Computing Generalized Collinearity Diagnostics for Categorical Variables Using Multivariate Regression

Mokshad P. Gaonkar **Amandiy N. Liwo** **T. Mark Beasley**
University of Alabama at Birmingham

Generalized Variance Inflation Factors (GVIFs) are a means of assessing multicollinearity for related sets of variables in regression models (Fox & Monette, 1992; Fox, 2016); however, they do not frequently appear in statistical literature. R is the only statistical software that has packages that report GVIFs. The main purpose of this paper is to demonstrate how GVIFs are a function of multivariate regression models that can be performed with other statistical software.

Two of the most used indices for assessing multicollinearity in linear regression models are the standard Variance Inflation Factor (VIF) and its reciprocal Tolerance (TOL=1/VIF). Fox and Monette (1992) contend that related sets of regressors (e.g., a set of coding scheme regressors that represent a categorical variable; polynomial terms) represent singular independent variables, and therefore, standard VIFs are not fully applicable. The reasoning underlying this stipulation is subtle but can been shown through the vector geometry of linear models (Fox, 2016, Ch. 10). Briefly, regression models are invariant to transformation and span the same subspace, regardless of the coding scheme used for the regressors. Therefore, any linear transformation of the regressors will produce the same predicted values. For example, orthogonal coding schemes that represent a categorical variable (e.g., Helmert) will have regressors that are uncorrelated with each other; thus, there is no correlation (i.e., collinearity) among this set of variables. A Reference Cell (i.e., Dummy; Indicator) coding scheme applied to the same data will produce the same predicted values but will impose "artificial" collinearity among the regressors that is not of concern (Fox, 2016, p. 357).

Fox and Monette (1992) introduced the Generalized Variance Inflation Factor (GVIF) as a means of assessing multicollinearity for terms in regression models that are based on related sets of regressors, and thus, have more than 1 degree-of-freedom (*df*). Despite this important development, the GVIF is rarely cited in statistical literature and is mentioned in only a few statistics books; most include Fox's (2016, 2020) own work. Furthermore, the R packages vif in the car library (Fox & Weisberg, 2011) and gvif in the glmtoolbox library (Vanegas, Rondón, & Paula, 2023) are the only statistical software that directly computes GVIFs. The purpose of this paper is to: review issues of multicollinearity; reintroduce GVIF using Fox & Monette's (1992) approach; conceptualize TOL, VIF, and GVIF as functions of multivariate regression models; and present approaches to computing GVIFs in popular statistical software that only report standard VIFs. Although this paper focuses on GVIFs for categorical variables in linear models, this method may be applied to other types of related regressors. This includes regressors that are functions of each other such as interaction and polynomial terms. The Appendix shows annotated SAS, R, STATA, and SPSS code to perform the major analyses presented.

Notation

Although scalar notation will be used throughout this paper, matrix notation for the general linear model (GLM) will also be employed for generalization to multivariate regression. The GLM parameter model in matrix notation is expressed as:

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} , \quad (1)$$

where N is the sample size, k=0 to K is a subscript for the regressors, K is the total number of regressor variables, p=K+1 is the number of parameters estimated by the regression model, $\mathbf{y}$ is a Nx1 vector for the dependent variable, $\mathbf{X}$ is a Nxp design matrix that contains the K regressor variables with a N-dimensional vector of ones adjoined to estimate the regression intercept, $\boldsymbol{\beta}$ is a px1 vector unknown population regression coefficients that must be estimated from the sample data, and $\boldsymbol{\varepsilon}$ is a Nx1 vector of population error terms. The parameter model in (1) has the following statistical model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} , \quad (2)$$

where $\mathbf{b}$ is a px1 vector of regression coefficients derived from the sample data and $\mathbf{e}$ is a Nx1 vector of sample residuals. The Ordinary Least Squares (OLS) solution for the sample regression coefficients is:

$$\mathbf{b} = (\mathbf{X'X})^{-1}(\mathbf{X'y}) , \quad (3)$$

where $(\mathbf{X'X})^{-1}$ is the inverse of the pxp uncorrected sum of squares and cross-product matrix for the x-variables and $\mathbf{X'y}$ is the px1 vector of uncorrected cross-products between the x-variables and dependent variable, y.

Collinearity and Multicollinearity

Collinearity (also known as multicollinearity) refers to a situation in which explanatory variables (x) in a multiple regression model are linearly related. Collinear variables contain much of the same information about the dependent variable (y), and therefore, suffer from redundancy.[1] Severe degrees of multicollinearity can lead to computational problems in that the inverse of the $\mathbf{X'X}$ matrix may be difficult to solve using the standard computer algorithms in many statistical software. Furthermore, if an approximate inverse is obtained it may be numerically inaccurate. Due to these computational issues, approaches such as ridge (Hoerl & Kennard, 1970), LASSO (Tibshirani, 1996), and elastic net (Zou & Hastie, 2005) regression have been developed.

Even when an accurate $(\mathbf{X'X})^{-1}$ is obtained, multicollinearity has several other deleterious effects on regression models including bias of the regression coefficients, instability of the regression solution, and difficulties in the interpretation of results (Johnson, Jones & Manley, 2018). Furthermore, multicollinearity not only affects the magnitude of the regression coefficients, but also inflates the variances (i.e., standard errors) of the coefficients and decreases the Full Model R^2 in linear models. This may lead to a failure to reject a false null hypothesis (i.e., Type 2 error). Thus, when there is substantial multicollinearity, loss of statistical power for tests performed in a regression model may be a concern. Yet, removing highly collinear variables from regression models is debatable (O'Brien, 2016).

Motivating Example

Suppose investigators regress scores from an actuarial certification exam (y) onto a set of continuous and categorical variables. The regressor variables include: scores on the Graduate Records Examinations Verbal (GREV) and Quantitative (GREQ) subscales; the examinee's self-reported Age, Gender, and coding scheme regressors to represent four Ethnicities (Asian, Black or African American; Hispanic; White). Also, suppose the examinees were randomized to two different educational preparation conditions: (a) a new experimental preparation training program and (b) a traditional preparation program. These hypothetical data, with a total sample size of N=100, are available for download in both Microsoft Excel and text formats. Table 1 gives definitions for each variable included in the dataset. The Excel sheet has a data dictionary that also describes the variables.

A K=8 regressor model will be investigated:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6B + b_7H + b_8W + e \quad (4)$$

where x_1 represents GREV scores centered at 150; x_2 represents GREQ scores centered at 150; x_3 is an indicator code representing membership of the new experimental training program; x_4 is an indicator code representing being Male; and x_5 represents Age centered at the mean of 30.05 years. The variables, B, H, and W are Indicator codes representing membership in the Black, Hispanic, or White Ethnic groups, respectively. Table 2 provides a summary of several string (character) variables[2] and coding schemes for Ethnicity variables. These centered continuous variables and coded categorical variables in model (4) yield a regression intercept (b_0) that can be interpreted as the predicted value for a 30-year-old Asian Female in the traditional preparation group that scored 150 on both the GREV and GREQ. The focus of this example is on the Ethnicity coding variables because this J=4 level between-subjects factor is represented by 3 dummy-coded regressors (i.e., 3 dfs) and therefore, GVIF is considered a more appropriate method of evaluating multicollinearity for the Ethnicity factor (Fox & Monette, 1992).

