The Effects of Multicolinearity on Accuricy of Prediction in Least Squares, Unit Weight and Absolute Value Prediction Systems

Daniel Segall and Jagdish N. Sheth
University of Illinois

The Effects of Multicolinearity on Accuricy of Prediction in Least Squares, Unit Weight and Absolute Value Prediction Systems

Multiple lenear regression systems have typically served two main purposes in the social sciences: interpretation and prediction. The interpretative aspect of regression analysis involves the examination of the relative contribution of several variables in explaining the variance of a dependent variable. Here the researcher is interested in the size of each beta weight assigned to each of the independent variables, and uses these beta values to draw inferences about their relative importance in explaining the dependent measure. For this purpose of multiple linear regression, much has been written about the effects of multicolinearity. In general, the greater the intercorrelations between independent variables, the more unstable the beta weights become. This instability of beta weights has led to the suggestion that highly colinear predictor variables should not be used, and if they are, the relative sizes of beta values may be meaningless.

But what if one is intersted in prediction rather than explaination? That is, suppose we are interested in using least squares regression to predict some criterion in one sample using the equation developed in a different sample. Do we still need

to worry about multicolinearity? This paper explores the effects of multicolinearity on the ability of three prediction systems to make accurate predictions. These three systems include: least squares, least absolute value, and unit weighting prediction analyses.

Method

The genral design is that of a monte carlo simulation involving three factors: (a) multicolinearity, (b) sample size, and (c) prediction method. There were three levels of the multicolinearity factor (0.10, 0.45, 0.65), three levels of sample size (20, 60, 100), and three levels of prediction method: least squares regression, unit weighting and least absolute value criterion (Barrodale & Roberts, 1973). The design involved a total of 27 (3x3x3) conditions. Observations were created as follows:

Stepl. The first step involved the specification of three matricies of predictor - predictor and predictor - criterion correlations. These three matricies are displayed in table 1, thabe 2, and table 3.

Insert Tables 1, 2, and 3 about here

Notice that these correlation matricies specify the the level of multicolinearity (one matrix for each level of multicolinearity). Each matrix specifies the interrelationships between the predictors and between the criterion and predictors in three

hypothetical populations. Notice also that these three matricies were specified in such a way that in each population there exists a multiple R of 0.6500.

Step2. Given these three population matricies, the next step was to sample observations. This was performed using a triangular factorization method (GGNSM) available in IMSL. This technique draws observations form a multivariate normal population with a specified variance-covariance matrix. From each of the three populations correlation matricies, 3 different size samples were drawn: N = 20, N = 60, N = 100, totaling 9 different samples.

Step3. Beta weights were then calculated for the least-squares technique and least absolute value technique for each of the three sample sizes. Thus, 3 sets of beta weights (corresponding to the three samples of N = 20, 60, and 100) were calculated for each technique. These calculations were repeated for each of the three population matricies, bringing the total number of beta weight sets to 18 (3 populations x 3 different sample sizes x 2 techniques).

Step4. Next, a set of validation samples were generated in a similar manner to the proceedure described in Step 2. Three different sample sizes (N = 20, 60 and 100) were generated from each of the three population matricies, totaling 9 samples.

Step5 The final stage involved the calculation of the dependent measure to be used in this study: the correlation (squared) between the actual scores in the validation sample and

those predicted on a basis of the weights derived in Step 3. In the case of the third prediction method, unit weighting, the correlation (squared) between the sum of equally weighted predictors and the actual criterion values was used as the dependent measure. Thus for each of the prediction methods (except unit weighting) a cross validation R-squared was calculated which constituted the measure of interest.

When calculating the cross validation measures for the least-squares technique and the least absolute technique, beta weights derived from a particular population matrix and sample size (step3) were used to predict a set of observations drawn from the same population covariance structure with a corresponding sample size. For example, least squares beta weights derived from a sample generated from population A with N = 20 (step 3) were used to predict observations in the cross validation sample drawn form the sampe population A with the cross validation sample N = 20.

Thus, at the end of Step 5, 27 dependpent measures will have been derived, each being the correlation (squared) between the actual and predicted criterion using the three prediction methods, on 3 different sample sizes, drawn from 3 different populations.

Step6. Next, the process described in Steps 2 through Steps 5 was repeated 60 times. Thus in each of the 27 different conditions described above (3 populations x 3 techniques x 3 sample sizes) there were 60 measures of validation

(correlation-squared).

Step7 In order to summarize the results obtained in Step 6, the validation measures were analyzed by a 3 factor analysis of variance model. The particular design used was a mixed design, involving two between subject factors (population type and sample size) and one within subject factor (Technique). The third factor was considered a within subject variable because the calculation of the 2 sets of betas for each technique were derived form a single sample. These betas (including unit weights) were then used to predict criterion values of a single validation sample, again the same for the the three methods.

Results and Discussion

A summary of the analysis of variance performed on the validation measures are presented in Table 4.

