Condition Indexes and Variance Decompositions for Diagnosing Collinearity in Linear Model Analysis of Survey Data

Dan Liao[1] **and Richard Valliant**[2]

[1]RTI International, 701 13th Street, N.W., Suite 750, Washington DC, 20005, dliao@rti.org

[2]University of Michigan and University of Maryland, Joint Program in Survey Methodology,
1218 Lefrak Hall, College Park, MD, 20742, rvalliant@survey.umd.edu

Abstract
Collinearities among explanatory variables in linear regression models affect estimates from survey data just as they do in non-survey data. Undesirable effects are unnecessarily inflated standard errors, spuriously low or high t-statistics, and parameter estimates with illogical signs. The available collinearity diagnostics are not generally appropriate for survey data because the variance estimators they incorporate do not properly account for stratification, clustering, and survey weights. In this article, we derive condition indexes and variance decompositions to diagnose collinearity problems in complex survey data. The adapted diagnostics are illustrated with data based on a survey of health characteristics.

Keywords: diagnostics for survey data; multicollinearity; singular value decomposition; variance inflation.

1 Introduction

When predictor variables in a regression model are correlated with each other, this condition is referred to as collinearity. Undesirable side effects of collinearity are unnecessarily high standard errors, spuriously low or high t-statistics, and parameter estimates with illogical signs or ones that are overly sensitive to small changes in data values. In experimental design, it may be possible to create situations where the explanatory variables are orthogonal to each other, but this is not true with observational data. Belsley (1991) noted that: "... in nonexperimental sciences, ..., collinearity is a natural law in the data set resulting from the uncontrollable operations of the data-generating mechanism and is simply a painful and unavoidable fact of life." In many surveys, variables that are substantially correlated are collected for analysis. Few analysts of survey data have escaped the problem of collinearity in regression estimation, and the presence of this problem encumbers precise statistical explanation of the relationships between predictors and responses.

Although many regression diagnostics have been developed for non-survey data, there are considerably fewer for survey data. The few articles that are available concentrate on identifying influential points and influential groups with abnormal data values or survey weights. Elliot (2007) developed Bayesian methods for weight trimming of linear and generalized linear regression estimators in unequal probability-of-inclusion designs. Li (2007a,b) and Li & Valliant (2011, 2009) extended a series of traditional diagnostic techniques to regression on complex survey data. Their papers cover residuals and leverages, several diagnostics based on case-deletion (DFBETA, DFBETAS, DFFIT, DFFITS, and Cook's Distance), and the forward search approach. Although an extensive literature in applied statistics provides valuable suggestions and guidelines for data analysts to diagnose the presence of collinearity (e.g., Belsley et al. 1980; Belsley 1991; Farrar & Glauber 1967; Fox 1986; Theil 1971), almost none of this research touches upon diagnostics for collinearity when fitting models with survey data. One prior, survey-related paper on collinearity problems is (Liao & Valliant, 2010) which adapted variance inflation factors for linear models fitted with survey data.

Suppose the underlying structural model in the superpopulation is $\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{e}$. The matrix $\boldsymbol{X}$ is an $n \times p$ matrix of predictors with n being the sample size; $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters. The error terms in the model have a general variance structure $\boldsymbol{e} \sim (0, \sigma^2\boldsymbol{R})$ where σ^2 is an unknown constant and $\boldsymbol{R}$ is a unknown $n \times n$ covariance matrix. Define $\boldsymbol{W}$ to be the diagonal matrix of survey weights. We assume throughout that the survey weights are constructed in such a way that they can be used for estimating finite population totals. The survey weighted least squares (SWLS) estimator is

$$\hat{\boldsymbol{\beta}}_{SW} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y} \equiv \boldsymbol{A}^{-\mathbf{1}}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y},$$

assuming $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X}$ is invertible. Fuller (2002) describes the properties of this estimator. The estimator $\hat{\boldsymbol{\beta}}_{SW}$ is model unbiased for $\boldsymbol{\beta}$ under the model $\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{e}$ regardless of whether $Var_M(\boldsymbol{e}) = \sigma^2\boldsymbol{R}$ is specified correctly or not, and is approximately design-unbiased for the census parameter $\boldsymbol{B}_U = (\boldsymbol{X}_U^T\boldsymbol{X}_U)^{-1}\ \boldsymbol{X}_U^T\boldsymbol{Y}_U$, in the finite population U of N units. The finite population values of the response vector and matrix of predictors are $\boldsymbol{Y}_U = (Y_1, ..., Y_N)^T$, and $\boldsymbol{X}_U = (\boldsymbol{X}_1, ..., \boldsymbol{X}_p)$ with $\boldsymbol{X}_k$ being the $N \times 1$ vector of values for covariate k.

The remainder of the paper is organized as follows. Section 2 reviews results on condition numbers and variance decompositions for ordinary least squares. These are extended to be appropriate for survey estimation in section 3. The fourth section gives some numerical illustrations of the techniques. Section 5 is a conclusion. In most derivations, we use model-based calculations since the forms of the model-variances are useful for understanding the effects of collinearity. However, when presenting variance decompositions, we use estimators that have both model- and design-based justifications.

2 Condition Indexes and Variance Decompositions in Ordinary Least Squares Estimation

In this section we briefly review techniques for diagnosing collinearity in ordinary least squares (OLS) estimation based on condition indexes and variance decompositions. These methods will be extended in section 3 to cover complex survey data.

2.1 Eigenvalues and Eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$

When there is an exact (perfect) collinear relation in the $n \times p$ data matrix $\boldsymbol{X}$, we can find a set of values, $\boldsymbol{v} = (v_1, \ldots, v_p)$, not all zero, such that

$$v_1\boldsymbol{X}_1 + \cdots + v_p\boldsymbol{X}_p = \mathbf{0}, \quad \text{or } \boldsymbol{X}\boldsymbol{v} = \mathbf{0}. \tag{1}$$

However, in practice, when there exists no exact collinearity but some near dependencies in the data matrix, it may be possible to find one or more non-zero vectors $\boldsymbol{v}$ such that $\boldsymbol{X}\boldsymbol{v} = \boldsymbol{a}$ with $\boldsymbol{a} \neq \mathbf{0}$ but close to $\mathbf{0}$. Alternatively, we might say that a near dependency exists if the length of vector $\boldsymbol{a}$, $\|\boldsymbol{a}\|$, is small. To normalize the problem of finding the set of $\boldsymbol{v}$'s that makes $\|\boldsymbol{a}\|$ small, we consider only $\boldsymbol{v}$ with unit length, that is, with $\|\boldsymbol{v}\| = 1$. Belsley (1991) discusses the connection of the eigenvalues and eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$ with the normalized vector $\boldsymbol{v}$ and $\|\boldsymbol{a}\|$. The minimum length $\|\boldsymbol{a}\|$ is simply the positive square root of the smallest eigenvalue of $\boldsymbol{X}^T\boldsymbol{X}$. The $\boldsymbol{v}$ that produces the $\boldsymbol{a}$ with minimum length must be the eigenvector of $\boldsymbol{X}^T\boldsymbol{X}$ that corresponds to the smallest eigenvalue. As discussed in the next section, the eigenvalues and eigenvectors of $\boldsymbol{X}$ are related to those of $\boldsymbol{X}^T\boldsymbol{X}$ and have some advantages when examining collinearity.

2.2 Singular-Value Decomposition, Condition Number and Condition Indexes

The singular-value decomposition (SVD) of matrix $\boldsymbol{X}$ is very closely allied to the eigensystem of $\boldsymbol{X}^T\boldsymbol{X}$, but with its own advantages. The $n \times p$ matrix $\boldsymbol{X}$ can be decomposed as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, where $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_p$ and $\boldsymbol{D} = diag(\mu_1, \ldots, \mu_p)$ is the diagonal matrix of singular values (or eigenvalues) of $\boldsymbol{X}$. Here, the three components in the decomposition are matrices with very special, highly exploitable properties: $\boldsymbol{U}$ is $n \times p$ (the same size as $\boldsymbol{X}$) and is column orthogonal; $\boldsymbol{V}$ is $p \times p$ and both row and column orthogonal; $\boldsymbol{D}$ is $p \times p$, nonnegative and diagonal. Belsley et al. (1980) felt that the SVD of $\boldsymbol{X}$ has several advantages over the eigen system of $\boldsymbol{X}^T\boldsymbol{X}$, for the sake of both statistical usages and computational complexity. For prediction, $\boldsymbol{X}$ is the focus not the cross-product matrix $\boldsymbol{X}^T\boldsymbol{X}$ since $\hat{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. In addition, the lengths $\|\boldsymbol{a}\|$ of the linear combinations (1) of $\boldsymbol{X}$ that are relate to collinearity are properly defined in terms of the square roots of the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$, which are the singular values of $\boldsymbol{X}$. A secondary consideration, given current computing power, is that the singular value decomposition of $\boldsymbol{X}$ avoids the additional computational burden of forming $\boldsymbol{X}^T\boldsymbol{X}$, an operation involving np^2 unneeded sums and products, which may lead to unnecessary truncation error.

