Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Collinearity diagnostics in generalized linear models fitted with survey data

by Dan Liao and Richard Valliant

Release date: December 23, 2025

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2025
Vol. 51, No. 2, pp. 561-588
Statistics Canada, Catalogue No. 12-001-X

Collinearity diagnostics in generalized linear models fitted with survey data

Dan Liao and Richard Valliant[1]

Abstract

The class of generalized linear models (GLM) is a flexible generalization of ordinary least squares regression that allows the linear model to be related to the response variable via a link function and assumes the magnitude of the variance of each measurement to be a function of its predicted value. Multicollinearity in GLMs can inflate variances of the estimated coefficients and cause poor prediction in certain regions of the regression space. It may also cause a nonsignificant Wald statistic even when the predictors are highly predictive in a model of the family of GLMs. Little previous research has closely investigated the diagnostics of multicollinearity in GLMs, especially when complex survey data are used. In this paper, we develop variance inflation factors (VIFs) that measure the amount that the variance of a parameter estimator is increased due to multicollinearity in GLMs. We also extend VIFs and condition indexes to apply to complex survey data, accounting for design features, e.g. weights, clusters, and strata. Illustrations of these methods are given using data from a household survey of health and nutrition.

Key Words: Condition indexes; Diagnostics for survey data; Multicollinearity; Regression with complex survey data; Variance inflation factors.

1. Introduction

The class of generalized linear models (GLMs) (McCullagh and Nelder, 1989; McCulloch and Searle, 2001) is a flexible generalization of ordinary least squares regression that allows the linear model to be related to the response variable via a link function and assumes the magnitude of the variance of each measurement to be a function of its predicted value. Collinearity in GLMs can inflate variances of the estimated coefficients and cause poor prediction in certain regions of the regression space. Hauck and Donner (1977) also pointed out that collinearity may cause a nonsignificant Wald statistic even when the predictors are highly predictive in a logistic model, and Væth (1985) noted that this effect of collinearity can happen for the other members of the family of the generalized linear models.

In prior decades, several authors have discussed collinearity problems in logistic models (see Schaefer, Roi and Wolfe, 1984; Schaefer, 1986; Marx and Smith, 1990b). Others explored this problem in the framework of GLMs (see Mackinnon and Puterman, 1989; Marx and Smith, 1990a; Weissfeld and Sereika, 1991; Lesaffre and Marx, 1993). All of these papers remark that collinearity in GLMs is not due simply to collinear relations among the explanatory variables (or, equivalently, among the columns of the design matrix $\mathbf{X}$), but rather to ill-conditioning of the observed information matrix. Lesaffre and Marx (1993), following a in ill-conditioning suggestion of Mackinnon and Puterman (1989), investigated the dependence of the ill-conditioning problems on the particular value of the regression parameter $\boldsymbol{\beta}$. They concluded that in GLMs the response and the choice of the model also play a role in the degree of collinearity (ill-conditioning) of the information matrix. The term, ML-collinearity, is given in their paper to describe the

1. Dan Liao, RTI International, 701 13th Street, N.W., Suite 750, Washington DC, 20005; Richard Valliant, Prof. Emeritus, Universities of Michigan and Maryland, 4620 N Park Ave, #1406W, Chevy Chase MD 20815. E-mail: valliant@umich.edu.

scenario when, at the maximum likelihood estimate of the model parameter, there is collinearity among the weighted explanatory variables; the weights are specific to the form of the link function and model-variance in the GLM and are different from survey weights.

The variance inflation factor (VIF) is one of the more popular conventional collinearity diagnostic techniques for ordinary or weighted least squares linear regressions. Although the idea is fairly old (e.g., see Theil, 1971, page 166), VIFs are still widely used. In the applied sciences (e.g., economics, soil science, sociology, marketing, and environmental science), recent articles that use VIFs in linear regression analyses are Abaidoo and Agyapong (2024), Alhakim-Naser, Hasan and Zaifoglu (2024), Dobos and Sasvari (2024), Gao, Xu and Pan (2024), Parasyris, Stankovic and Stankovic (2024), and Zhang, Li, Li and Yin (2024).

A VIF measures the inflation of the variance of a slope estimate, $\hat{\beta}_k$, due to non-orthogonality of the predictors by comparing the model variance of $\hat{\beta}_k$ in the multivariate regression with the one when the kth predictor is orthogonal to all other predictors, or equivalently, if there were only the kth predictor in the regression. In ordinary least squares regression, the VIF is $1/(1-R_k^2)$, where R_k^2 is the square of the multiple correlation from the regression of the kth predictor on the other predictors. Fox and Monette (1992) presented a generalization of linear model VIFs that is available in the `car` (Fox, Weisberg, Price, Adler, Bates, Baud-Bovy, Bolker, Ellison, Firth, Friendly, Gorjanc, Graves, Heiberger, Krivitsky, Laboissiere, Maechler, Monette, Murdoch, Nilsson, Ogle, Ripley, Short, Venables, Walker, Winsemius and Zeileis, 2023) and `glmtoolbox` (Vanegas, Rondón and Paula, 2024) R packages. Condition indexes of a predictor matrix are another tool that can be used in diagnosing collinearity problems in linear models (see, e.g., Belsley, Kuh and Welsch, 1980, Chapter 3) and can be calculated in the `klaR` (Roever, Raabe, Luebke, Ligges, Szepannek, Zentgraf and Meyer, 2023) and `olsrr` (Hebbali, 2024) packages. Liao and Valliant (2012b) and Liao and Valliant (2012a) extended VIFs and condition indexes for linear models to complex survey data.

When first introduced, VIF and condition indexes were not only used as tools to detect collinearity issues in models but also often employed for model selection by eliminating predictors that exhibited high correlations with other covariates in a model. However, there have since been advances in model selection, particularly with the rise of machine learning algorithms, that call into question the continued use of VIFs as an essential tool for model selection. Nevertheless, VIFs and condition indexes remain valuable for exploratory analysis, helping analysts detect collinearity problems and understand the relationships between predictors and their individual effects on a model. Both diagnostics identify multicollinearity, which can distort interpretation of the effect of a single predictor in regression models. In contrast, machine learning algorithms are better suited for dimension reduction when the primary goal is prediction rather than inference. Techniques like regularization that impose penalties for overfitting (e.g., lasso, ridge regression) or feature selection focus on reducing feature redundancy to improve model performance and generalization, rather than understanding the specific effects of predictors. Thus, VIFs and condition indexes are best applied for understanding the role of specific predictors, while machine learning excels at improving prediction.

In practice, analysts also employ VIFs as indicators of the degree of variance inflation and diagnosis for collinearity issues when a model in the family of GLMs is used. However, as will be revealed in section 2, the term $1/(1-R_k^2)$ is neither an accurate way to measure the variance inflation in GLMs, nor adequate to reflect the underlying collinearity pattern in GLMs. We derive VIFs for GLMs based on a linearization variance estimator for GLMs that is commonly used in statistical analysis. The newly-derived VIF is determined by the relationship of the weighted explanatory variables in the observed information matrix and reflects the degree of variance inflation caused by the co-existence of other predictors in the model.

To extend VIFs and condition indexes for GLMs to complex survey data, a number of factors must be considered. In general, complex survey data arise from complex sampling schemes or reflect underlying complex population structures as defined in Skinner, Holt and Smith (1989). The sampling design, for instance, can involve multi-stage sampling and lead to positive intra-cluster correlation for a given study variable among units in the sample. Or, stratified sampling may be adopted, for which different mean and error variance-covariance structures may be appropriate in different strata. Unequal weights are often unavoidable in survey data due to varying inclusion probabilities and reweighting that accounts for nonresponse or noncontact adjustments, poststratification, and calibration adjustments. To account for the various complexities, analysts may use specialized methods in regression analysis for parameter and standard error estimation, as discussed in *Analysis of Complex Surveys* by Skinner et al. (1989) and *Analysis of Survey Data* by Chambers and Skinner (2003). If the regression model for the data in the sample is the same as for the full population, the sample design is often called *ignorable* and the model could be legitimately fitted without using survey weights. On the other hand, if the sample model deviates from the population model, the design is *nonignorable* or *informative* (Pfeffermann and Sverchkov, 2003). In the latter case, sampling weights are used in model-fitting to ensure that the population model is being estimated. We cover both unweighted and weighted diagnostics in this paper.

The remainder of the paper is organized as follows. Section 2 derives model-based VIFs for non-survey data. In Section 3, we consider the case of an analyst who estimates GLM model parameters using survey weights (SWGLM) in order to account for the complexities of survey data, and correspondingly, we derive VIFs and condition indexes appropriate to SWGLM. Our approach includes developing estimators that have both a model and design interpretation as described in Section 3. The fourth section will give some numerical illustrations of the techniques, and the fifth is a conclusion.

2. Collinearity diagnostics in generalized linear model

In this section we review GLMs and present some results for model-based collinearity diagnostics for non-survey data. Section 2.1 gives definitions and the notation to be used which is mainly the same as that of McCullagh and Nelder (1989). Section 2.2 reviews the large sample, model variances of parameter estimators and derives the associated VIFs.

2.1 Definitions and notations

Assume that the data for a finite population of N elements are generated as the realization of a superpopulation model. The N-vector of response variables $\mathbf{y}_{N\times 1} = (y_1, \ldots, y_N)$ is assumed to contain measurements that are independent under a model with mean μ_i and density from the exponential family. Calculations made with respect to a model will be subscripted by M; design-based calculations are subscripted with π. The density of y_i can be written as:

$$f_{Y_i}(y_i; \theta_i, \tau) = \exp\left\{\left[y_i\theta_i - b(\theta_i)\right]/\tau^2 - c(y_i, \tau)\right\}, \quad i = 1, \ldots, N; \tag{2.1}$$

where for convenience, we have written the density in what is called *canonical form* with $b(\cdot)$ and $c(\cdot)$ being known functions and τ an unknown nuisance parameter that is constant for all y_i. As shown in McCullagh and Nelder (1989, Section 2.2.2), $E_M(y_i) = \partial b(\theta_i)/\partial\theta_i \equiv \mu_i$ and $\text{var}_M(y_i) = \tau^2\partial^2 b(\theta_i)/\partial\theta_i^2 \equiv \tau^2 v(\mu_i)$, where $v(\mu_i) = \partial^2 b(\theta_i)/\partial\theta_i^2$. Thus, the mean and variance of y_i involve only the natural parameter θ_i and the nuisance parameter τ.

Some examples of model (2.1) are the following. If y_i is normally distributed with mean μ and known variance σ_o^2, then $\theta_i = \mu$, $\tau^2 = \sigma_o^2$, $b(\theta_i) = \mu^2$, and $c(y_i, \tau) = -\left(\frac{y^2}{2\sigma_o^2} + \log\left(\sqrt{2\pi}\sigma_o\right)\right)$. If y_i is binomial with x successes in n trials with each trial having probability of success p, then $y_i = x$, $\theta_i = \log\left(p/(1-p)\right)$, $\tau^2 = 1$, $b(\theta_i) = n\log(1-p)$, and $c(y_i, \tau) = \log\binom{n}{x}$. Other distributions in the exponential family along with some of the notation to be used later that is associated with each distribution are listed in Table 2.1.

To account for auxiliary variables that are predictive of y, a *link* function is used that relates the auxiliaries to the mean μ_i as described below. Assume that p predictors (x's) are available for each element of the population. The $N \times p$ matrix of predictors for all elements in the population is $\mathbf{X}$. Row i of $\mathbf{X}$ is denoted by $\dot{\mathbf{x}}_i$, which is a p-vector. The kth column of $\mathbf{X}$ is an N-vector denoted by $\mathbf{x}_k$ and corresponds to predictor k. The matrix $\mathbf{X}$ can be written in two ways with this notation: $\mathbf{X} = (\dot{\mathbf{x}}_1, \ldots, \dot{\mathbf{x}}_N)^T$ or $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$. Define the linear predictor η_i, which is a transformation of the mean μ_i, by:

$$\eta_i = \dot{\mathbf{x}}_i\boldsymbol{\beta} = g(\mu_i) \tag{2.2}$$

where $\boldsymbol{\beta}$ is a p-vector of parameters, $\eta_i = \theta_i$ with the additional restriction that it is a linear combination of β's as in (2.2), and $g(\cdot)$ is a link function that is monotonic twice differentiable in the interior of an interval of real numbers. (Although we could use the θ_i notation only, the convention in the GLM literature, which we follow here, is to use θ_i in the definition of the exponential family density and η_i to specify the linear predictor.) Some examples of link functions shown in Table 2.1 are the identity for the normal distribution, log for the Poisson distribution, and logit or probit for the binomial distribution.