Table 3 reports the descriptive statistics for each Ethnic group. There are non-significant between-group mean differences on the continuous actuarial exam score outcome, y (p=0.2085). There are between-group Ethnic differences on the other variables in model (4), which indicate multicollinearity. Specifically, there are non-significant between-group mean differences on the continuous regressors (GREV, p=0.6005; Age, p=0.4199) and proportional differences on the categorical regressors (Training Condition, p=0.0678; Gender, p=0.7060). There are significant Ethnic groups differences on the GREQ score, p=0.0042. Although most of these between-group Ethnic differences are not statistically significant, any between-group differences for the other covariates in the model will result in some degree of collinearity.[3]

Table 1. Data Dictionary for Motiving Example

Continuous Variables		
y		Score on an Actuarial Certification Exam (Range 130-170)
GREV		Score GRE Verbal (Range 130 to 170)
GREV_C	(x_1)	Score GRE Verbal centered at 150 (Range -20 to 20)
GREQ		Score GRE Quantitative (Range 130 to 170)
GREQ_C	(x_2)	Score GRE Quantitative centered at 150 (Range -20 to 20)
Age		Years (Range 20 to 36)
Age_C	(x_5)	Years centered at mean age 30.05 (Range -5 to 11)
Age_2		Years Squared (Age^2)
Age_C2		Mean Centered Age Squared (Age_C^2)
Categorical Variables		
TRT	(x_3)	Dummy Code for Educational Program (Traditional = 0; Experimental = 1)
Tx		Effect Code for Educational Program (Traditional = -1; Experimental = 1)
Male	(x_4)	Dummy Code for Reported Gender (Female = 0; Male = 1)
Sex		Effect Code for Reported Gender (Female = -1; Male = 1)
Ethnicity		String Variable for Ethnicity (W =White; B = African American or Black; H = Hispanic; A = Asian)
EthGRP1		String Variable for Ethnicity (1A = Asian; 2B = African American or Black; 3H = Hispanic; 4W =White)
EthGRP2		String Variable for Ethnicity (1B = African American or Black; 2A = Asian; 3H = Hispanic; 4W =White)
EthGRP3		String Variable for Ethnicity (1H = Hispanic; 2A = Asian; 3B = African American or Black; 4W =White)
EthGRP4		String Variable for Ethnicity (1W =White; 2A = Asian; 3B = African American or Black; 4H = Hispanic)
EthGRP1R		String Variable for Ethnicity (Reverse Order of EthGRP1) (1W =White; 2H = Hispanic; 3B = African American or Black; 4A = Asian)
A		Indicator (Dummy) Code for Asian Ethnicity
B		Indicator (Dummy) Code for African American (AA) or Black Ethnicity
H		Indicator (Dummy) Code for Hispanic Ethnicity
W		Indicator (Dummy) Code for White Ethnicity
ETHN_H1		Helmert Contrast (Whites vs All Other Ethnic Groups)
ETHN_H2		Helmert Contrast (Hispanics vs Blacks/AA & Asians; Whites Excluded)
ETHN_H3		Helmert Contrast (Blacks/AA vs Asians; Whites & Hispanics Excluded)

Table 2. Summary of Coding Scheme for Ethnicity

Ethnicity	**N**	**EthGRP1**	**EthGRP2**	**EthGRP3**	**EthGRP4**	**EthGRP1R**	***A***	***B***	***H***	***W***	H_1	H_2	H_3
Asian	17	**1A**	2A	2A	2A	4A	1	0	0	0	–7	–4	–28
Black / AA	28	2B	**1B**	3B	3B	3B	0	1	0	0	–7	–4	17
Hispanic	20	3H	3H	**1H**	4H	2H	0	0	1	0	–7	9	0
White	35	4W	4W	4W	**1W**	**1W**	0	0	0	1	13	0	0

Note: For EthGRP1-EthGRP4 the **Reference Ethnic group** (first group is default Reference group in R `lm` package) is **bolded**.

The results for model (4) are reported in Table 4. The overall K=8 variable model accounts for 43.44% of the variance in y (p<0.0001). The regression results indicate that after adjusting for other variables in the model GREV (p=0.0315) and GREQ (p<0.0001) are positively related to y (i.e., higher GRE scores are associated with higher scores on the actuarial certification exam). The type of preparation program (TRT) was also significantly related to y (p=0.0153). Given the indicator coding scheme for TRT, the results indicate that after adjusting for the other covariates examinees that were in the new experimental preparation program scored 3.55 units higher than examinees that went through traditional preparation.

Table 3. Descriptive Statistics for Ethnicity.

Ethnicity	N	*y*	GREV	GREQ	Age	Condition	Male
Asian	17	153.06 (9.61)	149.35 (10.48)	150.24 (10.56)	30.53 (2.70)	5 (29.41%)	10 (58.8%)
Black or AA	28	148.68 (8.68)	148.04 (8.71)	141.43 (7.35)	30.18 (2.75)	17 (60.71%)	17 (60.7%)
Hispanic	20	153.90 (9.01)	151.75 (9.83)	150.10 (10.86)	30.55 (2.48)	8 (40.00%)	15 (75.0%)
White	35	150.80 (7.91)	148.69 (9.94)	146.17 (9.99)	29.43 (2.92)	22 (62.86%)	23 (65.7%)
***p*-values**		0.2085[W]	0.6005[W]	0.0042[W]	0.4199[W]	0.0678[P]	0.7060[P]
TOTAL	100	151.21 (8.74)	149.23 (9.63)	146.32 (10.10)	30.05 (2.75)	52 (52%)	65 (65%)

Note: W: p-value from Welch ANOVA *F*-test.
P: *p*-value from Pearson Chi-Square test.

Assuming that the actuarial exam, like the GRE, was scaled to have a standard deviation of 10, then these results would translate into a covariate-adjusted effect size of 0.355.[4] The regression coefficients (b_6-b_8) for the codes representing the Ethnicity factor (x_6-x_8) are differences in predicted values (i.e., least squares means) between the reference group (Asian; b_0) and each of the other Ethnic groups (Blacks, Hispanics, Whites, respectively). None of the regression coefficients tested by this particular coding scheme were statistically significant. Specifically, the predicted values for Blacks (b_6 = –2.0256, p=0.3782), Hispanics (b_7 = –0.1859, p=0.9533), and Whites (b_8 = –2.3467, p=0.2737) do not significantly differ from the predicted values for Asians.

