Coping With Collinearity

Greg Mason
University of Manitoba Research Ltd.

RÉSUMÉ

Les désordres co-linéaires sont très communs dans les analyses par régression. Les moyens courants pour évaluer la sévérité de ce désordes, tel que l'examination de la stabilité des paramètres quand on enlève des variables des l'équation, sont seulment suggestifs. Cet article passe en reveu les méthodes récents pour analyser ces désordres. Ces méthodes comprennent les indicateurs conditionels et la décompostion de la variation du groupe de variables et sont présentes dans plusiers ensembles statistiques pour ordinateurs. Des stratégies pour corriger ce problème sont examinées. En générale, on ne dit pas seulment laisser tomber les variables de l'équation à moins qu'ils soient rédondants à la théorie de base ou à la logique du programme.

ABSTRACT

Collinearity is very common in linear regression. The common methods for diagnosing this disturbance, such as evaluating parameter instability when variables are removed from the specification are only suggestive. Recent developments are reviewed which assist in diagnosing collinear disturbances. These include condition indexes and variance proportions decompositions and are available in a number of statistical packages. Some corrective strategies are also examined. In general, it is not correct to simply drop variables from the specification unless they are redundant to the program logic and underlying theory.

Introduction

The article by Eaton and Smith in this issue makes the essential point that dropping variables from regression models (or any statistical model) is wrong. In my review, I appear to recommend this as a practice. This is not the case. Dropping the intercept in the simple case cited, would probably have indicated the presence of collinearity by revealing a slope coefficient which was quite different in the two specifications. Although not an infallible guide to collinearity it can be suggestive. Under no conditions is dropping a variable a cure for collinearity.

Eaton and Smith are right to emphasize balancing the sample as the most appropriate remedy. Collinearity is the most troubling of regression disturbances since its diagnosis is not possible using conventional distribution based statistical tests. Furthermore in, simultaneous equations, it is likely that collinearity is not examined or evaluated. In practical evaluation situations, samples are often not randomly selected or balanced between treatments and controls. Evaluators must be specially careful to examine and present alternative specifications of the statistical models. Only in this way will data problems be revealed and correct advice be provided to the client.

Collinearity (multicollinearity) is the most important and intractable disturbance to linear regression models. Standard diagnoses such as examining the simple correlation or dropping variables/observations and judging the resultant variability of coefficients and associated statistics, may be inadequate.

This note reviews recent developments in diagnosing the presence of harmful collinearity and suggests some remedies. By necessity, the discussion is summary, with general rules of thumb replacing detailed proof. For this the reader is referred to the bibliography.

One caveat is important. The presentation here rests on the work of Belsley, Kuh, and Welsch (1980), and the reader should be cautioned that the entire area is under active research and debate. This note merely summarizes the present state-of-the-art; it may well change tomorrow.

What is Collinearity?

Collinearity is basically *not* a statistical problem. That is, according to the approach of BKW (1980), inferring the presence of "harmful" collinearity is not assisted by any statistical test in the classical research situation.

Rather, diagnosis of collinearity uses indicators developed by simulation and other trial and error processes, because collinearity is not due to failure in the least squares assumptions.

Intuitively, collinearity exists among two or more independent variables which are highly correlated. The effect of this is to produce regression estimates with inflated variances. In other words, the individual t values are unreliable, and it becomes difficult to test hypotheses on the regression parameters. A formal definition is supplied by Gunst (1984):

> "A collinearity is said to exist among the columns of $X = (X_1, X_2, \ldots X_p)$ if for a suitably predetermined $e_n > 0$ there exists constants $C_1, C_2, \ldots C_p$ not all zero, such that
> $C_1X_1 + C_2X_2 + \ldots C_pX_p = S$ with $||S|| < e_n||c||$"

A simple analogy is provided by Hocking and Pendleton (1983). A picket fence (where each picket represents an independent variable), has even spaces between the pickets. Collinearity exists when pickets overlap. The collinearity becomes more severe as individual pickets widen, and overlap other pickets, effectively hiding them from view. In other words, collinearity obscures the role of individual pickets (variables), and makes some pickets (variables) redundant.

Diagnosing Collinearity

3.1 *Conventional Indicators*

Conventional indications of collinearity are:

a. parameter instability when one or more variables are withdrawn from the regression;

b. instability in the estimates of standard error of the regression coefficients, when variables and/or observations are withdrawn;

c. large subsets of regressors (independent variables) statistically indistinguishable from zero, but high overall goodness of fit (R^2);

d. statistically significant pairwise correlations between independent variables.

The first indication that multicollinearity may be a problem in any specification occurs as variables are dropped/added to the equation. Previously significant (statistically) variables become insignificant, and may change sign. In simple cases, the pairwise correlations may provide straightforward indications of the offending variables. However, as a general rule, experimenting with specifications, and examining the correlation matrix are limited as collinearity diagnostics. Collinearity is frequently specific to model specification (what variables are included, and the algebraic form of the regression), and definition of the variables.

