Generalized Collinearity Diagnostics

Author(s): John Fox and Georges Monette

Source: *Journal of the American Statistical Association*, Mar., 1992, Vol. 87, No. 417 (Mar., 1992), pp. 178-183

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2290467

Generalized Collinearity Diagnostics

JOHN FOX and GEORGES MONETTE*

Working in the context of the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, we generalize the concept of variance inflation as a measure of collinearity to a subset of parameters in $\boldsymbol{\beta}$ (denoted by $\boldsymbol{\beta}_1$, with the associated columns of $\mathbf{X}$ given by $\mathbf{X}_1$). The essential idea underlying this generalization is to examine the impact on the precision of estimation—in particular, the size of an ellipsoidal joint confidence region for $\boldsymbol{\beta}_1$—of less-than-optimal selection of other columns of the design matrix ($\mathbf{X}_2$), treating still other columns ($\mathbf{X}_0$) as unalterable, even hypothetically. In typical applications, $\mathbf{X}_1$ contains a set of dummy regressors coding categories of a qualitative variable or a set of polynomial regressors in a quantitative variable; $\mathbf{X}_2$ contains all other regressors in the model, save the constant, which is in $\mathbf{X}_0$. If $\sigma^2\mathbf{V}$ denotes the realized variance of $\hat{\boldsymbol{\beta}}_1$ and $\sigma^2\mathbf{U}$ is the variance associated with an optimal selection of $\mathbf{X}_2$, then the corresponding scaled dispersion ellipsoids to be compared are $\mathcal{E}_V = \{\mathbf{x} : \mathbf{x}'\mathbf{V}^{-1}\mathbf{x} \leq 1\}$ and $\mathcal{E}_U = \{\mathbf{x} : \mathbf{x}'\mathbf{U}^{-1}\mathbf{x} \leq 1\}$, where $\mathcal{E}_U$ is contained in $\mathcal{E}_V$. The two ellipsoids can be compared by considering the radii of $\mathcal{E}_V$ relative to $\mathcal{E}_U$, obtained through the spectral decomposition of $\mathbf{V}$ relative to $\mathbf{U}$. We proceed to explore the geometry of generalized variance inflation, to show the relationship of these measures to correlation-matrix determinants and canonical correlations, to consider $\mathbf{X}$ matrices structured by relations of marginality among regressor subspaces, to develop the relationship of generalized variance inflation to hypothesis tests in the multivariate normal linear model, and to present several examples.

KEY WORDS: Canonical correlation; Joint confidence regions; Spectral decomposition; Variance inflation.

1. INTRODUCTION

Consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations on a response or dependent variable; $\mathbf{X}$ is an $n \times p$ full-rank design matrix of fixed constants, the first column of which consists of 1s; $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters to be estimated; and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of unobserved errors with $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $V(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. For some purposes we shall also assume that $\boldsymbol{\epsilon}$ is normally distributed $N_n(\mathbf{0}, \sigma^2\mathbf{I})$. The usual least squares estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, for which $V(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ (e.g., Fox 1984).

Since the methods developed in this article pertain to subspaces of the column space of $\mathbf{X}$, a deficient-rank parameterization of the model can be accommodated by selecting arbitrary bases for the subspaces. It is, likewise, possible to develop these methods for models without constants, simply by excluding the constant regressor from the $\mathbf{X}_0$ matrix defined subsequently.

If interest inheres in individual coefficients of the model, as is common in regression analysis, then the diagonal entries of $V(\hat{\boldsymbol{\beta}})$ reflect the (im)precision of estimation. Specifically, it may be shown that (Fox 1984; Theil 1971)

$$V(\hat{\beta}_j) = \frac{\sigma^2}{\Sigma(x_{ij} - \bar{x}_j)^2} \frac{1}{1 - R_j^2}, \tag{2}$$

where $\hat{\beta}_j$ is the jth entry in $\hat{\boldsymbol{\beta}}$ (but not the intercept estimate $\hat{\beta}_1$); x_{ij} is the entry in the ith row, jth column of $\mathbf{X}$; $\bar{x}_j$ is the mean of the jth column; and R_j^2 is the square of the multiple correlation from the regression of the jth column of $\mathbf{X}$ on the other columns. The second term in Equation (2) is called the *variance-inflation factor* (VIF) (Marquardt 1970), for it reflects the degree to which the sampling variance of $\hat{\beta}_j$ is increased as a consequence of correlations among the regressors: If the regressors are uncorrelated, then $R_j^2 = 0$ and VIF_j attains its minimum value of 1.

