Quality & Quantity (2007) 41:673–690
DOI 10.1007/s11135-006-9018-6

A Caution Regarding Rules of Thumb for Variance Inflation Factors

ROBERT M. O'BRIEN
Department of Sociology, University of Oregon, Eugene, OR 97408, USA
E-mail: bobrien@uoregon.edu

Abstract. The Variance Inflation Factor (VIF) and tolerance are both widely used measures of the degree of multi-collinearity of the *i*th independent variable with the other independent variables in a regression model. Unfortunately, several rules of thumb – most commonly the rule of 10 – associated with VIF are regarded by many practitioners as a sign of severe or serious multi-collinearity (this rule appears in both scholarly articles and advanced statistical textbooks). When VIF reaches these threshold values researchers often attempt to reduce the collinearity by eliminating one or more variables from their analysis; using Ridge Regression to analyze their data; or combining two or more independent variables into a single index. These techniques for curing problems associated with multi-collinearity can create problems more serious than those they solve. Because of this, we examine these rules of thumb and find that threshold values of the VIF (and tolerance) need to be evaluated in the context of several other factors that influence the variance of regression coefficients. Values of the VIF of 10, 20, 40, or even higher do not, by themselves, discount the results of regression analyses, call for the elimination of one or more independent variables from the analysis, suggest the use of ridge regression, or require combining of independent variable into a single index.

Key words: multi-collinearity, tolerance, variance inflation factors, variance of regression coefficients

1. Introduction

Collinearity can increase estimates of parameter variance; yield models in which no variable is statistically significant even though R_y^2 is large; produce parameter estimates of the "incorrect sign" and of implausible magnitude; create situations in which small changes in the data produce wide swings in parameter estimates; and, in truly extreme cases, prevent the numerical solution of a model (Belsley et al., 1980; Greene, 1993). These problems can be severe and sometimes crippling.

We use R_i^2 to represent the proportion of variance in the ith independent variable that is associated with the other independent variables in the model. It is an excellent measure of the collinearity of the ith independent variable with the other independent variables in the model. Tolerance for

the ith independent variable is 1 minus the proportion of variance it shares with the other independent variable in the analysis ($1 - R_i^2$). This represents the proportion of variance in the ith independent variable that is not related to the other independent variables in the model. The Variance Inflation Factor (VIF) is the reciprocal of tolerance: $1/(1 - R_i^2)$.[1] The VIF has an intuitive interpretation in terms of the effects of R_i^2 on the variance of the estimated regression coefficient for the ith independent variable.

Unfortunately practitioners often inappropriately apply rules or criteria that indicate when the values of VIF or tolerance have attained unacceptably high levels. Not uncommonly a VIF of 10 or even one as low as 4 (equivalent to a tolerance level of 0.10 or 0.25) have been used as rules of thumb to indicate excessive or serious multi-collinearity.[2] These rules for excessive multi-collinearity too often are used to question the results of analyses that are quite solid on statistical grounds.

When the VIF reaches these threshold levels, researchers may feel compelled to reduce the collinearity by eliminating one or more variables from their analysis; combining two or more independent variables into a single index;[3] resorting to Ridge Regression (a biased regression technique that can reduce the variance of the estimated regression coefficients); or, in the role of a manuscript reviewer, rejecting a paper because VIF exceeds a threshold value. This is the case, even when these remedies are not always suggested by the sources cited in note 2.

Below we address some interrelated problems that can arise when rules of thumb associated with particular levels of tolerance and/or VIF are used. (1) There is an exclusive focus on multi-collinearity when assessing the accuracy of regression coefficients and the "stability" of results from a regression analysis. (2) Cases in which the null hypothesis is rejected (or in which the accuracy of the estimated regression coefficient is sufficient) are treated in the same manner as cases in which the null hypothesis is not rejected (or in which the confidence interval for the regression coefficient is too wide to draw useful conclusions). (3) Researchers only examine those remedies to *the effects of multi-collinearity* that involve reducing multi-collinearity.

2. Exclusive Focus on Multi-collinearity

The VIF indicates how much the estimated variance of the ith regression coefficient is increased above what it would be if R_i^2 equaled zero: a situation in which the ith independent variable is orthogonal to the other independent variables in the analysis. VIF provides a reasonable and intuitive indication of the effects of multi-collinearity on the variance of the ith regression coefficient. Focusing on the effects of multi-collinearity on sample fluctuations of parameters, however, can seduce researchers into

ignoring other factors (besides multi-collinearity) that affect the stability of regression coefficients. These other factors can ameliorate or exacerbate the effects of multi-collinearity. Goldberger (1991), for example, notes the following quote from Johnston (1984: 250): "More data is no help in multi-collinearity if it is simply 'more of the same.' What matters is the structure of the X′X matrix, and this will only be improved by adding data which are less collinear than before." This statement is indisputable, if it refers to multi-collinearity by itself; but unless the collinearity is perfect, increasing the sample size (using more cases of the same sort) will reduce the variance of the regression coefficients. More of the same can certainly be helpful; a new analysis based on the additional cases of the same sort should provide more reliable point estimates and smaller confidence intervals.

