tiple Linear Regression Viewpoints
Volume 12, Number 2 Fall 1983

A PERSPECTIVE ON MULTICOLLINEARITY

John T. Pohlman
Southern Illinois University, Carbondale

Multicollinearity occurs when a set of predictors in a regression model are correlated. When one or more predictors are linear combinations of other predictors, exact multicollinearity exists. The most typical case encountered in practice is some imperfect, but possibly strong, association among predictors. The sampling variability of regression coefficients is, in part, a function of the correlations among predictors. High interpredictor correlations will lead to less stable estimates of regression weights. This relationship can be problematic if a regression weight's variability obscures some functional relationship of interest to a researcher. The multicollinearity problem, then, is an unacceptably high standard error of a regression weight that occurs because of high interpredictor correlations.

Presented at AERA 1983, MLR Special Interest Group
Not refereed by editorial staff

There are, however, meaningful research questions that require models with highly correlated predictors. A quadratic curve fitting model requires a predictor and its squared value as independent variables. If the predictor takes only positive integer values, the correlation between the predictor and its squared value will be extremely high. There are also instances in which a researcher, desiring to measure some construct reliably, uses a number of highly correlated, but fallible measures of the construct. Using both WISC and Binet scores as measures of intellectual ability would be an example of such multiple indicator models.

The impact of multicollinearity in any particular application depends on the roles taken by the variables affected. Predictors in a regression model may take one of two roles; they may be investigative variables or control variables. Investigative variables are those predictors whose influence on the dependent variable constitutes the primary interest of an analysis. Control variables or covariates are included in a model in order to statistically adjust both the dependent variable and investigative predictors for some contaminating influence.

Multicollinearity might occur within or between such classes of predictors. A model can be formed using multiple control variables measuring the same construct. Using a number of IQ tests to measure the construct mental ability exemplifies this approach. In these cases a researcher should not interpret the partial tests performed on the

separate weights assigned to the individual IQ measures, but instead should treat the collection of IQ tests as a set. In this context the researcher is interested in controlling for the influence of the IQ construct represented by the multiple test scores. The individual partial tests are irrelevent to this research problem.

Similarly, a set of investigative variables might be multiple measures of the same construct. If mental ability was an investigative construct of interest, the above comments would apply equally in this case. Setwise regression methods would be the appropriate response to this problem (Cohen and Cohen, 1975)

A third possibility exists; an investigative variable and a control variable are highly correlated. The researcher has been lead to this position in pursuit of an answer to a meaningful research question. Multicollinearity in such models is a legitimate and unavoidable consequence of posing the research question. Under these circumstances the researcher will either have to endure the problem, or with a slight amount of prior planning, correct for the impact of multicollinearity.

How can a researcher reconcile the legitimate need to estimate coefficients in models with highly correlated predictors and the inevitable imprecision associated with such problems? Ridge regression (Hoerl and Kennard, 1970) and similar methods (Smith and Campbell, 1980) have been proposed. Ridge regression methods tend to adjust coefficients closer to zero than the corresponding ordinary

least squares (OLS) estimates. Ridge regression methods yield estimators of coefficients that exhibit less sampling variability but can be biased. In theory the estimation bias is offset by smaller mean square errors of estimation. The mathematical justification for ridge estimation is based on an existence theorem (Draper and Smith, 1981, p.316) which simply states that there always exists an estimate of a regression weight that will deviate from the true weight less than an OLS estimate. Unfortuneately, in any specific application of ridge regression there is no guarantee that the sample estimate is a member of the class of more accurate estimates. The efficacy of ridge estimation depends upon the extent of prior information the researcher has regarding the population model. If the prior information is correct, and the ridge estimators accurately reflect this prior information, then the ridge estimates will be more precise than their OLS counterparts. A researcher in possesion of prior information could justifiably use ridge estimation as a vehicle for a Bayesian analysis. The use of ridge methods without a clear specification of the implied prior constraints is unwarranted. Another approach, the one advocated here, is to use OLS estimates, but plan for the potential problems of multicollinearity. This paper will focus on the consequenses of multicollinearity for the most common applications of statistical inference in linear models.