The variance-inflation factor is a useful diagnostic because it indicates directly the harm inflicted by collinearity on the precision of estimation: Indeed, a confidence interval around $\hat{\beta}_j$ has width proportional to $\text{VIF}_j^{1/2}$, which we term the *standard-error inflation factor* (SIF). But the VIF (or SIF) is only relevant when individual coefficients are of direct interest.

It is often natural to be concerned with sets of regressors and the column subspaces of $\mathbf{X}$ that they span rather than with individual regressors and their associated coefficients. This is the case, for example, when a qualitative independent variable gives rise to a set of dummy-variable regressors and when polynomial terms in a quantitative independent variable appear in a linear model. In each of these cases, the specific basis selected for a subspace of $\mathbf{X}$ is essentially arbitrary, though the subspace itself is not. Here relations among different sets of regressors and the subspaces that they generate are still of interest.

The plan of this article is as follows: In Section 2 we suggest a generalization of variance inflation to subsets of coefficients, and in Section 3 we consider a simple special case of generalized variance inflation that will often be of interest in applications and develop the relationship of variance inflation to correlation-matrix determinants. Section 4 presents some elementary examples. Section 5 develops the relationship of generalized variance inflation to angles between regressor subspaces and to multivariate test statistics. Section 6 shows how the notion of generalized variance inflation may be applied to models in which the design matrix is structured by relations of marginality among regressor subspaces. Section 7 presents an illustrative application, and Section 8 offers brief conclusions.

2. GENERALIZED VARIANCE INFLATION

We begin by rewriting model (1):

$$\mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{3}$$

* John Fox is Professor of Sociology and Mathematics and Statistics and Georges Monette is Associate Professor of Mathematics and Statistics at York University, Toronto, Ontario, Canada M3J 1P3. Both are associated with the Statistical Consulting Service of the York University Institute for Social Research.

Journal of the American Statistical Association
March 1992, Vol. 87, No. 417, Theory and Methods

where the $n \times p$ design matrix $\mathbf{X}$ has been partitioned into (a) $\mathbf{X}_0$ with p_0 columns, containing variables whose values could not have been selected differently by the investigator, even hypothetically (the archetypal example is the constant regressor); (b) $\mathbf{X}_1$ with p_1 columns, containing variables whose "effects" are of simultaneous interest—a set of dummy regressors, for example; (c) $\mathbf{X}_2$ with p_2 columns, containing variables that are "controlled" in defining the effects of the variables in $\mathbf{X}_1$ and whose values could potentially be selected by the investigator, at least in imagination.

We wish to assess the impact of the particular "selection" of values for $\mathbf{X}_2$ on the estimation of $\boldsymbol{\beta}_1$ in model (3). Let $\mathbf{W}_{ij}$ denote $\mathbf{X}_i'\mathbf{X}_j$. We can compare the variance of $\hat{\boldsymbol{\beta}}_1$,

$$V(\hat{\boldsymbol{\beta}}_1) = \sigma^2\left[\mathbf{W}_{11} - [\mathbf{W}_{10}, \mathbf{W}_{12}]\begin{bmatrix}\mathbf{W}_{00} & \mathbf{W}_{02}\\ \mathbf{W}_{20} & \mathbf{W}_{22}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{W}_{01}\\ \mathbf{W}_{21}\end{bmatrix}\right]^{-1} = \sigma^2\mathbf{V}, \quad (4)$$

with what the variance could have been had the $\mathbf{X}_2$ variables been selected to maximize the precision of estimation of $\boldsymbol{\beta}_1$:

$$\sigma^2\mathbf{U} = \sigma^2(\mathbf{W}_{11} - \mathbf{W}_{10}\mathbf{W}_{00}^{-1}\mathbf{W}_{01})^{-1}. \quad (5)$$

Expressions (4) and (5) are obtained using standard formulas (Graybill 1976) for the inverse of a partitioned matrix. That (5) minimizes (4) for fixed $\mathbf{X}_1$ and a fixed span for $\mathbf{X}_0$ is shown in Section 5. We will refer to (5) as the "utopian" variance of $\hat{\boldsymbol{\beta}}_1$, to emphasize that it is the best attainable variance for a selection of $\mathbf{X}_2$ that may not be feasible under the circumstances of the investigation. It is easily shown that the dispersion ellipsoid

$$\mathcal{E}_V = \{\mathbf{x} : \mathbf{x}'\mathbf{V}^{-1}\mathbf{x} \leq 1\} \quad (6)$$

contains the utopian ellipsoid

$$\mathcal{E}_U = \{\mathbf{x} : \mathbf{x}'\mathbf{U}^{-1}\mathbf{x} \leq 1\}. \quad (7)$$

The motivation for adopting an explicit point of comparison in assessing collinearity was well developed by Cook (1984).