Goldberger (1991) notes that while the number of pages in econometrics texts devoted to the problem of multi-collinearity in multiple regression is large the same books have little to say about sample size. Goldberger states: "Perhaps that imbalance is attributable to the lack of an exotic polysyllabic name for 'small sample size.' If so, we can remove that impediment by introducing the term *micronumerosity*" (Goldberger, 1991: 248–249).[4] His aim is not to introduce "micronumerosity" into the lexicon of statistics, but is to contest the predominant focus on multi-collinearity in relation to the inflation of the variance of regression coefficients and the "stability" of results from a regression analysis. When the focus is on the variance of a regression coefficient and the stability of result: the sample size, the proportion of the variance in the dependent variable associated with the independent variables, the variance of the independent variable whose coefficient is of concern, and the multi-collinearity of the independent variable of concern with the other independent variables are all important. Goldberger (1991: 251) notes, "In the classic regression (CR) model all of the consequences of multi-collinearity are reflected in $V(b) = \sigma^2 \mathrm{Q}^{-1}$ or in its unbiased estimator $\hat{V}(b) = \hat{\sigma}^2 \mathrm{Q}^{-1}$."[5]

3. Specifying the Effects of Several Factors on the Variance of the Regression Coefficients

Equation (1) represents the formula for the variance of the regression coefficient for the ith independent variable when the population residual variance is known (we show how this formula is derived in Appendix):

$$\sigma^2(b_i) = \frac{\sigma_\varepsilon^2}{\left(1 - R_i^2\right) \sum x_i^2}, \tag{1}$$

where $\sigma^2(b_i)$ the variance of the ith regression coefficient, σ_ε^2 the residual variance, R_i^2 the squared multiple correlation of the ith independent variable regressed on the other independent variables in the analysis, and

$\sum x_i^2$ is the sum of the squared deviations for the ith independent variable around its mean.

In (2), we substitute an unbiased estimate of the population residual variance for σ_ε^2 (see Appendix) yielding a formula for the *estimated* the variance of the ith regression coefficient:

$$\hat{\sigma}^2(b_i) = \frac{\left[\frac{(1-R_y^2)\times\sum(Y_i-\bar{Y})^2}{(n-k-1)}\right]}{(1-R_i^2)\times\sum x_i^2}, \tag{2}$$

where $\hat{\sigma}^2(b_i)$ is the estimated variance of the regression coefficient for the ith independent variable, R_y^2 is the squared multiple correlation of the dependent variable regressed on all of the independent variables in the model, $\sum(Y_i-\bar{Y})^2$ is the sum of the squared deviations of the dependent variable around its mean, n is the number of cases on which the analysis is based, and k is the number of independent variables in the analysis. As noted in the Appendix, as n increases the expected value of the term in square brackets in (2) remains an unbiased estimate of the residual variance. Below we use (2) to show the affects of both R_i^2 and R_y^2 on the estimated variance of the ith regression coefficient.

3.1. EFFECT OF THE VARIANCE INFLATION FACTOR

Multiplying the numerator and denominator of (2) by $1/(1-R_i^2)$ yields

$$\begin{aligned}\hat{\sigma}^2(b_i) &= \frac{\left[\frac{(1-R_y^2)\times\sum(Y_i-\bar{Y})^2}{(n-k-1)}\right]}{\sum x_i^2}\times\frac{1}{(1-R_i^2)}\\ &= \frac{\left[\frac{(1-R_y^2)\times\sum(Y_i-\bar{Y})^2}{(n-k-1)}\right]}{\sum x_i^2}\times \mathrm{VIF},\end{aligned} \tag{3}$$

where VIF, which is defined as $1/(1-R_i^2)$, indicates the multiplicative increase in the variance of the regression coefficient of the ith independent variable above what it would be if $R_i^2 = 0$.

There is a sense in which the VIF has a natural metric – comparing the effects of the proportion of variance a particular independent variable shares with the other independent variables to the situation in which it shares none of its variance with the other independent variables. This baseline, however, is likely to occur only in an experiment designed so that the correlation among the independent variables is zero.

3.2. EFFECT OF R_y^2

We use the situation in which the dependent variable is linearly unrelated to the independent variables in the model ($R_y^2 = 0$) as a "natural metric" for deriving a measure of the effects of R_y^2 on the variance of the estimated regression coefficients. This is the typical null hypothesis in most studies, and this measure highlights the relationship of explained variance (R_y^2) to the variance of the regression coefficients. Using the same general approach that we used to derive VIF. We divide the numerator and denominator of (2) by $1/(1-R_y^2)$ and simplify:

$$\hat{\sigma}^2(b_i) = \frac{\left[\frac{\sum(Y_i-\bar{Y})^2}{(n-k-1)}\right]}{\left(1-R_i^2\right)\times\sum x_i^2} \times \left(1-R_y^2\right) = \frac{\left[\frac{\sum(Y_i-\bar{Y})^2}{(n-k-1)}\right]}{\left(1-R_i^2\right)\times\sum x_i^2} \times \mathrm{VDF}\left(R_y^2\right). \quad (4)$$

Equation (4) indicates that increases in R_y^2 deflate the size of the estimated variance of the ith regression coefficient. We, therefore, label $(1-R_y^2)$, as a Variance Deflation factor (VDF): $\mathrm{VDF}(R_y^2)$. If $R_y^2 = 0.80$ then the proportion of the variance that is unexplained is 0.20, and the variance of estimated regression coefficient is 0.20 (one-fifth) as large as it would have been if the dependent variable had been linearly unrelated to the independent variables. The variance of the estimated regression coefficient has been reduced by 80%.