In terms of evaluating multicollinearity, Table 4 shows that Age (x_5) and the dummy variables representing preparation condition (TRT; x_3) and gender (Male; x_4) have very small VIFs (high tolerance). The GREV (x_1) and GREQ (x_2) scores have VIFs of 1.46 and 1.66, respectively, also indicating a small degree of multicollinearity. The Ethnicity variables using the model (4) reference cell coding scheme with Asians as the reference group, indicate larger amounts of multicollinearity with VIFs ranging from 1.78 to 2.24. The last two columns of Table 4 report GVIF-related values that can be produced with either the `vif` (Fox & Weisberg, 2011) or `gvif` (Vanegas, et al., 2023) R packages. For the single *df* regressors (x_1-x_5), the GVIFs are the same as the standard VIFs that would be reported by other statistical packages (e.g., SAS; STATA; SPSS). Unlike other statistical software, the `vif` and `gvif` R packages report a single GVIF with 3 *df*s for the Ethnicity factor. The last column reports $GVIF^{(1/(2df))}$, a scaling to "preserve comparability across subspaces of different dimension" (Fox & Monette, 1992, p. 180) and make the single *df* and the multiple *df* GVIFs "roughly comparable" (Fox, 2016, p. 635). Both GVIF and its scaled value indicate a much lesser degree of collinearity than the reference cell coding scheme for model (4). SAS, STATA, R, and SPSSS code to perform model (4) appear in Appendices A.1 through A.4, respectively. Appendices A.5 and A.6 show output from the R `lm` package along with output from `vif` and `gvif` packages. Note that if dummy coded variables that represent the Ethnicity factor are entered into `lm`, then the `vif` and `gvif` packages do not invoke the computation of GVIF (Appendix A.5). If a single string variable that represents the Ethnicity factor is entered into `lm`, then the `vif` and `gvif` packages compute and report GVIF (Appendix A.6).

To demonstrate issues with interpreting standard VIFs for related sets of regressors, three alternate version of the K=8 regressor model in (4) with different coding schemes were performed:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6A + b_7H + b_8W + e \quad (5)$$
$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6A + b_7B + b_8W + e \quad (6)$$
$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6A + b_7B + b_8H + e \quad (7)$$

The Refence Cell Coding schemes for Models 5, 6, and 7 force the Reference Ethnic group to be Blacks, Hispanics, and Whites, respectively. The results for the Ethnicity coding scheme variables for Models 5 through 7 appear in Tables 5 through 7, respectively. Because the models are invariant to transformation and span the same subspace, most of the results do not change with changes in the coding scheme for the Ethnicity categories. The Full Model and the associated global *F*-test are identical for models 4 through 7: R^2 = 0.4344; $F_{(8,91)}$=8.74, p <0.0001. The coefficients, standard errors (*SE*s), *p*-values, TOLs, and VIFs do not change for any of the single *df* variables (x_1-x_5), and thus, are not reported in Tables 5-7. For the Ethnicity results, values that differ across models 4-7 are in **bold blue** font in Tables 4-7. The intercepts and their *SE*s change because each coding scheme forces a different Ethnic group to be the Reference cell.

Table 4. Results of Model (4): Reference Ethnic Group (Asians).

Variable	*b*	SE(*b*)	p-value	TOL	VIF	GVIF	GVIF$^{(1/(2df))}$
GREV (x_1)	0.1891	0.0866	0.0315	0.68415	1.46168	1.461678	1.208999
GREQ (x_2)	0.3864	0.0879	<0.0001	0.60324	1.65772	1.657716	1.287523
TRT (x_3)	3.5526	1.4365	0.0153	0.91317	1.09509	1.095089	1.046465
Male (x_4)	1.6017	1.4785	0.2815	0.94575	1.05736	1.057364	1.028282
Age (x_5)	–0.4412	0.2618	0.0954	0.91415	1.09391	1.093911	1.045902
B (x_6)	**–2.0256**	**2.2872**	**0.3782**	**0.44594**	**2.24245**		
H (x_7)	**–0.1859**	**2.2900**	**0.9355**	**0.56055**	**1.78398**	**1.327977**	**1.048411**
W (x_8)	**–2.3467**	**2.1309**	**0.2737**	**0.45529**	**2.19640**		
Intercept	**153.5427**	**2.4893**	**<0.0001**	-	-	-	-
Full Model R^2 = 0.4344; $F_{(8, 91)}$ = 8.74, p <0.0001				**$F_{(3,91)}$ = 0.62, p = 0.6013**			

Note: Last Two Columns contain output from the R package vif.

Table 5. Results of Model (5): Reference Ethnic Group (Blacks).

Variable	b	SE(b)	p-value	TOL	VIF	GVIF	GVIF$^{(1/(2df))}$
A (x_6)	**2.0256**	**2.28724**	**0.3782**	**0.63715**	**1.56949**		
H (x_7)	**1.8397**	**2.15591**	**0.3957**	**0.63242**	**1.58122**	**1.327977**	**1.048411**
W (x_8)	**–0.3211**	**1.79936**	**0.8588**	**0.63852**	**1.56613**		
Intercept	**151.5171**	**2.4979**	**<0.0001**	-	-	-	-
Full Model R^2 = 0.4344; $F_{(8, 91)}$ = 8.74, p <0.0001				**$F_{(3,91)}$ = 0.62, p = 0.6013**			

Table 6. Results of Model (6): Reference Ethnic Group (Hispanics).

Variable	b	SE(b)	p-value	TOL	VIF	GVIF	GVIF$^{(1/(2df))}$
A (x_6)	**0.1859**	**2.2900**	**0.9355**	**0.63563**	**1.57324**		
B (x_7)	**–1.8397**	**2.1559**	**0.3957**	**0.50192**	**1.99234**	**1.327977**	**1.048411**
W (x_8)	**–2.1608**	**2.0053**	**0.2841**	**0.51411**	**1.94512**		
Intercept	**153.3568**	**2.5773**	**<0.0001**				
Full Model R^2 = 0.4344; $F_{(8, 91)}$ = 8.74, p <0.0001				**$F_{(3,91)}$ = 0.62, p = 0.6013**			

Table 7. Results of Model (7): Reference Ethnic Group (Whites).

Variable	b	SE(b)	p-value	TOL	VIF	GVIF	GVIF$^{(1/(2df))}$
A (x_6)	**2.3467**	**2.1309**	**0.2737**	**0.73408**	**1.36225**		
B (x_7)	**0.3211**	**1.7994**	**0.8588**	**0.72055**	**1.38783**	**1.327977**	**1.048411**
H (x_8)	**2.1608**	**2.0053**	**0.2841**	**0.73099**	**1.36800**		
Intercept	**151.1960**	**2.2425**	**<0.0001**				
Full Model R^2 = 0.4344; $F_{(8, 91)}$ = 8.74, p <0.0001				**$F_{(3,91)}$ = 0.62, p = 0.6013**			

The regression coefficients (b_6-b_8) for the codes representing the Ethnicity variable are differences in expected values between each group and the reference group (intercept; b_0). Consequently, some of the regression coefficients and their *SE*s change with changes in the coding scheme. Of importance to the goal of this paper, the values for VIF and TOL change with changes in the coding scheme.

