Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 13, Number 5, June 2008 ISSN 1531-7714

Revisiting the Collinear Data Problem: An Assessment of Estimator 'Ill-Conditioning' in Linear Regression

Karen Callaghan, *Texas Southern University*
Jie Chen, *University of Massachusetts, Boston*

Linear regression has gained widespread popularity in the social sciences. However, many applications of linear regression have been in situations in which the model data are collinear or 'ill-conditioned.' Collinearity renders regression estimates with inflated standard errors. In this paper, we present a method for precisely identifying coefficient estimates that are ill-conditioned, as well as those that are not involved, or only marginally involved in a linear dependency. Diagnostic tools are presented for a hypothetical regression model with ordinary least squares (OLS). It is hoped that practicing researchers will more readily incorporate these diagnostics into their analyses.

The linear regression model is at the core of social scientific research. Analysts estimate these models with the aim of interpreting the coefficient estimates as measures of the 'true characterstics' of a population. However, when collinearity is present, the value of the estimated coefficients in the sample may differ markedly from the true value in the population.[1] Unfortunately for social scientists, collinearity is the normal state of the world; independent variables are often linearly related to another independent variable or a subset of variables. Furthermore, collinearity is not simply present or not present, it occurs in degrees.[2]

Surprisingly, although most multivariate statistics texts address collineariy and the techniques for assessing collinearity are available in most statistical software (SPSS, SAS, STATA, S-Plus), many analysts fail to give serious consideration to the possibility of collinear data. Alternatively, researchers who find coefficients with large standard errors often incorrectly seize on collinearity as the reason. Consequently, faulty conclusions about the way the world works are inevitable.

The purpose of this research is to illustrate a useful, reliable method for evaluating collinearity in a multivariate model. Diagnostics are calculated for a hypothetical regression model with the aim of identifying the degree of collinearity and the variables that are involved (or not involved) in a strong collinear relationship. This article focuses on the detection of collinearity rather than on the procedures for combating it.[3] Our goal is to quantify the risks of ignoring collinearity for the practicing researcher.

IDENTIFYING THE PROBLEM OF COLLINEAR DATA

In a regression model, the coefficients are descriptive characteristics of the population from which the sample was taken. The estimated standard errors of the β coefficients are used for hypothesis testing. For instance, in regression analysis, one asks: "Does *x,* the regression variable truly influence *y,* the response?" The hypothesis of interest is often formulated as H_o: $B_1 = 0$ and H_1: $B_1 \neq$

[1] By collinearity we mean the case in which at least one variable is (practically) completely correlated with other predictors. We use the term synonymously with ill-conditioning.

[2] Perfect collinearity is quite rare, however, and usually attributed to data coding errors.

[3] Different remedies have been proposed including omitting variables, grouping variables in blocks, collecting additional data, and Ridge Regression, among others (see Fox, 1997; Weisberg, 2005 or Gujarti 1988). However, these remedies may be time consuming, costly, impossible to achieve or controversial (e.g., Maddala, 1992); thus diagnostic tools that signal the presence of collinearity are crucial.

0. If H_o is true, the implication is that the model reduces to $E(y) = B$, suggesting that *x* the regressor variable does not influence the response variable, at least not through the type of relationship implied by the model. If, however, H_o is rejected in favor of H_1, the implication is that *x* significantly influences the response *y*.

Population inferences depend on the accuracy of the estimate of the value of the population parameter. Large standard errors (low *t*-tests) and unstable coefficients (with implausible signs or magnitude) provide a red flag that interpretations of the relative importance of those parameter estimates are unreliable. Still, collinearity may be present in a model without these warning signs. When coefficient estimates are degraded, hypothesis tests do not possess the accuracy attributed to them. Unusually large standards errors generate the possibility of a *Type II* error. This reduction in statistical power increases the researcher's inability to replicate her findings with an independently drawn random sample from the same population. Two major methods researchers use to gain confidence in their findings are significance tests and randomly divided samples from the same population.