The condition number of $\boldsymbol{X}$ is defined as $\kappa(\boldsymbol{X}) = \mu_{max}/\mu_{min}$, where μ_{max} and μ_{min} are the maximum and minimum singular values of $\boldsymbol{X}$. Condition indexes are defined as $\eta_k = \mu_{max}/\mu_k$. The closer that μ_{min} is to zero, the nearer $\boldsymbol{X}^T\boldsymbol{X}$ is to being singular. Empirically, if a value of κ or η exceeds a cutoff value of, say, 10 to 30, two or more columns of $\boldsymbol{X}$ have moderate or strong relations. The simultaneous occurrence of several large η_k's is always remarkable for the existence of more than one near dependency.

One issue with the SVD is whether the $\boldsymbol{X}$'s should be centered around their means. Marquardt (1980) maintained that the centering of observations removes nonessential ill conditioning. In contrast, Belsley (1984) argues that mean-centering typically masks the role of the constant term in any underlying near-dependencies. A typical case is a regression with dummy variables. For example, if gender is one of the independent variables in a regression and most of the cases are male (or female), then the dummy for gender can be strongly collinear with the intercept. The discussions following Belsley (1984) illustrate the differences of opinion that occur among practitioners (Wood, 1984; Snee & Marquardt, 1984; Cook, 1984). Moreover, in linear regression analysis, Wissmann et al. (2007) found that the degree of multicollinearity with dummy variables may be influenced by the choice of reference category. In this article, we do not center the $\boldsymbol{X}$'s but will illustrate the effect of the choice of reference category in Section 4.

Another problem with the condition number is that it is affected by the scale of the x measurements (Steward, 1987). By scaling down any column of $\boldsymbol{X}$, the condition number can be made arbitrarily large. This situation is known as *artificial ill-conditioning*. Belsley (1991) suggests scaling each column of the design matrix $\boldsymbol{X}$ using the Euclidean norm of each column before computing the condition number. This method is implemented in SAS and the package *perturb* of the statistical software R (Hendrickx, 2010). Both use the root mean square of each column for scaling as its standard procedure. The condition number and condition indexes of the scaled matrix $\boldsymbol{X}$ are referred to as the *scaled condition number* and *scaled condition indexes* of the matrix $\boldsymbol{X}$. Similarly, the variance decomposition proportions relevant to the scaled $\boldsymbol{X}$ (which will be discussed in next section) will be called the *scaled variance decomposition proportions*.

2.3 Variance Decomposition Method

To assess the extent to which near dependencies (i.e., having high condition indexes of $\boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{X}$) degrade the estimated variance of each regression coefficient, Belsley et al. (1980) reinterpreted and extended the work of Silvey (1969) by decomposing a coefficient variance into a sum of terms each of which is associated with a singular value. In the remainder of this section, we review the results of ordinary least squares (OLS) under the model $E_M(\boldsymbol{Y}) = \boldsymbol{X\beta}$ and $Var_M(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. These results will be extended to survey weighted least squares in section 3. Recall that the model variance-covariance matrix of the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ under the model with $Var_M(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n$ is $Var_M(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Using the SVD, $\boldsymbol{X} = \boldsymbol{UDV}^T$, $Var_M(\hat{\boldsymbol{\beta}})$ can be written as:

$$Var_M(\hat{\boldsymbol{\beta}}) = \sigma^2[(\boldsymbol{UDV}^T)^T(\boldsymbol{UDV}^T)]^{-1} = \sigma^2\boldsymbol{VD}^{-2}\boldsymbol{V}^T \tag{2}$$

and the k^{th} diagonal element in $Var_M(\hat{\boldsymbol{\beta}})$ is the estimated variance for the k^{th} coefficient, $\hat{\beta}_k$. Using (2), $Var_M(\hat{\beta}_k)$

can be expressed as:

$$Var(\hat{\beta}_k) = \sigma^2 \Sigma_{j=1}^{p} \frac{v_{kj}^2}{\mu_j^2} \tag{3}$$

where $\boldsymbol{V} = (v_{kj})_{p \times p}$. Let $\phi_{kj} = v_{kj}^2/\mu_j^2$, $\phi_k = \Sigma_{j=1}^{p}\phi_{kj}$ and $\boldsymbol{Q} = (\phi_{kj})_{p\times p} = (\boldsymbol{V}\boldsymbol{D}^{-1}) \cdot (\boldsymbol{V}\boldsymbol{D}^{-1})$, where $\cdot$ is the Hadamard (elementwise) product. The variance-decomposition proportions are $\pi_{jk} = \phi_{jk}/\phi_k$, which is the proportion of the variance of the k^{th} regression coefficient associated with the j^{th} component of its decomposition in (3). Denote the *variance decomposition proportion matrix* as $\boldsymbol{\Pi} = (\pi_{jk})_{p\times p} = \boldsymbol{Q}^T \bar{\boldsymbol{Q}}^{-1}$, where $\bar{\boldsymbol{Q}}$ is the diagonal matrix with the row sums of $\boldsymbol{Q}$ on the main diagonal and 0 elsewhere.

If the model is $E_M(\boldsymbol{Y}) = \boldsymbol{X\beta}$, $Var_M(\boldsymbol{Y}) = \sigma^2 \boldsymbol{W}^{-1}$ and weighted least squares is used, then $\hat{\boldsymbol{\beta}}_{WLS} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$ and $Var_M(\hat{\boldsymbol{\beta}}_{WLS}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$. The decomposition in (3) holds with $\tilde{\boldsymbol{X}} = \boldsymbol{W}^{1/2}\boldsymbol{X}$ being decomposed as $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$. However, in survey applications, it will virtually never be the case that the covariance matrix of $\boldsymbol{Y}$ is $\sigma^2\boldsymbol{W}^{-1}$ if $\boldsymbol{W}$ is the matrix of survey weights. Section 3 covers the more realistic case.

In the variance decomposition (3), other things being equal, a small singular value μ_j can lead to a large component of $Var(\hat{\beta}_k)$. However, if v_{kj} is small too, then $Var(\hat{\beta}_k)$ may not be affected by a small μ_j. One extreme case is when $v_{kj} = 0$. Suppose the k^{th} and j^{th} columns of $\boldsymbol{X}$ belong to separate orthogonal blocks. Let $\boldsymbol{X} \equiv [\boldsymbol{X}_1, \boldsymbol{X}_2]$ with $\boldsymbol{X}_1^T\boldsymbol{X}_2 = \boldsymbol{0}$ and let the singular-value decompositions of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ be given, respectively, as $\boldsymbol{X}_1 = \boldsymbol{U}_1\boldsymbol{D}_{11}\boldsymbol{V}_{11}^T$ and $\boldsymbol{X}_2 = \boldsymbol{U}_2\boldsymbol{D}_{22}\boldsymbol{V}_{22}^T$. Since $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are the orthogonal bases for the space spanned by the columns of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ respectively, $\boldsymbol{X}_1^T\boldsymbol{X}_2 = \boldsymbol{0}$ implies $\boldsymbol{U}_1^T\boldsymbol{U}_2 = \boldsymbol{0}$ and $\boldsymbol{U} \equiv [\boldsymbol{U}_1, \boldsymbol{U}_2]$ is column orthogonal. The singular value decomposition of $\boldsymbol{X}$ is simply $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}_2^T$, with:

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_{22} \end{bmatrix} \tag{4}$$

and

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{22} \end{bmatrix}. \tag{5}$$

Thus $\boldsymbol{V}_{12} = \boldsymbol{0}$. An analogous result clearly applies to any number of mutually orthogonal subgroups. Hence, if all the columns in $\boldsymbol{X}$ are orthogonal, all the $v_{kj} = 0$ when $k \neq j$ and $\pi_{kj} = 0$ likewise. When v_{kj} is nonzero, this is a signal that predictors *k* and *j* are not orthogonal.