The two ellipsoids (6) and (7) can be compared by considering the radii of $\mathcal{E}_V$ relative to $\mathcal{E}_U$, obtained through the spectral decomposition of $\mathbf{V}$ relative to $\mathbf{U}$. By a version of the spectral decomposition theorem (Rao 1973), there exists a nonsingular matrix $\mathbf{G}$ of order p_1 such that $\mathbf{U} = \mathbf{G}\mathbf{G}'$ and $\mathbf{V} = \mathbf{G}\boldsymbol{\Lambda}^2\mathbf{G}'$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_{p_1})$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p_1}$. Because $\mathbf{V} \geq \mathbf{U}$, the smallest eigenvalue $\lambda_{p_1}^2 \geq 1$. The λ's are all 1 when the ellipsoids $\mathcal{E}_V$ and $\mathcal{E}_U$ coincide.

The sequence of λ's is invariant under the following transformations:

1. Nonsingular linear transformation of the variables in $\mathbf{X}_0$.
2. Nonsingular linear transformation of the variables in $\mathbf{X}_1$ and addition of linear combinations of variables in $\mathbf{X}_0$ to those in $\mathbf{X}_1$.
3. Nonsingular linear transformation of the variables in $\mathbf{X}_2$ and addition of linear combinations of variables in $\mathbf{X}_0$ to those in $\mathbf{X}_2$.

Properties 1–3 follow easily by observing that $\mathbf{V}^{-1} = \mathbf{W}_{11\cdot 02}$, the residual sum of squares and cross-products (SSCP) matrix for $\mathbf{X}_1$ regressed on $\mathbf{X}_0$ and $\mathbf{X}_2$, whereas $\mathbf{U}^{-1} = \mathbf{W}_{11\cdot 0}$, the residual SSCP matrix for $\mathbf{X}_1$ regressed on $\mathbf{X}_0$ alone. We shall also show in Section 5 that the set of λ's exceeding 1 remains the same when the roles of $\mathbf{X}_1$ and $\mathbf{X}_2$ are interchanged.

A consequence of Property 2 is that the variables in $\mathbf{X}_1$ can be individually rescaled and centered (if $\mathbf{X}_0$ contains a constant regressor) without changing the λ's. The same, of course, is true (from Property 3) for the variables in $\mathbf{X}_2$.

Measuring the "loss" due to the choice of $\mathbf{X}_2$ is a question of assessing how much larger $\mathcal{E}_V$ is than $\mathcal{E}_U$. Although this assessment can be made by considering the entire sequence of λ's, it is convenient in practice to have a single index as a summary. The following possibilities are suggested in analogy to test statistics used in multivariate analysis of variance (MANOVA) (Eaton 1983) to compare the marginal dispersion of the dependent variables with their conditional dispersion under a hypothetical model:

1. $\Pi\lambda_i$ measures the ratio of volumes of the ellipsoids and corresponds to Wilks's criterion in MANOVA. We prefer this measure, which we term a *generalized standard-error inflation factor* (GSIF), for its straightforward interpretation if for no deeper reason. Thus the *generalized variance-inflation factor* is $\text{GVIF} = \text{GSIF}^2$. Similar measures comparing confidence-ellipsoid volumes were suggested by Andrews and Pregibon (1978) and Belsley, Kuh, and Welsch (1980) for assessing the influence of observations on the precision of regression estimates.
2. $\Sigma\lambda_i^2$ corresponds to Pillai's trace criterion in MANOVA.
3. λ_1^2 corresponds to Roy's maximum-root criterion in MANOVA.

3. GENERALIZED VARIANCE-INFLATION FACTORS AND CORRELATION-MATRIX DETERMINANTS

In many, likely most, applications, the investigator is interested in assessing the effect of a particular set of regressors (perhaps a set of dummy regressors) adjusting for all other variables in the model. In this case $\mathbf{X}_0$ of model (3) contains only the constant regressor. We find it convenient here to scale each column of $\mathbf{X}_1$ and $\mathbf{X}_2$ to 0 mean and length 1 which, as we have shown, does not affect generalized variance inflation. Under these transformations sums of cross-products are correlations and thus (specializing Eq. 4),

$$V(\hat{\boldsymbol{\beta}}_1) = \sigma^2(\mathbf{R}_{11} - \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21})^{-1}, \quad (8)$$

where $\mathbf{R}_{ij}$ is the matrix of correlations between $\mathbf{X}_i$ and $\mathbf{X}_j$. More generally, when $\mathbf{X}_0$ contains more than the constant regressor, $\mathbf{R}_{ij}$ represents the correlations between $\mathbf{X}_i$ and $\mathbf{X}_j$ partialing for $\mathbf{X}_0$.