For model (5), regression coefficients (b_6-b_8) for the codes representing the Ethnicity variable (x_6-x_8) are differences in predicted values between the reference group (Blacks; b_0) and each of the other Ethnic group (Asians, Hispanics, Whites, respectively). Specifically, the predicted values for Asians (b_6 = 2.0256, p=0.3782), Hispanics (b_7 = –1.8397, p=0.3957), and Whites (b_8 = –0.3211, p=0.8588) do not significantly differ from Blacks (Table 5). For model (5), the Ethnicity variables using the Reference Cell coding scheme with Blacks as the Reference group, indicate smaller amounts of multicollinearity with VIFs of approximately 1.57 compared to the VIFs from model (4). These changes in the standard VIFs due to changes in coding schemes are problematic for interpretating multicollinearity and its effects on statistical significance and power for the Ethnicity factor. For example, the results of model (4) and model (5) both indicate that the pairwise comparison between Asians and Blacks (b_6 in both models) have the same covariate adjusted mean difference (b_6 = 2.0256; in absolute value), *SE* = 2.2872, and p-value = 0.3782. Yet, these regression coefficients have different indices for multicollinearity (VIF = 2.24245 in model 4;

VIF = 1.56949 in model 5). In principle, for two variables with the same coefficient in absolute value, the variable with the larger VIF might be expected to have a larger *SE*, because in a homoscedastic model the *SE*s uses the same Mean Square Error (*MSE*). This is not the case, however, for the Asian vs Black pairwise difference because the *B* and *A* dummy codes have different variances.[5] By examining Tables 4 through 7, several examples of regression coefficients that lead to the same pairwise difference, *SE*, and *p*-value but quite different values for VIF and TOL can be found.

For single *df* regressors (x_1-x_5), the $GVIF^{(1/(2df))}$ transformation is simply taking the square root of the standard VIF. Depending on the chosen reference group the standard VIFs for the Ethnicity dummy codes range from 1.368 to 2.24245 with square roots ranging from 1.1696 to 1.4975 (see Tables 4-7). The scaled $GVIF^{(1/(2df))}$ = 1.048411 for Ethnicity falls below the square root of the two extreme standard VIFs using Reference Cell coding. Thus, scaled GVIF indicates that Ethnicity as a singular independent variable based on set of related coding vectors has a lesser degree of collinearity than would be indicated by the standard VIFs from any of the reference cell coding schemes. This highlights that "artificial" or at least "inconsistent" collinearity can be imposed by a coding scheme. The fact that the VIFs and TOLs change based on changes in coding scheme supports Fox and Monette's (1992) contention that standard VIFs are not applicable to related sets of regressors and GVIF is more appropriate.

Omnibus Test of the Ethnicity Factor

An omnibus *F*-test for whether the four ethnic groups have the same means after adjusting for the other covariates (H_0: $\mathbf{L\beta}=\mathbf{c}$) can be formed using General Linear Hypothesis (GLH) Testing of the form:

$$F = [(\mathbf{Lb}-\mathbf{c})'(\mathbf{LVL}')^{-1}(\mathbf{Lb}-\mathbf{c})] , \tag{8}$$

where **L** is a *q*x*p* vector of contrast values; **c** is *q*x1 vector of null values (in this case and typically, a vector of zeros), and *q* is the *df*s for the hypothesis tested. Under the assumption of homoscedasticity, $\mathbf{V} = MSE(\mathbf{X'X})^{-1}$. In this case, testing that all regression coefficients associated with Ethnic group differences is equal to $\mathbf{c}=\mathbf{0}$ uses a *q*x*p* (3x9) **L** matrix:

$$\mathbf{L} = \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{matrix};$$

which results in testing the joint null hypothesis, H_0: $\beta_6 = 0$; $\beta_7 = 0$; $\beta_8 = 0$. This 3-*df* omnibus test for Ethnicity, does not change with changes in the coding scheme; it is invariant to transformation. Therefore, "artificial" collinearity imposed by some coding schemes will have no effect on the omnibus test; only collinearity with other covariates in the model will affect the magnitude and power of the omnibus test. Applying this **L** matrix to models 4 through 7 (or any valid coding scheme for Ethnicity) results in the same value for the *F*-test in (8), $F_{(3,91)} = 0.62$, $p = 0.6013$, indicating no statistically significant differences among the Ethnic groups. Code to perform the test in (8) appear in Appendices A.1 (SAS `proc reg`), A.2 (STATA `regress`) A.3 (R `lm` and `glh.test` from the `gmodels` library), and A.4 (SPSS `UNIANOVA`).

Standard Variance Inflation Factors and Auxiliary Tolerance Models

Values for tolerance can be shown to be equal to $1\text{-}R^2_{kk\prime}$ of auxiliary "tolerance" models in which each x_k variable is regressed on to the other $x_{k\prime}$ variables in the model. For example, to find the TOL and VIF for x_1 (GREV), the following K=7 regression model can be used:

$$x_1 = a_0 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6A + a_7B + a_8W + e \tag{9}$$

As with regressing *y* onto the K=8 regressor models 4 through 7, the choice of coding scheme for Ethnicity is inconsequential to the R^2 of model (9). This model with x_1 representing GREV scores yields an $R^2_{kk\prime}$ = 0.315855, and therefore, TOL = 1–0.315855 = 0.684145. The standard VIF is the reciprocal of TOL, VIF = 1/(1–0.315855) = 1.461678. As can be seen, these values match the results reported in Table 4.

The "tolerance" model concept provides a heuristic for explaining TOL and VIF; however, it is not necessary to perform *K* auxiliary models to obtain these indices. As can be seen in equation (37) in the Discussion section, VIFs are a transformation of the $(\mathbf{X'X})^{-1}$ matrix used in the OLS regression solution (3). Because $(\mathbf{R_{XX}})^{-1}$ is also a transformation of $(\mathbf{X'X})^{-1}$, standard VIFs can be directly calculated from the matrix solution for a standardized regression model:

$$\mathbf{b^*} = (\mathbf{R_{xx}})^{-1}\mathbf{r_{Xy}} \tag{10}$$

where **b*** is the *K*x1 vector of standardized regression coefficients, $(\mathbf{R_{XX}})^{-1}$ is the inverse of the *K*x*K* correlation matrix among the *x*-variables, and $\mathbf{r_{xy}}$ is the *K*x1 vector of correlations between the *x*-variables and *y*. Computationally, VIFs are the diagonal of $(\mathbf{R_{XX}})^{-1}$:

$$\text{VIF} = \text{diag}[(\mathbf{R_{xx}})^{-1}] \quad (11)$$

(Marquardt, 1970). Table 8 shows **Rxx** for the variables used in models 4-7. To obtain a valid inverse and calculate VIF, one row and column for the Ethnicity Indicator variables from Table 9 needs to be removed. Table 9 shows $(\mathbf{R_{XX}})^{-1}$, for model (4) (row 6 and column 6 removed from Table 8, Asians are the reference group). As can be seen the bolded diagonal matches the VIF values reported in Table 4. Appendices A.7 through A.9 show SAS, SPSS, and STATA matrix language code to perform these VIF calculations, respectively.

Table 8. Correlation Matrix for the Variables included in Models 4 through 7.