Since at least one v_{kj} must be nonzero in (3), this implies that a high proportion of any variance can be associated with a large singular value even when there is no collinearity. The standard approach is to check a high condition index associated with a large proportion of the variance of two or more coefficients when diagnosing collinearity, since there must be two or more columns of $\boldsymbol{X}$ involved to make a near dependency. Belsley et al. (1980) suggested showing the matrix $\boldsymbol{\Pi}$ and condition indexes of $\boldsymbol{X}$ in a variance decomposition table as below. If two or more elements in the j^{th} row of matrix $\boldsymbol{\Pi}$ are relatively large and its associated condition index η_j is large too, it signals that near dependencies are influencing regression estimates.

Condition	Proportions of variance			
Index	$Var_M(\hat{\beta}_1)$	$Var_M(\hat{\beta}_2)$	$\cdots$	$Var_M(\hat{\beta}_p)$
η_1	π_{11}	π_{12}	$\cdots$	π_{1p}
η_2	π_{21}	π_{22}	$\cdots$	π_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
η_p	π_{p1}	π_{p2}	$\cdots$	π_{pp}

3 Adaptation in Survey-Weighted Least Squares

3.1 Condition Indexes and Variance Decomposition Proportions

In survey-weighted least squares (SWLS), we are more interested in the collinear relations among the columns in the matrix $\tilde{\boldsymbol{X}} = \boldsymbol{W}^{1/2}\boldsymbol{X}$ instead of $\boldsymbol{X}$, since $\hat{\boldsymbol{\beta}}_{SW} = (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{Y}}$. Define the singular value decomposition of $\tilde{\boldsymbol{X}}$ to be $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, where $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{D}$ are usually different from the ones of $\boldsymbol{X}$, due to the unequal survey weights.

The condition number of $\tilde{\boldsymbol{X}}$ is defined as $\kappa(\tilde{\boldsymbol{X}}) = \mu_{max}/\mu_{min}$, where μ_{max} and μ_{min} are maximum and minimum singular values of $\tilde{\boldsymbol{X}}$. The condition number of $\tilde{\boldsymbol{X}}$ is also usually different from the condition number of the data matrix $\boldsymbol{X}$ due to unequal survey weights. Condition indexes are defined as

$$\eta_k = \mu_{max}/\mu_k, \quad k = 1, ..., p \tag{6}$$

where μ_k is one of the singular values of $\tilde{\boldsymbol{X}}$. The scaled condition indexes and condition numbers are the condition indexes and condition numbers of the scaled $\tilde{\boldsymbol{X}}$.

Based on the extrema of the ratio of quadratic forms (Lin, 1984), the condition number $\kappa(\tilde{\boldsymbol{X}})$ is bounded in the range of:

$$\frac{w_{min}^{1/2}}{w_{max}^{1/2}}\kappa(\boldsymbol{X}) \leq \kappa(\tilde{\boldsymbol{X}}) \leq \frac{w_{max}^{1/2}}{w_{min}^{1/2}}\kappa(\boldsymbol{X}), \tag{7}$$

where w_{min} and w_{max} are the minimum and maximum survey weights. This expression indicates that if the survey weights do not vary too much, the condition number in SWLS resembles the one in OLS. However, in a sample with a wide range of survey weights, the condition number can be very different between SWLS and OLS. When SWLS has a large condition number, OLS might not. In the case of exact linear dependence among the columns of $\boldsymbol{X}$, the columns of $\tilde{\boldsymbol{X}}$ will also be linearly dependent. In this extreme case at least one eigenvalue of $\boldsymbol{X}$ will be zero, and both $\kappa(\boldsymbol{X})$ and $\kappa(\tilde{\boldsymbol{X}})$ will be infinite. As in OLS, large values of κ or of the η_k's of 10 or more may signal that two or more columns of $\boldsymbol{X}$ have moderate to strong dependencies.

The model variance of the SWLS parameter estimator under a model with $Var_M(\boldsymbol{e}) = \sigma^2\boldsymbol{R}$ is:

$$\begin{aligned} Var_M(\hat{\boldsymbol{\beta}}_{SW}) &= \sigma^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{R}\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1} \\ &= \sigma^2(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\boldsymbol{G}, \end{aligned} \tag{8}$$

where

$$\boldsymbol{G} = (g_{ij})_{p\times p} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{R}\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1} \tag{9}$$

is the *misspecification effect* (MEFF) that represents the inflation factor needed to correct standard results for the effect of intracluster correlation in clustered survey data and for the fact that $Var_M(\boldsymbol{e}) = \sigma^2\boldsymbol{R}$ and not $\sigma^2\boldsymbol{W}^{-1}$ (Scott & Holt, 1982).

Using the SVD of $\tilde{\boldsymbol{X}}$, we can rewrite $Var_M(\hat{\boldsymbol{\beta}}_{SW})$ as

$$Var_M(\hat{\boldsymbol{\beta}}_{SW}) = \sigma^2\boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^T\boldsymbol{G}. \tag{10}$$

The k^{th} diagonal element in $Var_M(\hat{\boldsymbol{\beta}})$ is the estimated variance for the k^{th} coefficient, $\hat{\beta}_k$. Using (10), $Var_M(\hat{\beta}_k)$ can be expressed as:

$$Var(\hat{\beta}_k) = \sigma^2\Sigma_{j=1}^{p}\frac{v_{kj}}{\mu_j^2}\lambda_{kj} \tag{11}$$

where $\lambda_{kj} = \Sigma_{i=1}^{p}v_{ij}g_{ik}$. if $\boldsymbol{R} = \boldsymbol{W}^{-1}$, then $\boldsymbol{G} = \boldsymbol{I}_p$, $\lambda_{kj} = v_{kj}$, and (11) reduces to (3). However, the situation is more complicated when $\boldsymbol{G}$ is not the identity matrix, i.e., when the complex design affects the variance of an estimated regression coefficient. If predictors k and j are orthogonal, $v_{kj} = 0$ for $k \neq j$ and the variance in (11) depends only on the k^{th} singular value and is unaffected by g_{ij}'s that are non-zero. If predictor k and several j's are not orthogonal, then λ_{kj} has contributions from all of those eigenvectors and from the off-diagonal elements of the MEFF matrix $\boldsymbol{G}$. The term λ_{kj} then measures both non-orthogonality of x's and effects of the complex design.

Consequently, we can define variance decomposition proportions and analogous to those for OLS but their interpretation is less straightforward. Let $\phi_{kj} = v_{kj}\lambda_{kj}/\mu_j^2$, $\phi_k = \Sigma_{j=1}^{p}\phi_{kj}$ and $\boldsymbol{Q} = (\phi_{kj})_{p\times p} = (\boldsymbol{V}\boldsymbol{D}^{-2}) \cdot (\boldsymbol{V}^T\boldsymbol{G})^T$. The variance-decomposition proportions are $\pi_{jk} = \phi_{jk}/\phi_k$, which is the proportion of the variance of the kth regression coefficient associated with the j^{th} component of its decomposition in (11). Denote the variance decomposition proportion matrix as

$$\boldsymbol{\Pi} = (\pi_{jk})_{p\times p} = \boldsymbol{Q}^T\bar{\boldsymbol{Q}}^{-1}, \tag{12}$$

where $\bar{\boldsymbol{Q}}$ is the diagonal matrix with the row sums of $\boldsymbol{Q}$ on the main diagonal and 0 elsewhere. The interpretation of the proportions in (12) is not as clear-cut as for OLS because the effect of the MEFF matrix. Section 3.2 discusses the interpretation in more detail in the context of stratified cluster sampling.

Analogous to the method for OLS regression, a variance decomposition table can be formed like the one at the end of section 2. When two or more independent variables are collinear (or “nearly dependent"), one singular value should make a large contribution to the variance of the parameter estimates associated with those variables. For example, if the proportions π_{31} and π_{32} for the variances of $\hat{\boldsymbol{\beta}}_{SW1}$ and $\hat{\boldsymbol{\beta}}_{SW2}$ are large, this would say that the third singular value makes a large contribution to both variances and that the first and second predictors in the regression are, to some extent, collinear. As shown in section 2.3, when the kth and jth columns in $\boldsymbol{X}$ are orthogonal, $v_{kj} = 0$ and the jth singular value's decomposition proportion π_{jk} on $Var(\hat{\beta}_k)$ will be 0.