Insert Table 4 about here

First, notice that the main effect of Multicolinearity was non-significant. As expected the main effect of sample size did have a significant effect, thus displaying a significant impact on accuracy of prediction.

In addition there was a significant main effect for Technique. The mean percent variance explained accross all conditions for each technique is shown in Table 5.

Insert Table 5 about here

Notice that the least-squares and unit weighting technique did roughly equal, while the absolute value technique did significantly worse. This may be due to an artifact in the "goodness of fit" measure. That is, in this analysis we are using the correlation between observed and predicted as the criterion, one which the least-square technique maximizes. This may give an unfair evaluation to the absolute value technique since it maximuzes the goodness of fit in terms of an alternate measure, the amount of agreement between observed and predicted in terms of absolute value units.

Next, notice that there exists a significant interaction between technique and sample size. The nature of this interaction can be oberved by examination of Table 6.

Insert Table 6 about here

Both, the least squares and absolute value technique appear to be effected by sample size while unit weighting appears unaffected. This effect is probably due to the extreme capitalization on chance that occurs in the least-squares and absolute value techniques when samples are small. Since beta weight estimates in unit weighting are data free (all are equal), capitalization on chance does not occurr, and thus no effect of sample size is

found. Notice that for small samples of N = 20 (see Table 6) that unit weighting actually does a great deal better then least-squares and least absolute value techniques. Even at N = 60 unit weighting does relatively well (almost equal) compared to the other two techniques.

Finally there exists a significant interaction between sample size and multicolinearity. There are probably two contributing factors to this effect. First, as can be seen from Tables 1, 2, and 3, as the level of multicolinearity changes between the three conditions so does the pattern of population beta weights. By examination of Table 7 the pattern makes sence in light of the above observation.

Insert Table 7 about here

That is, as the level of multicolinearity increases, the greater the heterogeneity of beta weights, and thus the poorer the approximation of the unit weight method becomes. This explains the downward trend of percent variance explained of the unit weight method, but what about the apparent upward trend shown by the other two methods? Why should an increase in multicolinearity actually be advantageous to these methods?

One likely reason is that as correlations become larger, their sampling error becomes smaller. This is brought about be the restricted range (+1.0 to -1.0) inherant in the pearson product moment correlation. That is, as the population r value

approached +1 or -1, the range around which sample estimates may vary is truncated on one side. This truncation reduces the sampling error. Given that the estimates of beta weights are a direct function of these correlation estimates, it follows that a reduction in th error of estimate for the correlations will also mean a reduction in the error of estimate of the beta values, thus resulting in more accurate predictions as indicated in Table 7.

In summary there appears to be little reason for concern about multicolinearity when using least squares for prediction purposes only. In fact, higher levels of multicolinearity can be expected to increase accuracy of prediction. One should instead be concerned with the effects of sample size on least-squares predition. In extremely low sample size to number of predictors ratios (as in our simulation) the researcher should consider the use of equal weighting, for equal weighting may yield significantly better predictions under these conditions.

Table 1

Population A

Correlation Matrix					
	X_1	X_2	X_3	X_4	Y
X_1	1.0000	0.1000	0.1000	0.1000	0.2895
X_2	0.1000	1.0000	0.1000	0.1000	0.3895
X_3	0.1000	0.1000	1.0000	0.1000	0.4895
X_4	0.1000	0.1000	0.1000	1.0000	0.5895
Y	0.2895	0.3895	0.4895	0.5895	1.0000

Beta Weights				
	X_1	X_2	X_3	X_4
B	0.1714	0.2826	0.3937	0.5048

$R = 0.6500$ $R^2 = 0.4225$

Table 2

Population B

Correlation Matrix					
	X_1	X_2	X_3	X_4	Y
X_1	1.0000	0.4500	0.4500	0.4500	0.4231
X_2	0.4500	1.0000	0.4500	0.4500	0.5231
X_3	0.4500	0.4500	1.0000	0.4500	0.6231
X_4	0.4500	0.4500	0.4500	1.0000	0.7231
Y	0.4231	0.5231	0.6231	0.7231	1.0000

Beta Weights				
	X_1	X_2	X_3	X_4
B	-0.0288	0.1530	0.3348	0.5166

$R = 0.6500$ $R^2 = 0.4225$

Table 3

Population C

Correlation Matrix					
	X_1	X_2	X_3	X_4	Y
X_1	1.0000	0.6500	0.6500	0.6500	0.4616
X_2	0.6500	1.0000	0.6500	0.6500	0.5616
X_3	0.6500	0.6500	1.0000	0.6500	0.6616
X_4	0.6500	0.6500	0.6500	1.0000	0.7616
Y	0.4616	0.5616	0.6616	0.7616	1.0000

Beta Weights				
	X_1	X_2	X_3	X_4
B	-0.1971	-0.0212	0.3744	0.6601

R = 0.6500 R^2 = 0.4225