Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture 2012 - 24th Annual Conference Proceedings

VARIANCE INFLATION FACTORS IN REGRESSION MODELS WITH DUMMY VARIABLES

Leigh Murray

Hien Nguyen

Yu-Feng Lee

Marta D. Remmenga

David W. Smith

See next page for additional authors

Follow this and additional works at: https://newprairiepress.org/agstatconference

 Part of the Agriculture Commons, and the Applied Statistics Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.

Recommended Citation

Murray, Leigh; Nguyen, Hien; Lee, Yu-Feng; Remmenga, Marta D.; and Smith, David W. (2012). "VARIANCE INFLATION FACTORS IN REGRESSION MODELS WITH DUMMY VARIABLES," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1034

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.

Author Information

Leigh Murray, Hien Nguyen, Yu-Feng Lee, Marta D. Remmenga, and David W. Smith

This is available at New Prairie Press: https://newprairiepress.org/agstatconference/2012/proceedings/12

VARIANCE INFLATION FACTORS IN REGRESSION MODELS WITH DUMMY VARIABLES

Leigh Murray[1], Hien Nguyen[2], Yu-Feng Lee[2], Marta D. Remmenga[3], and David W. Smith[1]

[1]Department of Statistics, Kansas State University, Manhattan, KS 66505; [2]Department of Economics and International Business, Las Cruces, NM 88003; [3]Centers for Epidemiology and Animal Health, USDA-APHIS-Veterinary Services, Fort Collins, CO 80526

Abstract

Variance Inflation Factors (VIFs) are used to detect collinearity among predictors in regression models. Textbook explanation of collinearity and diagnostics such as VIFs have focused on numeric predictors as being "co-linear" or "co-planar", with little attention paid to VIFs when a dummy variable is included in the model. This work was motivated by two regression models with high VIFs, where "standard' interpretations of causes of collinearity made no sense. The first was an alfalfa-breeding model with two numeric predictors and two dummy variables. The second was an economic model with one numeric predictor, one dummy and the numeric x dummy cross-product. This paper gives formulas for VIFs for several regression models with a dummy variable which indicate that these VIFs are functions of the numeric predictors' means, sums of squares and sample sizes within the two dummy groups. The economic regression model is also presented to illustrate how high VIFs occurred in this data. Researchers should be cautious in using high VIFs as a reason for deleting predictors in general but especially if dummy variables are involved. It is recommended that collinearity diagnostics be applied to the numeric predictors first to check for collinearity without the influence of any dummies, then add dummy variables in one at a time to see their effect on VIFs.

Keywords: collinearity, multicollinearity, indicator variable

1. Introduction

Variance Inflation Factors (VIFs) are used to detect collinearity (also called multicollinearity) among predictors in a multiple linear regression model (Belsley, et al. 1980). High VIFs reflect an increase in the variances of estimated regression coefficients due to collinearity among predictor variables, over variances obtained when predictors are orthogonal. [Note that in the context of this paper, by "collinearity" we specifically mean that the columns of the regression X-matrix have approximate or near linear dependencies; we are not considering other potential definitions of collinearity that may be defined with respect to unbalanced analysis of variance models.] Models with collinearity thus have estimators with lower precision, with consequent problems in testing hypotheses and forecasting (Marquardt 1970; Belsley et al. 1980; Fox and Monette 1992; Kutner, et al. 2004).

Many familiar regression textbooks (e.g., Mendenhall and Sincich 2004, Kutner et al. 2004, Kleinbaum et al., 1998, Graybill and Iyer 1994) illustrate collinearity and discuss diagnostics only in the context of numeric predictors, with an implicit assumption that results for dummy variables (or indicator variables) are equivalent. Indeed, in many regression examples where there are both numeric and dummy variables, little distinction is made between the two with respect to variable selection, collinearity diagnostics, and residual diagnostics like DFFITS and DFBETAS. This is in contrast to related Analysis of Covariance models where there is a clear hierarchy of importance or focus between the treatment factor (typically of most interest) and the covariate(s) (often of interest only as they make inferences on treatments more precise and powerful).

This work described in this paper was motivated by two statistical consulting projects where relatively simple regression models containing both numeric and dummy predictors were found to display high VIFs that didn't make obvious sense in the context of the two situations.

The first example involves data from an alfalfa breeding program at New Mexico State University (personal communication, Hem Bhandari and Ian Ray, Dept of Plant and Environmental Sciences). The overall experiment was a ½ diallel with nine parents and 36 crosses. The regression model of interest was developed to predict yield of the crosses based on two numeric predictors and two dummy variables. The two numerics were mid-parent heterosis (MPH, based on the parental yield data) and genetic distance (GD, measured by AFLP). The two dummies variables were based on whether parents agreed (1) or differed (0) with respect to respect to the categorical characteristics of winter hardiness and fall regrowth. Collinearity diagnostics, performed by a very conscientious Hem Bhandari, showed high VIFs for all predictors in the full model, but a plot of MPH versus GD did not reveal much "co-linear" behavior. Bivariate plots of other pairs of predictors were likewise unhelpful.