In the utopian situation, where $\mathbf{X}_1$ and $\mathbf{X}_2$ are uncorrelated, Equation (8) reduces to $V(\hat{\boldsymbol{\beta}}_1) = \sigma^2\mathbf{R}_{11}^{-1}$, and the generalized variance-inflation factor may, therefore, be expressed in the following simplified form:

$$\text{GVIF}_1 = \frac{\det\mathbf{R}_{11}}{\det(\mathbf{R}_{11} - \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21})}. \quad (9)$$

Now, let $\mathbf{R}$ denote the correlation matrix among all of the columns of $\mathbf{X}$ excluding the constant (or correlations among all columns in $\mathbf{X}_1$ and $\mathbf{X}_2$ partialing for $\mathbf{X}_0$). Then $\det \mathbf{R} = \det \mathbf{R}_{22} \times \det(\mathbf{R}_{11} - \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21})$ and, consequently, Equation (9) simplifies further to

$$\mathrm{GVIF}_1 = \frac{\det \mathbf{R}_{11} \times \det \mathbf{R}_{22}}{\det \mathbf{R}}. \tag{10}$$

This last result establishes the equality of GVIF_1 and GVIF_2.

The determinant of $\mathbf{R}$, incidentally, is a common ad hoc global measure of collinearity (e.g., Kmenta 1986), usually justified by noting that $\det \mathbf{R} = 1$ for orthogonal data and $\det \mathbf{R} = 0$ for perfectly collinear data. Alternatively, some researchers no doubt have noted that $\det \mathbf{R}$ is inversely proportional to the generalized variance of $\hat{\boldsymbol{\beta}}$ (omitting the constant) in a model in which the X variables are standardized. A deeper, and related, justification of this measure is provided by examining the size of the joint confidence region for $\boldsymbol{\beta}$ (excluding the constant) relative to that obtained for orthogonal $\mathbf{X}$ (again, excluding the constant): $\mathrm{GVIF}^* = \det \mathbf{R}^{-1}/\det \mathbf{I}$; thus $\det \mathbf{R} = \mathrm{GVIF}^{*-1}$. (Here we use GVIF^* to represent a ratio of squared ellipsoid volumes, but not of the ellipsoids defined in Section 2—hence the asterisk.)

Admitting $\det \mathbf{R}$ as a global index of collinearity suggests the following interpretation of Equation (10): The generalized variance-inflation factor $\mathrm{GVIF}_1 = \mathrm{GVIF}_2$ represents the global collinearity in $\mathbf{X}$ scaled by the product of the collinearity internal to each of $\mathbf{X}_1$ and $\mathbf{X}_2$. This scaling in effect adjusts for the arbitrary selection of bases for the spans of $\mathbf{X}_1$ and $\mathbf{X}_2$.

These observations explain why $\det \mathbf{R}$ is an unreasonable global measure of collinearity when the columns of $\mathbf{X}$ partition naturally into sets, some of which contain more than one member: To the extent that these sets have arbitrary bases for their column spans, some of the collinearity detected is artifactual. Instead, given a partition of $\mathbf{X}$ into k sets, it is natural to compute $\det \mathbf{R}_{11} \times \det \mathbf{R}_{22} \times \cdots \times \det \mathbf{R}_{kk}/\det \mathbf{R}$ as a global index of collinearity invariant with respect to changes of basis within sets.

To preserve comparability across subspaces of different dimension, we suggest examining $\mathrm{GVIF}^{1/2p_1} = \mathrm{GSIF}^{1/p_1}$ (the geometric mean of the λ's) in place of GVIF. Notice that when $p_1 = 1$ the generalized variance-inflation factor reduces to the usual VIF (Eq. 2), since then $\mathbf{R}_1^2 = \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}$.

4. SOME SIMPLE ILLUSTRATIONS

Table 1 shows several two-way contingency tables, each of which, we imagine, relates two qualitative independent variables in a linear model. From the contingency tables we constructed dummy-variable regressors for rows and columns, employing a 0/1 coding scheme and treating the last row and column as "baseline" categories (the coding scheme implicitly employed by the SAS GLM procedure, for example); then we computed correlations among the dummy regressors; and, finally, we evaluated the generalized variance-inflation factors for row and column regressors (shown as GVIF_R and GVIF_C) using Equation (9).