Maximum likelihood is used to estimate the regression parameters $\boldsymbol{\beta}$ and their functions, the linear predictors and the fitted values in succession (McCullagh and Nelder, 1989, McCulloch and Searle, 2001). As McCullagh and Nelder (1989, Section 2.5) show, the maximum likelihood estimates (MLE) of $\boldsymbol{\beta}$ can be obtained by a type of iterative reweighted least squares. The population log-likelihood of the generalized linear model is the sum of the natural logarithm of each of the components defined by (2.1) over the N observations:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N}\left[y_i\theta_i - b(\theta_i)\right]\big/\tau^2 - \sum_{i=1}^{N}c(y_i,\tau). \tag{2.3}$$

By setting the first-order partials for $\boldsymbol{\beta}$ equal to zero, the maximum likelihood estimating (MLE) equations for $\boldsymbol{\beta}$ can be written in matrix notation as

$$\frac{\partial l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \frac{1}{\tau^2}\mathbf{X}^T\boldsymbol{\Gamma}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) = \mathbf{0}^T, \tag{2.4}$$

with $\boldsymbol{\Gamma} = \operatorname{diag}(\gamma_i)$, $\gamma_i = \left\{v(\mu_i)[g'(\mu_i)]^2\right\}^{-1}$, $\boldsymbol{\Delta} = \operatorname{diag}(g'(\mu_i))$, $\mathbf{y} = (y_i)_{N\times 1}$, and $\boldsymbol{\mu} = (\mu_i)_{N\times 1}$. The solution to (2.4) is the population MLE of $\boldsymbol{\beta}$ and is denoted by $\hat{\boldsymbol{\beta}}$.

Table 2.1
Characteristics of some common univariate distributions in the exponential family

	Normal	**Poisson**	**Binomial**		**Gamma**	**Inverse Gaussian**
Notation	$N(\mu_i,\sigma^2)$	$P(\mu_i)$	$B(m,\mu_i)/m$		$G(\mu_i,v)$	$IG(\mu_i,\sigma^2)$
Range of y	$(-\infty,\infty)$	$(0(1),\infty)$	$\frac{(0(1),m)}{m}$		$(0,\infty)$	$(0,\infty)$
τ^2	σ^2	1	$1/m$		v^{-1}	σ^2
Link Function	identity	log	logit	probit	reciprocal	μ_i^{-2}
Canonical link $g(\mu_i)$	μ_i	$\log(\mu_i)$	$\log(\frac{\mu_i}{1-\mu_i})$	[a] $\Phi^{-1}(\mu_i)$	μ_i^{-1}	μ_i^{-2}
Variance function $v(\mu_i)$	1	μ_i	$\mu_i(1-\mu_i)$	$\mu_i(1-\mu_i)$	μ_i^2	μ_i^3
[b] $g'(\mu_i)$	1	μ_i^{-1}	$[\mu_i(1-\mu_i)]^{-1}$	[c] $\varphi^{-1}(\dot{\mathbf{x}}_i^T\boldsymbol{\beta})$	$-\mu_i^{-2}$	$-2\mu_i^{-3}$
[d] $\gamma_i = \{v(\mu_i)[g'(\mu_i)]^2\}^{-1}$	1	μ_i	$\mu_i(1-\mu_i)$	$\frac{\varphi^2(\dot{\mathbf{x}}_i^T\boldsymbol{\beta})}{\mu_i(1-\mu_i)}$	μ_i^2	$\frac{1}{4}\mu_i^3$
[e] $\gamma_i g'(\mu_i)$	1	1	1	$\frac{\varphi(\dot{\mathbf{x}}_i^T\boldsymbol{\beta})}{\mu_i(1-\mu_i)}$	-1	$-\frac{1}{2}$

Notes: [a] Φ is the cumulative distribution function of the standard normal distribution.
[b] $\boldsymbol{\Delta} = \operatorname{diag}[g'(\mu_i)]$.
[c] $\varphi = \Phi'$ is the probability density function of the standard normal distribution and $\mathbf{x}_i^T\boldsymbol{\beta} = g(\mu_i)$.
[d] $\boldsymbol{\Gamma} = \operatorname{diag}(\gamma_i)$.
[e] $\boldsymbol{\Gamma\Delta} = \operatorname{diag}[\gamma_i g'(\mu_i)]$.

There are cases where the MLE cannot be found. If $\boldsymbol{\beta}$ belongs to the boundary of P, there is at least one i, where $g^{-1}(\dot{\mathbf{x}}_i\boldsymbol{\beta}) = \mu_i = \mu_{\min}$ or $\mu_{\max}$, as defined earlier. Then, $g'(\mu_i) = 0$, because g is a monotone and twice differentiable function Therefore, γ_i in (2.4) will be infinite and the MLE does not exist. This situation can occur when y_i is binary (0, 1) if all units are 0 or 1 for a particular configuration of $\mathbf{X}$'s – a condition known as complete separation. Another drawback of MLEs is that they may not perform well in small samples.

2.2 Asymptotic variance of parameter estimator and its variance inflation factor

In model-based inference, the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is (see McCulloch and Searle, 2001):

$$\operatorname{avar}_M(\hat{\boldsymbol{\beta}}) = \tau^2(\mathbf{X}^T\boldsymbol{\Gamma}\mathbf{X})^{-1} \tag{2.5}$$

where avar_M stands for the limiting or asymptotic model variance. In generalized linear models, the parameter of interest, $\boldsymbol{\beta}$, is estimated and is used, in turn, to estimate the expectation of y_i, μ_i, conditional on $\mathbf{x}_i$. Denote the estimated μ_i as $\hat{\mu}_i$, $\hat{\gamma}_i = \{v(\hat{\mu}_i)[g'(\hat{\mu}_i)]^2\}^{-1}$ with $g'(\hat{\mu}_i) = \partial g(\hat{\mu}_i)/\partial\hat{\mu}_i$, and $\hat{\boldsymbol{\Gamma}} = \operatorname{diag}(\hat{\gamma}_i)$.

With $\hat{\mathbf{A}} = \mathbf{X}^T\hat{\boldsymbol{\Gamma}}\mathbf{X}$, the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ can be estimated by:

$$v_M(\hat{\boldsymbol{\beta}}) = \hat{\tau}^2\hat{\mathbf{A}}^{-1}, \tag{2.6}$$

and its kth element on the main diagonal is the asymptotic model variance of $\hat{\beta}_k$, which, equivalently, can be estimated by:

$$v_M(\hat{\beta}_k) = \hat{\tau}^2 a^{kk}, \tag{2.7}$$

where a^{kk} is the kth element on the main diagonal of matrix $\hat{\mathbf{A}}^{-1}$. For the Bernoulli and Poisson families, τ is identically 1. For gamma, normal, and inverse-Gaussian exponential families, τ is an unknown parameter. The estimate of τ may depend on which covariates are included in the model and may or may not be needed, depending on the model. For the analysis here, no estimate is needed, because the VIF given below does not include τ.

Derivation of the VIF in expression (2.10) below parallels the steps in Liao and Valliant (2012b, Appendix A) closely. Appendix A.1 in this paper sketches some of the details for GLMs and SWGLMs. For the case with no survey weights, a^{kk} can be expressed in terms of the coefficient of determination of a weighted least squares regression of $\mathbf{x}_k$ on the other x's:

$$a^{kk} = \mathbf{i}_k^T\hat{\mathbf{A}}^{-1}\mathbf{i}_k = \mathbf{i}_k^T(\mathbf{X}^T\hat{\boldsymbol{\Gamma}}\mathbf{X})^{-1}\mathbf{i}_k = \frac{1}{(1-R^2_{\Gamma(k)})\mathbf{x}_k^T\hat{\boldsymbol{\Gamma}}\mathbf{x}_k} \tag{2.8}$$

where $\mathbf{i}_k$ is a p-vector with 1 in position k and zeroes elsewhere; and $R^2_{\Gamma(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{\Gamma(k)}\mathbf{X}^T_{(k)}\hat{\boldsymbol{\Gamma}}\mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{\Gamma(k)}}{\mathbf{x}_k^T\hat{\boldsymbol{\Gamma}}\mathbf{x}_k}$ with $\hat{\beta}_{\Gamma(k)} = (\mathbf{X}^T_{(k)}\hat{\boldsymbol{\Gamma}}\mathbf{X}_{(k)})^{-1}\mathbf{X}^T_{(k)}\hat{\boldsymbol{\Gamma}}\mathbf{x}_k$ and $\mathbf{X}_{(k)}$ being the $\mathbf{X}$ matrix with the kth column omitted. The term $R^2_{\Gamma(k)}$ is a coefficient of determination and ranges from 0 to 1.

When there is only $\mathbf{x}_k$ in the regression, according to (2.5), the estimator of the asymptotic variance of $\hat{\beta}_k$ is:

$$v_M(\hat{\beta}_k) = \hat{\tau}^2(\mathbf{x}_k^T\hat{\boldsymbol{\Gamma}}\mathbf{x}_k)^{-1}. \tag{2.9}$$

Comparing (2.8) and (2.9) and noting $v_M(\widehat{\beta}_k)$ in (2.6), we can see that the asymptotic model-based variance estimator of $\widehat{\beta}_k$ is inflated

$$\mathrm{VIF}_{M(k)} = \frac{1}{1-R^2_{\Gamma(k)}} \tag{2.10}$$

times in (2.8) by taking all the other $p-1$ explanatory variables into the regression with $\mathbf{x}_k$. Note that the regression of $\mathbf{x}_k$ on the other $\mathbf{x}$-predictors uses $\widehat{\mathbf{\Gamma}}$ weights, but these are not the survey weights. This variance inflation factor is always larger than or equal to 1, because of the range of $R^2_{\Gamma(k)}$. For non-survey data Özkale (2021, Section 3.1) suggests a VIF measure for GLMs, similar to (2.10), based on the diagonal elements of the inverse of the correlation matrix for the x predictors, analogous to the proposal of Chatterjee and Price (1991) for linear models. Her VIF is based on the centered and scaled full $\mathbf{X}$ matrix, following a suggestion from Smith and Marx (1990). In addition, Özkale extended the Relative Efficiency of Design (RED) indicator – originally developed for linear regression in Kovács, Petres and Tóth (2005) – to GLMs. She also proposed corrected VIFs based on ridge regression estimates, which help stabilize variance estimates in the presence of multicollinearity by introducing a regularization parameter.

Under model (2.1), the variance of y_i in the GLM is proportional to $v(\mu_i)$. For example, the variance of y_i in a logistic model is $p_i(1-p_i)$ under the Bernoulli distribution assumption. This assumption is violated when the data are clustered in the population, which results in an *over-dispersion*, i.e., the variance of y_i exceeds variance $v(\mu_i)$. A typical way to handle clustering under model-based inference is using generalized linear mixed models (GLMMs) by treating clusters as random effects. Agresti (2002, Chapter 12) and Shao (2003, Chapter 4) discuss more details about this approach. In this article, we only treat the general GLMs and leave appropriate diagnostics for GLMMs to future research. In the next section, we do consider a different type of clustering – one induced by a sample design.

3. Adaptations to survey-weighted GLMs

The results for non-survey data in Section 2 can be adapted to apply to complex surveys as described in this section. In this section, we present both model-based and model-design results for VIFs. By “model-design” estimators, we refer to ones that maintain efficiency and interpretability associated with the model-based methods, while also ensuring robustness to complex survey design features. This approach is akin to the Kott (2018) concept of design-sensitive inference, which retains techniques from design-based analysis (such as weighted estimating equations and sandwich variance estimators) within a model-based framework. In this approach, estimates are treated as superpopulation parameters, with weights used to robustify model fitting, whereas design-based methods view them as finite-population quantities. Regardless of whether estimates are viewed through a design-based or design-sensitive lens, our model-design VIF estimators remain interpretable to assess multicollinearity within the corresponding conceptual framework.