Several special cases are worth noting. If $\boldsymbol{R} = \boldsymbol{W}^{-1}$ as assumed in WLS, then $\boldsymbol{G} = \boldsymbol{I}$. The variance decomposition in (11) has the same form as (2) in OLS. However, having $\boldsymbol{R} = \boldsymbol{W}^{-1}$ in survey data would be unusual since survey weights are not typically computed based on the variance structure of a model. Note that $\boldsymbol{V}$ is still different from the one in OLS and is one component of the SVD of $\tilde{\boldsymbol{X}}$ instead of $\boldsymbol{X}$. Another special case here is when $\boldsymbol{R} = \boldsymbol{I}$ and the survey weights are equal, in which case the OLS results can be used. However, when the survey weights are unequal, even when $\boldsymbol{R} = \boldsymbol{I}$, the variance decomposition in (11) is different from (2) in OLS since $\boldsymbol{G} \neq \boldsymbol{I}$. In the next section, we will consider some special models that take the population features such as clusters and strata into account when estimating this variance decomposition.

3.2 Variance Decomposition for A Model with Stratified Clustering

The model variance of $\hat{\boldsymbol{\beta}}_{SW}$ in (8) contains the unknown $\boldsymbol{R}$ that must be estimated. In this section, we present an estimator for $\hat{\boldsymbol{\beta}}_{SW}$ that is appropriate for a model with stratified clustering. The variance estimator has both model-based and design-based justification. Suppose that in a stratified multistage sampling design, there are strata $h = 1, ..., H$ in the population, clusters $i = 1, ..., N_h$ in stratum h and units $t = 1, ..., M_{hi}$ in cluster hi. We select clusters $i = 1, ..., n_h$ in stratum h and units $t = 1, ..., m_{hi}$ in cluster hi. Denote the set of sample clusters in stratum h by s_h and the sample of units in cluster hi as s_{hi}. The total number of sample units in stratum h is $m_h = \sum_{i\in s_h} m_{hi}$, and the total in the sample is $m = \sum_{h=1}^{H} m_h$. Assume that clusters are selected with varying probabilities and with replacement within strata and independently between strata. The model we consider is:

$$\begin{aligned} &E_M(Y_{hit}) = \boldsymbol{x}_{hit}^T\boldsymbol{\beta} \quad h = 1, \dots, H, \quad i = 1, \dots, N_h, \quad t = 1, \dots, M_{hi} \\ &Cov_M(\varepsilon_{hit}, \varepsilon_{hi't'}) = 0 \quad \text{where } \varepsilon_{hit} = Y_{hit} - \boldsymbol{x}_{hit}^T\boldsymbol{\beta}, \quad i \neq i' \\ &Cov_M(\varepsilon_{hit}, \varepsilon_{h'i't'}) = 0 \quad h \neq h'. \end{aligned} \tag{13}$$

Units within each cluster are assumed to be correlated but the particular form of the covariances does not have to be specified for this analysis. The estimator $\hat{\boldsymbol{\beta}}_{SW}$ of the regression parameter can be written as:

$$\hat{\boldsymbol{\beta}}_{SW} = \sum_{h=1}^{H} \sum_{i \in s_h} (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \boldsymbol{X}_{hi}^T \boldsymbol{W}_{hi} \boldsymbol{Y}_{hi} \tag{14}$$

where $\boldsymbol{X}_{hi}$ is the $m_{hi} \times p$ matrix of covariates for sample units in cluster hi, $\boldsymbol{W}_{hi} = diag(w_t),\ t \in s_{hi}$, is the diagonal matrix of survey weights for units in cluster hi and $\boldsymbol{Y}_{hi}$ is the $m_{hi} \times 1$ vector of response variables in cluster hi. The model variance of $\hat{\boldsymbol{\beta}}_{SW}$ is:

$$Var_M(\hat{\boldsymbol{\beta}}_{SW}) = (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \boldsymbol{G}_{st} \tag{15}$$

where

$$\begin{aligned} \boldsymbol{G}_{st} &= \left[\sum_{h=1}^{H} \sum_{i \in s_h} \boldsymbol{X}_{hi}^T \boldsymbol{W}_{hi} \boldsymbol{R}_{hi} \boldsymbol{W}_{hi} \boldsymbol{X}_{hi} \right] (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \\ &= \left[\sum_{h=1}^{H} \boldsymbol{X}_h^T \boldsymbol{W}_h \boldsymbol{R}_h \boldsymbol{W}_h \boldsymbol{X}_h \right] (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \end{aligned} \tag{16}$$

with $\boldsymbol{R}_{hi} = Var_M(\boldsymbol{Y}_{hi})$, $\boldsymbol{W}_h = diag(\boldsymbol{W}_{hi})$, and $\boldsymbol{R}_h = Blkdiag(\boldsymbol{R}_{hi})$, $\boldsymbol{W}_h = diag(\boldsymbol{W}_{hi})$, $\boldsymbol{X}_h^T = (\boldsymbol{X}_{h1}^T, \boldsymbol{X}_{h2}^T, ..., \boldsymbol{X}_{h,n_h}^T),\ i \in s_h$. Expression (16) is a special case of (9) with $\boldsymbol{X}^T = (\boldsymbol{X}_1^T, \boldsymbol{X}_2^T, ..., \boldsymbol{X}_H^T)$, where $\boldsymbol{X}_h$ is the $m_h \times p$ matrix of covariates for sample units in stratum h, $\boldsymbol{W} = diag(\boldsymbol{W}_{hi})$, for $h = 1, ..., H$ and $i \in s_h$ and $\boldsymbol{R} = Blkdiag(\boldsymbol{R}_h)$.

Based on the development in Scott & Holt (1982, sec. 4), the MEFF matrix $\boldsymbol{G}_{st}$ can be rewritten for a special case of $\boldsymbol{R}_h$ in a way that will make the decomposition proportions in (12) more understandable. Consider the special case of (13) with

$$Cov_M(\boldsymbol{e}_{hi}) = \sigma^2(1-\rho)\boldsymbol{I}_{m_{hi}} + \sigma^2\rho \boldsymbol{1}_{m_{hi}} \boldsymbol{1}_{m_{hi}}^T$$

where $\boldsymbol{I}_{m_{hi}}$ is the $m_{hi} \times m_{hi}$ identity matrix and $\boldsymbol{1}_{m_{hi}}$ is a vector of m_{hi} 1's. In that case,

$$\boldsymbol{X}_h^T \boldsymbol{W}_h \boldsymbol{R}_h \boldsymbol{W}_h \boldsymbol{X}_h = (1-\rho)\boldsymbol{X}_h^T \boldsymbol{W}_h^2 \boldsymbol{X}_h + \rho \sum_{i \in s_h} m_{hi} \boldsymbol{X}_{Bhi}^T \boldsymbol{W}_{hi}^2 \boldsymbol{X}_{Bhi}$$

where $\boldsymbol{X}_{Bhi} = m_{hi}^{-1} \boldsymbol{1}_{m_{hi}} \boldsymbol{1}_{m_{hi}}^T \boldsymbol{X}_{hi}$. Suppose that the sample is self-weighting so that $\boldsymbol{W}_{hi} = w\boldsymbol{I}_{m_{hi}}$. After some simplification, it follows that

$$\boldsymbol{G}_{st} = w[\boldsymbol{I}_p + (\boldsymbol{M} - \boldsymbol{I}_p)\rho]$$

where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix and $\boldsymbol{M} = (\sum_{h=1}^{H} \sum_{i \in s_h} m_{hi} \boldsymbol{X}_{Bhi}^T \boldsymbol{X}_{Bhi})(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1}$. Thus, if the sample is self-weighting and ρ is very small, then $\boldsymbol{G}_{st} \approx w\boldsymbol{I}_p$ and $Var_M(\hat{\boldsymbol{\beta}}_{SW})$ in (15) will be approximately the same as the OLS variance. If so, the SWLS variance decomposition proportions will be similar to the OLS proportions. In regression problems, ρ often is small since it is the correlation of the errors, $\varepsilon_{hit} = Y_{hit} - \boldsymbol{x}_{hit}^T \boldsymbol{\beta}$, for different units rather than for $\boldsymbol{Y}_{hit}$'s. This is related to the phenomenon that design effects for regression coefficients are often smaller than for means-a fact first noted by Kish & Frankel (1974). In applications where ρ is larger, the variance decomposition proportions in (12) will still be useful in identifying collinearity although they will be affected by departures of the model errors from independence.