Table 1. Illustrative Cell Frequencies for Two-Way Classifications, Showing Generalized Variance-Inflation Factors for Row and Column Effects Employing Dummy-Variable Coding

(A) Row and Column Classifications Independent

10	20	30
20	40	60
5	10	15
15	30	45

$\mathrm{GVIF}_R = .694/.694 = 1$
$\mathrm{GVIF}_C = .9/.9 = 1$
$\lambda_1 = \lambda_2 = 1$

(B) Weak Association

20	15	10	5
15	20	15	10
10	15	20	15
5	10	15	20

$\mathrm{GVIF}_R = .556/.490 = 1.14$
$\mathrm{GVIF}_C = .556/.490 = 1.14$
$\mathrm{GVIF}^{1/2\times 3} = 1.02$
$\lambda_1 = 1.06, \lambda_2 = 1.00, \lambda_3 = 1.00$

(C) Stronger Association

10	0	0	0
2	10	0	0
0	2	10	0
0	0	2	10
0	0	0	10

$\mathrm{GVIF}_R = .514/.00421 = 122$
$\mathrm{GVIF}_C = .476/.00390 = 122$
$\mathrm{GVIF}_R^{1/2\times 4} = 1.82$
$\mathrm{GVIF}_C^{1/2\times 3} = 2.23$
$\lambda_1 = 2.61, \lambda_2 = 1.41, \lambda_3 = 1.08$

(D) A Perfect Dependency

10	10	10	0
10	10	10	0
10	10	10	0
0	0	0	10

$\mathrm{GVIF}_R = .292/0 = \infty$
$\mathrm{GVIF}_C = .292/0 = \infty$
$\lambda_1 = \infty, \lambda_2 = \lambda_3 = 1$

Note that the specific bases selected for the regressor subspaces generated by rows and columns are immaterial, affecting the determinants in the numerator and denominator of Equation (9) but not their ratio. Furthermore, note that (as is clear from Eq. 10, as well as the results given in Sec. 5) GVIF_R and GVIF_C are necessarily equal since R, C, and the constant comprise the entire design. The sequence of λ's is also shown for each example. In Section 6, we consider analysis-of-variance designs that include interactions.

In Example A, the row and column classifications are independent, yielding GVIF's of 1. In Example B, there is a weak dependency between rows and columns, producing GVIF's slightly in excess of 1. Example C illustrates a stronger dependency between rows and columns, giving rise to larger variance-inflation factors (of 122), though even here the one-dimensional indices ($\mathrm{GVIF}^{1/2p_1}$) are not particularly large. Finally, Example D illustrates a perfect dependency (produced by the coincidence of the last row and last column), which yields infinite GVIF's. Note though

that two of the three λ's are 1, reflecting the balance of the first three rows and columns of the design.

5. GVIF's, CANONICAL ANGLES, AND MULTIVARIATE TESTS

The ellipsoids $\mathcal{E}_V$ and $\mathcal{E}_U$ [given in (6) and (7)] occur in two problems that are closely related to variance inflation: Fitting a multivariate linear model and canonical correlations between two sets of variables adjusted for a third. These relationships illuminate the essential character of variance inflation.

A multivariate linear model with $\mathbf{X}_1$ as dependent variables and $\mathbf{X}_0$ and $\mathbf{X}_2$ as regressors may be written as $\mathbf{X}_1 = \mathbf{X}_0\Xi_0 + \mathbf{X}_2\Xi_2 + \Delta$, where Ξ_0 and Ξ_2 are $p_0 \times p_1$ and $p_2 \times p_1$ matrices of regression coefficients, respectively, and Δ is an $n \times p_1$ matrix of "errors."

In considering the hypothesis $H_0 : \Xi_2 = \mathbf{0}$ we are led to compare the residual SSCP matrices resulting from fitting a full model and from fitting a restricted model in which $\Xi_2 = \mathbf{0}$. The full model yields the residual SSCP matrix $\mathbf{W}_{11\cdot02} = \mathbf{V}^{-1}$; the restricted model yields the residual SSCP matrix $\mathbf{W}_{11\cdot0} = \mathbf{U}^{-1}$. Expression (5) is a minimum for (4) if $\mathbf{X}_2$ is free to vary because $\mathbf{W}_{11\cdot0}$ is a maximum for $\mathbf{W}_{11\cdot02}$.

If the rows of Δ are independent and normally distributed, then the usual tests of H_0 are based on the eigenvalues in the spectral decomposition of $\mathbf{U}^{-1}$ relative to $\mathbf{V}^{-1}$ (Eaton 1983). These eigenvalues are the same as those of $\mathbf{V}$ relative to $\mathbf{U}$. In the present context, we can make no assumptions about the distribution of Δ, but the various test criteria (enumerated in Sec. 2) nevertheless provide measures of the association of $\mathbf{X}_1$ with $\mathbf{X}_2$ adjusting for $\mathbf{X}_0$.