3.1 Sample design

Suppose that a stratified multistage sampling design is used. There are $h=1,\ldots,H$ strata in the population, $i=1,\ldots,M_h$ clusters in stratum h and $t=1,\ldots,N_{hi}$ elements in cluster hi. We select $i=1,\ldots,m_h$ clusters in stratum h and $t=1,\ldots,n_{hi}$ elements in cluster hi. Denote the set of sample clusters in stratum h by s_h and the sample of elements in cluster hi as s_{hi}. The total number of sample clusters is $m=\sum_h m_h$, the total number of sample elements in stratum h is $n_h=\sum_{i\in s_h} n_{hi}$ and the total in the sample is $n=\sum_{h=1}^{H} n_h$. Assume that probability samples are selected of clusters and elements within clusters; the particular methods of selection do not need to be specified for the subsequent analyses. Clusters are assumed to be selected with replacement within strata and independently between strata. The with-replacement assumption for selection of clusters is common for theoretical analyses because it simplifies variance formulas.

3.2 Survey-weighted generalized linear models

In this section, for convenience we simplify the notation compared to that in Section 3.1 but will apply that more elaborate notation in later sections. In a complex survey design, denote the sampled data set as s and suppose there are n elements in s. The population log-likelihood of the generalized linear model in (2.3) is estimated by summing the natural logarithm of each of the components defined by (2.1) over the n observations weighted by the survey weights w_i:

$$\hat{l}(\boldsymbol{\beta}) = \sum_{i\in s} w_i\left[y_i\theta_i - b(\theta_i)\right]/\tau^2 - \sum_{i\in s} w_i c(y_i,\tau). \tag{3.1}$$

Assume that the weights, $\{w_i\}_{i\in s}$, are constructed to provide unbiased or approximately unbiased estimators of population totals.

By setting the first-order partials for $\boldsymbol{\beta}$ equal to zero and extending the result in (2.4), the pseudo-maximum likelihood estimating equations for $\boldsymbol{\beta}$ are given by:

$$\begin{aligned}\frac{\partial\hat{l}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} &= \frac{1}{\tau^2}\sum_{i\in s} w_i \frac{(y_i-\mu_i)}{v(\mu_i)\,g'(\mu_i)}\mathbf{x}_i^T \\ &= \frac{1}{\tau^2}\sum_{i\in s} w_i (y_i-\mu_i)\gamma_i g'(\mu_i)\mathbf{x}_i^T \\ &= \mathbf{0}^T\end{aligned} \tag{3.2}$$

where $\gamma_i = \left\{v(\mu_i)[g'(\mu_i)]^2\right\}^{-1}$ as in (2.4). In matrix notation this is

$$\frac{\partial\hat{l}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \frac{1}{\tau^2}\mathbf{X}^T\mathbf{W}\boldsymbol{\Gamma}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) = \mathbf{0}^T, \tag{3.3}$$

with $\mathbf{X}$ being the $n\times p$ matrix of predictors for the sample elements, $\mathbf{W}=\text{diag}(w_i)_{i\in s}$, $\boldsymbol{\Gamma}=\text{diag}(\gamma_i)_{i\in s}$, $\boldsymbol{\Delta}=\text{diag}\left[g'(\mu_i)\right]$ and $\boldsymbol{\mu}=(\mu_i)_{i\in s}$. The value of $\boldsymbol{\beta}$ that solves the survey-weighted estimating equations will be denoted by $\hat{\boldsymbol{\beta}}_{SW}$. To distinguish the generalized linear models in the complex survey design from the ordinary generalized linear models, we will refer to *survey-weighted generalized linear models* (SWGLM) since the survey weights are incorporated in the pseudo-maximum likelihood estimation.

The model-based VIF results for survey-weighted regression parallel those in Section 2.2 for non-survey GLMs and are sketched here. The asymptotic, model-based variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{SW}$ is:

$$\operatorname{avar}_M(\hat{\boldsymbol{\beta}}_{SW}) = \tau^2(\mathbf{X}^T\mathbf{W}\boldsymbol{\Gamma}\mathbf{X})^{-1} \equiv \tau^2\mathbf{A}^{-1}. \tag{3.4}$$

Even though this estimator does involve the survey weights, noninformative sampling is required for it to be consistent under model (2.1). The kth element on the main diagonal, $\tau^2 a^{kk}$, is the asymptotic model variance of $\hat{\beta}_{SWk}$. Expression (3.4) can be estimated by $\hat{\tau}^2\widehat{\mathbf{A}}^{-1}$ where $\widehat{\mathbf{A}} = \mathbf{X}^T\mathbf{W}\widehat{\boldsymbol{\Gamma}}\mathbf{X}$ is the estimator of the information matrix with $\widehat{\boldsymbol{\Gamma}} = \operatorname{diag}(\hat{\gamma}_i)_{i\in s}$. An estimator of the model variance of $\hat{\beta}_{SWk}$ is:

$$\operatorname{avar}_M(\hat{\beta}_{SWk}) = \hat{\tau}^2\hat{a}^{kk}, \tag{3.5}$$

where $\hat{a}^{kk}$ is the kth element on the main diagonal of matrix $\widehat{\mathbf{A}}^{-1}$. In some models, τ must be estimated, but for the analysis here, no estimate is needed, because the VIF given below does not include τ.

Using steps similar to those needed to derive (2.8),

$$\hat{a}^{kk} = \frac{1}{(1-R^2_{W\Gamma(k)})\,\mathbf{x}_k^T\mathbf{W}\widehat{\boldsymbol{\Gamma}}\mathbf{x}_k} \tag{3.6}$$

where $R^2_{W\Gamma(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{W\Gamma(k)}\mathbf{X}^T_{(k)}\mathbf{W}\hat{\boldsymbol{\Gamma}}\mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)}}{\mathbf{x}_k^T\mathbf{W}\hat{\boldsymbol{\Gamma}}\mathbf{x}_k}$ with $\hat{\beta}_{W\Gamma(k)} = (\mathbf{X}^T_{(k)}\mathbf{W}\widehat{\boldsymbol{\Gamma}}\mathbf{X}_{(k)})^{-1}\mathbf{X}^T_{(k)}\mathbf{W}\widehat{\boldsymbol{\Gamma}}\mathbf{x}_k$. $R^2_{W\Gamma(k)}$ is the coefficient of determination in the linear regression of $\mathbf{x}_k$ on the $p-1$ other explanatory variables with weight matrix $\mathbf{W}\widehat{\boldsymbol{\Gamma}}$. The asymptotic model variance of $\hat{\beta}_{SWk}$ is inflated by

$$\mathrm{VIF}_{MWk} = \frac{1}{1-R^2_{W\Gamma_{(k)}}} \tag{3.7}$$

compared to a weighted regression in which all predictors are orthogonal.

3.3 Model-design linearization variance estimator for $\hat{\boldsymbol{\beta}}_{SW}$

To compute a model-design VIF in a complex design, an estimator of the asymptotic variance of $\hat{\boldsymbol{\beta}}_{SW}$ is needed. One model-based choice is the estimator of the information matrix found by substituting estimators of τ and $\boldsymbol{\Gamma}$ into (3.4). A model-design choice, i.e., one that has good design and model properties, is the linearization variance estimator for $\hat{\boldsymbol{\beta}}_{SW}$ via the "Binder" method (Binder, 1983). The Binder sandwich estimator is convenient because it is available in widely used software like Stata (https://www.stata.com/manuals/svyvarianceestimation.pdf) and the R survey package (Lumley, 2024). The sandwich estimator is model-consistent if the specification of the GLM is correct and design-consistent under the conditions in Binder (1983, Appendix).

Binder and Roberts (2003, Section 3.3.1) also observe that when the sampling fraction is small, a design-consistent variance estimator of some point estimator $\hat{t}$ is suitable for estimating the combination model-design variance, $V_{M\pi}(\hat{t}) = V_M E_\pi(\hat{t}) + E_M V_\pi(\hat{t})$. The Binder-Roberts result requires that the design-consistent variance estimator also be asymptotically model-unbiased for $E_M V_\pi(\hat{t})$ – a requirement that the Binder sandwich estimator fulfills.

The estimating equations for $\hat{\boldsymbol{\beta}}_{SW}$ in generalized linear models are obtained by setting $\partial \hat{l}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ equal to zero as shown in (3.2) and (3.3). We can rewrite (3.3) by defining $\mathbf{z}_i = (y_i - \mu_i)\gamma_i g'(\mu_i)\mathbf{x}_i^T = e_i \gamma_i g'(\mu_i)\mathbf{x}_i^T$, where $e_i = y_i - \mu_i$ is the residual for element i, and $\hat{\mathbf{z}}_i^* = w_i \mathbf{z}_i$, $\hat{\mathbf{z}}^* = \sum_{i \in s} \mathbf{z}_i^* = \frac{1}{\hat{\tau}^2} \mathbf{X}^T \mathbf{W} \boldsymbol{\Gamma} \boldsymbol{\Delta} (\mathbf{y} - \boldsymbol{\mu})$ with $\boldsymbol{\mu} = (\mu_i)_{n \times 1}$ and $\boldsymbol{\Delta} = \operatorname{diag}[g'(\mu_i)]$. The estimating equations in (3.3) can then be briefly written as $\mathbf{z}^* = \mathbf{0}$. Note that to evaluate $\hat{\mathbf{z}}_i^*$, sample estimates of the μ_i must be used.

Subject to conditions listed in Binder (1983, Appendix), the approximate design variance-covariance matrix for $\hat{\boldsymbol{\beta}}_{SW}$ is:

$$V_\pi(\hat{\boldsymbol{\beta}}_{SW}) = \mathbf{J}(\boldsymbol{\beta})^{-1} V_\pi(\hat{\mathbf{z}}^*) \mathbf{J}(\boldsymbol{\beta})^{-1} \tag{3.8}$$

where V_π denotes variance with respect to the sample design, $\mathbf{J}(\boldsymbol{\beta})$ is the partial derivative (Jacobian) matrix of the estimating equations for $\hat{\boldsymbol{\beta}}_{SW}$ evaluated at the population $\boldsymbol{\beta}$. The Jacobian is $E_M\left[\frac{\partial^2 \hat{l}(\boldsymbol{\beta}_{SW})}{\partial \boldsymbol{\beta}_{SW} \partial \boldsymbol{\beta}_{SW}^T}\right] = \tau^{-2} \mathbf{X}^T \boldsymbol{\Gamma} \mathbf{X}$. The linearization variance estimator of the variance-covariance matrix for $\hat{\boldsymbol{\beta}}_{SW}$ is found by substituting design-based variance estimators for $\mathbf{J}$ and $V_\pi(\hat{\mathbf{z}}^*)$, leading to

$$\begin{aligned} v_L(\hat{\boldsymbol{\beta}}_{SW}) &= \hat{\tau}^4 [\mathbf{X}^T \mathbf{W} \hat{\boldsymbol{\Gamma}} \mathbf{X}]^{-1} v_\pi(\hat{\mathbf{z}}^*) [\mathbf{X}^T \mathbf{W} \hat{\boldsymbol{\Gamma}} \mathbf{X}]^{-1} \\ &= \hat{\tau}^4 \hat{\mathbf{A}}^{-1} v_\pi(\hat{\mathbf{z}}^*) \hat{\mathbf{A}}^{-1} \end{aligned} \tag{3.9}$$

with $\hat{\mathbf{A}} = \mathbf{X}^T \mathbf{W} \hat{\boldsymbol{\Gamma}} \mathbf{X}$ as in (3.4). Note that this variance estimator does not require the assumption for model (2.1) that the model variance of y is a function of the model mean of y. Estimator (3.9) can also be applied to non-survey data where no survey weights are used.

When a design-unbiased or consistent estimate of $V_\pi(\hat{\mathbf{z}}^*)$ is substituted in (3.9), the linearization variance estimator for $\hat{\boldsymbol{\beta}}_{SW}$ will also be approximately design-unbiased and consistent. To derive a model-design VIF, we need to integrate our diagnostic method with a particular sampling design, because the estimation of the design-based variance estimator $v_\pi(\hat{\mathbf{z}}^*)$ is related to the sampling design. Appendix A.2 presents the estimator, $v_\pi(\hat{\mathbf{z}}^*)$, in stratified, multistage sampling.