How do we know which parameter estimates are influenced by collinear relations, and which are unaffected and thus are reliable for further analyses? There are many statistical tests to guide us. These include, for example, (1) inspection of the correlation matrix of the x or explanatory" variables for pairwise correlations, (2) inspection of the correlation between various combinations of regression coefficients (see Ferrar & Glauber 1967), and (3) inspection of the tolerance levels and the variance inflation factors (VIFs). Method (1) has a significant drawback: one can examine only two variables at a time. Methods (2) and (3) consider the magnitude of the R^2 that results when X is regressed on the other independent variables. VIFs which measure the increase in the variability of the coefficient estimates over the orthogonal case (i.e., the case in which no collinearity exists). Although these are fairly reliable methods, it is difficult to determine the exact number of variables involved in near linear dependencies especially when there are several complex linear associations.

Other tests for assessing collinearity include (4) inspection of the "eigenvalues," and 5) a broader "eigensystem" analysis of the corresponding condition indexes and variance-decomposition proportions (VDPs). Methods (4) and (5) are generally considered best practices for assessing linear dependencies in model data. These methods, first proposed by Kendall (1957), have been more recently expanded in the field of applied econometrics (see Belsley, Kuh & Welsch 1980; Belsley, 1991a, b).

An *eigenvalue* (denoted by λ) is simply a number that characterizes in a single value the essential properties and numerical relationships within a matrix, hence the term "characteristic equation" (Coombs 1995). Table 1 presents guidelines for interpreting these values. A rule of thumb is that the greater the number of eigenvalues near zero, the greater the number of linear dependencies among the variables.

Table 1 Guidelines for Interpreting Collinearity Based on Eigenvalues

Degree of Collinearity	Form of Matrix	Magnitude of Eigenvalues
No Collinearity	Nonsingular	Not equal to zero
Near perfect Collinearity	Near singular	Close to zero
Perfect Collinearity	Singular (not positive definite)	Equal to zero (estimation terminated)

What constitutes a "small" eigenvalue? In other words, how close to zero must the values be? To address this question, researchers often analyze the *spectrum* of eigenvalues. This measure, called the *condition number,* is the ratio of the largest to the smallest eigenvalue ($\lambda_{max}/\lambda_{min}$). A related diagnostic, the *condition index*, provides another yardstick against which smallness can be measured. Condition indexes (CI) are calculated as follows:

$$CI_i = \left(\frac{\lambda_{\max}}{\lambda_i} \right)^{1/2} \quad for \quad i = 1,2. \qquad (1)$$

Condition indexes, often called the "complaint number" (Maddala,197 reveal the number and relative strength of the near dependencies. A high condition index indicates the presence of collinearity. A low condition index indicates near perfect collinearity. The guidelines for assessing condition numbers and indexes are shown in Table 2. These thresholds, however, are not akin to a classical significance level (e.g., $p \leq .05$) that must be chosen *a priori*. Instead, they are selected relativistically, depending on the patterns of the

condition indexes that arise (Belsley, 1991a, p. 38), a point to be explained shortly.

Table 2 Guidelines for Interpreting Collinearity Based on Condition Numbers and Indexes[a]

Condition Number $(\lambda_{max}/\lambda_j)$[b]	Degree of Collinearity
If (CN < 100)	Weak
If (100< CN < 1000)	Moderate to Strong
If (CN > 1000)	Severe
Condition Index $(\lambda_{max}/\lambda_j)^{1/2}$	
If (CI < 10)	Weak
If (10 < CI < 30)	Moderate to Strong
If (CI > 30)	Severe

Notes: [a]Based on values reported in Gujariti (2002). [b]Other programs (e.g., *SAS* and *S-plus*) define the condition number as the *square root* of this ratio. For this quantity, the rough cut offs are as shown below in the Condition Index subsection.