Multicollinearity and Model Assumptions

The illustrations provided in the remainder of this

paper will make use of a three predictor regression problem. A three predictor regression model for a dependent variable may be expressed as

$$(1) \qquad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$$

where Y is a vector of values for the dependent variable , (X_1 ,X_2 ,X_3) are the predictor vectors, (β_1, β_2 , and β_3) are the least squares weights and E , the residual vector. Researchers then usually interrogate the functional relationships implied in their model via statistical inference directed at the model's coefficients. From a random sample of observations and the resulting estimate of the model, confidence intervals may be placed about the coefficient estimates or tests of null hypotheses may be performed. A test of $H: \beta_1 = c$ is conducted as follows:

1. A full model estimating (1) is fitted to sample data using the least squares criterion

$$(2) \qquad Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + E(f)$$

2. A restriction is imposed on (2) by setting the weight under test equal to the value specified in the hypothesis, $\beta_1 = c$. The restricted model then is estimated

$$(3) \qquad Y = b_0 + cX_1 + b_2 X_2 + b_3 X_3 + E(r)$$

$$(4) \quad F_{\{m(f)-m(r)\},\{n-m(f)\}} = \frac{\{E'(r)E(r)-E'(f)E(f)\}\ /\{m(f)-m(r)\}}{E'(f)E(f)/\{n-m(f)\}}$$

where E'(f)E(f) and E'(r)E(r) are the error sums of squares for the full and restricted models respectively, m(f) and m(r) are the number of weights estimated in the full and restricted models and n is the sample size.

In order for the F statistic given in (4) to follow a central F distribution, it is assumed that

1. The elements of E(f) are $n(0, \sigma^2)$

2. The weights estimated in (1) and (2) satisfy the least squares criterion.

3. The variance of the residuals is homoscedastistic.

4. The sample is randomly drawn from the population.

5. The null hypothesis is true.

It is important to note that no assumption is made regarding the correlations among the predictors. The type I error rate for such tests is unaffected by multicollinearity. Hence a researcher rejecting H: need not be concerned about the multicollinearity problem for that test. With the current interest in multicollinearity it was inevitable that some authors (Loether and McTavish, 1980; p. 331) would include uncorrelated predictors as an inferential

assumption. Recently, Hawkes and Mosely (1982) have questioned this extreme interpretation. Loether and McTavish are simply wrong and have thereby misled their readers. We must now be prepared for the onslaught of papers that will use all sorts of suboptimal methods citing Loether and McTavish.

Multicollinearity, Interval Estimation and Power

The correlations among predictors can affect the type II error rate of tests and increase the range of confidence intervals. These effects obtain as a result of the increased imprecision in the estimation of weights that is associated with multicollinearity. This can be demonstrated with the expression for the standard error of a regression coefficient. Scaling all the variables in a model so that their variances are equal, the expression for the variance error of the regression weight assigned to X_1 is given by (Winer, 1971)

$$(5) \qquad \sigma^2_{b_1} = \frac{(1 - \rho^2_{y.123})}{n(1 - \rho^2_{1.23})}$$

where $\rho^2_{y.123}$ is the coefficient of determination for the model in (1), $\rho^2_{1.23}$ is the coefficient of determination for the model

$$(6) \qquad X_1 = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + E \ .$$

and n is the sample size. An unbiased estimate of $\sigma^2_{b_1}$ is obtained by replacing the coefficients of determination in (5) with the respective sample R squares and using (n-m(f)) in place of n. Inspection of (5) shows that as $\rho^2_{1\cdot23}$ increases so does $\sigma^2_{b_1}$. In the extreme case where $\rho^2_{1\cdot23}=1$, $\sigma^2_{b_1}$ is undefined. The variance of a regression coefficient is therefore a function of three factors: $\rho^2_{y\cdot123}$, $\rho^2_{1\cdot23}$, and n. High interpredictor correlations are not sufficient evidence of a problem. Sample size and model validity can compensate for multicollinearity. Table 1 illustrates the interrelationships among the terms in (5).

Table 1

Three Combinations of $\rho^2_{y\cdot123}$, $\rho^2_{1\cdot23}$ and n That Yield Equal Values of $\sigma^2_{b_1}$

	$\rho^2_{y\cdot123}$	$\rho^2_{1\cdot23}$	n	$\sigma^2_{b_1}$
case 1	.50	.00	10	.05
case 2	.50	.90	100	.05
case 3	.95	.90	10	.05

The three examples given in Table 1 show that many combinations of sample size, $\rho^2_{y\cdot123}$,and $\rho^2_{1\cdot23}$ yield the same standard error for a regression weight. One could not assert a multicollinearity problem for cases 2 or 3 in contrast to case 1, since all three cases produce the same variance error. In order for a problem to exist, the variance error of a weight must be greater that some desired value. Hence a researcher must specify, before the analysis, a desired value of the standard error of a coefficient of interest.

As noted earlier, multicollinearity does not affect the type I error rate of tests on linear models. Type II error rates and the width of confidence intervals will, however, be influenced by interpredictor correlations. A researcher can correct for these effects when planning an analysis by selecting a sample size that can compensate for any degree of multicollinearity. When interval estimation of β_1 is planned, the researcher can specify $\rho^2_{y\cdot123}$, $\rho^2_{1\cdot23}$ and the desired value of $\sigma^2_{b_1}$. Given these values, (5) can be used to solve for the sample size that will give the desired degree of precision.