3.4 Variance inflation factor in SWGLM

Let $\widehat{\mathbf{V}}$ be the estimator, $v_\pi(\hat{\mathbf{z}}^*)$, appropriate to whatever the sample design is. The linearization variance estimator $v_L(\hat{\boldsymbol{\beta}}_{SW})$ can then be rewritten as:

$$v_L(\hat{\boldsymbol{\beta}}_{SW}) = \hat{\tau}^4 \widehat{\mathbf{A}}^{-1} \mathbf{X}^T \mathbf{W} \widehat{\boldsymbol{\Gamma}} \widehat{\boldsymbol{\Delta}} \widehat{\mathbf{V}} \widehat{\boldsymbol{\Delta}} \widehat{\boldsymbol{\Gamma}} \mathbf{W} \mathbf{X} \widehat{\mathbf{A}}^{-1}. \tag{3.10}$$

Denote $\mathbf{V}_\Delta = \widehat{\boldsymbol{\Delta}}\widehat{\mathbf{V}}\widehat{\boldsymbol{\Delta}}$, $\mathbf{W}_\Gamma = \mathbf{W}\widehat{\boldsymbol{\Gamma}} = \operatorname{diag}(w_i\hat{\gamma}_i)$, $\mathbf{B} = \mathbf{X}^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma\mathbf{X}$, $\mathbf{G} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ and we have $v_L(\widehat{\boldsymbol{\beta}}_{SW}) = \hat{\tau}^4\mathbf{G}$. The linearization variance estimator for $\widehat{\beta}_{SWk}$ is the kth element on the main diagonal of $v_L(\widehat{\boldsymbol{\beta}}_{SW})$, so it is directly related to the kth element on the main diagonal of matrix $\mathbf{G}$ as:

$$v_L(\widehat{\beta}_{SWk}) = \hat{\tau}^4 g_{kk}, \tag{3.11}$$

upon defining $\mathbf{G} = (g_{ij})$. As shown in Appendix A.1, g_{kk} is

$$g_{kk} = \frac{\hat{\zeta}_k\hat{\varrho}_k}{1 - R^2_{W\Gamma(k)}}\frac{\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma\mathbf{x}_k}{(\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{x}_k)^2} \tag{3.12}$$

where $R^2_{\mathbf{W}\Gamma(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{W\Gamma(k)}\mathbf{X}^T_{(k)}\mathbf{W}_\Gamma\mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)}}{\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{x}_k}$ with $\widehat{\beta}_{\mathbf{W}\Gamma(k)} = (\mathbf{X}^T_{(k)}\mathbf{W}_\Gamma\mathbf{X}_{(k)})^{-1}\mathbf{X}^T_{(k)}\mathbf{W}_\Gamma\mathbf{x}_k$, $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_{(k)}\widehat{\boldsymbol{\beta}}_{W\Gamma(k)}$ is the residual from the survey-weighted regression of $\mathbf{x}_k$ on $\mathbf{X}_{(k)}$ using weight matrix $\mathbf{W}_\Gamma$,

$$\hat{\zeta}_k = \frac{(\mathbf{x}_k - \mathbf{X}_{(k)}\widehat{\boldsymbol{\beta}}_{W\Gamma(k)})^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma(\mathbf{x}_k - \mathbf{X}_{(k)}\widehat{\boldsymbol{\beta}}_{W\Gamma(k)})}{(\mathbf{x}_k - \mathbf{X}_{(k)}\widehat{\boldsymbol{\beta}}_{W\Gamma(k)})^T\mathbf{W}_\Gamma(\mathbf{x}_k - \mathbf{X}_{(k)}\widehat{\boldsymbol{\beta}}_{W\Gamma(k)})} = \frac{\mathbf{e}_{xk}^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma\mathbf{e}_{xk}}{\mathbf{e}_{xk}^T\mathbf{W}_\Gamma\mathbf{e}_{xk}},$$

and

$$\hat{\varrho}_k = \frac{\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{x}_k}{\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma\mathbf{x}_k}.$$

Note that if $\mathbf{V}_\Delta = \mathbf{W}_\Gamma^{-1}$, the VIF for the design-based approach will be $\left(1 - R^2_{W\Gamma(k)}\right)^{-1}$, which is similar to the non-survey VIF for linear models. However, this assumption is unlikely to hold in practice.

Consider a model with only $\mathbf{x}_k$. Using (3.10), the linearization variance estimator for $\widehat{\beta}_{SWk}$ is:

$$v_{0L}(\widehat{\beta}_{SWk}) = \hat{\tau}^4\frac{\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma\mathbf{x}_k}{(\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{x}_k)^2}. \tag{3.13}$$

Hence, $v_{0L}(\widehat{\beta}_{SWk})$ is inflated by

$$\text{VIF}_{SWk} = \frac{\hat{\zeta}_k\hat{\varrho}_k}{1 - R^2_{W\Gamma(k)}} \tag{3.14}$$

times when including other explanatory variables in the model, comparing $v_{0L}(\widehat{\beta}_{SWk})$ in (3.13) to $v_L(\widehat{\beta}_{SWk})$ in (3.12). This assumes that τ is the same in the model that includes only $\mathbf{x}_k$ and the one that includes all $\mathbf{x}$'s.

The term $\hat{\zeta}_k\hat{\varrho}_k \big/ 1 - R^2_{W\Gamma(k)}$ is called the VIF for the kth explanatory variable in the model. The VIF in (3.14) is analogous to ones produced from non-survey data for linear models by packages like SAS® and Stata® and will be called the *no-intercept* VIF here. In the special case of a linear model with homogeneous variance, i.e., $\operatorname{var}_M(\mathbf{y}) = \sigma^2\mathbf{I}$, we have $\mathbf{W}_\Gamma = \mathbf{W}$ and VIF_{SWk} will reduce to the VIF for a linear model in Liao and Valliant (2012b, equation (6)).

The VIF in (3.14) uses a model with only $\mathbf{x}_k$ as the comparison for determining how much the variance of $\widehat{\beta}_{SWk}$ is inflated by nonorthogonality of the x's. An alternative is to compare to a model that has both $\mathbf{x}_k$ and an intercept. The argument for using this as the comparison model is that a typical analyst will always include an intercept in a model. Thus, the appropriate comparison is between a model with an intercept and one x and a model with an intercept and all of the x's. In that case, using (3.12), the linearization variance estimator for $\widehat{\beta}_{SWk}$ can be written as:

$$v_{1L}(\widehat{\beta}_{SWk}) = \hat{\tau}^4 \frac{(\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma (\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)}{(\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k - \widehat{\widehat{N}}\bar{\tilde{x}}_k^2)^2} \tag{3.15}$$

where $\widehat{\widehat{N}} = \sum_{i \in s} w_i \hat{\gamma}_i$ and $\bar{\tilde{x}}_k = \sum_{i \in s} w_i \hat{\gamma}_i x_{ki} / \widehat{\widehat{N}}$.

The variance $v_L(\widehat{\beta}_{SWk})$ in (3.15) can be decomposed into several factors correspondingly:

$$v_L(\widehat{\beta}_{SWk}) = \tau^4 g_{kk}$$

with (3.16)

$$\begin{aligned} g_{kk} &= \frac{\hat{\zeta}_k}{1 - R^2_{W\Gamma m(k)}} \frac{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k - \widehat{\widehat{N}}\bar{\tilde{x}}_k^2}{(\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma (\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)} \times \frac{(\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma (\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)}{(\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k - \widehat{\widehat{N}}\bar{\tilde{x}}_k^2)^2} \\ &= \frac{\hat{\zeta}_k \hat{\varrho}_{km}}{1 - R^2_{W\Gamma m(k)}} v_{1L}(\widehat{\beta}_{SWk}) \end{aligned}$$

where $R^2_{W\Gamma m(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{W\Gamma(k)} \mathbf{X}^T_{(k)} \mathbf{W}_\Gamma \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} - \widehat{\widehat{N}}\bar{\tilde{x}}_k^2}{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k - \widehat{\widehat{N}}\bar{\tilde{x}}_k^2}$ is the coefficient of determination corrected for the mean in the WLS regression when the weight matrix is $\mathbf{W}_\Gamma$, and $\hat{\varrho}_{km} = \frac{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k - \widehat{\widehat{N}}\bar{\tilde{x}}_k^2}{(\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma (\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)}$. The term

$$\text{VIF}_{SWm(k)} = \frac{\hat{\zeta}_k \hat{\varrho}_{km}}{1 - R^2_{W\Gamma m(k)}} \tag{3.17}$$

is called the *intercept-adjusted* VIF for the kth explanatory variable in the model.

Various rules-of-thumb have been proposed as signs of harmful collinearity in linear regression. Marquardt (1970) and Kutner, Nachtsheim and Neter (2004) treat a VIF of 10 or greater as a cutoff for serious collinearity. Chatterjee and Price (1991) suggest a more elaborate two-part rule: (1) the largest VIF is 10 or greater and (2) the mean of all of the VIF's across all predictors is considerably larger than 1. Of course, these rules are somewhat arbitrary and may be adapted, depending on the analysis. If, for example, prediction is the goal, collinearity may be a problem only in some domains of prediction (Snee and Marquardt, 1984). In causal inference, collinearity is a general problem since it may obscure which predictors explain the phenomenon being modeled. Thus, in causal inference a more stringent cutoff may be warranted than when predicting or forecasting.

The design-based approach for fitting GLMs discussed above is one of the typical methods of using the survey weights directly in the estimating equations. Some alternative design-based approaches have also been developed for the fitting of GLMs under informative sampling. Pfeffermann and Sverchkov (2003) considered two other approaches, which require the modeling and estimation of the expectation of the sampling weights, either as a function of the outcome and the explanatory variables, or as a function of only the explanatory variables. The expectations of the sampling weights are used as the weights in the estimating equations. Thus, when these two approaches are used, we can simply treat the expectations of the weights as the $\mathbf{W}$ matrix in (3.14) or (3.17) to obtain the VIFs.

Some common univariate distributions in the exponential family and their special link functions are listed in Table 2.1. Substituting their corresponding elements of matrix $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$ in (3.14) and (3.17), we can obtain their respective VIF formulas. To demonstrate the techniques in a special GLM model, we will take a logistic model as an example in Section 4.

3.5 Condition indexes

Another method of diagnosing collinearity is the use of condition indexes as suggested in Belsley (1984) for linear models and extended to complex survey data by Liao and Valliant (2012a). To detect problematic collinearities in other members of the class of GLMs, Mackinnon and Puterman (1989), Weissfeld and Sereika (1991) and Lesaffre and Marx (1993) modified this approach using a scaled version of the information matrix in a GLM to detect instability of the estimates of $\boldsymbol{\beta}$.

In the survey-weighted generalized linear model, we will use the information matrix evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{SW}$. As shown in (3.4) the estimated information matrix is $\hat{\mathbf{A}} = \mathbf{X}^T \mathbf{W} \hat{\boldsymbol{\Gamma}} \mathbf{X}.$ Denote the weighted matrix $\mathbf{W}^{1/2} \hat{\boldsymbol{\Gamma}}^{1/2} \mathbf{X} \equiv \mathbf{W}_{\Gamma}^{1/2} \mathbf{X}$ as $\widetilde{\mathbf{X}}$. The kth column of $\widetilde{\mathbf{X}}$ is $\tilde{\mathbf{x}}_k$. The singular value decomposition (SVD) of $\widetilde{\mathbf{X}}$ is $\widetilde{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U}$ is an $n \times p$ orthogonal matrix, $\mathbf{V}$ is a $p \times p$ orthogonal matrix, and $\mathbf{D} = \operatorname{diag}(\mu_k)$ is the $p \times p$ diagonal matrix of singular values.

The condition number of $\widetilde{\mathbf{X}}$ is defined as $\kappa(\widetilde{\mathbf{X}}) = \mu_{\max} / \mu_{\min}$, where $\mu_{\max}$ and $\mu_{\min}$ are maximum and minimum singular values of $\widetilde{\mathbf{X}}$. The condition number of $\widetilde{\mathbf{X}}$ is usually different from the condition number of the data matrix $\mathbf{X}$ due to the unequal survey weights and $\hat{\gamma}_i$'s. Condition indexes are defined as

$$\psi_k = \mu_{\max} / \mu_k, \quad k = 1, \ldots, p \tag{3.18}$$

where μ_k is one of the singular values of $\widetilde{\mathbf{X}}$. As noted in Belsley et al. (1980), the size of a condition index depends on the scales of the columns of $\mathbf{X}$. For linear regressions, they recommend scaling the columns of $\mathbf{X}$ to have equal length – a procedure that we adopt here for $\widetilde{\mathbf{X}}$. The kth column of $\widetilde{\mathbf{X}}$ is divided by $\left(\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k\right)^{1/2}$ so that the length of each column is 1. This, in turn, would lead to all condition indexes being 1 when the columns are orthogonal to each other. In the results below, the terms "scaled condition indexes" and "scaled condition numbers" will refer to the condition indexes and condition numbers of the scaled $\widetilde{\mathbf{X}}$.