Variance-decomposition proportions (denoted by π) are closely related to the concept of eigenvalues however they give us more detailed information. The variance of the OLS regression coefficients can be shown to be equal to the residual variance multiplied by the sum of the variance proportions of all eigenvalues.[4] The criteria for a high VDP vary among researchers. The most common threshold is a VDP of .50 or greater for two or more variables associated with a high condition index.

In sum, the suggested procedure for diagnosing collinearity is a high condition index, which is also associated with a high variance-decomposition proportion for two or more regression coefficient variances. With this information in hand, in the next section, we apply diagnostic methods to a hypothetical regression model. Fortunately for the researcher, diagnosing any given data set for the presence of near linear dependencies and assessing their impact on regression estimates is a straightforward process.

DEMONSTRATING THE DIAGNOSTIC APPROACH

Suppose we wish to analyze the following regression model where *y* is an interval-level response variable, *Xi, i= 1, ... ,13* represents 13 independent variables, and ε is the error term:

$$Y = \beta_0 + \Sigma\beta_i X_i + \varepsilon \quad i=1,2,3, \ldots 13 \qquad (2)$$

To permit direct comparison of the variable coefficients, all variables were rescaled to range from 0 to 1. Using the rescaled variables, the ordinary least square linear regression modeling procedure of *SPSS* version 15 (Chicago, SPSS) provided the following equation:

$$\hat{Y} = 4.24 + .01X_1 + .20X_2 - .24X_3 - 1.72X_4{}^* + 1.50X_5{}^* + 2.06X_6 + .15X_7 + 1.92X_8 - .11X_9 + 2.21X_{10} + 4.63X_{11} + 1.19X_{12}{}^* + 1.76X_{13} \qquad (2)$$

Coefficient estimates β_4, β_5, and β_{12} are significant at the $p < .05$ level (two-tailed test). The standard errors for each of the coefficients are shown in Table 3.

Table 3: Regression Coefficient Standard Errors

Variable	0	1	2	3	4	5	6
Coefficient	10.48	.08	.17	.19	.73	.72	1.63
Variable	7	8	9	10	11	12	13
Coefficient	.10	1.18	.98	2.80	2.92	.49	1.24

While not statistically significant, the relative magnitudes of the coefficients β_8, β_{10}, β_{11} and β_{13} are also quite large (standard errors aside). Furthermore, the intercept β_0 has an aberrantly large standard error providing a clue to a variance inflation problem. As noted before, in the presence of collinearity, parameter estimates become very unstable: that is, sensitive to random error, as reflected in large standard errors of β. Do some parameter estimates have insignificant *t*-ratios

[4] Let $\mathbf{v}_i = (v_{i1}, \ldots, v_{ip})$ be the eigenvector associated with eigenvalue λ_i. Also, let $\Phi_{ij} = v^2_{ij}/\lambda_i$ and $\Phi_j = \sum_{i=1}^{P} \Phi_{ij}$ The VDP for the *j*th regression coefficient asasociated with the *i*th component is defined as $\pi_{ij} = \Phi_{ij} / \Phi_j$.

