interpretation of specific values of R_i^2. Clearly, variances as well as correlation coefficients must be taken into account. In addition, equation (10) shows that even holding s_i^2 constant, it cannot be inferred from the value of R_i^2 whether the variance of any estimate is "high" or "low". If there are lots of data

---

*Consider two nonsingular transformations A and $\bar{A}$ with $AA' = \bar{A}\bar{A}'$, and define $P = \bar{A}^{-1}A$. Then $P^{-1}\bar{Z}'\bar{Z}P = A^{-1}\bar{\bar{A}}\bar{\bar{A}}'X'X\bar{\bar{A}}\bar{\bar{A}}^{-1}A = A'X'XA = Z'Z$ so that $\bar{Z}'\bar{Z}$ and Z'Z are similar.

and/or a low disturbance variance, very accurate estimates are consistent with highly multicollinear data. Conversely, the estimates may have high variances even if the data are orthogonal. Nor does it seem reasonable to label a variance "high" or "low" without reference to the parameter being estimated. The scale of a variance depends, of course, on the scale of the parameter. The fact that the variance on $(10\hat{\beta}_i)$ is one hundred times that on $\hat{\beta}_i$ does not mean that estimating $10\beta_i$ rather than β_i diminishes accuracy.

A reliable accuracy assessment must take into account the researcher's objectives. For example, a high variance can be tolerated on a parameter that is uninteresting, perhaps because the model is intended for use in situations in which the associated variable will change little. The placement of the model in a forecasting context provides a parameterization, $AA' = n(\tilde{X}'\tilde{X})^{-1}$ such that the scale of trace $(Z'Z)^{-1}$ or $\sum_{i=1}^{p} 1/\mu_i$ can be meaningfully interpreted. The assessment of whether MSFE is large or small is still necessarily subjective.

Some Two-Variable Analytics

Consider the simple p = 2 case with

$$Z'Z = t\begin{bmatrix} s_1^2 & rs_1s_2 \\ rs_1s_2 & s_2^2 \end{bmatrix}, \quad \tilde{Z}'\tilde{Z} = n\begin{bmatrix} \tilde{s}_1^2 & \tilde{r}\tilde{s}_1\tilde{s}_2 \\ \tilde{r}\tilde{s}_1\tilde{s}_2 & \tilde{s}_2^2 \end{bmatrix}.$$

The mean squared forecasting error is

$$MSFE = n\tilde{\sigma}^2 + (n\sigma^2/t)(h_1^2 + h_2^2 - 2r\tilde{r}h_1h2)/(1 - r^2)$$

where $h_i = \tilde{s}_i/s_i$. Even given n, t, σ, and $\tilde{\sigma}$ the MSFE cannot be inferred from r. It is necessary to specify the variances as well as covariances, and to also specify the out-of-sample characteristics of the data. Even if r = 0, the forecasts may be unreliable if the h_i are large. On the other hand, a

value of r^2 close to one need not be serious if the correlation persists out-of-sample ($\tilde{r}$ 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 are many sample data.

Indeed, an increase in r^2 may even be beneficial. For example, when $h_1 = h_2 = h$,

$$MSFE = n\tilde{\sigma}^2 + \frac{n\sigma^2}{t} \; 2h^2 \left[\frac{1 - r\tilde{r}}{1 - r^2}\right]$$

and

$$\frac{\partial MSFE}{\partial r} = \frac{n\sigma^2}{t} \; 2h^2 \; \frac{2(r - \tilde{r}) + (1 - r^2)\tilde{r}}{(1 - r^2)^2} \; .$$

For a given positive (negative) out-of-sample correlation, the MSFE will decline at the point r = 0 as r rises (falls). The MSFE will turn back up at the point

$$r = \frac{1 - \sqrt{1 - \tilde{r}^2}}{r} \; ,$$

which is between 0 and $\tilde{r}$. For example, with $\tilde{r} = .95$, an increase in r up to .72 will reduce MSFE.