3.3. EFFECT OF THE SAMPLE SIZE

To facilitate our discussions of the effects of the sample size, we examine the expected value of $\sum x_i^2$ in (2), where $x_i = \left(X_i - \bar{X}_\cdot\right)$. We note that $E\left[\sum\left(X_i-\bar{X}_\cdot\right)^2/(n-1)\right] = \sigma_i^2$, where σ_i^2 is the population variance for the ith variable. Therefore, the expected value of $\sum x_i^2$ in (2) is $\sigma_i^2 \times (n-1)$. Conceptually, we may write the denominator for (2) as:[6]

$$\left(1-R_i^2\right)\times\sigma_i^2\times(n-1). \quad (5)$$

All other things being equal, increases in the value of n or σ_i^2 decrease the value the estimated variance of the ith regression coefficient.

There is no "natural metric" to use in describing the effects of shifts in the sample size on the variance of the regression coefficients, so we will compare shift in the sample size to a "baseline" sample size that we label as n_b. The comparison is then between a sample size that is n_b in comparison to a sample that is of size n_p. We define the ratio of $(n_p-1)/(n_b-1)$ as p. If p is greater than 1, the result is variance deflation; if p is less

than one the result is variance inflation. The Variance (inflation or deflation) Factor based on sample size can then be written as follows:

$$\mathrm{VF}(n) = \frac{1}{(n_p - 1)/(n_b - 1)} = \frac{1}{p}, \tag{6}$$

where $(1/p)$ indicates the effect of the change in sample size on the variance of the regression coefficient. If the baseline sample is 100 and the comparison sample is 397, then the estimated variance of the regression coefficients is expected to be one fourth $[(397-1)/(100-1)]$ as large as in the comparison sample as it is in the baseline sample (all other things being equal). Decreasing the sample size from 101 to 26 increases the variance of the regression coefficient by a factor of 4.

3.4. EFFECT OF THE VARIANCE OF THE INDEPENDENT VARIABLE

For the variance of the independent variable there again is no "natural metric," so we compare shifts in the variance of the ith independent variable to a "baseline" variance that we label as σ_{ib}^2. The comparison is to a variance that is that is q times as large as the baseline variance. If the comparison variance is denoted as $q \times \sigma_{ib}^2$, we can write the VF based on the variance of the independent variable as follows:

$$\mathrm{VF}\left(\sigma_i^2\right) = \frac{1}{q\,\sigma_{ib}^2/\sigma_{ib}^2} = \frac{1}{q}. \tag{7}$$

If the variance of the independent variable were four times larger than the baseline variance, the variance of the regression coefficient would be one fourth as large as the variance based on the baseline model. If the variance were half the size of the baseline variance, the variance of the regression coefficient would be increased by a factor of 2.

3.5. THE VALUE OF t AND $\hat{\sigma}^2(b_i)$

The value of t certainly does not affect the variance of the ith regression coefficient, but we will use the value of t and its dependence of the variance of the ith regression coefficient in the next section. The test for the statistical significance of a regression coefficient is typically a t-test based on the difference between the point estimate of the ith regression coefficient (b_i) and the null hypothesis value of $b_i[b_i(H_0)]$ divided by the estimated standard error of the ith regression coefficient.

$$t = \frac{b_1 - b_1(H_0)}{\sqrt{\hat{\sigma}^2(b_i)}}. \tag{8}$$

The difference between the null hypothesis value and the observed value does not affect the estimated variance of the regression coefficient or the confidence interval for an estimate – but it can affect the decision to reject or fail to reject the null hypothesis at a particular level of statistical significance (or whether the confidence interval includes the null hypothesis value).

If the variance of the ith regression coefficient is s times as large as in (8), the square root of its variance is the square root of s times as large as it is in (8). Therefore, if the variance of the ith regression coefficient is s times as large as it is in (8), t is $1/\sqrt{s}$ as large as it is in (8). If $\hat{\sigma}^2(b_i)$ is quadrupled, t is 1/2 $[=1/\sqrt{4}]$ as large as it is in (8); and if $\hat{\sigma}^2(b_i)$ is one-fourth as large, then t is doubled.

4. Variance Inflation Factors in Context

The previous section describes several factors that influence the estimated variance of the ith regression coefficient and shows how changes in these factors are reflected in the estimated variance of the regression coefficients. Knowing these factors and their effects indicates the degree to which they can be used to ameliorate the impact of multi-collinearity on the accuracy of estimated regression coefficients. In any given situation, however, it may or may not be possible to change these factors. Sometimes, for example, the size of the sample is set (all U.S. cities of size 100,000 or more in 1980). Dramatic changes in the amount of variance in the dependent variable explained by the independent variables typically are not easy to accomplish using well motivated models. Whatever the situation, however, the discussion in the previous section provides background on the degree to which several factors can offset the effects of multi-collinearity on the variance of regression coefficients.

Table I describes three hypothetical analyses to provide two of examples of how these effects can work in practice. The main goal of these examples is to put the interpretation of VIF into a context that does more than focus on the size of VIF and on the rules of 4 or 10 or some other arbitrary number designed to indicate "excessive multi-collinearity."

The situations we could depict are far more varied than those shown in Table I. We use a baseline analysis in which $R_i^2 = 0.20$, a VIF of 1.25; $R_y^2 = 0.40$, a VDF(R_y^2) of 0.60, the sample size is 100, and the t-value for the ith regression coefficient is 3.0 ($p < 0.01$).[7] We will assume that there are three independent variables and that $\sigma_i^2$is the same across all of the analyses.[8] Our purpose is to show that the rules of thumb associated with VIF (and tolerance), which focus on the variance of the ith regression

Table I. Baseline and comparison analyses and the effects of various factors on the estimated variance of regression coefficients.