The λ's are also interpretable as "angle cosecants" of the column spaces of $\mathbf{X}_1$ and $\mathbf{X}_2$ adjusted for $\mathbf{X}_0$. Let $\mathbf{X}_{1\cdot0}$ be the residuals of $\mathbf{X}_1$ on regression on $\mathbf{X}_0$, and let $\mathbf{X}_{2\cdot0}$ be defined similarly. Let $\mathcal{L}_{i\cdot0} = \text{span}(\mathbf{X}_{i\cdot0})$ for $i = 1, 2$. Finding canonical correlations between $\mathbf{X}_{1\cdot0}$ and $\mathbf{X}_{2\cdot0}$ is equivalent to finding the cosines of angles between pairs of "canonical vectors" in $\mathcal{L}_{1\cdot0}$ and $\mathcal{L}_{2\cdot0}$ (Eaton 1983). Let $\rho_1, \ldots, \rho_{p_1}$ be the ordered angle cosines (correlations) between $\mathcal{L}_{1\cdot0}$ and $\mathcal{L}_{2\cdot0}$. Since $\rho_1, \ldots, \rho_{p_1}$ are the nonnegative roots of the determinantal equation $\det(\mathbf{W}_{12\cdot0}\mathbf{W}_{22\cdot0}\mathbf{W}_{21\cdot0} - \rho^2\mathbf{W}_{11\cdot0}) = 0$ (Rao 1973), it follows that $\lambda_i^2 = (1 - \rho_i^2)^{-1}$, for $i = 1, \ldots, p_1$. Hence the λ_i^{-1} are angle sines and the λ_i are angle cosecants between the two spaces. This result also shows the equality of the λ's greater than 1 when the roles of $\mathbf{X}_1$ and $\mathbf{X}_2$ are interchanged and, hence, the equality of the GVIF's for $\mathbf{X}_1$ and $\mathbf{X}_2$ (a property that was established in an alternative manner in Sec. 3).

6. MARGINAL SUBSPACES

In certain linear models—for example, those containing both main-effect and interaction regressors—some subspaces are marginal to others when the model is "overparameterized" (e.g., McCullagh and Nelder 1983). A simple, and common, example is the two-way analysis-of-variance model $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$, for $i = 1, \ldots, r$; $j = 1, \ldots, c$; $k = 1, \ldots, n_{ij}$. Here, the regressors for the interactions γ_{ij} span the full column space of the design matrix, and the regressors for the main effects α_i and β_j of rows and columns span subspaces that are marginal to this space. A simple approach to the model is to employ a convenient reparameterization that yields a full-rank design and permits direct tests of hypotheses of interest concerning cell and marginal means. One such reparameterization is produced by constraining $\Sigma\alpha_i = 0$; $\Sigma\beta_j = 0$, $\Sigma_i\gamma_{ij} = 0$ ($j = 1, \ldots, c$), and $\Sigma_j\gamma_{ij} = 0$ ($i = 1, \ldots, r$). The resulting design matrix is of full column rank as long as all cells of the design are filled. The essential point here is not to argue for full-rank as opposed to deficient-rank parameterizations of the model, since the two are equivalent if treated properly, but rather to demonstrate how collinearity of the subspace for interactions with the subspaces for main effects. may reasonably be assessed despite the fact that the latter subspaces are marginal to the former.

For our purposes, the reparameterization has the convenient property that for "balanced" data (i.e., all $n_{ij} = n/rc$), the main-effect and interaction subspaces are mutually orthogonal. For unbalanced data, the generalized variance-inflation factor for the interactions after main effects exceeds 1 and is a useful index of the extent to which imbalance compromises our ability to estimate interactions accurately.

Since we are typically not interested in separately estimating main effects in the presence of interactions, we would usually find generalized variance-inflation factors for main effects adjusted for each other (as in Table 1) but ignoring interactions (corresponding, for example, to Type II sums of squares in SAS). Our approach is more general, however, and permits assessing collinearity of main effects adjusting for other main effects and interactions (corresponding to Type III sums of squares).

The full-rank methods developed in this article will function properly as long as all cells in the design are filled. If there is an empty cell, then some interaction contrasts are not estimable, and the reduced design matrix based on the usual constraints is of deficient rank. In this case the generalized variance-inflation factor is infinite. Equation (9) may fail, however, since in a sparse design the numerator as well as the denominator may be 0; this situation occurs when the regressors coding the interactions are themselves not of full column rank.

Generalized variance-inflation factors for interactions in the two-way classifications given in Examples (A) and (B) of Table 1 are as follows: (A) GVIF = .0209/.00704 = 2.96, $\text{GVIF}^{1/2\times6} = 1.10$; (B) GVIF = .0128/.00628 = 2.03, $\text{GVIF}^{1/2\times9} = 1.04$. Note that Example (A) is unbalanced for the interactions even though the main-effect subspaces are orthogonal to one another.