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

$$\frac{(w \cdot \hat{\gamma})_{\min}^{1/2}}{(w \cdot \hat{\gamma})_{\max}^{1/2}} \kappa(\mathbf{X}) \leq \kappa(\widetilde{\mathbf{X}}) \leq \frac{(w \cdot \hat{\gamma})_{\max}^{1/2}}{(w \cdot \hat{\gamma})_{\min}^{1/2}} \kappa(\mathbf{X}), \tag{3.19}$$

where $(w \cdot \hat{\gamma})_{\min}$ and $(w \cdot \hat{\gamma})_{\max}$ are the minimum and maximum values of $w_i \hat{\gamma}_i$ across different observed units.

Expression (3.19) indicates that if the working weights used in the SWGLM (i.e. $w_i \hat{\gamma}_i$) do not vary too much, the condition number in SWGLM resembles the one in ordinary least squares (OLS). If the design and estimation procedures lead to wide range of survey weights or the $\hat{\gamma}_i$ have a wide range, the condition number can be very different. When SWGLM has large condition number, OLS might not. There are no clear-cut rules for deciding when a condition index is "large". If the estimated information matrix is singular, then at least one singular value will be 0 and a condition index will be infinite. In practice, condition indexes of 5 to 10 are considered to be associated with weak dependencies while any values over 30 signal moderate to strong relationships.

Using the SVD of $\widetilde{\mathbf{X}}$, the approximate model variance in (3.4) is

$$\operatorname{avar}_M\left(\hat{\boldsymbol{\beta}}_{SW}\right) = \tau^2\left[(\mathbf{UDV})^T(\mathbf{UDV})\right]^{-1} = \tau^2\mathbf{VD}^{-2}\mathbf{V}^T. \tag{3.20}$$

The approximate model variance of $\widehat{\beta}_{SWk}$ is the kth diagonal element of (3.20):

$$\operatorname{avar}_M\left(\widehat{\beta}_{SWk}\right) = \tau^2\sum_{j=1}^{p}\frac{v_{kj}^2}{\mu_j^2} \tag{3.21}$$

where $\mathbf{V} = (v_{kj})_{p\times p}$. Let $\phi_{kj} = v_{kj}^2 / \mu_j^2$ and $\phi_k = \sum_{j=1}^{p}\phi_{kj}$. The variance decomposition proportions are $\pi_{jk} = \phi_{jk}/\phi_k$. Note that by setting $\mathbf{W} = \mathbf{I}$, (3.20) becomes the decomposition of the approximate variance for the GLM without survey weights in (2.5).

Since at least one v_{kj} must be nonzero in (3.21), 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 whether a high condition index is 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 $\mathbf{X}$ involved to make a near dependency. Liao and Valliant (2012a), following Belsley et al. (1980), suggested showing a matrix of decomposition proportions and condition indexes of $\widetilde{\mathbf{X}}$ in a variance decomposition table – a path we follow in the empirical example in Section 4. Note that the approximate, model-design variance in (3.8) could also be decomposed and variance proportions computed analogously to the above steps, but we will not pursue that alternative here.

4. Empirical example

We illustrate the foregoing techniques and investigate their behavior using dietary intake data from the US National Health and Nutrition Examination Survey (NHANES). Two cycles of NHANES, 2015-2016 and 2017-2018, were combined to increase the sample size for analysis as suggested in Johnson, Paulose-Ram and Ogden (2012). 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 (Curtin, Mohadjer and Dohrmann, 2012). Oversampling of certain population subgroups is done to increase the reliability and precision of health status estimates for these groups. Response rates for various years of NHANES are at https://wwwn.cdc.gov/nchs/nhanes/ResponseRates.aspx. The survey weights were constructed by taking initial sample weights and further adjusting for the household nonresponse, nonresponse to the dietary questions, and the differential allocation by day of the week for the dietary intake data collection. Weights were also poststratified to population totals for a set of demographic groups to account for under- and overcoverage. Fields available in the public use files do not permit these estimation complexities to be accounted for fully when estimating variances. The data set used in our study is a subset of 2015-2018 data composed of respondents aged between 18 and 65. Observations with missing values in the selected variables are excluded from the sample which finally contains 6,682 complete respondents. The final weights in our sample range from 1,358 to 345,570, with a ratio of 254:1. Following the NHANES documentation, we treat the sample as a stratified selection of 60 PSUs from 30 strata, with 2 PSUs sampled with replacement within each stratum.

For this empirical study, a binary logistic model is considered. The explanatory variables include a categorical variable for race-ethnicity (Black, White, and Other with White being the reference group), gender (male = 1, female = 0), age, the interaction between gender and age, and nine daily total nutrition intake variables – calorie(kcal), protein(gm), carbohydrate(gm), sugar(gm), total fat(gm), total saturated fatty acids(gm), total monounsaturated fatty acids(gm), total polyunsaturated fatty acids(gm) and alcohol(gm). Logistic regression models are fit with obese/non-obese as a binary response variable. Obesity is defined as the respondent's Body Mass Index (BMI) being 30 or higher. The food intake variables are correlated, so collinearity problems are expected when all are used together as predictors of obesity.

Three regression methods are applied and compared in this study. Each method uses maximum likelihood estimation implemented via iterative reweighted least squares but differ in how survey features are handled. The first is denoted *generalized least squares* (GLS) but ignores all sampling complexities including stratification, clustering, and weighting. The second one is referred to as *generalized weighted least squares* (GWLS), which incorporates the survey weights but assumes persons are independent and ignores strata and clustering. Both GLS and GWLS are misspecified in the sense of ignoring any over-dispersion caused by clustering. Omission of important predictors is another type of misspecification but is not considered in this example. The third is the new method, described in Section 3.1, that uses the actual complex sampling design and is denoted in tables as SWGLM. The formulas for the weight matrices, coefficient variance estimators and collinearity diagnostics of these three methods are listed in Table 4.1.

The coefficient and standard error (SE) estimates from the R `survey` package (Lumley, 2024) for the fitted models using the three different regression methods are presented in Table 4.2. Note the SEs for GLS and GWLS have a model-based interpretation (assuming the model is correctly specified). The SWGLM SEs have both a design-based and model-based interpretation. The *full* models with all the explanatory variables are in the upper tier of the table. Then, a *reduced* model with fewer near-dependency problems is fitted with one dummy variable for race (Black with all other races as the reference group), age, total saturated fatty acids, total monounsaturated fatty acids, and alcohol and is shown in the lower tier. The reduced model is not an attempt to find the best model for predicting obesity but, rather, is used to illustrate the effect that collinearity can have on the size of estimated coefficients and standard errors (SEs).

Table 4.1
Regression models and their collinearity diagnostic statistics used in the experimental study

Model	Y	Regression Type	$\mathbf{W}_\Gamma$	Variance Estimation of $\hat{\boldsymbol{\beta}}$	VIF fomula
Logistic	Obese	GLS	$\hat{\boldsymbol{\Gamma}}$ [a]	$(\mathbf{X}^T\hat{\boldsymbol{\Gamma}}\mathbf{X})^{-1}$ [b]	$\mathrm{VIF} = \dfrac{1}{1-R^2_{\Gamma(k)}}$
(Binary)	/Non-obese	GWLS	$\mathbf{W}\hat{\boldsymbol{\Gamma}}$ [c]	$(\mathbf{X}^T\mathbf{W}_\Gamma\mathbf{X})^{-1}$	$\mathrm{VIF} = \dfrac{1}{1-R^2_{W\Gamma(k)}}$
		(model-based) SWGLM	$\mathbf{W}\hat{\boldsymbol{\Gamma}}$	$(\mathbf{X}^T\mathbf{W}_\Gamma\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_\Gamma\mathbf{V}\boldsymbol{\Delta}\mathbf{W}_\Gamma\mathbf{X}(\mathbf{X}^T\mathbf{W}_\Gamma\mathbf{X})^{-1}$	$\mathrm{VIF} = \dfrac{\hat{\zeta}_k\hat{\varrho}_k}{1-R^2_{W\Gamma(k)}}$ with
		(design-based)		$\mathbf{V}_\Delta = \hat{\boldsymbol{\Delta}}\hat{\mathbf{V}}\hat{\boldsymbol{\Delta}}$ [d] $= \hat{\boldsymbol{\Delta}}\left\{\sum_{h=1}^{H}\dfrac{n_h}{n_h-1}\left[\mathrm{Blkdiag}\left(\mathbf{e}_{hi}\mathbf{e}_{hi}^T\right)-\dfrac{1}{n_h}\mathbf{e}_h\mathbf{e}_h^T\right]\right\}\hat{\boldsymbol{\Delta}}$	$\hat{\zeta}_k = \dfrac{\mathbf{e}_{xk}^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma\mathbf{e}_{xk}}{\mathbf{e}_{xk}^T\mathbf{W}_\Gamma\mathbf{e}_{xk}}$, $\hat{\varrho}_k = \dfrac{\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{x}_k}{\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma\mathbf{x}_k}$

Notes: [a] $\hat{\boldsymbol{\Gamma}} = \mathrm{diag}[\hat{p}_i(1-\hat{p}_i)]$.
[b] $\tau^2 = 1$ in logistic model.
[c] Diagonal matrix with survey weights $w_i\gamma_i$ on the main diagonal.
[d] $\hat{\boldsymbol{\Delta}} = \mathrm{diag}\{[\hat{p}_i(1-\hat{p}_i)]^{-1}\}$.
- GLS = generalized least squares; GWLS = generalized weighted least squares; SWGLM = survey-weighted generalized linear models; VIF = variance inflation factor.

Table 4.2
Parameter estimates with their associated standard errors (in parentheses) for the full and reduced models using three different regression methods

Variable	GLS		GWLS		SWGLM	
Full model						
(Intercept)	-0.887	**	-1.043	*	-1.043	*
	(0.297)		(0.456)		(0.470)	
Other race	-0.125	*	0.111		0.111	
	(0.060)		(0.082)		(0.108)	*
Black	0.276	***	0.396	***	0.396	**
	(0.069)		(0.088)		(0.108)	
Gender	0.080		0.049		0.049	
	(0.173)		(0.256)		(0.250)	
Age in years	0.002		0.012		0.012	
	(0.006)		(0.010)		(0.009)	
Gender*Age	0.005		0.001		0.001	
	(0.004)		(0.006)		(0.005)	
Calories	0.006	***	0.007	***	0.007	***
	(0.001)		(0.002)		(0.002)	
Protein	-0.027	***	-0.031	***	-0.031	***
	(0.005)		(0.008)		(0.007)	
Carbohydrates	-0.026	***	-0.029	***	-0.029	***
	(0.005)		(0.007)		(0.006)	
Sugar	0.001	*	0.001		0.001	
	(0.001)		(0.001)		(0.001)	
Total fat	0.007		0.023		0.023	
	(0.016)		(0.024)		(0.024)	
Total saturated fatty acids	-0.057	***	-0.075	***	-0.075	**
	(0.014)		(0.021)		(0.023)	
Total monounsaturated fatty acids	-0.068	***	-0.101	***	-0.101	***
	(0.013)		(0.020)		(0.021)	
Total polyunsaturated fatty acids	-0.060	***	-0.086	***	-0.086	**
	(0.013)		(0.019)		(0.023)	
Alcohol	-0.045	***	-0.052	***	-0.052	***
	(0.008)		(0.013)		(0.011)	

Notes: - p-values: *** $0 < p \le 0.001$; ** $0.001 < p \le 0.01$; * $0.01 < p \le 0.05$.
- GLS = generalized least squares; GWLS = generalized weighted least squares; SWGLM = survey-weighted generalized linear models.

Table 4.2 (continued)
Parameter estimates with their associated standard errors (in parentheses) for the full and reduced models using three different regression methods

Variable	GLS		GWLS		SWGLM	
Reduced model						
(Intercept)	-0.855	***	-0.966	***	-0.966	***
	(0.099)		(0.149)		(0.187)	
Black	0.407	***	0.408	***	0.408	***
	(0.059)		(0.077)		(0.091)	
Age in years	0.008	***	0.012	***	0.012	**
	(0.002)		(0.003)		(0.004)	
Total saturated fatty acids	0.013	***	0.022	***	0.022	***
	(0.003)		(0.004)		(0.005)	
Total monounsaturated fatty acids	-0.008	**	-0.017	***	-0.017	**
	(0.003)		(0.004)		(0.005)	
Alcohol	-0.004	***	-0.004	*	-0.004	*
	(0.001)		(0.002)		(0.001)	

Notes: - p -values: *** $0 < p \leq 0.001$; ** $0.001 < p \leq 0.01$; * $0.01 < p \leq 0.05$.
- GLS = generalized least squares; GWLS = generalized weighted least squares; SWGLM = survey-weighted generalized linear models.