Analysis	R_i^2	VIF	R_y^2	VDF (R_y^2)	n	VF(n)	t_i
Baseline	0.20	1.25	0.40	0.60	100	1.00	3.00
Comparison 1	0.95	20.00	0.90	0.10	397	0.25	3.67
Comparison 2	0.975	40.00	0.95	0.05	1585	0.0625	7.35

coefficient, *should be put into the context of the effects of other factors that affect the variance of the ith regression coefficient*.

The results from the baseline analysis in Table I that relate to the ith regression coefficient would not, by themselves, trouble most researchers. The ith regression coefficient is statistically significant ($p < 0.01$) and VIF is only 1.25 (well below the rules of thumb of 4 or 10). The sample size is not large, but seems adequate given the number of independent variables in the analysis. Researchers in most areas would not quibble with an R_y^2 of 0.40.

On the other hand, questions might be raised about the results from a regression analysis in comparison 1, since the VIF for that analysis is 20 (well above the level suggested by the rules of 4 or 10). Note that in this analysis, since the $R_y^2 = 0.90$, the VDF(R_y^2) is 0.10 rather than the 0.60 baseline model. All other things being equal, the variance of the regression coefficient in analysis 1 would be one-sixth ($= 0.10/0.60$) as large as in the baseline analysis. The sample size is $4 \times (n-1)_b$ larger in comparison 1 than in the baseline analysis, this factor alone would result in a variance of the regression coefficient in analysis 1 that is one-fourth as large as that for the baseline model (all other things being equal). These two factors are multiplicative in their reduction of the variance of the regression coefficients. The R_y^2 in analysis 1 would reduce the variance to one-sixth of what it is in the baseline model and that variance would be reduced by one-fourth by the sample size. If these two factors were the only differences between the baseline and comparison 1 analyses, the variance of the regression coefficients in analysis 1 would be 0.0417 ($= 1/6 \times 1/4 = 1/24$) as large as those in the baseline model.

We know from Table I, however, that all other things are not equal, since VIF for the comparison group is 20 while it is only 1.25 in the baseline analysis. VIF increases the variance of the ith regression coefficient for this comparison analysis, relative to the baseline analysis, by a factor of 16 ($= 20/1.25$). Overall, considering R_i^2, R_y^2, and the sample size, the variance of the ith regression coefficient in the comparison model is smaller by a factor of 0.67 ($= 16/24$) than that of the baseline model. If all other things were equal, the value of t in this case would be 3.67 [$= 3.00 \times (1/\sqrt{0.67})$].

Has the VIF of 20, which is based on the collinearity of the ith independent variable with the other independent variables, made us question the results in comparison 1 even though we are comfortable with the results from the baseline model? The t-value for the ith regression coefficient in the comparison 1 analysis is larger than in the baseline analysis and, correspondingly, has a confidence interval that is narrower for this coefficient. We should be more confident about the value of the ith regression coefficient in analysis 1 than in the baseline model. We could just as easily question the statistical significance of the relationship in the baseline model because it is based on an analysis in which R_y^2 is only 0.40 or in which the sample size is only 100. This would, of course, be inappropriate. These factors have already been taken into account in (1).

Comparison 2 is more extreme than comparison 1, but I should note that a colleague and I routinely analyze data with R_y^2 values in excess of 0.95. In comparison 2 VIF is 40 and the variance deflation factor due to R_y^2 is 0.05. The R_y^2 of comparison 2 reduces the variance of the regression coefficients in this analysis to 1/12 of the variance of the baseline analysis ($=0.05/0.60=1/12$). The number of cases in comparison 2 is $16\times(n-1)_b$, so that this factor by itself deflates the variance of the regression coefficients to 1/16 of the variance in the baseline analysis. Together these two effects result in a variance of the estimated regression coefficient in analysis 2 that is 0.0052 [$=(1/12)\times(1/16)=1/192$] the size of the variance in the baseline analysis (if all other things were equal). But again, not all other things are equal. In the baseline analysis VIF is only 1.25, while the VIF for the comparison analysis is 40. This inflates the variance of the ith regression coefficient in comparison 2 by a factor of 32 [$=40/1.25$] relative to the baseline model. Still the variance of the ith regression coefficient in comparison 2 is only 16.67% ($=32/192$) as large as it is in the baseline model. If all other things were equal, the t-value would be 7.35 for the comparison analysis. Again, we might ask whether we should be more comfortable with the results from the baseline model compared to those from comparison analysis, because VIF in the comparison 2 analysis is 40.[9]

Even with VIF values that greatly exceed the rules of 4 or 10, one can often confidently draw conclusions from regression analyses. How confident one can be depends upon the t-values and/or confidence intervals, which the variance of the regression coefficients help generate. The practice of automatically questioning the results of studies when the variance inflation factor is greater than 4, 10, or even 30 is inappropriate. Using VIF for this purpose is as inappropriate as questioning studies that are based on R_y^2 values that are less than "the rule of 0.40." Or as inappropriate as questioning the results of studies based on sample sizes less than 200, because they do not meet "the rule of 200."

