
References

Morrow-Howell, N. (1994). The M word: Multicollinearity in multiple regression. *Social Work Research, 18*(4), 247-251.

Section:
RESEARCH NOTE

THE M WORD: MULTICOLLINEARITY IN MULTIPLE REGRESSION

The existence of a substantial correlation between two or more independent variables creates problems of multicollinearity in multiple regression. In social work research, the independent variables are usually intercorrelated, sometimes strongly; thus, multicollinearity must be understood and handled correctly. Social work researchers learn early that multicollinearity is a problem that should be avoided in developing explanatory models. Social workers learn the definition of multicollinearity and learn to avoid placing two highly correlated independent variables in the same regression equation. However, there are a few points that never seem to be fully articulated. How high a correlation is "too high"? How does one know if a model has a multicollinearity problem? How is the model jeopardized if there is multicollinearity? Often, answers to these questions remain elusive in textbooks, and teachers gloss over the details. The aim of this research note is to clarify the problems created by, the detection of, and ways to deal with multicollinearity.

PROBLEMS CREATED BY MULTICOLLINEARITY

Multicollinearity concerns only the relationships of the independent variables. Independent variables are highly correlated when information is redundant. Although there may be good reason for a researcher to collect redundant information (for example, two measures of depression or three measures of socioeconomic status), redundancy is not desirable in multiple regression. Researchers should not put two repetitive measures, such as two measures of depression, in the same model (Gordon, 1968). More likely, the redundancy occurs because two conceptually distinct phenomena are related, for example, depression and self esteem. When redundancy is high, several estimation problems result.

When a correlation between independent variables is one (r = 1), and there is perfect collinearity, the parameter cannot be calculated; perfect collinearity makes the necessary calculations mathematically impossible. Researchers could make this mistake if a categorical variable is converted to a series of dummy variables for inclusion in a regression model and all categories are retained in the model, or if a total score is computed from several subscores and the total and subscores are included in the same model. Computer printouts include an explanation that the matrix is singular and that an estimate cannot be produced.

Large correlations do not prevent the inversion of the matrix, but the regression coefficients produced have two problems: (1) inflated standard errors on the parameter estimates and (2) reduced magnitude of the parameter estimates.

Inflated Standard Errors

The greatest problem is that the standard errors calculated on the partial slopes of highly correlated independent variables are inflated. As the equation below shows, R^2_i in the formula for the standard error of the estimate.

$$S_{by1.2...k} = \text{Square root of } S^2_{y.12...k}/\text{Sigma } x^2_1 (1 - R^2_{1.2...k}$$

where $S_{by1.2...k}$ is the standard error of the parameter estimate, $S^2_{y.12...k}$ is the error variance in the model; X^2 is the variance of the independent variable on which the estimate is being made; and $R^2_{1.2...k}$ is the coefficient of determination or the R^2 that results when X_1 is regressed on the other independent variables. (In this research note, the coefficient of determination resulting from X_1 being regressed on X_2, $X3_3$, . . . and X_n or from X_2 being regressed on X_1, X_3, . . . and X_n will be referred to as R^2_i.) A large R^2 is associated with a large error term.

Larger error terms lead to difficulty in achieving a large t ratio in the significance test on the parameter as well as large confidence limits around the parameter. Thus, statistical power is reduced, and there is a lesser probability of claiming statistical significance. Because of this effect, when there is multicollinearity in the model, a researcher can have a significant model R^2 with no statistically significant variables, because none of the t ratios reach significance.

Table 1 contains two correlation matrices taken from Pedhazur (1982). The correlation between the three independent variables and the dependent variables is the same, as are the relationships between the independent variables, with one exception: In matrix A, variable 3 and variable 2 have a correlation of .10; in matrix B, that correlation is .85, and clearly multicollinearity is introduced in the model. The parameter estimates in model B are reduced because highly correlated variables are being partialed from each other. The standard errors are greatly increased in model B. Whereas all three independent variables are statistically significant in model A, only the first one is in model B. Thus, even though variables 2 and 3 have the same relationship to the dependent variable in each matrix, the high correlation between two of the independent variables changes the estimates dramatically.

Inflated standard errors cause the partial slopes to be unstable from sample to sample (Hays, 1981; Neter, Wasserman, & Whitmore, 1978). If an estimate has a large standard error, by definition it is more likely to vary between samples. This standard error of the estimate will be

expressed from sample to sample. Mathematically, large differences in parameter estimates can occur between samples, even if the samples are only slightly different. Small changes in the relationship between the dependent and independent variables can lead to greatly different parameter estimates, even when the interrelationships of the independent variables are constant (although multicollinear). Pedhazur (1982) demonstrated that in the presence of highly correlated predictors, the estimates generated on two samples are quite different. In both samples, the interrelationships of the independent variables are identical, and all but one of the correlations between the independent and dependent variables are identical. One X Gamma correlation is changed from .50 to .52, and the parameter estimates, error terms, and significance test scores are. different, even to the point that an estimate is significant in one model and not the other.