Denote the cluster-level residuals as a vector, $\boldsymbol{e}_{hi} = \boldsymbol{Y}_{hi} - \boldsymbol{X}_{hi}\hat{\boldsymbol{\beta}}_{SW}$. The estimator of (15) that we consider was originally derived from design-based considerations. A linearization estimator, appropriate when clusters are selected with replacement, is:

$$var_L(\hat{\boldsymbol{\beta}}_{SW}) = (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \hat{\boldsymbol{G}}_L \tag{17}$$

with the estimated misspecification effect as

$$\hat{\boldsymbol{G}}_L = (\hat{g}_{ij})_{p\times p} = \left[\sum_{h=1}^{H} \frac{n_h}{n_h-1} \sum_{i\in s_h} (\boldsymbol{z}^*_{hi} - \bar{\boldsymbol{z}}^*_h)(\boldsymbol{z}^*_{hi} - \bar{\boldsymbol{z}}^*_h)^T\right] (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1}, \tag{18}$$

where $\bar{\boldsymbol{z}}^*_h = \frac{1}{n_h}\sum_{i\in s} \boldsymbol{z}^*_{hi}$ and $\boldsymbol{z}^*_{hi} = \boldsymbol{X}^T_{hi}\boldsymbol{W}_{hi}\boldsymbol{e}_{hi}$ with $\boldsymbol{e}_{hi} = \boldsymbol{Y}_{hi} - \boldsymbol{X}_{hi}\hat{\boldsymbol{\beta}}_{SW}$, and the variance-covariance matrix $\boldsymbol{R}$ can be estimated by $\hat{\boldsymbol{R}} = \sum_{h=1}^{H} \frac{n_h}{n_h-1}\left[Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}^T_{hi}) - \frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}^T_h\right]$.

Expression (17) is used by the Stata and SUDAAN packages, among others. The estimator $var_L(\hat{\boldsymbol{\beta}}_{SW})$ is consistent and approximately design-unbiased under a design where clusters are selected with replacement (Fuller, 2002). The estimator in (17) is also an approximately model-unbiased estimator of (15) (see Liao, 2010). Since the estimator $var_L(\hat{\boldsymbol{\beta}}_{SW})$ is also currently available in software packages, we will use it in the empirical work in section 4.

Using (12) derived in section 2, the variance decomposition proportion matrix $\boldsymbol{\Pi}$ for $var_L(\hat{\boldsymbol{\beta}}_{SW})$ can then be written as

$$\boldsymbol{\Pi} = (\pi_{jk})_{p\times p} = \boldsymbol{Q}^T_L \bar{\boldsymbol{Q}}^{-1}_L \tag{19}$$

with $\boldsymbol{Q}_L = (\phi_{kj})_{p\times p} = (\boldsymbol{U}_2\boldsymbol{D}^{-2}) \cdot (\boldsymbol{U}^T_2\hat{\boldsymbol{G}}_L)^T$ and $\bar{\boldsymbol{Q}}_L$ is the diagonal matrix with the row sums of $\boldsymbol{Q}_L$ on the main diagonal and 0 elsewhere.

4 Numerical Illustrations

In this section, we will illustrate the collinearity measures described in section 3 and investigate their behaviors using the dietary intake data from 2007-2008 National Health and Nutrition Examination Survey (NHANES).

4.1 Description of the Data

The dietary intake data are used to estimate the types and amounts of foods and beverages consumed during the 24-hour period prior to the interview (midnight to midnight), and to estimate intakes of energy, nutrients, and other food components from those foods and beverages. NHANES uses a complex, multistage, probability sampling design; oversampling of certain population subgroups is done to increase the reliability and precision of health status indicator estimates for these groups. Among the respondents who received the in-person interview in the mobile examination center (MEC), around 94% provided complete dietary intakes. The survey weights were constructed by taking MEC sample weights and further adjusting for the additional nonresponse and the differential allocation by day of the week for the dietary intake data collection. These weights are more variable than the MEC weights. The data set used in our study is a subset of 2007-2008 data composed of female respondents aged 26 to 40. Observations with missing values in the selected variables are excluded from the sample which finally contains 672 complete respondents. The final weights in our sample range from 6,028 to 330,067, with a ratio of 55:1. The U.S. National Center for Health Statistics recommends that the design of the sample is approximated by the stratified selection with replacement of 32 PSUs from 16 strata, with 2 PSUs within each stratum.

4.2 Study One: Correlated Covariates

In the first empirical study, a linear regression model of respondent's body mass index (BMI) was considered. The explanatory variables considered included two demographic variables, respondent's age and race (Black/Non-black), four dummy variables for whether the respondent is on a special diet of any kind, on a low-calorie diet, on a low-fat diet, and on a low-carbohydrate diet (when he/she is on diet, value equals 1, otherwise 0), and ten daily total nutrition intake variables, consisting of total calories (100kcal), protein (100gm), carbohydrate (100gm), sugar (100gm), dietary fiber (100gm), alcohol (100gm), total fat (100gm), total saturated fatty acids (100gm), total monounsaturated fatty acids (100gm), and total polyunsaturated fatty acids (100gm). The correlation coefficients among these variables are displayed in Table 2. Note that the correlations among the daily total nutrition intake variables are often high. For example, the correlations of the total fat intakes with total saturated fatty acids, total monounsaturated fatty acids and total polyunsaturated fatty acids are 0.85, 0.97 and 0.93.

Three types of regressions were fitted for the selected sample to demonstrate different diagnostics. More details about these three regression types and their diagnostic statistics are displayed in Table 1.
TYPE1: OLS regression with estimated σ^2; the diagnostic statistics are obtained using the standard methods reviewed in section 2;
TYPE2: WLS regression with estimated σ^2 and assuming $\boldsymbol{R} = \boldsymbol{W}^{-1}$; the scaled condition indexes are estimated using (6) and the scaled variance decomposition proportions are estimated using (12). With $\boldsymbol{R} = \boldsymbol{W}^{-1}$, these are the variance decompositions that will be produced by standard software using WLS and specifying the weights to be the survey weights;
TYPE3: SWLS with estimated $\hat{\boldsymbol{R}}$, when $\sigma^2\boldsymbol{R}$ is unknown; the scaled condition indexes are estimated using (6); the scaled variance decomposition proportions are estimated using (12).

Table 1: Regression Models and their Collinearity Diagnostic Statistics used in this Experimental Study

Type	Regression Method	Weight matrix $\boldsymbol{W}^a$	$var(\hat{\boldsymbol{\beta}})$	$var(\hat{\beta}_k)$	Matrix for Condition Indexes ^{b}	Variance Decomposition Proportion π_{jk}
TYPE1	OLS	$\boldsymbol{I}$	$\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$	$\sigma^2\Sigma_{j=1}^p \frac{u_{2kj}^2}{\mu_j^2}{}^c$	$\boldsymbol{X}^T\boldsymbol{X}$	$\frac{u_{2kj}^2}{\mu_j^2}/\Sigma_{j=1}^p \frac{u_{2kj}^2}{\mu_j^2}$
TYPE2	WLS	$\boldsymbol{W}$	$\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$	$\sigma^2\Sigma_{j=1}^p \frac{u_{2kj}^2}{\mu_j^2}{}^d$	$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$	$\frac{u_{2kj}^2}{\mu_j^2}/\Sigma_{j=1}^p \frac{u_{2kj}^2}{\mu_j^2}$
TYPE3	SWLS	$\boldsymbol{W}$	$\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{R}}\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$	$\sigma^2\Sigma_{j=1}^p \frac{u_{2kj}\Sigma_{i=1}^p \hat{g}_{ik}u_{2ij}}{\mu_j^2}{}^e$	$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$	$\frac{u_{2kj}\Sigma_{i=1}^p \hat{g}_{ik}u_{2ij}}{\mu_j^2}/\Sigma_{j=1}^p \frac{u_{2kj}\Sigma_{i=1}^p \hat{g}_{ik}u_{2ij}}{\mu_j^2}$
			$\hat{\boldsymbol{R}} = \sum_{h=1}^H \frac{n_h}{n_h-1}\left[Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) - \frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right]$			

[a] In all the regression models, the parameters are estimated by: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$.
[b] The eigenvalues of this matrix will be used to compute the Condition Indexes for the corresponding regression model.
[c] The terms u_{2kj} and μ_j are from the singular value decomposition of the data matrix $\boldsymbol{X}$.
[d] The terms u_{2kj} and μ_j are from the singular value decomposition of the weighted data matrix $\tilde{\boldsymbol{X}} = \boldsymbol{W}^{1/2}\boldsymbol{X}$.
[e] The terms u_{2kj} and μ_j are from the singular value decomposition (SVD) of the weighted data matrix $\tilde{\boldsymbol{X}}$. The term $\hat{g}_{ik}$ is the unit element of misspecification effect matrix $\hat{\boldsymbol{G}}$.