5. Treating Rejection and Non-Rejection in the Same Manner

To the extent that a researcher's concern is the null hypothesis: the situation in which the null hypothesis is rejected and the situation in which it is not rejected are not the same in terms of our concern about the values of R_i^2, R_y^2, n, and σ_i^2. Similarly, to the extent that a researcher's concern is with the confidence interval: the situation in which the confidence interval is small enough for the researcher's purposes is not the same as the situation in which it is too wide for the researcher's purposes. When the confidence interval is small enough or the null hypothesis has been rejected even though R_y^2 is small and/or R_i^2 is large and/or σ_i^2 is small and/or n is small; we have "dodged a bullet." In this situation, we have rejected the null hypothesis or have an accurate enough estimate of the ith regression coefficient in the face of: a high degree of multi-collinearity, a model that explains little of the variance in the dependent variable, modest variance of the ith independent variable, and/or a small sample size.

The reason for this asymmetry of concern can be seen clearly if we focus on the power of a statistical test. Whether or not the null hypothesis is rejected, a large R_i^2, small R_y^2, small n, and small σ_i^2 reduce the statistical power of an analysis. Since power is the probability of rejecting the null hypothesis when the null hypothesis is false, it depends not only on the variance of the ith regression coefficient (which R_i^2, R_y^2, n, and σ_i^2 affect), but also on the level of statistical significance chosen and value of the population parameter being estimated (the difference between this population value and the null hypothesis value). When the null hypothesis is not rejected or the confidence interval is too wide, then anything that increases the variance of the regression coefficient is crucially important and one may argue that the power of the hypothesis test was too low because of one or more of these factors.

Belsley et al. (1980) also note the asymmetry between these situations. After more than 30 pages of discussing various ways of detecting and assessing the damage done by multi-collinearity they state (1980: 116),

> If s^2 [our $\hat{\sigma}_\varepsilon^2$] is sufficiently small, it may be that particular var(b_k)'s [our $\hat{\sigma}^2(b_i)$] are small enough for specific testing purposes in spite of ... near collinearity... Thus, for example, if an investigator is only interested in whether a given coefficient is significantly positive, and is able, even in the presence of collinearity, to accept that hypothesis on the basis of the relevant *t-test*, then collinearity has caused no problem. Of course, the resulting forecasts or point estimates may have wider confidence intervals than would be needed to satisfy a more ambitious researcher, but for the limited purpose of the test of significance initially proposed, collinearity has caused no practical harm

> ... These cases serve to exemplify the pleasantly pragmatic philosophy that collinearity does not hurt so long as it does not bite.

We subscribe to this philosophy and have taken the pragmatic approach of focusing our concern on the effects of collinearity (and other factors) on the variance of the estimated regression coefficients. If one is interested in the effect of multi-collinearity on the ith regression coefficient, then VIF is a perfectly adequate indicator of the degree of multi-collinearity. If a regression coefficient is statistically significant even when there is a large amount of multi-collinearity – it is statistically significant in the "face of that collinearity." It is no more appropriate to question its statistical significance because there is multi-collinearity than to question a statistically significant relationship (at a specified level) because the variance explained by the model is low.

6. Suggested Remedies for Reducing Collinearity

One of the problems with focusing predominantly on the inflation of the variance of the regression coefficient due to multi-collinearity (VIF) is that it leads researchers to focus their attention on reducing multi-collinearity as the solution to the problem of an inflated variance for a regression coefficient. The literature includes several suggestions that directly address the problem of reducing multi-collinearity. Such attempts, however, may do more harm than good.

One suggestion is to respecify the model by eliminating one or more of the independent variables that are highly correlated with the other independent variables. The problem with this solution is that dropping X_j from the equation means that the ith regression coefficient no longer represents the relationship between the Y and X_i controlling for X_j and any other independent variables in the model. The model being tested has shifted, and this often means that the theory being tested by the model has changed. Simply dropping X_j because it is highly correlated with the X_i or using step-wise regression to select variables (typically the selected variables are not "too highly correlated" with each other) leads to a model that is not theoretically well motivated.

At times, however, it may be reasonable to eliminate or combine highly correlated independent variables, but doing this should be theoretically motivated. For example if we were to regress student GPA on whether students come from an intact home, their sex, their score on the Iowa Achievement Test (IAT) and their score on the Oregon Achievement Test (OAT), we might find that much of the variance in the IAT was shared with the other independent variables in the analysis and that much of the variance in the OAT was shared with the other independent variable – primarily based on a strong

relationship between the IAT and the OAT. In this case, these two variables probably are both measuring academic achievement and to have them both in the same regression model is to commit the partialling fallacy (Gordon, 1968); that is, controlling the relationship between an independent variable and a dependent variable for itself.[10] In this example, the specification of the equation that includes both the IAT and the OAT would mean that we essentially were asking about the relationship was between GPA and Achievement Test Scores controlling for Achievement Test Scores. If both of these measures of achievement were conceptually similar, we could combine them into a single measure and use the newly created variable in the analysis. This would solve the collinearity problem created by the high correlation between these two variables and typically give us a more reliable estimate of student achievement. Alternatively, we could drop one of these variables from the analysis (if they measure the same thing) and still control the other relationships in the analysis for achievement.

Ridge regression is yet another technique used to reduce collinearity. This technique biases the estimated regression coefficients, but reduces the level of multicollinearity. Obenchain (1977) notes, however, that under the assumptions of the linear model the confidence intervals centered at the least-squares estimate, which are unbiased estimates of the regression coefficients, are the narrowest possible estimates at their selected levels of confidence.