TABLE 3: Collinearity Diagnostics for the Hypothetical Regression Model

x	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Eigenvalue, λ	.001	.002	.010	.013	.030	.052	.069	.095	.174	.288	.354	.499	.769
Condition index	138	78	34	29	19	15	13	11	8	6	6	5	4
Variable	Variance Decomposition Proportions												
Intercept	**.883**	.111	.004	.001	.001	.000	.000	.000	.000	.000	.000	.000	.000
X_1	.003	.093	**.603**	.258	.028	.012	.000	.003	.000	.000	.000	.000	.000
X_2	.046	**.399**	**.027**	**.477**	.016	.019	.013	.002	.001	.001	.000	.000	.000
X_3	.006	.205	.016	.155	.364	.069	.035	.001	.088	.010	.042	.009	.000
X_4	**.752**	.247	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
X_5	.006	.263	.013	.050	.000	.000	**.568**	.165	.007	.003	.002	.009	.010
X_6	.008	.345	.087	.012	.282	.210	.000	.054	.000	.000	.000	.000	.000
X_7	.002	.041	.355	.135	.019	.001	.008	.000	.286	.003	.144	.158	.100
X_8	.012	.271	.084	.215	.004	.019	.061	.058	.008	.199	.058	.006	.004
X_9	.000	.004	.114	.022	.131	.060	.014	.069	.422	.032	.097	.013	.001
X_{10}	.001	.239	.008	.094	.328	.182	.051	.000	.039	.006	.000	.040	.022
X_{11}	.222	**.555**	.176	.038	.005	.001	.000	.002	.000	.000	.000	.000	.000
X_{12}	.022	.304	.002	.034	.134	.001	.163	.287	.002	.011	.027	.005	.000
X_{13}	.146	.002	.123	.044	.001	.108	.011	.208	.027	.205	.005	.021	.091

due to excessive linear dependencies? To address this question, we used *SPSS,* selecting Collinearity Diagnostics in the Linear Regression dialog box.

The results in Table 4 lead to the following observations. First, at least one eigenvalue represents a near serious linear dependency. In fact, there are eight very small (near zero) eigenvalues symptomatic of seriously ill-conditioned data: λ_1 =.001, λ_2 =.002, λ_3 = .010, λ_4 = .013, λ_5 =.030, λ_6 =.052, λ_7 = .069, and λ_8 = .095. Moreover, three CIs exceed 30 indicating moderate to severe collinearity (see Table 2). These include CI_1 = 138, CI_2 = 78, and CI_3 = 34. A fourth condition index, CI_4= 29, should also be considered; it is close in the order of magnitude to 30 and reveals a gap in the numerical progression (between 29 and 19). The relative strengths of the CIs are determined by their position along the progression.

Which coefficient estimates are adversely affected by the presence of those near dependencies? To address this question, we examine the variance-decomposition proportions. VDPs are arranged in a 13 × 13 matrix in Table 4. The rows of the matrix represent the 13 variances of the regression estimates, thus each row must sum to one. A variable is considered involved in (and its corresponding regression coefficient degraded by) at least one near dependency if the total proportion of its variance associated with a CI, or a set of CIs in a numerical progression, exceeds 0.5.

Clearly, two variables in the first column of Table 4, i.e., the Intercept and X_2 are involved in a severe linear dependency. In fact, this condition index accounts for 88.3% of the variance of the Intercept as shown by the value of 0.883 in column 1. The linear dependency also damages the coefficient for X_4 accounting for 75.2% of its variance (VDP = 0.752. The other coefficients in that column are not affected.

The next strongest dependency CI 78 (column 2) involves X_{11}, accounting for 55.5% of its variance. This could also involve the Intercept and X_4 due to the dominance of CI =138. A *dominating* dependency occurs when the CI is in an order of magnitude larger than the other CIs.

A third condition index CI 34 (column 3) seems to involve X_1, accounting for 60.3%, although no other

variables in that column seem to be affected.

Variance-decomposition proportions can also be used to identify *competing* dependencies. A competing dependency exists when the sum of the VDPs have a set of condition indexes of the same order of magnitude (in this case, greater than 19) exceeds the value of .50. The aggregate proportions for X_2 = .903 (from .399 + .027 + .477) suggest that its regression coefficient may be degraded as well.[5]

An equally important question is which variables are unaffected by the collinear relations. A variable is not involved in a linear dependency if the total proportion of its variance associated with a set of *small* condition indexes exceeds .50. Table 4 highlights the information that indicates that X_5 with a VDP of 0.568 is associated with the smaller condition indexes. Furthermore, it is only weakly involved in the stronger near dependencies (CIs above 19). Therefore, the lack of statistical significance for β_5 in Equation 2 is not due to ill-conditioned data; this variable likely has no real impact on the dependent variable.