For example, a researcher might be planning a path analysis of school achievement in the 5th grade (Y) with a structural model that includes attitudes toward teachers (X_1), attitudes toward school in general (X_2) and achievement in the 4th grade (X_3). The model that would be estimated is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$$

If we assume all of the variables in this model have been scaled to a common variance, then the coefficient β_1 may be treated as a beta weight and (5) gives its variance error. The researcher might decide that an acceptable value of $\sigma^2_{b_1}$ is .01. A review of the literature or a pilot study might yield estimates of $\rho^2_{y \cdot 123}$ and $\rho^2_{1 \cdot 23}$ to be .7 and .5 respectively. Substituting these values in (5) and solving for n, the requisite sample size is found to be 60. If the researcher wanted to be more cautious, a larger value of $\rho^2_{1 \cdot 23}$ could be used to accommodate stronger multicollinearity effects. If $\rho^2_{1.23}$ were set equal to .8 the requisite sample size becomes 150. With this sample size the researcher would know that the estimation of β_1 is sufficiently precise even allowing for stronger than anticipated multicollinearity effects.

A similar approach might be taken if the researcher is using hypothesis testing methods. A power analysis can be performed to determine a requisite sample size that will provide a desired power for a test of H: $\beta_1 = 0$.

The least squares method along with the assumption $e_1 \sim N(0, \sigma^2)$ (Winer, 1971) insures that

$$b_1 \sim N(\beta_1, \sigma^2_{b_1})$$

The standard error of b_1 that will yield the desired

power for this test when $\beta_1 = \Delta$ is

$$(7) \qquad \sigma_{b_1} = \frac{\Delta}{z_{(1-\alpha)} + z_{(1-\beta)}}$$

where α = the significance level, β = the type II error rate desired and z_p = the unit normal deviate at the pxl00 percentile.

Substituting (7) for σ_{b_1} in (5) and solving for n gives

$$(8) \qquad n = \frac{\{z_{(1-\alpha)} + z_{(1-\beta)}\}^2 \ (1 - \rho^2_{y\cdot 123})}{\Delta^2 \ (1 - \rho^2_{1\cdot 23})}$$

Formula (8) may then be used to determine the requisite sample size so that a test of H: $\beta_1 = 0$ at significance level α will have of power of $(1-\beta)$ if $\beta_1 = \Delta$, given the values of $\rho^2_{y\cdot 123}$ and $\rho^2_{1\cdot 23}$. For example, letting $\alpha = .05$ (one tailed), power = .80, $\Delta = .30$, $\rho^2_{y\cdot 123} = .70$ and $\rho^2_{1\cdot 23} = .50$ yields an n of 42. If the researcher chose to accommodate a larger value of $\rho^2_{1\cdot 23}$, (8) could be applied to determine the necessary sample size to compensate for stronger multicollinearity effects. If in the previous example $\rho^2_{1\cdot 23}$ is changed to .8 the required sample size becomes 104. A sample of this size would then allow for the effects of multicollinearity and provide a test with the desired power.

These examples have shown that it is possible to use OLS estimation and still allow for multicollinearity. Sample size was used as the compensating parameter, but it

is possible to use $\rho^2_{y \cdot 123}$ in a similar fashion. Hence collinearity effects could be countered by adding covariates to a model that are highly correlated with the dependent variable. It is extremely important that the research question not be compromised if this approach is employed.

Summary

Multicollinearity effects can result in imprecise estimation of regression coefficients, but small sample sizes and low model coefficients of determination can produce the same effects. Multicollinearity does not result in a violation of the assumptions that underly statistical inference on linear models. This implies that a researcher wishing only to protect against type I errors need not be concerned about high interpredictor correlations. The researcher concerned about the power of tests should incorporate a consideration of multicollinearity into the planning of an analysis. It has been shown here that it is possible to compensate for any degree of multicollinearity by increasing sample sizes or model validity. The use of these methods insures that the researcher will be able to benifit from the valuable sampling characteristics of least squares estimation; the sample coefficients are unbiased estimators of their respective parameters.

Reference Notes

Hawkes, R. and Mosely, R. Demystifying Multicollinearity. SPSS Issues Conference, October, 1982.

References

Cohen, J. and Cohen, Patricia. Applied multiple regression/correlation analysis for the behavioral sciences. New York:Wiley, 1975.

Draper, N. and Smith, H. Applied regression analysis (2nd ed.). New York: Wiley, 1981.

Hoerl, A. and Kennard, R. Ridge regression: biased estimation for non-orthogonal problems. Technometrics, 1970, 12, 55-67.

Loether, H. and McTavish, D. Descriptive and inferential statistics: an introduction. Boston: Allyn and Bacon, 1980.

Smith, G. and Campbell, F. A critique of some ridge regression methods. Journal of the American Statistical Association, 1980, 75, 74-81.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1971.