7. Discussion

The VIF (and tolerance) is based on the proportion of variance the ith independent variable shares with the other independent variables in the model. This is a measure of the ith independent variable's collinearity with the other independent variables in the analysis and is connected directly to the variance of the regression coefficient associated with this independent variable. One reason for the popularity of VIF as a measure of collinearity is that it has a clear interpretation in terms of the effects of collinearity on the estimated variance of the ith regression coefficient: a VIF of 10 indicates that (all other things being equal) the variance of the ith regression coefficient is 10 times greater than it would have been if the ith independent variable had been linearly independent of the other independent variable in the analysis. Thus, it tells us how much the variance has been inflated by this lack of independence.

Rules of thumb for values of VIF have appeared in the literature: the rule of 4, rule of 10, etc. When VIF exceeds these values, these rules often are interpreted as casting doubts on the results of the regression analysis. With high values of VIF (inflated standard errors of regression

coefficients), it is possible to have no independent variable that is statistically significant even though R_y^2 is relatively large. Given large standard errors, parameter estimates may be so variable (have such large standard errors) that they are in the opposite direction of established theory or they may be of implausible magnitude. Large standard errors mean that small changes in the data can produce wide swings in parameter estimates. All of these conditions are associated with increases in the estimated variance of the regression coefficients.[11]

We demonstrate that the rules of thumb associated with VIF (and tolerance) need to be interpreted in the context of other factors that influence the stability of the estimates of the ith regression coefficient. These effects can easily reduce the variance of the regression coefficients far more than VIF inflates these estimates even when VIF is 10, 20, 40, or more. Importantly, concern with the effects of variance inflation is different in situations in which we reject the null hypothesis or when the confidence interval around the ith regression coefficient is sufficiently small for the purposes of the study than in situations in which the null hypothesis is not rejected or when the confidence interval around the ith regression coefficient is too broad for the purposes of the study. In the former case, we have found a statistically significant result, or found a narrow enough confidence interval, in the face of variance inflation. In the latter case, we may well have been hurt by the increased variance associated with the ith regression coefficient.

Appendix: Derivation of the Formula for the Variance of the Regression Coefficients

The standard formula for the variance–covariance matrix of the regression coefficients is:

$$\sigma\left(b_i\, b_j\right) = \sigma_\varepsilon^2 \left(X'X\right)^{-1}, \tag{9}$$

where $\mathbf{X}$ is an n by $k+1$ matrix with the first column consisting of ones and the next k columns consisting of the values of k independent variables, and σ_ε^2 is the population variance of the residuals. The derivation of (9) can be found in Greene (1993). Below we use (9) to derive (1) and (2) in the main text. These equations are then used to describe the effects of multi-collinearity and other factors on the variance of the ith regression coefficient.

Our concern is with the main diagonal elements of the inverse matrix in (9), specifically the final k elements along the main diagonal that correspond to the k regression coefficients (the first element represents the variance of the intercept term). Let i represent the ith independent variable and a repre-

sent the remaining independent variables. Greene (1993: 246) shows that the variance of the ith regression coefficient can be represented as:

$$\sigma^2 (b_{i.a}) = \sigma_\varepsilon^2 \left(\mathbf{X}_i' \mathbf{M}_a \mathbf{X}_i\right)^{-1}, \tag{10}$$

where $\mathbf{X}_i$ is the n by 1 vector representing the values of the n observations on the ith independent variable and

$$\mathbf{M}_a = \mathbf{I} - \mathbf{X}_a \left(\mathbf{X}'_a \mathbf{X}_a\right)^{-1} \mathbf{X}'_a, \tag{11}$$

where $\mathbf{X}_a$ is an n by k matrix with the first row consisting of a column vector of n ones to represent the intercept and the $k-1$ remaining columns containing all of the independent variables except for the ith independent variable. $\mathbf{M}_a$ is the "residual maker" for ith independent variable; that is, when $\mathbf{M}_a$ is post-multiplied by the $\mathbf{X}_i$ vector (or pre-multiplied by $\mathbf{X}_i'$), it produces a vector of the n residuals of ith independent variable regressed on the other independent variables and the constant term. When $\mathbf{M}_a$ pre-multiplied by $\mathbf{X}'_i$ and post-multiplied by $\mathbf{X}_i$ it produces the variance of the ith independent variable conditional on the other independent variables in the model. σ_ε^2 is the residual variance from the regression analysis of the dependent variable on all of the independent variables vectors and the constant vector.

Eliminating the $(\cdot a)$ notation after b_i in (10), since the control for all of the other independent variables in the model is understood we write:

$$\sigma^2 (b_i) = \sigma_\varepsilon^2 \left(\mathbf{X}'_i \mathbf{M}_a \mathbf{X}_i\right)^{-1}. \tag{12}$$

For convenience of presentation, we allow $\mathbf{X}_i$ to be centered around its mean and write this mean centered vector as $\mathbf{x}_i$.[12] Letting $\mathbf{xx}^{ii}$ represent the inverse of the ith diagonal element of the inverse matrix, we may write from (10) and (11):

$$\mathbf{xx}^{ii} = \left(\mathbf{x}_i' \mathbf{x}_i - \mathbf{x}_i' \mathbf{X}_a \left(\mathbf{X}_a' \mathbf{X}_a\right)^{-1} \mathbf{X}_a' \mathbf{x}_i\right)^{-1}. \tag{13}$$