Their diagnostic statistics, including the scaled condition indexes and variance decomposition proportions are reported in Tables 3, 4 and 5, respectively. To make the tables more readable, only the proportions that are larger than 0.3 are shown. Proportions that are less than 0.3 are shown as dots. Note that some terms in decomposition (12) can be negative. This leads to the possibility of some "proportions" being greater than 1. This occurs in five cases in Table 5. Belsley et al. (1980) suggest that a condition index of 10 signals that collinearity has a moderate effect on standard errors; an index of 100 would indicate a serious effect. In this study, we consider a scaled condition index greater than 10 to be relatively large, and ones greater than 30 as large and remarkable. Furthermore, the large scaled variance-decomposition proportions (greater than 0.3) associated with each large scaled condition index will be used to identify those variates that are involved in a near dependency.

In Tables 3, 4 and 5, the weighted regression methods, WLS and SWLS, used the survey-weighted data matrix $\tilde{\boldsymbol{X}}$ to obtain the condition indexes while the unweighted regression method, OLS, used the data matrix $\boldsymbol{X}$. The largest scaled condition index in WLS and SWLS is 566, which is slightly smaller than the one in OLS, 581. Both of these values are much larger than 30 and, thus, signal a severe near-dependency among the predictors in all three regression models. Such large condition numbers imply that the inverse of the design matrix, $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$, may be numerically unstable, i.e., small changes in the x data could make large changes in the elements of the inverse.

The values of the decomposition proportions for OLS and WLS are very similar and lead to the same predictors being identified as potentially collinear. Results for SWLS are somewhat different as sketched below. In OLS and WLS, six daily total nutrition intake variables–calorie, protein, carbohydrate, alcohol, dietary fiber and total fat–are involved in the dominant near-dependency that is associated with the largest scaled condition index. Four daily fat intake variables, total fat, total saturated fatty acids, total monounsaturated fatty acids and total polyunsaturated fatty acids, are involved in the secondary near-dependency that is associated with the second largest scaled condition index. A moderate near-dependency between intercept and age is also shown in all three tables. The associated scaled condition index is equal to 38 in OLS and 37 in WLS and SWLS. However, when SWLS is used, sugar, total saturated fatty acids and total polyunsaturated fatty acids also appear to be involved in the dominant near-dependency as shown in Table 5. While, only three daily fat intake variables, total saturated fatty acids, total monounsaturated fatty acids and total polyunsaturated fatty acids, are involved in the secondary near-dependency that is associated with the second largest scaled condition index. Thus, when OLS or WLS is used, the impact of near-dependency among sugar, total saturated fatty acids, total polyunsaturated fatty acids and the six daily total nutrition intake variables is not as strong as the ones in SWLS. If conventional OLS or WLS diagnostics are used for SWLS, this near-dependency might be overlooked.

Rather than using the scaled condition indexes and variance decomposition method (in Tables 3, 4 and 5), an analyst might attempt to identify collinearities by examining the unweighted correlation coefficient matrix in Table 2. Although the correlation coefficient matrix shows that almost all the daily total nutrition intake variables are highly or moderately pairwise correlated, it cannot be used to reliably identify the near-dependencies among these variables when used in a regression. For example, the correlation coefficient between "on any diet" and "on low-calorie diet" is relatively large (0.73). This near dependency is associated with a scaled condition index equal to 11 (larger than 10, but less than the cutoff of 30) in OLS and WLS (shown in Table 3 and 4) and is associated with a scaled condition index equal to 2 (less than 10) in SWLS (shown in Table 5). The impact of this near dependency appears to be not very harmful not matter which regression method is used. On the other hand, alcohol is weakly correlated with all the daily total nutrition intake variables but is highly involved in the dominant near-dependency shown in the last row of Tables 3-5.

Table 2: Correlation Coefficient Matrix of the data matrix $\boldsymbol{X}$

	age	black	on any diet	on low-calorie diet	on low-fat diet	on low-carb diet [a]	calorie	protein	carbohydrate	sugar	fiber	alcohol	total.fat	sat.fat	mono.fat	poly.fat
age	1															
black	.[b]	1														
on any diet	.	.	1													
on low-calorie diet	.	.	**0.87**[c]	1												
on low-fat diet	.	.	.	.	1											
one low-carb diet	.	.	.	.	.	1										
calorie	.	.	.	.	.	.	1									
protein	.	.	.	.	.	.	**0.75**	1								
carb	.	.	.	.	.	.	**0.84**	0.45	1							
sugar	.	.	.	.	.	.	**0.58**	.	**0.84**	1						
fiber	.	.	.	.	.	.	**0.57**	**0.52**	**0.54**	.	1					
alcohol	.	.	.	.	.	.	.	.	.	.	.	1				
total.fat	.	.	.	.	.	.	**0.86**	**0.72**	**0.54**	.	0.48	.	1			
sat.fat[d]	.	.	.	.	.	.	**0.74**	**0.56**	0.47	.	0.46	.	**0.85**	1		
mono.fat[e]	.	.	.	.	.	.	**0.83**	**0.68**	**0.51**	.	0.46	.	**0.97**	**0.82**	1	
poly.fat[f]	.	.	.	.	.	.	**0.81**	**0.71**	**0.51**	.	0.43	.	**0.93**	**0.63**	**0.87**	1

[a] The term “carb" stands for carbohydrate.

[b] The correlation coefficient less than 0.3 is omitted in this table.

[c] The correlation coefficient larger than 0.5 is highlighted in this table.

[d] Total Saturated Fatty Acids

[e] Total Monounsaturated Fatty Acids

[f] Total Polyunsaturated Fatty Acids

Table 3: Scaled Condition Indexes and Variance Decomposition Proportions: using TYPE1: OLS

Scaled	Scaled Proportion of the Variance of								
Condition Index	Intercept	Age	Black	on any Diet	on Low-Calorie Diet	on Low-fat Diet	on Low-carb Diet	Calorie	Protein
1	.[a]	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	0.574	.	.
3	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	0.379	.	.	.
4	.	.	0.794	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.	.
6	.	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.
9	.	.	.	.	.	.	.	.	.
11	.	.	.	0.842	0.820	.	.	.	.
12	.	.	.	.	.	.	.	.	.
22	.	.	.	.	.	.	.	.	.
26	.	.	.	.	.	.	.	.	.
38	0.970	0.960	.	.	.	.	.	.	.
157	.	.	.	.	.	.	.	.	.
581	.	.	.	.	.	.	.	0.993	0.966
Scaled Condition Index	Carbohydrate	Sugar	Dietary Fiber	Alcohol	Total Fat	Sat.fat[b]	Mono.fat[c]	Poly.fat[d]	
1	.	.	.	.	.	.	.	.	
2	.	.	.	.	.	.	.	.	
3	.	.	.	.	.	.	.	.	
3	.	.	.	.	.	.	.	.	
3	.	.	.	.	.	.	.	.	
4	.	.	.	.	.	.	.	.	
5	.	.	.	.	.	.	.	.	
6	.	.	.	.	.	.	.	.	
8	.	.	.	.	.	.	.	.	
9	.	.	.	.	.	.	.	.	
11	.	.	.	.	.	.	.	.	
12	.	.	.	.	.	.	.	.	
22	.	.	.	.	.	.	.	.	
26	.	0.633	.	.	.	.	.	.	
38	.	.	.	.	.	.	.	.	
157	.	.	.	.	0.304	0.866	0.890	0.904	
581	0.988	.	0.482	0.986	0.696	.	.	.	

[a] The scaled variance decomposition proportions smaller than 0.3 is omitted in this table.

[b] Total Saturated Fatty Acids

[c] Total Monounsaturated Fatty Acids

[d] Total Polyunsaturated Fatty Acids

Table 4: Scaled Condition Indexes and Variance Decomposition Proportions: using TYPE2: WLS

Scaled	Scaled Proportion of the Variance of								
Condition Index	Intercept	Age	Black	on any Diet	on Low-Calorie Diet	on Low-fat Diet	on Low-carb Diet	Calorie	Protein
1	.[a]	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	0.609	.	.
3	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	0.347	.	.	.
4	.	.	0.711	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.	.
7	.	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.	.	.
11	.	.	.	0.902	0.878	.	.	.	.
13	.	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.	.
26	.	.	.	.	.	.	.	.	.
37	0.959	0.940	.	.	.	.	.	.	.
165	.	.	.	.	.	.	.	.	.
566	.	.	.	.	.	.	.	0.992	0.963
Scaled Condition Index	Carbohydrate	Sugar	Dietary Fiber	Alcohol	Total Fat	Sat.fat[b]	Mono.fat[c]	Poly.fat[d]	
1	.	.	.	.	.	.	.	.	
2	.	.	.	.	.	.	.	.	
3	.	.	.	.	.	.	.	.	
3	.	.	.	.	.	.	.	.	
3	.	.	.	.	.	.	.	.	
4	.	.	.	.	.	.	.	.	
5	.	.	.	.	.	.	.	.	
7	.	.	.	.	.	.	.	.	
8	.	.	.	.	.	.	.	.	
10	.	.	.	.	.	.	.	.	
11	.	.	.	.	.	.	.	.	
13	.	.	.	.	.	.	.	.	
21	.	.	.	.	.	.	.	.	
26	.	0.630	.	.	.	.	.	.	
37	.	.	.	.	.	.	.	.	
165	.	.	.	.	0.342	0.871	0.909	0.919	
566	0.987	.	0.486	0.981	0.658	.	.	.	