7. AN ILLUSTRATIVE APPLICATION

Heberlein and Baumgartner (1978) developed a 10-regressor model for predicting response rates in mail surveys conducted principally in the United States. When they failed to replicate certain of Heberlein and Baumgartner's findings in a similar study of surveys conducted in Europe, Eichner and Habermehl (1981) suggested that the discrepancy might be partly due to collinearity, producing unstable

estimates of regression coefficients. Further work by Goyder (1982), employing a different sample of mail surveys mostly from the United States, yielded results similar to those in the initial study. Goyder (1985) provided a comparison between surveys conducted in the United States and Canada.

The issue of interstudy coefficient differences should, of course, be addressed directly—for the coefficient vector as a whole and for individual coefficients—but none of the reports cited includes the necessary information (for example, coefficient standard errors). On the basis of Table 2, calculated from Goyder's (1985) data, it seems likely that some of the coefficient differences between the Heberlein and Baumgartner (1978) and Eichner and Habermehl (1981) studies are statistically significant. Nevertheless, several coefficients are imprecisely estimated.

Table 2. Regression Coefficients and Variance-Inflation Factors for Regression of Final Percent Response Rates in U.S. Mail Surveys on Sponsorship, Type of Population, Saliency of Topic, Length in Pages, Number of Contacts, Special Third Contact, and Incentive

(A) Regression Results

Regressor	Coefficient	Standard Error	SIF
Constant	29.3	3.15	
Sponsorship[a]			
Market research	−4.34	2.78	1.28
Government	12.4	3.25	1.12
Type of population[b]			
General	−5.73	2.65	1.13
Employee	−1.00	4.69	1.04
School or army	4.42	3.45	1.03
Saliency of topic[c]	11.8	1.75	1.31
Length in pages	−0.343	0.199	1.07
Number of contacts	7.00	1.09	1.59
Special third contact[d]	3.88	1.66	1.56
Incentive[e]	7.27	1.49	1.06
Standard error of regression		15.34	
R^2		.618	
Number of cases[f]		270	

(B) Generalized Variance-Inflation Factors

Sponsorship

ρ_i	λ_i
.639	1.30
.380	1.08

GVIF = 1.97
$GVIF^{1/2\times 2}$ = 1.19

Type of population

ρ_i	λ_i
.448	1.12
.231	1.03
.130	1.01

GVIF = 1.34
$GVIF^{1/2\times 3}$ = 1.05

NOTE: The source of the data is personal communication from John Goyder. See Goyder (1985) and Heberlein and Baumgartner (1978) for more detailed information about the definition of variables.

[a]Baseline category for sponsorship: neither government nor market research.

[b]Baseline category for type of population: other type of population.

[c]Coding of saliency: (0) not salient; (1) possibly salient; (3) salient.

[d]Coding of special third contact: (0) no third contact; (1) regular mail; (2) special mail; (3) telephone or personal.

[e]Coding of incentive (on first contact): (0) no incentive; (1) less than $.25; (2) $.25; (3) $.50; (4) $1.00 or more.

[f]The reported results are based on imputing missing values, as in Goyder (1985); there are 102 complete cases.

Eichner and Habermehl (1981), Goyder (1982), and Heberlein and Baumgartner (1981) (in response to criticism) furnished multiple correlations between each regressor and all others in the model. The largest of these multiple correlations (less than .8) is not indicative of serious collinearity (corresponding to inflation in coefficient standard errors of less than 2). As well, two of the variables employed in the model—sponsor and type of population—give rise to more than one dummy regressor. In the case of sponsorship, the baseline category (neither government nor market research) provides an arguably natural comparison, but in the case of type of population, the selected baseline (neither general, employee, nor school or army) is clearly arbitrary. Consequently, the methods of this article are relevant for addressing the issue of collinearity here.

The results shown in Table 2 are for a data set generously provided by John Goyder (Goyder 1985). There were 270 U.S. mail surveys included in the data set, 102 of which had valid data for all of the variables employed in the analysis. To be consistent with Goyder (1985), missing data were imputed, primarily by substituting means for missing values. An analysis based only on complete cases provides essentially similar results. With this substantial quantity of missing data, however, a more sophisticated approach than mean imputation or complete-case analysis would be desirable [see, for example, Little and Rubin (1987)].

Part A of Table 2 shows estimated regression coefficients, standard errors, and individual standard error inflation factors. Multiple-degree-of-freedom tests for sponsorship and type of population yield the following results: Sponsorship, $F(2,259) = 9.15$, $p < .0001$; Type of Population, $F(3,259) = 2.41$, $p = .066$. Part B of the table shows generalized variance-inflation factors for sponsorship and type of population. It is apparent that collinearity is not a substantial problem here.