Dividing and multiplying (13) by $\mathbf{x}'_i \mathbf{x}_i$ and rearranging the terms we write:

$$\mathbf{xx}^{ii} = \left(\mathbf{x}_i' \mathbf{x}_i \left[1 - \frac{\mathbf{x}_i' \mathbf{X}_a \left(\mathbf{X}_a' \mathbf{X}_a\right)^{-1} \mathbf{X}_a' \mathbf{x}_i}{\mathbf{x}_i' \mathbf{x}_i}\right]\right)^{-1}. \tag{14}$$

The expression within the square brackets is equal to 1 minus the variance of the ith independent variable associated with the regression divided by the total variance of the ith independent variable; that is, proportion of

variance in the ith independent variable that is associated with the other independent variables. We can write (14) without using matrix notation as:

$$xx^{ii} = \frac{1}{\left(1 - R_i^2\right) \sum x_i^2}, \tag{15}$$

where R_i^2 is the proportion of the variance in the ith independent variable that is associated with the other independent variables in the analysis and $\sum x_i^2$ is the sum of squares for the ith independent variable. Using (12) and (15):

$$\sigma^2 (b_i) = \frac{\sigma_\varepsilon^2}{\left(1 - R_i^2\right) \sum x_i^2}. \tag{16}$$

To estimate the variance of the regression coefficients, we need an unbiased estimate of σ_ε^2. We choose a form of the standard unbiased estimate of σ_ε^2, which shows the dependence of the residual variance on the proportion of the sums of squares for the dependent variable that is explained by the regression equation. The estimate we use is:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_i \left(Y_i - \hat{Y}_i\right)^2}{n - k - 1} = \frac{\left(1 - R_y^2\right) \times \sum \left(Y_i - \bar{Y}\right)^2}{n - k - 1}, \tag{17}$$

where n is the sample size, k the number of independent variables in the analysis, and R_y^2 is the squared multiple correlation of the dependent variable regressed on all of the other independent variables in the analysis.

Substituting $\hat{\sigma}_\varepsilon^2$ in (17) for σ_ε^2 in (16) yields (18). This equation provides the unbiased estimate of the variance of the ith regression coefficient:[13]

$$\hat{\sigma}^2 (b_i) = \frac{\left[\frac{\left(1 - R_y^2\right) \times \sum \left(Y - \bar{Y}\right)^2}{n - k - 1}\right]}{\left(1 - R_i^2\right) \times \sum x_i^2}. \tag{18}$$

Importantly, we note that as n increases the expected value of the term in square brackets in (18) does not change. It remains an unbiased estimate of the residual variance.

Notes

1. We define VIF (and tolerance) in terms R_i^2 (based on the variance of the ith independent variable around its mean that is explained by the other independent variables in the model). This defines a measure of collinearity for the centered values of the independent variables. This is standard practice, but see Belsley (1984), Cook (1984), Gunst (1984), Snee and Marquardt (1984), and Wood (1984) for an interesting exchange on this topic. For most work in the social sciences the centered VIF (and tolerance) are the

most appropriate ones for making inferences *about the relationships between the independent and dependent variables in the range for which we have data*. Extrapolation outside of that range (including extrapolation to the intercept of the raw data) must, as always, be done with much caution.

2. Menard (1995: 66) states "A tolerance of less than 0.20 is cause for concern; a tolerance of less than 0.10 almost certainly indicates a serious collinearity problem." Since VIF is the inverse of tolerance a tolerance of 0.20 corresponds to the rule of 5 and a tolerance of 0.10 to the rule of 10. Neter et al. (1989: 409) state "A maximum VIF value in excess of 10 is often taken as an indication that multi-collinearity may be unduly influencing the least square estimates." Hair et al. (1995) suggest that a VIF of less than 10 are indicative of inconsequential collinearity. Marquardt (1970) uses a VIF greater than 10 as a guideline for serious multi-collinearity. Mason et al. (1989) cite a VIF of greater than 10 as reason for concern. The STATA manual (StataCorp 1997: 390) notes: "However, most analysts rely on informal rules of thumb applied to VIF (see Chaterjee and Price 1991). According to these rules, there is evidence of multi-collinearity if 1. The largest VIF is greater than 10 (some chose the more conservative threshold value of 30). 2. The mean of all of the VIF's is considerably larger than 1." Kennedy (1992: 183) states that "for standardized data $\mathrm{VIF}_i > 10$ indicates harmful collinearity."
3. This decision might at times justifiable on theoretical grounds, for example, if the variables measure the same underlying concept.
4. In the same spirit, we could introduce the terms *hypoheterogeneity* to denote situations where the independent variable of interest has a too small a variance and *hyperscedasticity* to denote situations when the regression equation explains too little variance in the dependent variable and leaves us with too large a standard error of estimate. As noted below, each of these factors affects the reliability of estimates of individual regression coefficients in a regression analysis.
5. In Appendix, $\sigma_\varepsilon^2 (X'X)^{-1}$ corresponds to $V(b) = \sigma^2 \mathrm{Q}^{-1}$ and $\hat{\sigma}_\varepsilon^2 (X'X)^{-1}$ corresponds to $\hat{V}(b) = \hat{\sigma}^2 \mathrm{Q}^{-1}$.
6. We say "conceptually," because we have substituted the expected value of $\sum x_i^2$ for $\sum x_i^2$ in the denominator of (2), while leaving all of the other terms in (2) in the form of their sample realizations.
7. We could have used a baseline model in which R_i^2 and R_y^2 are both zero and it would not interfere with our demonstration. An $R_i^2 = 0$ yields a VIF that is the standard one for comparisons (an independent variable that is orthogonal to all of the other independent variables in the analysis). An $R_y^2 = 0$ also fits with our "natural metric" for R_y^2. We choose the baseline values of $R_i^2 = 0.20$ and $R_y^2 = 0.40$, because they are probably more typical of the values we find in social research.
8. We hold σ_i^2 constant between the baseline and comparison models. We think it is important to keep this factor in mind, especially in experimental situations where the variance of the independent variable is determined by the experimental manipulation or in situations in which the variance of the independent variable might be reduced due to sampling choices. Often the variance of the independent variable is beyond the control of the researchers.
9. Some authors note some or part of these tradeoffs, especially between multi-collinearity and the variance explained in the dependent variable. For example, Freund and Wilson (1998) suggest comparing the values of R_i^2 and R_y^2. Judge et al. (1985) note that some people use as a criterion that a problem exists if the simple correlation between two independent variables exceeds the R_y^2. Note that we have not included the problem of reduced variance in the ith independent variable, $\mathrm{VF}(\sigma_i^2)$, in our comparisons in Table I. Such reductions can occur in the context of restrictive sampling, for example,