[a] The scaled variance decomposition proportions smaller than 0.3 is omitted in this table.

[b] Total Saturated Fatty Acids

[c] Total Monounsaturated Fatty Acids

[d] Total Polyunsaturated Fatty Acids

Table 5: Scaled Condition Indexes and Variance Decomposition Proportions: using TYPE3: SWLS

Scaled	Scaled Proportion of the Variance of								
Condition Index	Intercept	Age	Black	on any Diet	on Low-Calorie Diet	on Low-fat Diet	on Low-carb Diet	Calorie	Protein
1	.[a]	.	.	.	.	.	.	.	.
2	.	.	.	0.717	1.278	0.553	.	.	.
3	.	.	.	.	.	.	0.697	.	.
3	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.	.
7	0.766	1.686	0.461	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.	.
13	.	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.	.
26	.	.	.	.	.	.	.	.	.
37	.	.	.	.	.	.	.	.	.
165	.	.	.	.	.	.	.	.	.
566	0.318	.	.	.	.	.	.	1.095	1.190

Scaled Condition Index	Carbohydrate	Sugar	Dietary Fiber	Alcohol	Total Fat	Sat.fat[b]	Mono.fat[c]	Poly.fat[d]
1	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.
7	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.
13	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.
26	.	0.379	.	.	.	.	.	.
37	.	.	.	.	.	.	.	.
165	.	.	.	.	.	0.651	0.749	0.615
566	1.008	1.509	0.740	1.036	0.805	0.486	.	0.390

[a] The scaled variance decomposition proportions smaller than 0.3 is omitted in this table.

[b] Total Saturated Fatty Acids

[c] Total Monounsaturated Fatty Acids

[d] Total Polyunsaturated Fatty Acids

After the collinearity patterns are diagnosed, the common corrective action would be to drop the correlated variables, refit the model and reexamine standard errors, collinearity measures and other diagnostics. Omitting $\boldsymbol{X}$'s one at a time may be advisable because of the potentially complex interplay of explanatory variables. In this example, if the total fat intake is one of the key variables that an analyst feels must be kept, sugar might be dropped first followed by protein, calorie, alcohol, carbohydrate, total fat, dietary fiber, total monounsaturated fatty acids, total polyunsaturated fatty acids and monounsaturated fatty acids. Other remedies for collinearity could be to transform the data or use some specialized techniques such as ridge regression and mixed Bayesian modeling, which require extra (prior) information beyond the scope of most research and evaluations.

To demonstrate how the collinearity diagnostics can improve the regression results in this example, Table 6 presents the SWLS regression analysis output of the original models with all the explanatory variables and a reduced model with fewer explanatory variables. In the reduced model, all of the dietary intake variables are eliminated except total fat intake. After the number of correlated offending variables is reduced, the standard error of total fat intake is only the one forty-sixth of its standard error in the original model. The total fat intake becomes significant in the reduced model. The reduction of correlated variables appears to have substantially improved the accuracy of estimating the impact of total fat intake on BMI. Note that the collinearity diagnostics do not provide a unique path toward a final model. Different analysts may make different choices about whether particular predictors should be dropped or retained.

Table 6: Regression Analysis Output using TYPE3: SWLS

	Original Model		Reduced Model	
Variable	Coefficient	SE[a]	Coefficient	SE
Intercept	24.14***[b]	2.77	24.20***	2.69
Age	0.06	0.08	0.06	0.08
Black	3.19***	1.04	3.67***	0.98
on any Diet[c]	1.79	1.52	1.28	1.80
on Low-calorie Diet	4.09**	1.50	4.59**	1.69
on Low-fat Diet	3.67	2.86	3.87	3.76
on Low-carb Diet	0.46	3.51	0.87	3.86
Calorie	-0.88	2.36		
Protein	7.05	9.59		
Carbohydrate	3.69	9.62		
Sugar	-0.31	1.11		
Dietary Fiber	-14.52*	5.89		
Alcohol	2.09	16.47		
Total Fat	29.34	31.37	1.47*	0.68
Total Saturated Fatty Acids	-15.90	20.18		
Total Monounsaturated Fatty Acids	-22.40	23.01		
Total Polyunsaturated Fatty Acids	-27.69	21.10		
Intracluster Coefficient ρ	0.0366		0.0396	

[a] standard error

[b] p-value: *, 0.05; **, 0.01; ***, 0.005

[c] The reference category is "not being on diet" for all the on-diet variables here.

4.3 Study Two: Reference Level for Categorical Variables

As noted earlier, using non-survey data, dummy variables can also play an important role as a possible source for collinearity. The choice of reference level for a categorical variable may affect the degree of collinearity in the data. To be more specific, choosing a category that has a low frequency as the reference and omitting that level in order to fit the model may give rise to collinearity with the intercept term. This phenomenon carries over to survey data analysis as we now illustrate.

We employed the four on-diet dummy variables used in the previous study, which we denote this section as “on any diet" (DIET), “on low-calorie diet" (CALDIET), “on low-fat diet" (FATDIET) and “one low-carbohydrate diet"

(CARBDIET). The model considered here is:

$$
\begin{aligned}
\mathrm{BMI}_{hit} = \beta_0 + \beta_{\mathrm{black}} * \mathrm{black}_{hit} + \beta_{\mathrm{TOTAL.FAT}} * \mathrm{TOTAL.FAT}_{hit} + \beta_{\mathrm{DIET}} * \mathrm{DIET}_{hit} + \\
\beta_{\mathrm{CALDIET}} * \mathrm{CALDIET}_{hit} + \beta_{\mathrm{FATDIET}} * \mathrm{FATDIET}_{hit} + \\
\beta_{\mathrm{CARBDIET}} * \mathrm{CARBDIET}_{hit} + \varepsilon_{hit}
\end{aligned}
\tag{20}
$$

where subscript hit stands for the tth unit in the selected PSU hi, black is the dummy variable of black (black=1 and non-black=0), and $\mathrm{TOTAL.FAT}$ is the variable of daily total fat intake. According to the survey-weighted frequency table, 15.04% of the respondents are “on any diet", 11.43% of them are “on low-calorie diet", 1.33% of them are “on low-fat diet" and 0.47% of them are “on low-carbohydrate diet". Being on a diet is, then, relatively rare in this example. If we choose the majority level, “not being on the diet", as the reference category for all the four on-diet dummy variables, we expect no severe collinearity between dummy variables and the intercept, because most of values in the dummy variables will be zero. However, when fitting model (20), assume that an analyst is interested to see the impact of “not on any diet" on respondent's BMI and reverses the reference level of variable DIET in model (20) into “being on the diet". This change may cause a near dependency in the model because the column in $\boldsymbol{X}$ for variable DIET will nearly equal the column of ones for the intercept. The following empirical study will illustrate the impact of this change on the regression coefficient estimation and how we should diagnose the severity of the resulting collinearity.

Table 7 and 8 present the regression analysis output of the model in (20) using the three regression types, OLS, WLS and SWLS, listed in Table 1. Table 7 is modeling the effects of on-diet factors on BMI by treating “not being on the diet" as the reference category for all the four on-diet variables. While Table 8 changes the reference level of variable DIET from “not on any diet" into “On any diet" and models the effect of “not on any diet" on BMI. The choice of reference level effects the sign of the estimated coefficient for variable DIET but not its absolute value or standard error. The size of the estimated intercept and its SE are different in Tables 7 and 8, but the estimable functions, like predictions, will of course, be the same with either set of reference levels. The SE of the intercept is about three times larger when “on any diet" is the reference level for variable DIET (Table 8) than when it is not (Table 7).