	GREV	GREQ	Age	TRT	Male	A	B	H	W
GREV	1	0.5272	-0.0717	-0.0250	0.0461	0.0058	-0.0777	0.1315	-0.0417
GREQ	0.5272	1	0.1037	0.0246	-0.0622	0.1764	-0.3036	0.1881	-0.0109
Age	-0.0717	0.1037	1	-0.0263	-0.1703	0.0792	0.0293	0.0912	-0.1664
TRT	-0.0250	0.0246	-0.0263	1	-0.0336	-0.2046	0.1088	-0.1201	0.1595
Male	0.0461	-0.0622	-0.1703	-0.0336	1	-0.0586	-0.0560	0.1048	0.0110
A	0.0058	0.1764	0.0792	-0.2046	-0.0586	1	-0.2822	-0.2263	-0.3321
B	-0.0777	-0.3036	0.0293	0.1088	-0.0560	-0.2822	1	-0.3118	-0.4576
H	0.1315	0.1881	0.0912	-0.1201	0.1048	-0.2263	-0.3118	1	-0.3669
W	-0.0417	-0.0109	-0.1664	0.1595	0.0110	-0.3321	-0.4576	-0.3669	1

Table 9. Inverse of Correlation Matrix $[(\mathbf{R_{xx}})^{-1}]$ for the Variables included in Model 4.

	GREV	GREQ	Age	TRT	Male	B	H	W
GREV	**1.46168**	-0.84809	0.18481	0.08572	-0.08349	-0.25556	-0.14736	-0.10127
GREQ	-0.84809	**1.65772**	-0.19258	-0.16865	0.12635	0.63421	0.09155	0.29989
Age	0.18481	-0.19258	**1.09391**	0.01310	0.16885	-0.00819	-0.04684	0.16279
TRT	0.08572	-0.16865	0.01310	**1.09509**	0.01905	-0.36874	-0.10495	-0.37816
Male	-0.08349	0.12635	0.16885	0.01905	**1.05736**	0.03010	-0.13653	-0.02498
B	-0.25556	0.63421	-0.00819	-0.36874	0.03010	**2.24245**	1.11266	1.48773
H	-0.14736	0.09155	-0.04684	-0.10495	-0.13653	1.11266	**1.78398**	1.16899
W	-0.10127	0.29989	0.16279	-0.37816	-0.02498	1.48773	1.16899	**2.19640**

Generalized Collinearity

As previously stated, categorical factors with $J > 2$ levels (i.e., have more than 1 *df*) represent singular independent variables and the set of related variables (i.e., coding scheme) that represent these categories can impose "artificial" and inconsequential collinearity (Fox, 2016, p. 357). There are a multitude of coding schemes that can validly represent the $q=(J-1)$ *df*s of a categorical predictor. Effect, Polynomial, Cell Mean, and Helmert[6] coding schemes are employed in many research applications. Reference Cell (i.e., Indicator; Dummy) coding, however, is the most commonly used approach and is the focus in this paper.

The correlations among a set of dummy coded variables are affected by the choice of reference category when the categories have unequal sample sizes. As demonstrated in Tables 4-7, the "artificial" collinearity can depend on which set of $q=(J-1)$ dummy codes are used. In practice, data analysts often want a regression model with VIFs below some acceptable threshold (Buteikis, 2020). Consequently, researchers are sometimes advised to pick the category with the most cases to serve as the reference group for a set of dummy regressors in order to reduce multicollinearity (Allison, 2012; Fox, 2016. p. 357). In Table 4, the Ethnic group with the smallest sample size, Asians (n_j=17) were the Reference group in model (4), and this led to the highest standard VIFs. By contrast, in Table 7, Whites (n_j=35) were the Reference group, which led to the lowest standard VIFs. Using the group with the smallest sample size as the reference category may be a poor computational choice, resulting in elevated levels of multicollinearity and potentially unstable results. But because the models are invariant to transformation, the choice of coding

scheme is not fundamental to model fit or calculating predicted values. As shown in Tables 4-7, these practices can alter estimates of multicollinearity (i.e., standard VIFs). GVIFs, however, are independent of the bases selected for the subspaces spanned by the columns of **X**, and therefore, are invariant to the choice of coding scheme for categorical variables.[7]

Computing Generalized Variance Inflation Factors

Using the matrix formulation in (1) and (2), the Full Design matrix, **X**, for model (4) is formed as:

$$\mathbf{X} = \mathbf{1}_N \mid x_1 \mid x_2 \mid x_3 \mid x_4 \mid x_5 \mid B \mid H \mid W \tag{12}$$

where $\mathbf{1}_N$, N-dimensional vector of ones adjoined to K=8 regressor variables to estimate the regression intercept (b_0). To calculate GVIF, Fox and Monette (1992) partition the design matrix, **X**, and re-express the model as:

$$\mathbf{y} = \mathbf{1}_N b_0 + \mathbf{Z}\mathbf{b}_\mathbf{Z} + \mathbf{W}\mathbf{b}_\mathbf{W} + \mathbf{e} \ , \tag{13}$$

where **W** is a Nxq matrix containing the matrix for the variable(s) being evaluated for multicollinearity and **Z** is a Nx(K-q) matrix containing the remaining regressors, excluding the constant (b_0). Fox and Monette (1992) proposed a statistic based on Wilks' Lambda (λ_W) defined as:

$$\text{GVIF} = \frac{\det(\mathbf{R_{WW}})\det(\mathbf{R_{ZZ}})}{\det(\mathbf{R_{XX}})} \tag{14}$$

where $\mathbf{R_{XX}}$ is the correlation matrix for all K variables in the full design matrix, **X**, $\mathbf{R_{WW}}$ is a partition of **Rxx** that is the qxq correlation matrix for the variable(s) being evaluated for multicollinearity, **W**, and $\mathbf{R_{ZZ}}$ is a partition of **Rxx** is the correlation matrix for the M=(K-q) remaining regressors (covariates), **Z**.

For any single *df* variable (continuous or dummy code for a binary variable), GVIF reduces to the standard VIF (Fox & Monette, 1992). To demonstrate for GREV (x_1), **X** for models (4) and (12) is partitioned into **W**, a Nx1 matrix containing x_1, and **Z**, a Nx7 matrix containing the remaining regressors (x_2-x_8), excluding the constant:

$$\mathbf{W} = x_1 \text{ and} \tag{15}$$

$$\mathbf{Z} = x_2 \mid x_3 \mid x_4 \mid x_5 \mid B \mid H \mid W \ . \tag{16}$$

$\mathbf{R_{XX}}$ (Table 8) is partitioned in to (a) $\mathbf{R_{WW}} = \mathbf{R_{11}}$, a 1x1 matrix with only the correlation for x_1 (GREV) and itself, a scalar value of $\mathbf{R_{WW}} = \mathbf{R_{11}} = 1$, and (b) $\mathbf{R_{zz}}$, a 7x7 correlation matrix for the other 7 regressor variables in Models 4-8. The determinant of $\mathbf{R_{WW}}$ is also 1: det($\mathbf{R_{WW}}$) = det(**1**) = 1. For Model 4, the determinant of the 7x7 $\mathbf{R_{ZZ}}$ is det($\mathbf{R_{zz}}$) = 0.5398237, and the determinant of 8x8 $\mathbf{R_{xx}}$ is det($\mathbf{R_{xx}}$) = 0.3693177. Thus,

$$\text{GVIF}_1 = \frac{\det(\mathbf{R_{WW}})\det(\mathbf{R_{ZZ}})}{\det(\mathbf{R_{XX}})} = \frac{(\mathbf{1})(\mathbf{0.5398237})}{(\mathbf{0.3693177})} = 1.461678, \tag{17}$$

matching the standard VIF reported for GREV (x_1) in Table 4.