Thus, the magnitude of the partial slope estimate can vary from sample to sample not because the samples are different, but because of the mathematics involved with large error terms. Thus, the terms "unreliable" and "unstable" are used to describe the estimates on samples in which multicollinearity problems exist. The researcher does not know if the estimates on the sample at hand reflect what would be seen on the next sample or the true relationship of the variables in the population.

Reduced Magnitude of Parameter Estimates

A second problem concerns the regression coefficients themselves. The simultaneous analysis of two highly correlated variables reduces the magnitudes of the partial coefficients. The variables involved will lay claim to largely the same portion of the Gamma variance. Thus, neither can make much of a unique contribution, and interpretation of the partial coefficients of such a set of variables will be misleading (Cohen & Cohen, 1983). The reduction in the magnitude of the parameter estimate as well as the lack of the possibility of unique determination was illustrated by Kleinbaum, Kupper, and Muller (1988), who suggested using two X_1 s in a model to represent two perfectly correlated or redundant variables:

$$Gamma_i = B_0 + B_1 X_{i1} + E_i.$$

The calculation of the parameter estimate can be expressed as

$$B_j = C'_j (1/0),$$

which shows that the estimate is indeterminant because of a denominator of zero. However, this model can also be rewritten as

$$Gamma_i = B_0 + (B_1 + B_2) X_{i1} + E_i,$$

which shows that an infinite number of values add to the same coefficient value. Thus, if the estimate of the slope of X_1 is really 10, then B_1 and B_2 could each be 5; or B_1 could be 2 and B_2 could be 8, or 4 and 6, and so forth. These various combinations would be seen from sample to sample, clearly demonstrating the meaning of unstable. This phenomenon can lead to parameter estimates in the wrong direction or of an unexpected magnitude. The removal of one of the redundant variables would allow the unique estimation of the parameter (which would be 10 in this example).

DETECTION OF MULTICOLLINEARITY

There is no agreement about what constitutes "too high" of a correlation between independent variables, and there is no magic number. The point at which a problem is created varies according to the data at hand. Although a correlation of .80 is probably the most frequently cited guideline (Lewis-Beck, 1980), this is misleading, because problems can occur at lower levels of correlation, even at a moderate correlation of .40 in some samples (personal communication with E. Spitznagel, professor, Math Department, Washington University, St. Louis, January 20, 1992). Unfortunately, Pedhazur (1982) concluded that the existence of too high a correlation remains a judgment on the part of the researcher.

Most often, researchers determine the magnitude of the correlation among independent variables by inspecting a zero-order correlation matrix, or each independent variable with one other independent variable. Yet this approach ignores the fact that multicollinearity exists when one variable is a linear combination of any of the other independent variables. Thus, when a researcher regresses one variable on the other three, four, or five variables in the model, a substantial r^2_i presents the same problems. Thus, a researcher must inspect each independent variable as linear combinations of all others.

It may be simpler to examine each variable as a linear combination of the others after parameter estimates have been calculated, because a researcher can figure out the R^2_i for each independent variable by using the following formula:

$$R^2_1 = 1 - (1 - R^2) F_i/(N - k - 1) B^2_1$$

where R^2_i is the squared multiple correlation of X' with the other variables, R^2 is the squared multiple correlation of the model, F_i is the F ratio for testing the significance of the regression coefficient for X_i, B_i is the standardized regression coefficient for X_i, N is the sample size, and k is the number of independent variables. (The researcher may need to standardize the coefficients and transform t ratios to F ratios.)

The variance inflation factor (VIF) is also used to measure the multicollinearity in a specific model (Kleinbaum et al., 1988). The VIF = 1/1 - R_{2i}, (The tolerance is 1/VIF or, simply, 1 - R^2_i. Some researchers prefer to use the tolerance, which contains the same information as the VIF.) This quantity is part of the estimate of the variance for a particular parameter, and that standard error is proportional to the VIF, in that

$$S^2/_{bi} = S^2/(n - 1)S^2_i \times 1/1 - R^2_i$$

Note that the standard error of the estimate includes a measure of the estimated error variance in the model (S^2), as well as a measure of variance of the independent variable of interest (S[sup 2$_i$]). Thus, the R^2_i part of a factor that inflates other error measures as it produces the standard error of the estimate. The larger the VIF, the larger the standard error or the more inflated the standard error.