This conclusion has implications for how one might proceed to obtain improved estimates. Were the level of collinearity high, then the ideal procedure would be to collect additional, noncollinear data, perhaps experimentally. Because collinearity is not seriously problematic, however, it would be more fruitful to increase the number of observations, especially in sparse categories of qualitative independent variables (such as surveys conducted on employee populations), to increase the variability of quantitative independent variables (such as incentive for responding to the survey), and to reduce the error variance by improving the specification of the model. It might also be of interest to consider further the role of cultural differences among populations in determining survey nonresponse. Indeed, it is our impression that social scientisits are too quick to ascribe imprecise and (upon replication) unstable coefficient estimates to collinearity when more likely culprits are large error variances, small effects, poorly specified models, nontrivial differences among studies, and other substantive problems.

8. CONCLUDING COMMENTS

When our interest in collinearity is not for its numerical consequences [the approach emphasized, for example, by

Belsley et al. (1980)] but for its impact on the variability of estimates (an admittedly related, though distinguishable phenomenon), it is important to realize that we essentially wish to compare two sets of X values—the ones that we have and the ones that we wish we had. This point was made clearly by Cook (1984); also see Leamer (1973, 1983).

A reviewer of an earlier version of this article suggested that "collinearity is not so much a problem as a state of nature—like the law of gravity—and that railing against collinearity is rather like complaining about not being able to fly by flapping your arms." Although we have some sympathy with this point of view, we believe that it overstates the case: The identification of specific sources of imprecision in estimation may, in certain instances, suggest how estimates can be improved, for example, by collecting additional data (abandoning arm-flapping and trying an airplane). In other instances, the discovery of collinearity may motivate respecification of a statistical model or reorientation of the goals of a study—as when a variable selection method is employed for prediction.

The method introduced in this article extends collinearity diagnostics to subsets of coefficients. This method is easy to apply because it involves familiar computations; it is flexible because it permits distinctions between intrinsically fixed variables and those that could be selected or sampled differently; and it is cogent because it speaks directly to the harm produced by collinearity.

[*Received January 1990. Revised January 1991.*]

REFERENCES

Andrews, D. F., and Pregibon, D. (1978), "Finding Outliers That Matter," *Journal of the Royal Statistical Society*, Ser. B, 40, 85–93.

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

Cook, R. D. (1984), Comment on "Demeaning Conditioning Diagnostics Through Centering," by D. A. Belsley, *The American Statistician*, 38, 78–79.

Eaton, M. L. (1983), *Multivariate Statistics: A Vector Space Approach*, New York: John Wiley.

Eichner, K., and Habermehl, W. (1981), "Predicting Response Rates to Mailed Questionnaires," Comment on "Factors Affecting Response Rates to Mailed Questionnaires" by T. A. Heberlein and R. Baumgartner (1978), *American Sociological Review*, 46, 361–363.

Fox, J. (1984), *Linear Statistical Models and Related Methods, With Applications to Social Research*, New York: John Wiley.

Goyder, J. (1982), "Factors Affecting Response Rates to Mailed Questionnaires," *American Sociological Review*, 47, 550–553.

——— (1985), "Nonresponse on Surveys," *Canadian Journal of Sociology*, 10, 231–251.

Graybill, F. A. (1976), *Theory and Applications of the Linear Model*, North Scituate, MA: Duxbury Press.

Heberlein, T. A., and Baumgartner, R. (1978), "Factors Affecting Response Rates to Mailed Questionnaires," *American Sociologial Review*, 43, 447–462.

——— (1981), "The Effectiveness of the Heberlein–Baumgartner Models for Predicting Response Rates to Mailed Questionnaires: European and U.S. Examples" (Reply to comment by Eichner and Habermehl), *American Sociological Review*, 46, 363–367.

Kmenta, J. (1986), *Elements of Econometrics* (2nd ed.), New York: Macmillan.

Leamer, E. E. (1973), "Multicollinearity: A Bayesian Interpretation," *Review of Economics and Statistics*, 55, 371–380.

——— (1983), "Model Choice and Specification Analysis," in *Handbook of Econometrics, Volume 1*, eds. Z. Griliches and M. D. Intriligator, Amsterdam: North-Holland, pp. 285–330.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data*, New York: John Wiley.

Marquardt, D. W. (1970), "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation," *Technometrics*, 12, 591–612.

McCullagh, P., and Nelder, J. A. (1983), *Generalized Linear Models*, London: Chapman & Hall.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), New York: John Wiley.

Theil, H. (1971), *Principles of Econometrics*, New York: John Wiley.