The second example (even simpler than the alfalfa data) is from the research of the third author of this paper. The initial regression model contained one numeric predictor, one dummy variable and their cross-product. The response was annual per capita consumption in Taiwan (R.O.C.) from 1976 to 2004. The numeric predictor was annual per capita income, and the dummy variable was defined as 1 for the "pre-reform" stage (1996-1999) and 0 for the "post-reform" stage (2000-2004). Initial model fitting showed a highly significant model and very high R-square but with all three regression coefficients being nonsignificant. Doing the obvious thing—to check for collinearity—revealed that the three VIFs for this model were all larger than 5800! To understand what was causing the high VIFs in this data+model, the second author obtained formulas for the VIFs for two models as her Experimental Statistics masters project (Nguyen 2008; models 1 and 2 below). Her work was then extended to try to illuminate what might cause high VIFs in more complicated models with dummy variables, like the alfalfa yield model.

This paper presents VIF formulas for four simple regression models with both numeric and dummy variables: 1) one numeric and one dummy; 2) one numeric, one dummy, and their cross-product; 3) Two numerics (the standard simple collinearity example) and 4) two numerics and one dummy. The outline is as follows. Section 2 provides a brief review of Variance Inflation Factors. Section 3 provides formulas for the X -matrix, the correlation matrix R, and the VIFs for the four models given above. In addition, Section 3 gives conditions for VIFs being "large" or equal to 1 and for $R = \mathrm{I}$. Section 4 examines the economic example in more detail, to

illustrate the VIF formulas for Models 1 and 2. Finally, Section 5 provides some general conclusions about using VIFs in regression models with dummy variables.

2. A Brief Review of Variance Inflation Factors

As mentioned previously, many regression textbooks illustrate collinearity using only numeric predictors, showing the observed values of two predictors being literally "co-linear" or three predictors being "co-planar". Diagnostics include pairwise correlation coefficients (for a collinearity involving only two predictors) and condition indices and VIFs (Marquardt 1970) for multi-variable relationships. In this paper, we focus on VIFs, because VIF formulas are given for each predictor, which is supposed to identify that predictor's contribution to a collinearity problem. It is, however, likely that condition indices suffer from similar issues as VIFs.

For the multiple regression model with p predictors, X_i, $i=1,\ldots,p$, VIFs are the diagonal elements (r^{ii}) of the inverse of the correlation matrix R_{pxp} of the p predictors (Chatterjee and Price 1977; Belsley et al. 1980). The VIF for the i^{th} predictor variable can be expressed by

$$VIF_i = r^{ii} = \frac{1}{1 - R_i^2}, i = 1,..., p,$$

where R_i^2 is the multiple correlation coefficient of the regression between X_i and the remaining $p-1$ predictors.

Belsley et al. (1980) pointed out that there is not a clear cutoff point to distinguish between "high" and "low" VIFs. Several researchers (e.g., Hocking and Pendelton 1983; Craney and Surles 2002) have suggested that the "typical" cutoff values (or rules of thumb) for "large" VIFs of 5 or 10 are based on the associated R_i^2 of 0.80 or 0.90, respectively. O'Brien (2007) recommended that well-known VIF rules of thumb (e.g., VIFs greater than 5 or 10 or 30) should be treated with caution when making decisions to reduce collinearity (like eliminating one or more predictors) and indicated that researchers should also consider other factors (e.g., sample size) which influence the variability of regression coefficients.

3. Four Regression Models and Their VIFs

3.1 General Notation Used for All Four Models

In this section, we define general notation and statistics that will be used with each model. First, the scalars y_{0j}, x_{0j}, and w_{0j}, $j = 1,..., n_0$, denote observations for the response and first (x) and second (w) numeric predictors, respectively, for observation j in the set of the n_0 observations for which the dummy variable is zero. Similarly, y_{1j}, x_{1j}, and w_{1j}, $j = 1,..., n_1$, denote the analogous values of the n_1 observations for which the dummy variable is one. The total number of observations is denoted by $n = n_0 + n_1$. Necessary scalar quantities, for the predictor x, are sample means and sums of squares, for each group and overall, denoted as:

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i},\ i = 0,1,$$

$$\bar{x} = \frac{\sum_{i=0}^{1}\sum_{j=1}^{n_i} x_{ij}}{n} = \frac{n_0\bar{x}_0 + n_1\bar{x}_1}{n},$$

$$SS_{x_i} = \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2,\ i = 0,1,$$

$$SS_x = \sum_{i=0}^{1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2 = SS_{x_0} + SS_{x_1} + \frac{n_0 n_1}{n}(\bar{x}_1 - \bar{x}_0)^2,$$

with analogous quantities being defined for the second predictor, *w*. In addition, the sum of cross-products between *x* and *w* is denoted as

$$SS_{xw} = \sum_{i=0}^{1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})(w_{ij} - \bar{w}) = SS_{x_0w_0} + SS_{x_1w_1} + \frac{n_0 n_1}{n}(\bar{x}_1 - \bar{x}_0)(\bar{w}_1 - \bar{w}_0),$$

where $SS_{x_0w_0}$ and $SS_{x_1w_1}$ are defined analogously to SS_{x_0} and SS_{x_1}.

The $n \times 1$ response vector *y* (to establish the order of the data) is given by

$$y = \begin{bmatrix} y_{01} & y_{02} & \cdots & y_{0n_0} & y_{11} & y_{12} & \cdots & y_{1n_1} \end{bmatrix}' .$$

The $n_i \times 1$ vector of predictor values for *x* when the dummy is either zero or one is therefore

$$X_i = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{in_i} \end{bmatrix}', i = 0,1 ,$$

with analogous vectors defined for *w*. The symbols J and 0 denote a column vector of ones or zeroes, respectively, of appropriate dimension.

For all models, we give the model matrix, denoted as $X_{n\times(p+1)}$ (which includes the intercept column of ones), the $p \times p$ R matrix of correlations among the predictors, and the *p* VIF formulas, obtained as the diagonal elements of R^{-1}.

3.2 Regression Model 1: One Numeric Predictor and a Dummy Variable

With Model 1, we are modeling two simple linear regressions with a common slope and different intercepts. The $n \times 3$ model matrix is

$$X_{n\times 3} = \begin{bmatrix} J_{n_0} & X_0 & 0 \\ J_{n_1} & X_1 & J_{n_1} \end{bmatrix},$$

with columns corresponding to intercept, *x*, and dummy variable, in order. The 2×2 correlation matrix is

$$R = \begin{bmatrix} 1 & \dfrac{\sqrt{n_0 n_1}(\bar{x}_1 - \bar{x}_0)}{\sqrt{nSS_x}} \\ symmetric & 1 \end{bmatrix}.$$

Taking the inverse of R, we obtain the following VIF, which is, of course, the same for x and the dummy variable:

$$\begin{aligned} VIF_x = VIF_{dummy} &= \frac{SS_x}{SS_{x_0} + SS_{x_1}}, \\ &= 1 + \frac{(n_0 n_1 / n)(\bar{x}_1 - \bar{x}_0)^2}{SS_{x_0} + SS_{x_1}} \\ &= \frac{1}{1 - r_{pb}^2} \\ &= 1 + t_x^2 / (n_0 + n_1 - 2), \end{aligned} \qquad (3.1)$$

where r_{pb} is the point-biserial correlation coefficient (Kendall and Stuart 1961)

$$r_{pb} = \frac{(\bar{x}_1 - \bar{x}_0)}{\sqrt{\dfrac{SS_x}{(n-1)}\left(\dfrac{1}{n_1} + \dfrac{1}{n_0}\right)}}$$

and t_x is the formula for the ordinary two independent-sample t with $df = n_0 + n_1 - 2$ for testing the equality of the two dummy group x-means when the group variances are equal:

$$t_x = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{\dfrac{SS_{x_0} + SS_{x_1}}{(n_0 + n_1 - 2)}\left(\dfrac{1}{n_0} + \dfrac{1}{n_1}\right)}} \quad .$$

Note that

$$\frac{r_{pb}^2}{1 - r_{pb}^2} = \frac{t_x^2}{n_0 + n_1 - 2}.$$

This VIF makes sense because $R_{xd}^2 = r_{pb}^2$. It is also invariant with respect to the dummy coding (i.e., which group is assigned the dummy value of zero). This VIF is clearly "large" if the two-sample t is large and the two sample sizes are small, that is, if the dummy group means are quite different from each other relative to the within-sample variability and sample size. Given this situation, a "large" VIF should not be naively taken to imply collinearity nor used as the basis for deleting either variable from the model.

Note also that if $\bar{x}_1 = \bar{x}_0$, $VIF_x = VIF_{dummy} = 1$ and $R = \mathrm{I}_2$. Does this condition really imply "no collinearity" in the sense of an approximate linear dependency between columns of the X matrix?