Kleinbaum et al.'s (1988) rule is that any VIF greater than 10 is problematic. Fox (1991) suggested using the square root of the VIF and demonstrates that a squareroot VIF greater than 2 can impair the precision of the parameter estimate. (This is a more conservative rule,

because the square root of 10 is more than 2.) In Fox's example, squareroot VIFs of this magnitude corresponded to zero-order correlation coefficients from .535 to .738. Thus, by calculating the VIF, which can be done easily through computerized statistical packages, the researcher has a direct index of the extent to which collinearity harms an estimation (Fox, 1991).

Multicollinearity can also be diagnosed through the use of eigenvalues computed on the correlation matrix of the predictor variables. Computerized regression diagnostic procedures will produce the eigenvalues directly. A principal components analysis is used, creating a set of new variables that are linear combinations of the original variables. The variances of these new variables are called "eigenvalues," and the larger the eigenvalue, the more important the principal component is in representing the information in the predictors (Kleinbaum et al., 1988). If an eigenvalue is close to 0, it is not important in representing information in the original matrix; all information is represented by the other components. Thus, an eigenvalue close to 0 indicates the presence of multicollinearity among the original predictors. An eigenvalue of 0 means a linear dependency exists.

The eigenvalue is used to produce several indexes, one of which will be reported here. The condition number (CN) associated with a particular eigenvalue is the square root of the largest eigenvalue divided by that particular eigenvalue. Thus,

CN_k = square root of $Lambda_1/Lambda_k$,

where $Lambda_1$ is the largest eigenvalue and $Lambda_k$ the eigenvalue of variable k. Belsley, Kuh, and Welsch (1980) suggested that a CN of 10 represents moderate collinearity, whereas 30 represents severe collinearity. Spitznagel (personal communication, January 20, 1992) warned that error inflation may be hurting the estimates when the CN approaches 6. Like the VIF, eigenvalues and CNs can be produced as a regular part of the regression output so that multicollinearity can be evaluated.

WAYS TO DEAL WITH MULTICOLLINEARITY

It is always best to reduce the number of independent variables for conceptual reasons before multicollinearity problems arise in the calculations of the model estimates. Researchers are continually reminded that models with fewer variables that are conceptually sound are better than models with more marginally related variables. Researchers can usually add a variable or two that reach significance and boost the model R_2 by a percentage or two, but their addition may be at the expense of introducing problems with multicollinearity and jeopardizing the calculations on the rest of the variables.

The simplest solution is to omit the offending variable if only one correlation is the problem, which may be desirable if the researcher finds appeal in a simpler model. For example, marital status and age are too highly related in some samples of elderly people. The researcher may believe that marital status is the critical variable, and age matters only because of its influence through that variable. Thus, a simple solution is to drop the age variable. However, this solution may lead to model specification error or the inability of the model to test an original research hypothesis.

Another solution is to combine information from two variables into one. If the shared variance is attributable to a single central property or latent variable, it is most appropriate to combine these variables into a single index (Cohen & Cohen, 1983). Factor analysis can be used to identify variables associated with a latent factor; a factor score can be used in place of the measures of the individual variables. Simply combining the variables, perhaps through standardization and then addition, is another possibility if there is evidence through factor analysis or alpha coefficients that this reduction is appropriated. For example, in a sample of caregivers to Alzheimer's patients, three items were related to stress created by the caregiving (Mud, 1992). An alpha coefficient approaching .80 suggested that they could be combined; after a t transformation, the three measures were added to create a single stress variable. Finally, several related variables may call for the use of a latent variable causal model (Cohen & Cohen, 1983).

Rescaling the data has been suggested as a way to reduce multicollinearity, although there is disagreement about when it is helpful (Kleinbaum et al., 1988). The process of subtracting a constant from a variable is a form of scaling, and if that constant is the mean of the variable, the process is called "centering." Especially in polynomial regression models (in which there is high correlation between different powers of the same variable), centering the original data can relieve multicollinearity problems.

Two alternative solutions allow for retaining all the independent variables as they are. One approach is not to interpret the individual parameters at all but only use the model information, the R^2. This number reflects the explanatory power of the variables taken all together and thus is not affected by multicollinearity. This is not likely to be an option that social work researchers choose, because predictive ability is usually not the sole interest.