Table 4.3 lists the VIFs computed from (3.7) and (3.14). VIFs labeled MB VIF are the model-based ones in (3.7) for regressions that use the survey weights but do not account for stratification and clustering. VIFs labeled GLS, GWLS, and SWGLM are from (3.14). Table 4.4 lists the intercept-adjusted VIFs computed from (3.17) for the full and reduced models. All VIFs were computed using the R package `svydiags` (Valliant, 2024). GLS VIFs are computed as if a simple random sample was used, i.e., survey weights, strata, and clusters were not used. GWLS incorporates only survey weights while SWGLM accounts for all design features.

Table 4.3
No-intercept VIF values for logistic regression model using three different estimation methods, MB VIF values are from (3.7), GLS GWLS, and SWGLM values are from (3.14)

Variable	MB VIF (3.7)	GLS VIF (3.14)	GWLS VIF (3.14)	SWGLM VIF (3.14)	SWGLM $\zeta_k \hat{\varrho}_k$ (3.14)
Full model					
Other race	1.5	2.6	2.3	5.3	3.42
Black	1.2	1.8	1.8	3.3	2.67
Gender	118.0	123.8	114.8	106.3	0.90
Age in years	115.9	119.9	113.4	113.4	0.98
Gender*Age	117.1	120.0	110.5	107.7	0.92
Calories	11,469.7	11,776.7	12,405.7	19,545.3	1.70
Protein	320.6	327.3	337.0	511.9	1.60
Carbohydrates	2,494.7	2,630.2	2,728.0	4,174.7	1.67
Sugar	12.0	11.8	11.5	16.2	1.35
Total fat	4,187.4	3,833.8	3,934.1	6,005.3	1.43
Total saturated fatty acids	348.6	311.2	323.4	380.3	1.09
Total monounsaturated fatty acids	346.0	322.3	333.2	427.1	1.23
Total polyunsaturated fatty acids	162.8	148.3	144.1	162.5	1.00
Alcohol	76.9	78.8	78.0	109.8	1.43
Reduced model					
Black	1.1	1.3	1.5	2.1	1.86
Age in years	11.4	11.6	10.7	13.9	1.22
Total saturated fatty acids	12.2	11.8	12.3	20.3	1.66
Total monounsaturated fatty acids	12.9	12.6	12.9	21.5	1.67
Alcohol	1.2	1.2	1.2	1.1	0.93

Note: GLS = generalized least squares; GWLS = generalized weighted least squares; MB = model-based; SWGLM = survey-weighted generalized linear models; VIF = variance inflation factor.

Table 4.4
Intercept-adjusted VIF values from (3.17) for logistic regression model using three different estimation methods

Variable	GLS	GWLS	SWGLM	
	VIF	VIF	VIF	$\zeta_k \hat{\varrho}_k$
Full model				
Other race	1.4	1.3	1.4	1.27
Black	1.4	1.3	1.5	1.41
Gender	11.7	10.8	8.7	0.76
Age in years	10.4	10.7	8.3	0.80
Gender*Age	20.3	19.3	16.1	0.77
Calories	2,140.7	2,225.3	3,390.9	1.77
Protein	73.2	76.2	117.3	1.75
Carbohydrates	529.9	539.7	754.0	1.52
Sugar	4.0	3.8	5.1	1.26
Total fat	956.3	980.8	1,334.2	1.39
Total saturated fatty acids	90.9	92.9	98.1	1.06
Total monounsaturated fatty acids	87.4	89.7	99.3	1.14
Total polyunsaturated fatty acids	46.1	44.2	41.8	0.86
Alcohol	69.0	67.7	100.1	1.51
Reduced model				
Black	1.0	1.0	1.0	0.99
Age in years	1.0	1.0	1.0	0.99
Total saturated fatty acids	3.4	3.5	5.1	1.57
Total monounsaturated fatty acids	3.4	3.5	4.8	1.48
Alcohol	1.0	1.0	1.0	1.01

Note: GLS = generalized least squares; GWLS = generalized weighted least squares; SWGLM = survey-weighted generalized linear models; VIF = variance inflation factor.

One rule-of-thumb cited earlier is that a no-intercept VIF of 10 or higher signals harmful collinearity (Kutner et al., 2004, Marquardt, 1970). Some of the VIFs are extraordinarily large in the full model. The intercept-adjusted SWGLM VIFs for calories, carbohydrates, and total fat are 3,391, 754, and 1,334, respectively in Table 4.4. The no-intercept VIFs in Table 4.3 for calories, carbohydrates, and total fat are even higher at 19,545, 4,175, and 6,005. The no-intercept VIFs in Table 4.3 are always larger than the intercept-adjusted VIFs in Table 4.4.

As shown in Table 4.1, the VIF in SWGLM can be obtained by multiplying the VIF from GWLS by the adjustment coefficient $\zeta_k \varrho_k$. For some coefficients, e.g., calories in the full model with $\zeta_k \varrho_k = 1.77$ in Table 4.4, this adjustment can be substantial, implying that the GWLS VIF would be too small. The SWGLM VIFs are often larger than those for MB, GLS, and GWLS, although if one is large, all are large. MB, GLS, and GWLS all convey about the same diagnostic information in this example since all three have similar sizes.

Collinearity can lead to anomalies in some coefficients. For example, in the full model the sign for total saturated fatty acids is negative and the coefficient estimate is highly significant with GLS, GWLS, and SWGLM even though high consumption of saturated fats is known to be positively associated with obesity (see, e.g. Phillips, Kesse-Guyot, McManus, Hercberg, Lairon, Planells and Roche, 2012).

Table 4.5 shows the condition indexes and variance decomposition proportions for the full SWGLM model. If all predictors in the model were orthogonal, the condition indexes would all equal 1, but this is far from the case for the full model where the largest condition index is 407.2. Proportions that are less than or equal to 0.01 are shown as dots to improve readability. The second through last columns each sum (approximately) to 1. The proportions are used to find collinearities by first locating the rows with "large" condition indexes then reading down the columns for predictors and looking for two or more proportions in the same row that are sizeable. For example, for carbohydrates, total saturated fatty acids, and total monounsaturated fatty acids, proportions of 0.21, 0.18, and 0.35 are associated with a condition index of 18.7. Thus, those three predictors are somewhat collinear. On the other hand, for the predictor, alcohol, 0.82 of its variance is associated with a condition index of 3.4, but no other predictor has a high variance proportion associated with that condition index. Thus, alcohol does not appear to be harmfully collinear with other covariates even though its VIFs are 109.8 and 100.1 in Tables 4.3 and 4.4.

How to form a reduced model by deleting collinear predictors is hardly a unique process and different analysts may make different decisions. For this example, we dropped most of the variables with the largest VIFs and recoded race-ethnicity to have two categories: Black and All Other race-ethnicities with the latter used as the reference. In the reduced model all intercept-adjusted SWGLM VIFs in Table 4.4 are less than 10 and the sign on total saturated fatty acids is positive. The lower tier of Table 4.5 shows the condition indexes and variance decomposition proportions for the reduced model. The largest condition index is 8.4 – another indicator of limited collinearity. The absence of the most correlated variables in the reduced model makes the SEs of estimated coefficients substantially smaller for the retained predictors. For example, in the full model the SE in Table 4.2 of age in SWGLM is 0.009 while it is 0.004 in the reduced model; its coefficient estimate is 0.012 in both the full and reduced models. However, collinearity affects the size of other coefficients. The SWGLM estimate for alcohol consumption is -0.052 in Table 4.2 in the full model but -0.004 in the reduced model. Because the SE drops from 0.011 in the full to 0.001 in the reduced, the alcohol coefficient stays significantly different from zero in the reduced model.

The fact that, in the reduced model, total saturated fat and total monounsaturated fat have intercept-adjusted SWGLM VIFs of 5.1 and 4.8 and no-intercept VIFs of 20.3 and 21.5 implies that the model might be further refined to reduce collinearity. The unweighted correlation of these two variables is 0.84, confirming that the two are somewhat redundant. We do not pursue the possibility of further model refinement here.

Table 4.5
Scaled condition indexes from (3.18) and variance decomposition proportions from SWGLM

Scaled condition index	Other race	Black	Gender	Age in years	Gender * Age	Calories	Protein	Carbo-hydrates	Sugar	Total fat	Total saturated fatty acids	Total mono-unsaturated fatty acids	Total poly-unsaturated fatty acids	Alcohol
Full model														
1.0	0.03	.	0.08	0.08	0.07	0.10	0.09	0.09	0.08	0.09	0.09	0.09	0.09	0.02
3.2	0.30	0.69	.	.	.	.	.	.	.	.	.	.	.	.
3.4	0.05	0.04	0.03	.	0.03	.	.	.	.	.	.	.	.	0.82
3.7	0.23	0.07	0.10	0.06	0.13	0.02	.	.	0.02	0.05	0.06	0.05	0.04	0.15
4.3	0.38	0.18	0.10	0.09	0.21	.	.	.	.	.	.	.	.	.
5.5	.	.	0.00	.	.	.	0.02	0.17	0.58	0.05	0.03	0.06	0.07	.
8.7	.	.	.	.	.	.	0.29	.	.	.	0.12	.	0.55	.
10.2	.	.	0.09	0.11	.	.	0.35	.	.	0.03	0.28	0.03	0.08	.
11.0	.	.	0.40	0.38	.	.	0.13	.	.	.	0.03	0.02	.	.
17.3	.	.	.	.	.	0.05	0.04	0.31	0.19	.	0.09	0.29	.	.
18.7	.	.	.	.	.	0.03	0.04	0.21	0.10	.	0.18	0.35	0.09	.
24.8	.	.	0.19	0.25	0.51	.	.	0.03	.	.	.	.	.	.
170.1	.	.	.	.	.	0.10	.	0.02	.	0.61	0.11	0.11	0.05	.
407.2	.	.	.	.	.	0.68	0.02	0.15	.	0.15	.	.	.	.

Notes: - Proportions that are less than or equal to 0.01 are shown as dots.
- SWGLM = survey-weighted generalized linear models.

Table 4.5 (continued)
Scaled condition indexes from (3.18) and variance decomposition proportions from SWGLM

Scaled condition index	Other race	Black	Gender	Age in years	Gender * Age	Calories	Protein	Carbo-hydrates	Sugar	Total fat	Total saturated fatty acids	Total monounsaturated fatty acids	Total polyunsaturated fatty acids	Alcohol
Reduced model														
	Black	**Age in years**	**Total fat**	**Total monounsaturated fatty acids**	**Alcohol**									
1.0	0.06	0.27	0.29	0.30	0.08									
1.8	0.77	.	0.01	0.00	0.23									
2.0	0.17	0.03	0.06	0.05	0.69									
3.7	0.00	0.71	0.16	0.13	0.00									
8.4	.	0.00	0.48	0.52	.									

Notes: - Proportions that are less than or equal to 0.01 are shown as dots.
- SWGLM = survey-weighted generalized linear models.

The example presented here is intended to simply illustrate the VIFs and condition indexes using a real survey dataset. Different analysts are likely to vary in how they arrive at a final model. In the example above, an analyst might be specifically interested in the relation to obesity of calories, protein, and carbohydrates. Thus, those covariates might be added one at a time in a reduced model rather than eliminating them and including only total saturated and monounsaturated fatty acids. There are also subtleties in survey data analysis that we do not attempt to cover here. In practice, accounting for the interplay of the regressors with the design variables may be important in formulating a model (e.g., see Korn and Graubard, 1999; Little, 2004). For example, population density used as a regressor may be highly correlated with strata membership in a design that stratifies by urban/rural. A regressor may be strongly correlated with weights (e.g. past years' revenue in a probability-proportional-to-size establishment survey design). In other designs, the relation between survey weights and potential predictors may be negligible. Specific applications should receive their own detailed consideration about how best to do the modeling.

5. Conclusion

As for ordinary linear regression models, generalized linear models (GLMs) can be affected by collinearity of the predictors. The literature on collinearity in linear models approaches the problem in either of two ways: calculation of variance inflation factors (VIFs) or of condition indexes for the matrix of predictors. Previous literature on GLMs mainly covers only condition indexes. In this paper we developed new formulations for VIFs appropriate for GLMs fitted from either survey or non-survey data. A VIF measures the amount by which the variance of a parameter estimator is inflated due to predictor variables being correlated with each other, rather than being orthogonal. The GLM VIF for non-survey data depends on a type of R^2 from a weighted regression of one x-predictor on the other x's with the weight depending on the model variance parameter and the derivative of the link function. For survey data, the weight in the regression additionally depends on the survey weight.