Related set of variables requiring more than 1 coefficient, and thus multiple *df*s, are evaluated using the GVIF. For Ethnicity (x_6-x_8), **X** for models (4) and (12) is partitioned into **W**, a Nx3 matrix containing coding scheme variables for Ethnicity, and **Z**, a Nx5 matrix containing the remaining regressors (x_1-x_5), excluding the constant:

$$\mathbf{W} = B \mid H \mid W \text{ and} \tag{18}$$

$$\mathbf{Z} = x_1 \mid x_2 \mid x_3 \mid x_4 \mid x_5 \ . \tag{19}$$

In this case, $\mathbf{R_{XX}}$ is partitioned in to (a) $\mathbf{R_{WW}}$ a 3x3 (qxq) correlation matrix for Ethnicity the coding scheme variables, **W**, and (b) $\mathbf{R_{zz}}$ 5x5 correlation matrix for the other 5 regressors, **Z**, (first 5 rows and columns of Table 8). For model (4), the determinant of $\mathbf{R_{WW}}$, det($\mathbf{R_{WW}}$) = 0.7320946. The determinant of $\mathbf{R_{ZZ}}$ is det($\mathbf{R_{zz}}$) = 0.6699208. Thus,

$$\text{GVIF}_{\text{ETHN}} = \frac{\det(\mathbf{R_{WW}})\det(\mathbf{R_{ZZ}})}{\det(\mathbf{R_{XX}})} = \frac{(\mathbf{0.7320946})(\mathbf{0.6699208})}{(\mathbf{0.3693177})} = 1.327977, \tag{20}$$

which matches the GVIF produced by the `vif` and `gvif` R packages (see Table 4). Note that the choice of coding scheme will affect the determinants of the $\mathbf{R_{XX}}$, $\mathbf{R_{ZZ}}$, and $\mathbf{R_{WW}}$ correlation matrices; however, it will not affect the value of GVIF. Table 10 reports the determinants of $\mathbf{R_{XX}}$, $\mathbf{R_{ZZ}}$, and $\mathbf{R_{WW}}$ for each choice of reference cell coding scheme for Ethnicity (Models 4-7) that comprise the GVIF computations demonstrated for GREV (17) and Ethnicity (20). Note regardless of coding scheme, the GVIF is invariant to transformation, even though the determinants differ by coding scheme. Appendices A.7, A.8, and A.9

Table 10. Determinants to Compute GVIF for GREV (17) and Ethnicity (20) for Each Refence Cell Coding Scheme Models 4-7.

	Model 4		**Model 5**		**Model 6**		**Model 7**	
	GREV (*q*=1)	**Ethnicity (*q*=3)**	**GREV (*q*=1)**	**Ethnicity (*q*=3)**	**GREV (*q*=1)**	**Ethnicity (*q*=3)**	**GREV (*q*=1)**	**Ethnicity (*q*=3)**
$\mathbf{R_{WW}}$	1	0.4540598	1	0.6487488	1	0.51488	1	0.732095
$\mathbf{R_{ZZ}}$	0.3348096	0.6699210	0.4783672	0.6699210	0.3796565	0.6699210	0.539824	0.669921
$\mathbf{R_{XX}}$	0.2290583	0.2290583	0.3272726	0.3272726	0.2597401	0.2597401	0.369318	0.369318
GVIF	1.461678	1.327977	1.461678	1.327977	1.461678	1.327977	1.461678	1.327977

Note: GVIF = [det($\mathbf{R_{WW}}$)*(det($\mathbf{R_{ZZ}}$)]/det($\mathbf{R_{XX}}$)

contain matrix language code to compute GVIF from the data via the Fox & Monette (1992) approach (14) for SAS/IML, SPSS `MATRIX`, and STATA `matrix`, respectively.

Generalized Variance Inflation Factors and Auxiliary Tolerance Models

Unfortunately, the `vif` (Fox & Weisberg, 2011) and `gvif` (Vanegas et al., 2023) packages in R are the only statistical software that compute and report GVIF. To demonstrate how to compute GVIF using other popular statistical software (e.g., SAS, SPSS, STATA), we return to the heuristic concept of tolerance being 1-$R^2_{kk'}$ from auxiliary "tolerance" models in which each x_k variable is regressed on to the other $x_{k'}$ variables in the model. For example, model (9) can be re-expressed as a multivariate regression model, which in scalar notation is:

$$x_2 + x_3 + x_4 + x_5 + B + H + W \;\; = a_0 + a_1 x_1 + e \; . \tag{21}$$

In general, matrix notation for a multivariate "tolerance" model can be expressed as:

$$\mathbf{Z} = \mathbf{WA} + \; \mathbf{E} \; , \tag{22}$$

where q is the number of regressors being evaluated for multicollinearity, M=(K-q) is the number of remaining covariates, **W** is an Nx(q+1) design matrix with the variable(s) being evaluated for multicollinearity from the partition in (13) with a vector of ones ($\mathbf{1}_N$) adjoined to estimate M intercepts, **Z** is an NxM matrix of the remaining covariates (dependent variables) from the partition in (13), **A** is a (q+1)xM matrix of regression coefficients from the auxiliary model, and **E** is a NxM matrix of sample residuals.

To obtain GVIF for x_1 (GREV) from model (21) and the partition in (16), **Z** is a horizontal concatenation of GREQ, TRT, Male, Age, B, H, and W:

$$\mathbf{Z} = \; x_2 \mid x_3 \mid x_4 \mid x_5 \mid B \mid H \mid W \; ; \tag{23}$$

and from the partitioning in (15), **W** is GREV with a vector of ones adjoined to estimate M = 7 intercepts:

$$\mathbf{W} = \mathbf{1}_N \mid x_1 \tag{24}$$

To compute Wilks' lambda from these data matrices, define $\mathbf{H}_0$, as a NxN projection "Hat" matrix constructed from the vector of ones:

$$\mathbf{H}_0 = \mathbf{1}_N[(\mathbf{1'}_N\mathbf{1}_N)^{-1}]\mathbf{1'}_N. \tag{25}$$

Define $\mathbf{H}_W$, as a NxN Hat matrix based on (24):

$$\mathbf{H}_W = \mathbf{W}[(\mathbf{W'W})^{-1}]\mathbf{W'} \; . \tag{26}$$

The MxM Total Sums of Squares matrix **T** is computed as:

$$\mathbf{T} = \mathbf{Z}'(\mathbf{I}_N - \mathbf{H}_0)\mathbf{Z} \; ; \tag{27}$$

where $\mathbf{I}_N$ is an N-dimensional Identity matrix. The MxM Residual Sums of Squares matrix **R** is computed as:

$$\mathbf{R} = \mathbf{Z}'(\mathrm{I}_N - \mathbf{H}_W)\mathbf{Z} \; . \tag{28}$$