Instead researchers want to understand the relationship of the independent variables to the dependent variable when controlling for the other variables; therefore, inspection of individual variables is required. Thus, researchers must depend on hierarchical rather than simultaneous regression. If the hierarchical sums of squares information (Type I in SAS) and the associated F ratios are used, the estimates under scrutiny are the increments to the R^2 in the hierarchical setting. The researcher will see the contribution of each independent variable as it is entered in a particular order. For example, living alone and widowhood may be highly related in a sample. If the researcher conceptualizes widowhood as the first event and considers the effects of living alone after the effects of being widowed, the researcher can see the effects of widowhood without controlling for living alone and see the effects of living alone with all the variance from widowhood removed. The parameter estimates and standard errors are still under the influence of multicollinearity, so they should not be interpreted.

Hierarchical regression can also be used in another way. When variables are subgrouped into related factors, such as demographic, social support, or health factors, and the multicollinearity exists within each factor, the variables can be used as they are, but the increment to the R^2 can be presented by blocks of factors. For example, after the demographic variables are entered, the social support variables are entered in a second run. The difference in the model R^2 is the amount contributed by the social support factors. The relative contribution within the block is still subject to multicollinearity problems, so interpreting individual variables' contribution should not

be done.

Another solution is to use regression techniques other than least squares. Ridge regression is becoming more popular, and this procedure was designed to overcome situations plagued by multicollinearity (Draper & Smith, 1981). Ridge regression is a biased estimation procedure in that a small amount of bias is introduced in the coefficient estimates for a trade-off of a substantial reduction in sampling variance (Fox, 1991). The use of ridge regression is still controversial and rare.

CONCLUSION

Social work researchers are likely to encounter problems with multicollinearity in regression analysis, because independent variables of interest are often highly correlated. With the computer software currently available, it is easy to examine regression diagnostics that indicate if a multicollinearity problem exists in a regression model. The researcher then has a number of options to remedy the situation. In presenting regression estimates, the researcher should indicate that multicollineanty has been addressed, so that the reader can be assured that the regression results are free from associated problems.

TABLE 1--Correlation Matrices and Models That Show Multicollinearity

Variable	1	2	3	Y	Beta	SE	
Matrix A							
1	1.00	.20	.20	.50	.34		.07
2	.20	1.00	.10	.50	.39		.07
3	.20	.10	1.00	.50	.39		.07
Gamma	.50	.50	.50	1.00		R[sup2] =	.56
Matrix B							
1	1.00	.20	.20	.50	.41		.08
2	.20	1.00	.85	.50	.22		.15
3	.20	.85	1.00	.50	.2		.15
Gamma	.50	.50	.50	1.00		R[sup2] =	.43

Variable	F
Matrix A	
1	24.025
2	32.360
3	32.360
Gamma	
Matrix B	
1	27.07
2	2.38

3 2.38
Gamma

SOURCE: Exceprts from Multiple Regression in Behavioral Research: Explanation and Prediction, 2nd edition, by Elazar J. Pedhazur, copyright [C] by Holt, Rinehart, and Winston, Inc., reprinted by permission of the publisher.

REFERENCES

Belsley, D., Kuh, E., & Welsch, R.(1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York: John Wiley & Sons.

Cohen, J., & Cohen, P.(1983). Applied multiple regression/ correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum.

Draper, N., & Smith, H. (1981). Applied regression analysis. New York: John Wiley & sons.

Fox, J. (1991). Regression diagnostics. Beverly Hills, CA: sage Publications.

Gordon, R.A. (1968). Issues in multiple regression. American Journal of Sociology, 73, 592-616.

Hays, W. (1981). Statistics. New York: Holt, Rinehart, & Winston.

Kleinbaum, D., Kupper, L., & Muller, K. (1988). Applied regression analysis and other multivariable methods. Boston: PWS-Kent.

Lewis-Beck, M. (1980). Applied regression: An introduction. Beverly Hills, CA: sage Publications.

Mui, A. C. (1992). Caregiver strain among black and white daughter caregivers. Gerontologist, 32, 203-212.

Neter, J., Wassermann, W., & Whitmore, G. (1978). Applied statistics. Boston: Allyn & Bacon.

Pedhazur, E. (1982). Multiple regression in behavioral research: Explanation and prediction (2nd ed.). New York: Holt, Rinehart, & Winston.

Original manuscript received September 11, 1992
Final revision received September 9, 1993
Accepted September 30, 1993

An earlier version of this research note was presented at the Annual Pro,gram Meeting of the Council on Social Work Education, March 1992, Kansas City, MO.

~~~~~~~~

By Nancy Morrow-Howell, PhD, ACSW, Associate Professor and Investigator, Center for Mental Health Services, George Warren Brown School of Social Work, Washington University Campus Box 1196, St. Louis, MO 63130
~~~~~~~~