In fitting a regression, the analyst ought to have developed a clear structural model, for which the regression serves as a step in confirming the hypotheses underlying the research. Exploratory analysis, comprising bivariate plots, summary statistics, and various graphical approaches (stem-leaf plots, etc.) ought to be undertaken prior to the specification of a confirmatory model. Regression models, although ill-suited for exploratory analysis, can be used to evaluate the degree to what model parameters are sensitive to variable replacement or sample truncation. The final regression model should be theoretically grounded for evaluation purposes since these seek to confirm the presence or absence of program effects.

Also, each variable must be clearly understood in terms of its construction and definition. Aside from aiding in the interpretation of the regression parameters, the variable definition is essential to unraveling measures which may "overlap".

Recent Developments in Diagnosing Collinearity

a. *Condition Index*

The condition index, developed by BKW (1980), measures the degree to which regression parameters are influenced by small perturbations in the data matrix. As this index rises toward 10 to 15, the sensitivity of regression parameters to such disturbance increases. In general, BKW indicate that a condition index in excess of 20 is a clear indication of harmful collinearity. As the condition index rises, the 'R^2' on the underlying linear relationship (within the subset of offending independent variables) will rise.

b. *Variance-Decomposition Proportions*

The condition index signals that a subset of the independent variable are collinear. The variance decomposition matrix signals which variables are involved. This is a p x p matrix (where p is the number of independent variables). Reading across each row, any set of two entries or more which have values in excess of .5 signals a variable which is likely to contribute to harmful collinearity. Note that aside from the first row (corresponding to the intercept term), one entry will often be relatively high (.6 or greater). It is also likely that for any high index, several variables will be involved, and the variance decomposition will be distributed across the row. Therefore, if the condition index is high (20 or more), look for several "high" entries. With only two variables (i.e., pairwise correlation), the

values will be relatively high, around .7, but as the number of variables involved increases, the typical value will fall and be spread across a number of (but not all) entries.

c. *Variance Inflation Factors*

Variance inflation factors (VIF) provide a simple measure of the susceptibility of a variable to collinearity from other terms included in the regression. Generally, values in excess of 10 indicate that the variable is redundant. However, there is no real basis for such a rule of thumb. Also, the set of VIFs does not reveal near dependencies; that is, subsets of independent variables which are weakly collinear and which can degrade statistical testing of a regression.

Example

Consider the following regression run on the MINCOME Baseline data. This data set has 2173 observations, and is cross-sectional. While collinearity is pervasive for time series data, cross-sectional data can also exhibit this problem. The object of this regression is to estimate the determinants of household total net worth as a function of

a. total household income (F35)
b. family size (number of persons) (F8)
c. family size index (F11)
d. number of children under 6 (F10)
e. net worth (less house value) (NTWRTH1)
f. age of male head (F6)
g. age of female head (F7)
h. number of years in work force—male head (F56)
i. wage income in 1974 (INCOME74)
j. wage income in 1973 (INCOME73)

Clearly certain subsets of these variables, namely (a,i,j), (f,g,h), and (b,c,d) are likely to be collinear. (Note this regression is run solely to illustrate the diagnostics.) The results (using SAS) are as follows (the dependent variable is total household net worth).

Table 1a

Variable	DF	Parameter Estimate	Standard Error	T for HO: Parameter = 0	Variance Inflation
INTERCEP	1	−1446.55	829.90	−1.7	0.00
F35	1	−0.34	0.08	−4.0	3.70
F8	1	−337.23	327.69	−1.0	12.67
F11	1	12.72	18.82	0.6	14.40
F10	1	−606.88	267.15	−2.2	1.34
NTWRTH1	1	1.33	0.02	49.3	1.06
F6	1	39.80	12.16	3.2	2.40
F7	1	114.01	11.69	9.7	1.30
F56	1	74.33	25.62	2.9	2.80
INCOME74	1	0.47	0.12	2.8	7.59
INCOME73	1	0.33	0.10	3.3	3.52

As expected, several variables give indications of collinearity, in particular F8 and F11, and possibly INCOME74, however, we have no idea of how these may be linearly related. The condition index and variance decompostions proportion matrix provides useful clues as shown in Table 1b.

Table 1b

Variable	Condition Index	Variance Proportions 1	2	3	4	5	6	7	8	9	10	11
INTERCEP(1)	1.0	.0007	.0010	.0003	.0002	.0036	.0028	.0026	.0027	.0020	.0007	.0015
F35(2)	2.8	.0034	.0036	.0000	.0001	.0308	.0035	.0470	.0086	.1886	.0000	.0000
F8(3)	2.9	.0001	.0000	.0003	.0000	.0817	.7301	.0002	.0054	.0082	.0001	.0003
F11(4)	3.5	.0012	.0135	.0004	.0000	.5371	.1714	.0008	.0083	.0003	.0023	.0032
F10(5)	4.4	.0047	.0535	.0033	.0017	.0428	.0583	.0203	.1498	.0038	.0143	.0142
NTWRTH1(6)	5.6	.0301	.0206	.0041	.0003	.0028	.0044	.4372	.0721	.1032	.0055	.0330
F6(7)	6.6	.0032	.0443	.0233	.0055	.1849	.0056	.0222	.3663	.2511	.0055	.0439
F7(8)	8.7	.1050	.0185	.0182	.0061	.0571	.0157	.3996	.3007	.3822	.0128	.1561
F56(9)	10.4	.1511	.0509	.0295	.0008	.0073	.0000	.0431	.0402	.0377	.0997	.5971
INCOME74(10)	17.7	.2740	.7209	.1186	.0000	.0143	.0047	.0211	.0140	.0176	.7992	.1403
INCOME73(11)	32.6	.4264	.0732	.8058	.9854	.0374	.0026	.0082	.0318	.0072	.0598	.0104