the variance of achievement test scores is likely to much greater in high school than in graduate school. This increases the standard error of the regression coefficient for (for example) grade point average regressed on achievement test scores in graduate school over high school (all other things being equal).

10. Gordon's article is literally a citation classic (Gordon, 1987). As Gordon notes (Gordon, 1987: 18) "Often my article is misconstrued as being about multi-collinearity – mainly a statistical issue – when it really takes up where that problem leaves off. Since my concerns were abstractly substantive, my critique should be viewed as dealing with a general form of what econometricians call specification error." He examines the effects on coefficient estimates that occur when several variables from one domain of content are contained in the analysis (repetitiveness), when they are highly correlated with each other (redundancy), when their correlations with the dependent variable are different. But, as he notes, these effects need to be addressed though theory that should help the researcher form the "correct" (or best) specification for the regression model.
11. If the tolerance is zero, we have a case of linear dependence between the ith independent variable and the other independent variables and the solutions to the regression equation will not be unique. It could happen that solutions will not be possible if the VIF is extremely high – but this condition is unlikely using real data.
12. Note that when we residualize a raw score by subtracting its predicted value (based on OLS regression on a set of other variables including a column of ones for the intercept), we obtain a set of residuals: $\mathbf{X}_1 - \hat{\mathbf{X}}_1 = \mathbf{e}_1$. This set of residuals is equivalent to the set of residuals we would obtain if we residualized the mean centered scores on the same variable on their predicted value (based on OLS regression on the same set of other variables including a column of ones for the intercept):$\mathbf{x}_1 - \hat{\mathbf{x}}_1 = \mathbf{e}_1$. In either case, the regression takes into consideration the mean of the dependent variable; in the case of the mean centered dependent variable this mean is zero.
13. This form of the equation for the variance of the ith regression coefficient can be found in Fox (1997: 121); Greene (1993: 268).

References

Belsley, D. A. (1984). Demeaning conditioning diagnostics through centering. *The American Statistician* 38: 73–82.

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

Chatterjee, S. & Price B. (1991). *Regression Analysis by Example*, 2nd edn. New York: Wiley.

Cook, R. D. (1984). Comment on demeaning conditioning diagnostics through centering. *The American Statistician* 38: 78–79.

Freund, R. J. & Wilson, W. J. (1998). *Regression Analysis: Statistical Modeling of a Response Variable*. San Diego: Academic Press.

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage.

Goldberger, A. S. (1991). *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Gordon, R. A. (1968). Issues in multiple regression. *American Journal of Sociology* 73: 592–616.

Gordon, R. A. (1987). Citation-classic – issues in multiple regression. *Current Contents/Social and Behavioral Sciences* 36: 18–18.

Greene, W. H. (1993). *Econometric Analysis*, 2nd edn. New York: Macmillan.

Gunst, R. F. (1984). Toward a balanced assessment of collinearity diagnostics. *The American Statistician* 38: 79–82.

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). *Multivariate Data Analysis*, 3rd edn. New York: Macmillan.

Johnston, J. (1984). *Econometric Methods*. New York: McGraw-Hill.

Judge, G. G., Griffths, W. E., Hill E. C., Lutkepöhl, H. & Lee, T. C. (1985). *The Theory and Practice of Econometrics*, 2nd edn. New York: Wiley.

Kennedy, P. (1992). *A Guide to Econometrics*. Oxford: Blackwell.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics* 12: 591–256.

Mason, R. L., Gunst, R. F. & Hess, J. L. (1989). *Statistical Design and Analysis of Experiments: Applications to Engineering and Science*. New York: Wiley

Menard, S. (1995). *Applied Logistic Regression Analysis: Sage University Series on Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: Sage.

Neter, J., Wasserman, W. & Kutner, M. H. (1989). *Applied Linear Regression Models*. Homewood, IL: Irwin.

Obenchain, R. L. (1977). Classical *F*-tests and confidence intervals for ridge regression. *Technometrics* 19: 429–439.

Snee, R. D. & Marquardt, D. W. (1984). Collinearity diagnostics depend on the domain of prediction, and model, and the data. *The American Statistician* 38: 83–87.

StataCorp (1997). *Reference Manual A-F* (*Release 5*). College Station, TX: Stata Press.

Wood, F. S. (1984). Effect of centering on collinearity and interpretation of the constant. *The American Statistician* 38: 88–90.