The MxM Model Sums of Squares matrix **M** is computed as:

$$\mathbf{M} = \mathbf{Z}'(\mathbf{H}_W - \mathbf{H}_0)\mathbf{Z} \; . \tag{29}$$

As in univariate ANOVA models: $\mathbf{T} = \mathbf{M} + \mathbf{R}$. Wilks' lambda is computed as:

$$\lambda_W = \frac{\det(\mathbf{R})}{\det(\mathbf{T})} \; . \tag{30}$$

For model (22), the determinant of the Total SS matrix is det(**T**) = 10.57217x10^{13} and the determinant of the Residual SS matrix is det(**R**) = 7.2328973x10^{12}. The Wilks' lambda of this model is λ_W = 0.68414503, which is the Tolerance for x_1 (GREV), and therefore, GVIF = $(1/\lambda_W)$ = 1.46168, which matches the TOL and VIF for GREV in Table 4. Therefore, by taking the reciprocal of λ_W in (30):

$$\text{GVIF} = \frac{\det(\mathbf{T})}{\det(\mathbf{R})} . \tag{31}$$

For related sets of regressor variables that have multiple *df*s (e.g., categorical variables), GVIF provides a single value to evaluate multicollinearity. Again, following the heuristic approach to explaining TOL and VIF, the auxiliary "tolerance" model for Ethnicity is a multivariate analysis of variance (MANOVA) regression model with the following scalar notation:

$$x_1 + x_2 + x_3 + x_4 + x_5 = a_0 + a_1B + a_2H + a_3W + e . \tag{32}$$

In matrix notation for model (32) and the partition in (20), **Z** is a horizontal concatenation of GREV, GREQ, TRT, Male, and Age:

$$\mathbf{Z} = x_1 \mid x_2 \mid x_3 \mid x_4 \mid x_5 \tag{33}$$

and from the partition in (18), **W** is the Reference Cell coding scheme for Ethnicity with a vector of ones adjoined to estimate M = 5 intercepts:

$$\mathbf{W} = \mathbf{1}_N \mid B \mid H \mid W \tag{34}$$

For model (32), the determinant of the Total SS matrix is det(**T**) = 26.456489x10^{13} and the determinant of the Residual SS matrix is det(**R**) = 19.922399x10^{13}. From (30), the Wilks' lambda is λ_W = 0.75302505, which could be thought of as a "Generalized Tolerance", and therefore, from (31), GVIF = $(1/\lambda_W)$ = 1.327977, which matches the GVIF for Ethnicity in Tables 4-7 from the `vif` and `gvif` R packages. To obtain the Wilks' λ_W for GREV and Ethnicity: Appendix A.10 contains annotated SAS `proc reg` and `proc glm` code; Appendix A.11 contains annotated STATA `manova` code; Appendix A.12 contains annotated SPSS `GLM` code; and Appendix A.13 contains annotated R `lm` and `anova` code. The user will need to take the take the reciprocal of λ_W to obtain GVIF (ie., GVIF=$1/\lambda_W$). Appendices 14 and 15 contain SAS/IML and STATA `matrix` code to compute GVIF via the multivariate regression approach (31), respectively. SPSS `MATRIX` code for this approach appears in Appendix A.8.

Discussion

In homoscedastic linear models, the covariance matrix for the regression coefficients, $\mathbf{V}_{(\mathbf{b})}$, is:

$$\mathbf{V}_{(\mathbf{b})} = MSE(\mathbf{X}'\mathbf{X})^{-1} ; \tag{36}$$

and the diagonal of $\mathbf{V}_{(\mathbf{b})}$ contains the squared Standard Errors (i.e., variance of the regression coefficients), which can be re-expressed as:

$$SE_k^2 = MSE(\mathbf{X}'\mathbf{X})_k^{-1} = \frac{MSE}{SS_k(1-\mathrm{R}^2_{kk\prime})} = \frac{MSE}{SS_k(\mathrm{TOL}_k)} = \frac{MSE(\mathrm{VIF}_k)}{SS_k} ; \tag{37}$$

where $(\mathbf{X}'\mathbf{X})_k^{-1}$ is the k^{th} diagonal element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix and SS_k is the mean corrected Sum of Squares for the k^{th} x-variable. As can be seen in (37), standard VIFs can be interpreted as the degree to which the square of the Standard Error (SE_k) for each regression coefficient (b_k) in the model is inflated by multicollinearity (Marquardt, 1970). Equation (37) also shows that for each x-variable and SE_k only differs by the diagonal value, $(\mathbf{X}'\mathbf{X})_k^{-1}$. Holding the regression coefficient (b_k), *MSE*, and SS_k constant, VIF_k inflates the SE_k for any one test, t_k=(b_k/SE_k). This in turn yields a larger p-value which is less likely to reject the null hypothesis (H_0: $\beta_k = 0$) and lead to a potential Type 2 error (i.e., loss of power)[8]. For Reference Cell coding, both VIF_k and SS_k change with a change in the Reference group; thus, a regression coefficient that represents a pairwise contrast will retain its *SE* and p-value but have a different index of collinearity. Again, this reflects on the dubious nature of interpreting standard VIFs that come from related sets of regressors (e.g., coding schemes for categorical factors).

In the context of between-subjects designs with covariates (i.e., ANCOVA), GVIFs estimate multicollinearity for the omnibus test of a categorical factor that is represented by a related set of regressors (i.e., coding scheme). This points to the issue that there are two types of "dependency" when analyzing categorical variables in ANCOVA models. First, there is multicollinearity with the other covariates in the model. As previously mentioned, this occurs because of between-group mean differences on continuous

covariates and proportional differences on categorical covariates. For example, if there is collinearity, then the expected value of any of the other covariates depends on the categorical variable.[2]

Secondly, the groups in a between-subjects design are independent because membership to each category is mutually exclusive; however, there may be a dependency among statistical tests if a set *a priori* or *post hoc* contrasts are performed. If the contrasts are orthogonal to each other, then corrections for multiple testing such as the Bonferroni or Dunn-Šidák, which are based on an assumption of independence, should be used. If all pairwise comparisons are performed, then there is a known dependency among the test statistics. Type 1 error rate inflation due to the dependency created by performing all pairwise contrasts should be corrected with a multiple comparison procedure such as the Tukey-Kramer studentized range test. This form of dependency due to multiple testing of dependent contrasts is a completely separate issue than linear dependency in the **X** design matrix due to multicollinearity.

Conclusion

Tables 4 through 7 show that standard VIFs depend on the coding scheme used for a multiple *df* categorical variable in a linear regression model, thus demonstrating that these coding schemes create "artificial" collinearity (Fox, 2016, p.357), which supports the use a single estimate of GVIF in these situations (Fox & Monette, 1992). GVIFs are invariant to the choice of coding scheme for categorical variables and estimate multicollinearity for the omnibus test of terms represented by a related set of regressors (i.e., coding scheme). Unfortunately, R is the only statistical software that has packages compute and report GVIF. We have demonstrated how to compute GVIF using multivariate regression models. GVIF is calculated for sets of related regressors, in this case a set of indicator regressors for a categorical variable, but this method may be applied to other types of related regressors, such as polynomial and interaction terms. It should be noted that for generalized linear models (e.g., logistic regression), a slightly different approach to computing GVIFs is necessary (Fox, 2020).