The condition index rises to 32.5, indicating the possible presence of collinearity. Further reading across row 11 reveals large proportions associated with F8 and F11. In row 10, with a condition index of 17.7, there are large proportions associated with F35 and INCOME74.

None of this is surpirising. The definition of the variables is such that F8 and F11 are very close, as are F35 and INCOME74. While these pairwise relationships are revealed in the correlation matrix, if more than two variables are involved, the correlation matrix is not helpful. Revising the specification produces the results in Tables 2a and 2b. Here, the *theoretical* specification is improved which leads to the exclusion of redundant variables.

Table 2a

Variable	Parameter Estimate	Standard Error	T for HO: Parameter=0	Variance Inflation
INTERCEP	794.4	542.5	1.5	.00
F35	−.2	.1	−3.2	1.51
F8	286.1	131.1	2.3	1.91
NTWRTH1	1.4	.0	51.2	1.00
F6	24.5	12.3	1.9	2.30
F56	103.3	26.2	3.9	2.71
INCOME73	.6	.1	7.5	2.51

Table 2b

Variable	Condition Index	Variance Proportions 1	2	3	4	5	6	7
INTERCEP(1)	1.0	.0045	.0082	.0055	.0087	.0078	.0066	.0053
F35(2)	2.3	.0139	.0186	.0000	.1720	.0336	.1840	.0002
F8(3)	2.4	.0071	.0177	.0032	.7895	.0054	.0216	.0036
NTWRTH1(6)	4.1	.0396	.2538	.1134	.0108	.1257	.1434	.0312
F6(7)	4.9	.0279	.0969	.1906	.0111	.3938	.0258	.1091
F56(9)	6.6	.2734	.0106	.0262	.0008	.3858	.5535	.4556
INCOME73(11)	8.7	.6331	.5972	.6606	.0073	.0470	.0651	.3949

Note that despite the condition index rising to only 8.7, a number of near linearities are signaled. Ideally, only one entry in the row ever should be relatively large and stand out. This regression equation still has plenty of problems, not the least of which is a severe simultaneous equation bias because net worth type of variables appear on both sides of the regression equation. Again, specification error underlies the collinearity problem.

Summary And Diagnosis

Collinearity reduces the hypotheses-testing power of linear regression by inflating the variance on parameter estimates. In diagnosing collinearity, the analyst must proceed carefully. Recommended are:

a. careful exploratory analysis on all variables and key bivariate relationships;

b. analysis of the simple correlation matrix;

c. careful development of the structural model, and definition of variables;

d. collinearity should be suspected if

—VIF is greater than 10

—condition index is greater than 15; *and*

—variance proportion are relatively high on more than one variable.

These are likely to be the offending terms, but note that others could be involved.

Remedies: A Parting Word

The corrective action taken in "fixing" the model was to drop the offending variable. Unless some silly mistake has been made (as in the example above), care should be taken not to drop variables which have a theoretical role in one's statistical analysis.

Another common remedy is to transform the data. Standardized regression, centering, logarithms, first differences (on times series data) are transformations which frequently eliminate collinearity. They can also trigger collinearity. Unless these transformations are indicated by the theory, it is ill-advised to undertake them as a cure of collinearity. For example, polynomial transformations often induce collinearity in certain ranges of the data.

Specialized techniques such as mixed Bayesian and ridge regression procedures can be used, but require extra (prior) information beyond the scope of most research and evaluations.

Regression models are part of many advanced procedures in the social sciences. LISREL, ANCOVA, and other approaches all may contain collinear relationships which are hard to detect, expecially if the researcher simply lets the machine do the analysis. In my view, this is a major weakness of these new procedures.

In the final analysis, the focus must be on the realtion between theory and estimation. Collinearity seems to occur so frequently, simply because performing regression analysis is trivial using modern software. Coping with collinearity is primarily accomplished through care in the specification of the model. It makes no sense to correct for collinearity if the underlying model has silly mistakes in specification (as the example above), or because the variables are poorly understood.

Note:Both SAS (mainframe) and StatPac (Wolonick Assoc.) have collineary diagnostics.

References

Belsley, D. A., Kuh, E., and Welsch, R. E. *Regression Diagnostics.* New York: John Wiley, 1980.

Gunst, R. F., "Toward a Balanced Assessment of Collinearity Diagnostics." *American Statistician.* May, Volume 38, 1984, pp. 79-82.

Hocking, R. R. and Pendleton, O. J., "The Regression Dilemma," *Communication in Statistics.* A 12, 1983, pp. 497-527.