If the data are parameterized so that $\tilde{Z}'\tilde{Z} = nI$, then

$$MSFE = n\tilde{\sigma}^2 + (n\sigma^2/t)(1/s_1^2 + 1/s_2^2)/(1 - r^2)$$

so that these estimate variances $^1/s_i^2 \; (1-r^2)$ do provide a guide to MSFE. Notice though, that there is still some interchangeability of r and s_i^2. The interpretation of correlation coefficients is ambiguous even with a forecasting parameterization.

Summary

In regression models there is seldom a compelling specific parameteri-

zation. Variables might be measured as deviations from zero, from their sample means, or from some other number. Each variable's units can be freely chosen. Variables are often added to or subtracted from one another. These are all examples of arbitrary nonsingular linear transformations which do not affect the substance of a model.

Multicollinearity measures that are sensitive to arbitrary parameterizations are themselves arbitrary. The variance on any single parameter can be equivalently analyzed in terms of either the correlations among certain variables or the variation of certain variables. Multicollinearity measures for the model as a whole depend also upon the specific parameters analyzed. I have argued here that an explicit forecasting context provides a natural scalar measure of estimation accuracy and a meaningful parameterization for multicollinearity measures based on eigenvalues.

Table 2. Some Normalizations of Friedman's Consumption Data

Equation	Correlation between Variables	Parameters Explicitly Estimated		Characteristic Roots (divided by σ^2)		Ratio of Characteristic Roots
(6)	.942	$.80\beta_1 + .61\beta_2$	$.61\beta_1 - .80\beta_2$	43.93	1.24	35
(6)scaled	.942	$3.8\beta_1 + 2.9\beta_2$	$3.8\beta_1 - 2.9\beta_2$	1.94	.06	33
(7)	.452	$-.71\beta_1 - .97\beta_2$	$.71\beta_1 - .26\beta_2$	17.86	2.99	6.0
(7)scaled	.452	$4.3\beta_1 + 2.9\beta_2$	$-1.5\beta_1 - 2.9\beta_2$	1.45	.55	2.6
(8)	.724	$-.96\beta_1 - .68\beta_2$	$-.28\beta_1 - 1.2\beta_2$	30.28	1.75	17.3
(8)scaled	.724	$3.8\beta_1 + 2.3\beta_2$	$3.8\beta_1 + 5.2\beta_2$	1.72	.28	6.2

References

[1] Belsley, D.A., E. Kuh and R.E. Welsch, Regression Diagnostics, New York: Wiley, 1980.

[2] Christ, C.F., "A Simple Macroeconomic Model with a Government Budget Restraint," Journal of Political Economy, 76 (February 1968), 53-67.

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

[4] Friedman, B., "Financial Flow Variables and the Short-run Determination of Long-term Interest Rates," Journal of Political Economy, 85 (1977), 661-89.

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

[6] Hoerl, A.E. and R.W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, 12 (1970), 55-67.

[7] Holmes, J.M., "A Direct Test of Friedman's Permanent Income Theory," Journal of the American Statistical Association, 65 (1971), 1159-62.

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

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

[10] Lovell, M., "Seasonal Adjustment of Economic Time Series," Journal of the American Statistical Association, 58 (1963), 993-1010.

[11] Mandell, J., "Use of the Singular Value Decomposition in Regression Analysis," The American Statistician, February 1982, 15-24.

[12] Mansfield, E.R. and B.P. Helms, "Detecting Multicollinearity," The American Statistician, (August 1982), 158-160.

[13] Marquardt, D.W., "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation," Technometrics, 12 (1970), 591-612.

[14] Marquardt, D.W. and R.D. Snee, "Ridge Regression in Practice," The American Statistician, 29 (1975), 3-20.

[15] Mosteller, F. and J.W. Tukey, Data Analysis and Regression, Reading, Mass.: Addison-Wesley, 1977.