Table 7: Regression Analysis Output: When “not on any diet" is the Reference Category for DIET variable in the Model

Regression Type	Intercept	black	total.fat	on any diet	on low-calorie diet	on low-fat diet	on low-carb diet
TYPE1	27.22***[a]	3.20***	0.95	3.03	1.75	2.75	-1.48
OLS	(0.61)[b]	(0.70)	(0.72)	(1.94)	(2.03)	(2.72)	(3.66)
TYPE2	26.13***	3.65***	1.44*	1.39	4.46*	3.86	0.94
WLS	(0.58)	(0.82)	(0.67)	(1.67)	(1.79)	(2.59)	(4.22)
TYPE3	26.13***	3.65***	1.44*	1.39	4.46**	3.86	0.94
SWLS	(0.64)	(0.99)	(0.63)	(1.80)	(1.70)	(3.73)	(3.87)

[a] p-value: *, 0.05; **, 0.01; ***, 0.005

[b] Standard errors are in parentheses under parameter estimates.

Table 8: Regression Analysis Output: When “on any diet" is the Reference Category for DIET variable in the Model

Regression Type	Intercept	black	total.fat	not on any diet	on low-calorie diet	on low-fat diet	on low-carb diet
TYPE1	30.25***[a]	3.20***	0.95	-3.03	1.75	2.75	-1.48
OLS	(2.00)[b]	(0.70)	(0.72)	(1.94)	(2.03)	(2.72)	(3.66)
TYPE2	27.52***	3.65***	1.44*	-1.39	4.46*	3.86	0.94
WLS	(1.71)	(0.82)	(0.67)	(1.67)	(1.79)	(2.59)	(4.22)
TYPE3	27.52***	3.65***	1.44*	-1.39	4.46**	3.86	0.94
SWLS	(1.75)	(0.99)	(0.63)	(1.80)	(1.70)	(3.73)	(3.87)

[a] p-value: *, 0.05; **, 0.01; ***, 0.005

[b] Standard errors are in parentheses under parameter estimates.

When choosing “not being on diet" as the reference category for all the four on-diet dummy variables in Table 9, the scaled condition indexes are relatively small and do not signify any remarkable near-dependency regardless of the type of regression. Only the last row for the largest condition index is printed in Tables 9 and 10. Often, the reference category for a categorical predictor will be chosen to be analytically meaningful. In this example, using “not being on diet" for each of the four diet variables would be logical.

In Table 10, when “on any diet" is chosen as the reference category for variable DIET, the scaled condition indexes are increased and show a moderate degree of collinearity (condition index larger than 10) between the on-diet dummy variables and the intercept. Using the table of scaled variance decomposition proportions, in OLS and WLS, dummy variable for “not on any diet"" and “on low-calorie diet" are involved in the dominant near-dependency with the intercept; however, in SWLS, only the dummy variable for "not on any diet" is involved in the dominant near-dependency with the intercept and the other three on-diet variables are much less worrisome.

Table 9: Largest Scaled Condition Indexes and Its Associated Variance Decomposition Proportions: When “not on any diet" is the Reference Category for variable DIET in the Model

Scaled	Scaled Proportion of the Variance of						
Condition Index	Intercept	gender	total.fat	on any diet	on low-calorie diet	on low-fat diet	on low-carb diet
TYPE1: OLS							
6	0.005	0.000	0.016	0.949	0.932	0.157	0.200
TYPE2: WLS							
6	0.013	0.008	0.020	0.938	0.926	0.189	0.175
TYPE3: SWLS							
6	0.006	0.007	0.013	0.686	0.741	0.027	0.061

Table 10: Largest Scaled Condition Indexes and Its Associated Variance Decomposition Proportions: When “on any diet" is the Reference Category for variable DIET in the Model

Scaled	Scaled Proportion of the Variance of						
Condition Index	Intercept	gender	total.fat	not on any diet	on low-calorie diet	on low-fat diet	on low-carb diet
TYPE1: OLS							
17	**0.982**	0.001	0.034	**0.968**	**0.831**	0.155	0.186
TYPE2: WLS							
17	**0.982**	0.011	0.029	**0.968**	**0.820**	0.182	0.160
TYPE3: SWLS							
17	**0.897**	0.018	-0.006	**0.971**	0.318	0.014	-0.019

5 Conclusion

Dependence between predictors in a linear regression model fitted with survey data affects the properties of parameter estimators. The problems are the same as for non-survey data: standard errors of slope estimators can be inflated and slope estimates can have illogical signs. In the extreme case when one column of the design matrix is exactly a linear combination of others, the estimating equations cannot be solved. The more interesting cases are ones where predictors are related but the dependence is not exact. The collinearity diagnostics that are available in standard software routines are not entirely appropriate for survey data. Any diagnostic that involves variance estimation needs modification to account for sample features like stratification, clustering, and unequal weighting. This paper adapts condition numbers and variance decompositions, which can be used to identify cases of less than exact dependence, to be applicable for survey analysis.

A condition number of a survey-weighted design matrix $\boldsymbol{W}^{1/2}\boldsymbol{X}$ is the ratio of the maximum to the minimum eigenvalue of the matrix. The larger the condition number the more nearly singular is $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$, the matrix which must be inverted when fitting a linear model. Large condition numbers are a symptom of some of the numerical problems associated with collinearity. The terms in the decomposition also involve "misspecification effects" if the model errors are not independent as would be the case in a sample with clustering. The variance of an estimator of a regression parameter can also be written as a sum of terms that involve the eigenvalues of $\boldsymbol{W}^{1/2}\boldsymbol{X}$. The variance decompositions for different parameter estimators can be used to identify predictors that are correlated with each other. After identifying which predictors are collinear, an analyst can decide whether the collinearity has serious enough effects on a fitted model that action should be taken. The simplest step is to drop one or more predictors, refit the model, and observe how estimates change. The tools we provide here allow this to be done in a way appropriate for survey-weighted regression models.

References

Belsley, D. A. (1984). Demeaning conditioning diagnostics through centering. *The American Statistician*, *38*(2), 73–77.

Belsley, D. A. (1991). *Conditioning Diagnostics, Collinearity and Weak Data in Regression*. New York: John Wiley and Sons.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. New York: Wiley Interscience.

Cook, R. D. (1984). Comment on demeaning conditioning diagnostics through centering. *The American Statistician*, *2*, 88–90.

Elliot, M. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, *33*, 23–34.

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis. *Review of Economics and Statistics*, *49*, 92–107.

Fox, J. (1986). *Linear Statistical Models and Related Methods, With Applications to Social Research*. New York: John Wiley.

Fuller, W. A. (2002). Regression estimation for survey samples. *Survey Methodology*, *28*(1), 5–23.

Hendrickx, J. (2010). *perturb: Tools for evaluating collinearity*. R package version 2.04.
URL `http://CRAN.R-project.org/package=perturb`

Kish, L., & Frankel, M. (1974). Inference from complex samples. *Journal Of the Royal Statistical Society B*, *36*(1), 1–37.

Li, J. (2007a). Linear regression diagnostics in cluster samples. *ASA Proceedings of the Section on Survey Research Methods*, (pp. 3341–3348).

Li, J. (2007b). Regression diagnostics for complex survey data. Unpublished doctoral dissertation, University of Maryland.

Li, J., & Valliant, R. (2009). Survey weighted hat matrix and leverages. *Survey Methodology*, *35*(1), 15–24.

Li, J., & Valliant, R. (2011). Linear regression influence diagnostics for unclustered survey data. *Journal of Official Statistics*, *27*.

Liao, D. (2010). *Collinearity Diagnostics for Complex Survey Data*. Ph.D. thesis, University of Maryland.

Liao, D., & Valliant, R. (2010). Variance inflation factors in the analysis of complex survey data. *submitted*.

Lin, C. (1984). Extrema of quadratic forms and statistical applications. *Communications in Statistics-Theory and Methods*, *13*, 1517 – 1520.

Marquardt, D. W. (1980). Comment on “a critique on some ridge regression methods" by G. smith and F. compbell: “You should standardize the predictor variables in your regression models". *Journal of the American Statistical Association*, *75*(369), 87–91.

Scott, A. J., & Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, *77*(380), 848–854.

Silvey, S. D. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society*, *31*(3), 539–552.

Snee, R. D., & Marquardt, D. W. (1984). Collinearity diagnostics depend on the domain of prediction, and model, and the data. *The American Statistician*, *2*, 88–90.

Steward, G. W. (1987). Collinearity and least squares regression. *Statistical Science*, *2*(1), 68–84.

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

Wissmann, M., Toutenburg, H., & Shalabh (2007). Role of categorical variables in multicollinearity in the linear regression model. Technical Report Number 008, Department of Statistics, University of Munich. Available at `http://epub.ub.uni-muenchen.de/2081/1/report008_statistics.pdf`.

Wood, F. S. (1984). Effect of centering on collinearity and interpretation of the constant. *The American Statistician*, *2*, 88–90.