The conventional VIF method usually uses a model with a single $\mathbf{x}_k$ as the comparison for determining how much the variance of an estimator of the slope, $\boldsymbol{\beta}_k$, in the comparison model can be inflated by nonorthogonality of the $\mathbf{x}$'s. But, in practice, researchers may want to use a model with both $\mathbf{x}_k$ and an intercept as the comparison model to examine potential collinearity issues when adding additional predictors in the model. We also derived intercept-adjusted VIF formulas to fit this purpose.

In addition, we developed theory for condition indexes related to the weighted information matrix associated with a GLM estimated from complex survey data. The information matrix is the main ingredient in the approximate model variance of the estimated regression slope in a survey-weighted GLM. The condition indexes supplement VIFs by helping to identify which pairs of predictors are leading to collinearity.

The theory for the VIFs and condition indexes was illustrated with a logistic regression predicting obesity using a dataset containing a number of correlated variables related to caloric-intake. As anticipated, collinearity can affect both the sizes of estimated coefficients and their standard errors. R functions for computing VIFs and condition indexes for either linear models or GLMs are in the R package `svydiags` (Valliant, 2024).

The methods presented here do have limitations and will not be equally useful in all applications. If prediction is the main goal, inclusion of collinear predictors may not be a concern as long as prediction error is minimized. In contrast is causal inference where the relative sizes of estimated coefficients may be critically important. A model with interactions is an interesting case since it will naturally have some degree of collinearity due to an interaction being correlated with the main effects that go into it. To decide whether to keep an interaction term with a high VIF, an analyst can consider the following:

(1) Retain the interaction if it is statistically significant, has a large effect size, or improves model fit and predictive performance.

(2) Consider removing it if it is non-significant, has minimal impact on the outcome, or if multicollinearity severely affects model interpretation.

(3) Regularization may also be used if an analyst feels that interactions may be important and having good predictions overrides concerns about multicollinearity.

Note that some fitting methods like random forests that use model averaging do not lend themselves to VIF or condition index analyses since all the models in the average do not necessarily include the same covariates or their interactions.

Misspecification is always a concern when fitting models, particularly with survey data. In such cases, the estimated model will be biased to some extent, but the analyst is still using a model that fits the data best given the form of model being used (e.g., linear, logistic, etc.) and the set of covariates that are input. When important covariates are omitted because they are either overlooked or unavailable, VIFs and condition indexes will still identify $\mathbf{x}$'s that are collinear with others in a model. Examination of VIFs and other diagnostics like residuals can sharpen the model being used even in the face of misspecification. Although the collinearity diagnostics have their limitations, they do serve a useful role in model refinement.

Potential future work in this area would be to extend the theory for condition indexes in GLMs to model-design variance estimators, extend the VIF theory to cover mixed effects superpopulation models, investigate how VIFs and other model diagnostics can be applied to study properties of estimators of descriptive statistics like totals and means, and, where necessary, to develop additional software to make these tools easily accessible to analysts.

Acknowledgements

We thank the Editor, Associate Editor, and referees for their comments which improved the presentation.

Appendix A

A.1 Partitioning the $\mathbf{A}$ matrix in GLM and SWGLM

Define $\widetilde{\mathbf{X}} = (\mathbf{W}\boldsymbol{\Gamma})^{1/2}\mathbf{X} = \mathbf{W}_{\Gamma}^{1/2}\mathbf{X}$ and $\tilde{\mathbf{x}}_k = \mathbf{W}_{\Gamma}^{1/2}\mathbf{x}_k$. The matrix $\mathbf{A} = \widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}$ in (3.9) can be partitioned as

$$\mathbf{A} = \begin{pmatrix} \tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k & \tilde{\mathbf{x}}_k^T\widetilde{\mathbf{X}}_{(k)} \\ \widetilde{\mathbf{X}}_{(k)}^T\tilde{\mathbf{x}}_k & \widetilde{\mathbf{X}}_{(k)}^T\widetilde{\mathbf{X}}_{(k)} \end{pmatrix} \tag{A.1}$$

after reordering the columns of $\widetilde{\mathbf{X}}$ so that $\widetilde{\mathbf{X}} = \left(\tilde{\mathbf{x}}_k \;\; \widetilde{\mathbf{X}}_{(k)}\right)$ with $\widetilde{\mathbf{X}}_{(k)}$ being the $n \times (p-1)$ matrix containing all columns except the kth column of $\widetilde{\mathbf{X}}$. Using the formula for the inverse of a partitioned matrix (see, e.g. Theil, 1971, Section 1.2) and setting $\mathbf{W} = \mathbf{I}_N$ leads directly to (2.8). For the survey-weighted version, we use the more general form in (3.6).

Derivation of g_{kk} in SWGLM, Section 3.4

To derive g_{kk}, we partition the components of $\mathbf{G} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ as

$$\mathbf{G} = \begin{pmatrix} a^{kk} & \mathbf{a}^{k(k)} \\ \mathbf{a}^{(k)k} & \mathbf{A}^{(k)(k)} \end{pmatrix} \begin{pmatrix} b_{kk} & \mathbf{b}_{k(k)} \\ \mathbf{b}_{(k)k} & \mathbf{B}_{(k)(k)} \end{pmatrix} \begin{pmatrix} a^{kk} & \mathbf{a}^{k(k)} \\ \mathbf{a}^{(k)k} & \mathbf{A}^{(k)(k)} \end{pmatrix}$$

where $\mathbf{A}^{-1} = [a^{lk}]$ $(l, k = 1, \ldots, p)$, $\mathbf{a}^{k(k)}$ is the kth row of $\mathbf{A}^{-1}$ excluding a^{kk}, $\mathbf{a}^{(k)k} = [\mathbf{a}^{k(k)}]^T$, and $\mathbf{A}^{(k)(k)}$ is the $(p-1)\times(p-1)$ part of $\mathbf{A}^{-1}$ excluding the kth row and column. Using the formula for the inverse of a partitioned matrix, the upper left-hand and lower left-hand components of $\mathbf{A}^{-1}$ are:

$$a^{kk} = \mathbf{i}_k^T\mathbf{A}^{-1}\mathbf{i}_k = \mathbf{i}_k^T\left(\mathbf{X}^T\mathbf{W}_{\Gamma}\mathbf{X}\right)^{-1}\mathbf{i}_k = \frac{1}{\left(1 - R_{\Gamma W(k)}^2\right)\mathbf{x}_k^T\mathbf{W}_{\Gamma}\mathbf{x}_k} \tag{A.2}$$

$$\mathbf{a}^{(k)k} = -a^{kk}\left(\mathbf{X}_{(k)}^T\mathbf{W}_{\Gamma}\mathbf{X}_{(k)}\right)^{-1}\mathbf{X}_{(k)}\mathbf{W}_{\Gamma}\mathbf{x}_k \equiv -a^{kk}\hat{\boldsymbol{\beta}}_{W\Gamma(k)} \tag{A.3}$$

where $R^2_{\Gamma W(k)}$ was defined in Section 3.4. The partitioned matrix $\mathbf{B}$ is:

$$\mathbf{B} = \begin{pmatrix} b_{kk} & \mathbf{b}_{k(k)} \\ \mathbf{b}_{(k)k} & \mathbf{B}_{(k)(k)} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_k & \mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{X}_{(k)} \\ \mathbf{X}_{(k)}^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_k & \mathbf{X}_{(k)}^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{X}_{(k)} \end{pmatrix}. \tag{A.4}$$

Multiplying the matrices in $\mathbf{G}$ and using the symmetry of $\mathbf{A}$ and $\mathbf{B}$ gives

$$\begin{aligned} g_{kk} &= a^{kk}(a^{kk} b_{kk} + 2\mathbf{b}_{k(k)} \mathbf{a}^{(k)k}) + \mathbf{a}^{(k)kT} \mathbf{B}_{(k)(k)} \mathbf{a}^{(k)k} \\ &= (a^{kk})^2 (b_{kk} - 2\mathbf{b}_{k(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} + \hat{\boldsymbol{\beta}}^T_{W\Gamma(k)} \mathbf{B}_{(k)(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)}). \end{aligned}$$

Using (A.2)-(A.4), and steps analogous to the derivation of VIF for linear regression in the appendix of Liao and Valliant (2012b), g_{kk} can be decomposed into a term with $R^2_{W\Gamma(k)}$ and two adjustment coefficients:

$$\begin{aligned} g_{kk} &= \left(\frac{1}{1 - R^2_{W\Gamma(k)}} \frac{1}{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k} \right)^2 \times \\ &\quad \left(\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_k - 2\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} + \hat{\boldsymbol{\beta}}^T_{W\Gamma(k)} \mathbf{X}_{(k)} \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} \right) \\ &= \frac{1}{1 - R^2_{W\Gamma(k)}} \frac{\left(\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} \right)^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \left(\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} \right)}{\left(\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} \right)^T \mathbf{W}_\Gamma \left(\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} \right)} \frac{1}{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k} \\ &= \frac{1}{1 - R^2_{W\Gamma(k)}} \frac{\mathbf{e}_{xk}^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{e}_{xk}}{\mathbf{e}_{xk}^T \mathbf{W}_\Gamma \mathbf{e}_{xk}} \frac{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k}{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_k} \frac{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_k}{\left(\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k \right)^2} \\ &= \frac{\hat{\zeta}_k \hat{\varrho}_k}{1 - R^2_{W\Gamma(k)}} \frac{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_k}{\left(\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k \right)^2} \end{aligned} \tag{A.5}$$

where $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)}$ is the residual from the survey-weighted regression of $\mathbf{x}_k$ on $\mathbf{X}_{(k)}$ using weight matrix $\mathbf{W}_\Gamma$, $\hat{\zeta}_k = \frac{(\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)})^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma (\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)})}{(\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)})^T \mathbf{W}_\Gamma (\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)})} = \frac{\mathbf{e}_{xk}^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{e}_{xk}}{\mathbf{e}_{xk}^T \mathbf{W}_\Gamma \mathbf{e}_{xk}}$, and $\hat{\varrho}_k = \frac{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k}{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_k}$.

A.2 Variance estimator in a stratified, multistage sample

Suppose that the design is the one specified in Section 3.1. Consider a GLM with a common intercept and slope vector across strata. The design-based estimating equations in a sample are:

$$\sum_{h=1}^{H} \sum_{i \in s_h} \sum_{t \in s_{hi}} w_{hit} (y_{hit} - \mu_{hit}) \hat{\gamma}_{hit} g'(\mu_{hit}) \mathbf{x}_{hit} = \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \hat{\boldsymbol{\Gamma}}_{hi} \hat{\boldsymbol{\Delta}}_{hi} (\mathbf{y}_{hi} - \boldsymbol{\mu}_{hi}) = \mathbf{0}; \tag{A.6}$$

where $\mathbf{y}_{hi} = (y_{hi1}, \ldots, y_{hin_{hi}})^T$, $\mu_{hi} = (\mu_{hi1}, \ldots, \mu_{hin_{hi}})^T$, $\mathbf{W}_{hi} = \operatorname{diag}(w_{hi1}, \ldots, w_{hin_{hi}})$, $\hat{\boldsymbol{\Gamma}}_{hi} = \operatorname{diag}(\gamma_{hi1}, \ldots, \gamma_{hin_i})$, $\hat{\boldsymbol{\Delta}}_{hi} = \operatorname{diag}(g'(\mu_{hi1}), \ldots, g'(\mu_{hin_i}))$ and $\mathbf{X}_{hi}(n_{hi} \times p) = (\mathbf{x}_{1hi}, \ldots, \mathbf{x}_{phi})$ with $\mathbf{x}_{khi} = (x_{khi1}, \ldots, x_{khin_{hi}})^T$.