At this point, we have confirmed the existence of several near dependencies and adequately identified the variables involved. However, the analyst may wish to develop an even more nuanced evaluation of the variables involved by forming a set of auxiliary regressions that displays the structure of these linear dependencies in greater detail (for this approach, see Belsley 1991a).

CONCLUSION

A decisive feature of multivariate models is their collinearity. In this case, the standard errors of the regression coefficients increase dramatically, thereby reducing *t* values. Such inflated variances preclude the use of regression as a basis for hypothesis testing. Moreover, standard errors become very sensitive to even the slightest change in the data, making it impossible to replicate the findings in an independent random sample from the same population—crucial for the day-to-day researcher.

In this paper, we used eigenvalues and condition indexes from a hypothetical regression model to illustrate the best practices for collinearity diagnostics. This procedure allowed us to identify the variables that were either ill-conditioned, only marginally involved in a linear dependency, or not adversely affected. Our findings revealed that collinearity resulted in poor efficiency in the estimation of the model Intercept and the β_4 and β_{11} coefficients. A less severe, though still important dependency may have obfuscated the true impact of β_1 and β_2. Although these variables may have a real impact on the dependent variable, collinearity clouds our assessment. For these coefficients, we can not make defensible population inferences.

By the same token, we were able to isolate the variable that was not involved in a linear dependency. Thus, as we demonstrated, knowing that collinearity may exist is not equivalent to knowing it is a debilitating problem for the investigation. The problem requires careful thought and subtle analyses.

To date, econometricians and biostatisticians are more likely to properly address the issue of collinearity than social scientists. However, social scientists should be held to the same standard. Without a precise understanding of the standard errors of model coefficients, there is no population one can reasonably infer. Moreover, the sample data may be compatible with a diverse set of hypotheses, thus the probability of accepting a false hypothesis increases. In applied fields like education, counseling and administration, the importance of the accuracy and interpretation of model coefficients looms especially large. The information presented here, if properly applied, can be used effectively to understand the reliability of a regression model for the purposes of research and policymaking. We hope this article will encourage researchers to adopt this more precise diagnostic approach.

REFERENCES

Belsley, D. A . (1991a). A guide to using the collinearity diagnostics. *Computer Science in Economics and Management*, 4, 33-50.

Belsley, D. A. (1991b). *Collinearity diagnostics: Collinearity and weak data in regression*. New York: John Wiley & Sons.

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

[5]If, however, the researcher's interest centers on whether a particular coefficient is significantly positive and, despite the presence of collinearity, is able to accept the hypothesis on that basis of the relevant *t*-test, then collinearity is not a problem.

Coombs, W. T. (1995). What are eigenvalues? D*isseminations of the International Statistical Applications Iinstitute,*1, 90-91.

Farrar, D. & Glauber, R. R. (1967). Multicollinearity in regression Analysis: The problem revisited. *Review of Economics and Statistics*, 49, 92-107.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage Publications.

Gujarati, D. N. (2002). *Basic econometrics*. New York: McGraw-Hill.

Kendall, M. G. (1957). *A course in multivariate analysis*. London: Griffin.

Maddalla, G.S. (1992). *Introduction to Economics*. New York: Macmillan.

Weisberg, S. (2005). *Applied linear regression*. New York: John Wiley & Sons.

Citation

Callaghan, Karen J. and Chen, Jie (2008). Revisiting the Collinear Data Problem: An Assessment of Estimator 'Ill-Conditioning' in Linear Regression. *Practical Assessment Research & Evaluation*, 13(5). Available online: http://pareonline.net/getvn.asp?v=13&n=5

Authors

Karen Callaghan (Correspondence Author)
Political Science Department
Texas Southern University
3100 Cleburne Street, Houston, Texas 77004
(713) 313-4803, Email: callaghank @ tsu.edu

Jie Chen
Office of Graduate Studies
University of Massachusetts, Boston MA 02115
(617) 287 5241, Email: jie.chen@umb.edu