Endnotes

1. Suppose a linear model with K=2 continuous regressors. In the presence of collinearity, the estimate of one variable's, x_1, impact on the dependent variable, y, while controlling for the other, x_2, tends to be less precise than if predictors were uncorrelated, r_{12}=0. If x_1 is highly correlated with x_2, then the usual interpretation of a regression coefficient is affected because x_1 and x_2 have a particular bivariate space. To elaborate, with a strong positive correlation (e.g., r_{12}=0.95), thre may not be large values of x_1 that coordinate with small values of x_2. Consequently, there may not be a set of observations for which all changes in x_1 are independent of changes in x_2, resulting in imprecise estimates of the effect of independent changes in x_1 and x_2.
2. There are several versions of string (character) variables for Ethnicity because SAS and SPSS make the last category, alphanumerically, the Reference group by default, whereas STATA and R make the first category the Reference group by default. These variables were created so that the reader can replicate the analyses in this article using the default settings in their preferred software.
3. To elucidate, in an Analysis of Covariance (ANCOVA) model, the categorial variable (G) and the continuous covariate (X) are orthogonal (G⊥X) if and only if the categories have identical means on the covariate, otherwise there is some degree of collinearity between G and X. This generalizes to proportional differences for categorical covariates.
4. There are other choices for the assumed standard deviation (SD) used as the denominator in Cohen's *d*. This includes the sample SD=8.74 or the root mean square error (RMSE) from a regression model. The unadjusted mean difference for the preparation program factor is 3.13. This would yield effect size estimates of d=0.313, using the assumed population SD of 10, of d=0.358, using the sample SD, and d=0.362, using the RMSE of 8.644 from an ANOVA. The covariate-adjusted mean difference of 3.55 could also yield effect sizes of d=0.406, using the sample SD of 8.74, or of d=0.518, using the ANCOVA RMSE of 6.858.
5. Suppose two regression coefficients with the same absolute value but different values for VIF as is the case with B in model (4) and A in model (5). In principle, the variable with the larger VIF might be expected to have a larger SE, because the SEs uses the same mean square error (MSE). This is not the case, however, for the Asian vs Black pairwise difference because the B and A dummy codes have different variances, and thus different SS_k (see eq. 37). Because linear regression models are invariant

to transformation, all *x*-variables could be standardized to have the same mean and variance, and thus the same SS_k. Performing analogous regression models, such as (4) vs (5), with the standardized *x*-variables will yield the same *MSE*, TOLs, VIFs, *t*-statistics, and *p*-values as original models; however, this transformation will change the scale of the variables, yielding different regression coefficients and *SE*s.

6. Although applying Helmert coding to the Ethnicity factor does not make a great deal of sense, it is used to demonstrate orthogonal coding. Table 2 shows one of $J!/2 = 12$ possible versions of a Helmert-type coding scheme for the Ethnicity factor that are orthogonal to each other based on a method shown in Pedhazur (1982, p. 324). The reader can verify that these vectors have means of zero and are uncorrelated, and thus, would have VIF and TOL of 1 in an ANOVA model.
7. Researchers often employ more complex regression terms to investigate whether a continuous variable has a non-linear relationship with the outcome by using polynomial regressors. For example, if the investigator suspect Age (*x*) has a non-linear relationship to the outcome, then quadratic and cubic polynomial terms could be constructed as x^2 and x^3, respectively. Since all values for Age are positive, *x*, x^2, and x^3 will be highly correlated creating a extremely large degree of collinearity. As mentioned, this could lead to a very unstable regression solution. As with choice of coding scheme with categorical factors, the correlations among a set of polynomial regressors in an explanatory variable *x* are affected by linear transformation of the *x*-values. It is common practice to subtract the mean from *x* prior to constructing polynomial regressors (i.e., centering) to lessen the collinearity among polynomial regressors. Fox and Monette (1992) contend that in this situation assessing collinearity for Age should be treated as a singular independent factor with 3 *df*s, instead of 3 separate variables with 1 *df* each.
   Similarly, researchers may choose to investigate whether the relationship of one factor is modified by another factor and construct interaction terms via cross-products of the variables. If two dummy coded binary variables (*g* and *z*) are used, then the cross-product (*gz*) will have some degree of collinearity with *g* and *z*, even if *g* and *z* are uncorrelated (orthogonal). If a cross-product interaction term is formed with a dummy coded binary variable (*g*) and a continuous variable (*x*), then *g*, *x*, and the cross-product (*gx*) will necessarily be correlated because when g=0, *gx* also equals 0. Although this "false" collinearity and the standard VIFs could be lessened by using a different coding scheme (i.e., Effect coding), Fox (2016) maintains that collinearity for such related sets of regressors should be evaluated with GVIFs.
8. In actuality, statistical power in a linear regression model is more complicated because multicollinearity, which leads to large VIFs, also tends to reduce the Full Model R^2 leading to a larger *MSE*.

References

Allison, P. D. (2012). *When Can You Safely Ignore Multicollinearity*? https://statisticalhorizons.com/multicollinearity/

Buteikis A. (2020). *Practical Econometrics and Data Science*. http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models*, 3rd Edition. Sage.

Fox, J. (2020). *Regression Diagnostics: An Introduction*, 2nd Edition. Sage.

Fox, J., & Monette, G. (1992). Generalized Collinearity Diagnostics. *Journal of the American Statistical Association*, *87*(*417*), 178–183.

Fox, J. & Weisberg, S. (2011) *An R Companion to Applied Regression*, 2nd Edition, Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. *12*(*1*), 55–67. doi:10.2307/1267351.

IBM SPSS Statistics for Windows, Version 28.0. (2021). IBM Corp., Armonk, NY.

Johnson, R., Jones, K., & Manley, D. (2018). Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. *Quality & Quantity*, *52*(*4*), 1957–1976.

Marquardt, D. W. (1970). Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. *Technometrics*, *12*(*3*), 591-612.

O'Brien, R. (2016). Dropping highly collinear variables from a model: why is it typically not a good idea? *Social Science Quarterly*, *98*(*1*), 360-375.

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction (2nd ed.). New York: Holt, Rinehart, & Winston.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

SAS/STAT 14.3. (2016). SAS Institute Inc., Cary, NC.

SAS/IML 14.3. (2016). SAS Institute Inc., Cary, NC.

Stata Statistical Software: Release 17. (2021) StataCorp., College Station, TX.

Tibshirani, R, (1996). Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B, 58*(*1*), 267–288.

Vanegas L. H., Rondón, L. M., & Paula, G. A. (2023). *glmtoolbox: Set of Tools to Data Analysis using Generalized Linear Models*. R package version 0.1.7. https://CRAN.R-project.org/package=glmtoolbox

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B, 67*(*2*), 301–320.

Send correspondence to:
T. Mark Beasley
University of Alabama at Birmingham
Email: mbeasley@uab.edu