Let $\mathbf{z}_{hi} = \mathbf{X}_{hi}^T \hat{\boldsymbol{\Gamma}}_{hi} \hat{\boldsymbol{\Delta}}_{hi} (\mathbf{y}_{hi} - \boldsymbol{\mu}_{hi})$. Define a weighted vector of residuals for cluster i as $\mathbf{z}^*_{hi} = \mathbf{X}_{hi}^T \mathbf{W}_{hi} \hat{\boldsymbol{\Gamma}}_{hi} \hat{\boldsymbol{\Delta}}_{hi} (\mathbf{y}_{hi} - \boldsymbol{\mu}_{hi})$. The design consistent variance estimator for $\mathbf{z}^*$ in this stratified multistage sampling design is:

$$\begin{aligned} v_\pi(\hat{\mathbf{z}}^*) &= \sum_{h=1}^{H} \frac{m_h}{m_h - 1} \sum_{i \in s_h} \left(\mathbf{z}_{hi}^* - \overline{\mathbf{z}}_h^*\right)\left(\mathbf{z}_{hi}^* - \overline{\mathbf{z}}_h^*\right)^T \\ &= \sum_{h=1}^{H} \frac{m_h}{m_h - 1} \left(\sum_{i \in s_h} \mathbf{z}_{hi}^* \mathbf{z}_{hi}^{*T} - m_h \overline{\mathbf{z}}_h^* \overline{\mathbf{z}}_h^{*T} \right) \\ &= \sum_{h=1}^{H} \frac{m_h}{m_h - 1} \mathbf{X}_h^T \mathbf{W}_h \hat{\mathbf{\Gamma}}_h \hat{\mathbf{\Delta}}_h \left[\text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^T) - \frac{1}{m_h} \mathbf{e}_h \mathbf{e}_h^T \right] \hat{\mathbf{\Delta}}_h \hat{\mathbf{\Gamma}}_h \mathbf{W}_h \mathbf{X}_h, \end{aligned} \tag{A.7}$$

where $\mathbf{X}_h^T = (\mathbf{X}_{h1}, \ldots, \mathbf{X}_{hm_h})^T$, $\mathbf{W}_h = \text{diag}(\mathbf{W}_{hi})_{i \in s_h}$, $\hat{\mathbf{\Gamma}}_h = \text{diag}(\hat{\mathbf{\Gamma}}_{hi})_{i \in s_h}$, $\hat{\mathbf{\Delta}}_h = \text{diag}(\hat{\mathbf{\Delta}}_{hi})_{i \in s_h}$, $\mathbf{e}_{hi} = \mathbf{y}_{hi} - \boldsymbol{\mu}_{hi}$ and $\mathbf{e}_h = (\mathbf{e}_{h1}, \mathbf{e}_{h2}, \ldots, \mathbf{e}_{hm_h})^T$ is a vector of element residuals in stratum h, and $\text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^T)$ is the $m \times m$ matrix with $\mathbf{e}_{hi}\mathbf{e}_{hi}^T$ on the main diagonal position and 0 elsewhere.

Correspondingly, the linearization variance estimator $v_L(\hat{\boldsymbol{\beta}}_{SW})$ is:

$$\begin{aligned} v_L(\hat{\boldsymbol{\beta}}_{SW}) &= \hat{\tau}^4 \sum_{h=1}^{H} \mathbf{A}^{-1} \frac{m_h}{m_h - 1} \mathbf{X}_h^T \mathbf{W}_h \hat{\mathbf{\Gamma}}_h \hat{\mathbf{\Delta}}_h \left[\text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^T) - \frac{1}{m_h} \mathbf{e}_h \mathbf{e}_h^T \right] \times \\ &\quad \hat{\mathbf{\Delta}}_h \hat{\mathbf{\Gamma}}_h \mathbf{W}_h \mathbf{X}_h \mathbf{A}^{-1} \\ &= \hat{\tau}^4 \mathbf{A}^{-1} \mathbf{X}^T \mathbf{W} \hat{\mathbf{\Gamma}} \hat{\mathbf{\Delta}} \, \text{Blkdiag} \left\{ \frac{m_h}{m_h - 1} \left[\text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^T) - \frac{1}{m_h} \mathbf{e}_h \mathbf{e}_h^T \right]_{i \in s_h} \right\}_{h=1,\ldots,H} \times \\ &\quad \hat{\mathbf{\Delta}} \hat{\mathbf{\Gamma}} \mathbf{W} \mathbf{X} \mathbf{A}^{-1}. \end{aligned} \tag{A.8}$$

When m_h is large, $\mathbf{e}_h \mathbf{e}_h^T / m_h \xrightarrow{p} \mathbf{0}$ under reasonable conditions and

$$v_\pi(\hat{\mathbf{z}}^*) = \sum_{h=1}^{H} \frac{m_h}{m_h - 1} \mathbf{X}_h^T \mathbf{W}_h \hat{\mathbf{\Gamma}}_h \hat{\mathbf{\Delta}}_h \, \text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^T) \, \hat{\mathbf{\Delta}}_h \hat{\mathbf{\Gamma}}_h \mathbf{W}_h \mathbf{X}_h. \tag{A.9}$$

Consequently,

$$v_L(\hat{\boldsymbol{\beta}}_{SW}) \doteq \hat{\tau}^4 \sum_{h=1}^{H} \mathbf{A}^{-1} \frac{m_h}{m_h - 1} \mathbf{X}_h^T \mathbf{W}_h \hat{\mathbf{\Gamma}}_h \hat{\mathbf{\Delta}}_h \, \text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^T) \, \hat{\mathbf{\Delta}}_h \hat{\mathbf{\Gamma}}_h \mathbf{W}_h \mathbf{X}_h \mathbf{A}^{-1}. \tag{A.10}$$

The formulas for simpler designs, like single stage sampling with unequal weights, are obtained by specializing (A.10).

References

Abaidoo, R. and Agyapong, E. (2024). Financial market uncertainty and the macro economy: The role of governance and institutional quality. *Economia*, 23. DOI: 10.1108/ECON-02-2023-0034.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, Inc, 2nd edition.

Alhakim-Naser, M., Hasan, G. and Zaifoglu, H. (2024). A geospatial analysis of flood risk zones in Cyprus: Insights from statistical and multi-criteria decision analysis methods. *Environmental Science and Pollution Research*, 31(2), 1-26. DOI: 10.1007/s11356-024-33391-x.

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

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

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Binder, D.A. and Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters. In *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner), Chapter 3, 29-48. Wiley Series in Survey Methodology.

Chambers, R.L. and Skinner, C.J. (2003). *Analysis of Survey Data*. Wiley Series in Survey Methodology. New York: John Wiley & Sons, Inc.

Chatterjee, S. and Price, B. (1991). *Regression Analysis by Example*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, Inc., 2nd edition.

Curtin, L.R., Mohadjer, L. and Dohrmann, S. (2012). The National Health and Nutrition Examination Survey: Sample design. *Vital and Health Statistics, National Center for Health Statistics*, 2(155).

Dobos, I. and Sasvari, P. (2024). Statistical analysis of QS World University Rankings 2021 university rankings using Scopus/SciVal databases. *Regional Statistics*, 14(4), 768-792. DOI: 10.15196/RS140407.

Fox, J. and Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417), 178-183.

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., Bolker, B., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Krivitsky, P., Laboissiere, R., Maechler, M., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Short, T., Venables, W., Walker, S., Winsemius, D. and Zeileis, A. (2023). `car`*: Companion to Applied Regression, R package version 3.1-2*. https://CRAN.R-project.org/package=car.

Gao, J., Xu, D. and Pan, Y. (2024). Configurational paths to arouse tourists' positive experiences in smart tourism destinations: A fsQCA approach. *SAGE Open*, 14(2). DOI: 10.1177/21582440241248226.

Hauck, W.W. and Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72(360), 851-853.

Hebbali, A. (2024). *olsrr: Tools for Building OLS Regression Models, R package version 0.6.0*. https://CRAN.R-project.org/package=olsrr.

Johnson, C.L., Paulose-Ram, R. and Ogden, C.L. (2012). National Health and Nutrition Examination Survey: Analytic guidelines, 1999-2010. *Vital and Health Statistics, National Center for Health Statistics*, 2(161).

Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

Kott, P.S. (2018). A design-sensitive approach to fitting regression models with complex survey data. *Statistics Surveys*, 12, 1-17. https://doi.org/10.1214/17-SS118.

Kovács, P., Petres, T. and Tóth, L. (2005). A new measure of multicollinearity in linear regression models. *International Statistical Review*, 73(3), 405-412.

Kutner, M., Nachtsheim, C. and Neter, J. (2004). *Applied Linear Statistical Models*. New York: McGraw-Hill/Irwin, 4th edition.

Lesaffre, E. and Marx, B.D. (1993). Collinearity in generalized linear regression. *Communications in Statistics – Theory and Methods*, 22(7), 1933-1952.

Liao, D. and Valliant, R. (2012a). Condition indexes and variance decompositions for diagnosing collinearity in linear model analysis of survey data. *Survey Methodology*, 38(2), 189-202. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2012002/article/11757-eng.pdf.

Liao, D. and Valliant, R. (2012b). Variance inflation factors in the analysis of complex survey data. *Survey Methodology*, 38(1), 53-62. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2012001/article/11685-eng.pdf.

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

Little, R. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of American Statistical Association*, 99(466), 546-556. DOI: 10.1198/016214504000000467.

Lumley, T. (2024). *survey: Analysis of Complex Survey Samples, R package version 4.4-2*. https://CRAN.R-project.org/package=survey.

Mackinnon, M. J. and Puterman, M. L. (1989). Collinearity in generalized linear models. *Communications in Statistics – Theory and Methods*, 18(9), 3463-3472.

Marquardt, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12, 591-612.

Marx, B.D. and Smith, E.P. (1990a). Principal component estimation for generalized linear regression. *Biometrika*, 77(1), 23-32.

Marx, B.D. and Smith, E.P. (1990b). Weighted multicollinearity in logistic regression: Diagnostics and biased estimation techniques with an example from lake acidification. *Canadian Journal of Fisheries and Aquatic Sciences*, 47(6), 1128-1135.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman and Hall, 2nd edition.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, Inc.

Özkale, M.R. (2021). The red indicator and corrected VIFs in generalized linear models. *Communications in Statistics – Simulation and Computation*, 50(12), 4144-4170. DOI: 10.1080/03610918.2019.1639740.

Parasyris, A., Stankovic, L. and Stankovic, V. (2024). A machine learning-driven approach to uncover the influencing factors resulting in soil mass displacement. *Geosciences*, 14(8), 220-241. DOI: 10.3390/geosciences14080220.

Pfeffermann, D. and Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. In *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner), Chapter 12, 175-195. Wiley Series in Survey Methodology.

Phillips, C.M., Kesse-Guyot, E., McManus, R., Hercberg, S., Lairon, D., Planells, R. and Roche, H.M. (2012). High dietary saturated fat intake accentuates obesity risk associated with the fat mass and obesity-associated gene in adults. *Journal of Nutrition*, 142(5). DOI: 10.3945/jn.111.153460.

Roever, C., Raabe, N., Luebke, K., Ligges, U., Szepannek, G., Zentgraf, M. and Meyer, D. (2023). `klaR`*: Classification and Visualization, R package version 1.7-3*. https://CRAN.R-project.org/package=klaR.

Schaefer, R.L. (1986). Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computations and Simulations*, 25, 75-91.

Schaefer, R.L., Roi, L.D. and Wolfe, R.A. (1984). A ridge logistic estimator. *Communications in Statistics – Theory and Methods*, 13(1), 99-113.

Shao, J. (2003). *Mathematical Statistics*. New York: Springer, 2nd edition.

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*. Wiley Series in Survey Methodology. Wiley Interscience.

Smith, E.P. and Marx, B.D. (1990). III-conditioned information matrices, generalized linear models and estimation of the effects of acid rain. *Environmetrics*, 1(1), 57-71. DOI: 10.1002/env.3170010107.

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

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

Væth, M. (1985). On the use of Wald's test in exponential families. *International Statistical Review*, 53(2), 199-214.

Valliant, R. (2024). *svydiags: Regression Model Diagnostics for Survey Data, R package version 0.7*. https://CRAN.R-project.org/package=svydiags.

Vanegas, L.H., Rondón, L.M. and Paula, G.A. (2024). *glmtoolbox: Set of Tools to Data Analysis using Generalized Linear Models*. https://CRAN.R-project.org/package=glmtoolbox.

Weissfeld, L.A. and Sereika, S.M. (1991). A multicollinearity diagnostic for generalized linear models. *Communications in Statistics – Theory and Methods*, 20(4), 1183-1198.

Zhang, K., Li, F., Li, H. and Yin, C. (2024). Revealing urban residents' intention to pay for the greening of farmland in the urban fringe by extending the theory of planned behavior: Insights from payment for ecosystem services. *Land Use Policy*, 141(2). DOI: 10.1016/j.landusepol.2024.107127.