3.3 Regression Model 2: One Numeric Predictor, a Dummy Variable and the Numeric × Dummy Cross-product

With the second model, we are modeling two simple linear regressions with different intercepts and different slopes. The $n \times 4$ model matrix is

$$X_{nx4} = \begin{bmatrix} J_{n_0} & X_0 & 0 & 0 \\ J_{n_1} & X_1 & J_{n_1} & X_1 \end{bmatrix},$$

with columns corresponding to intercept, *x*, dummy variable, and *x* × dummy cross-product, in that order. The 3×3 correlation matrix is

$$R = \begin{bmatrix} 1 & \dfrac{\sqrt{n_0 n_1 / n}\ (\bar{x}_1 - \bar{x}_0)}{\sqrt{SS_x}} & \dfrac{SS_{x_1} + (n_0 n_1 / n)[\bar{x}_1^2 - \bar{x}_1 \bar{x}_0]}{\sqrt{SS_x\,[SS_{x_1} + (n_0 n_1 / n)\,\bar{x}_1^2]}} \\ & 1 & \dfrac{\sqrt{n_0 n_1 / n}\ \bar{x}_1}{\sqrt{SS_{x_1} + (n_0 n_1 / n)\,\bar{x}_1^2}} \\ \textit{symmetric} & & 1 \end{bmatrix},$$

The VIF for *x* is:

$$VIF_x = \frac{SS_x}{SS_{x_0}}$$

$$= 1 + \frac{SS_{x_1}}{SS_{x_0}} + \frac{(n_0 n_1 / n)(\bar{x}_1 - \bar{x}_0)^2}{SS_{x_0}}$$

$$= 1 + \frac{SS_{x_1}}{SS_{x_0}} + \frac{(t_x^*)^2}{n_0 - 1}, \qquad (3.2)$$

where t_x^* is the formula of an alternative two independent-sample *t* statistic with equal variances, which uses only the sample variance for group 0 to construct the t statistic and hence has $df = n_0 - 1$:

$$t_x^* = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{\frac{SS_{x_0}}{(n_0 - 1)}\left(\frac{1}{n_0} + \frac{1}{n_1}\right)}} \quad .$$

VIF_x is "large" if t_x^* is large and n_0 is small, similar to the VIF for Model 1, but is also large if the sum of squares for the group dummy of 1 is large, relative to that in the group with dummy of 0. This means that this VIF is not, in general, invariant with respect to the coding of the dummy variable. However, if $\bar{x}_1 = \bar{x}_0$ and $SS_{x_0} = SS_{x_1}$, then VIF_x *is* invariant with respect to the dummy coding and is equal to 2.

Now the interesting result: VIF_x=1 if and only if $\bar{x}_1 = \bar{x}_0$ and $SS_{x_1} = 0$! The first condition obviously can be made to happen by separately centering X_1 and X_0, but the second requires that all the values of X_1 be a constant (!) which means that the $n \times 4$ X matrix (above) has an exact singularity between column 3 and column 4 and is hence not full-rank. This situation would never arise in "real life," which implies that VIF_x will always indicate some degree of "variance inflation".

The VIF for the dummy variable is:

$$VIF_{dummy} = 1 + \frac{n_0 n_1}{n}\left[\frac{\bar{x}_0^2}{SS_{x_0}} + \frac{\bar{x}_1^2}{SS_{x_1}}\right].$$

$$= 1 + \frac{n_1 t_{x_0}^2}{n(n_0 - 1)} + \frac{n_0 t_{x_1}^2}{n(n_1 - 1)}, \qquad (3.3)$$

where t_{x_i}, $i = 0,1$, is the one-sample *t* statistic for testing the true *x*-mean for group *i* is zero, which has $df = n_i - 1$:

$$t_{x_i} = \frac{\bar{x}_i}{\sqrt{\frac{SS_{x_i}}{n_i - 1}\left(\frac{1}{n_i}\right)}}, \; i = 0,1 \quad .$$

VIF_{dummy} is invariant with respect to the dummy coding. In addition, it is "large" if the one or both of the one-sample t_{x_i}'s are large (i.e., the group *x*-means are large, relative to their respective sums of squares), and is also large (all other things being equal) if n_0 and n_1 are close to $n/2$.

Note that VIF_{dummy} is equal to 1 when $\bar{x}_0 = \bar{x}_1 = 0$.

The third VIF, for the $x \times$ dummy cross-product, is

$$VIF_{x\times dummy} = \left[\frac{1}{SS_{x_0}} + \frac{1}{SS_{x_1}}\right] SS_{xd}$$

$$= 1 + \frac{SS_{x_1}}{SS_{x_0}} + \frac{n_0 n_1 \bar{x}_1^2}{n}\left[\frac{1}{SS_{x_0}} + \frac{1}{SS_{x_1}}\right], \qquad (3.4)$$

where SS_{xd} is the sum of squares calculated from column four of the model matrix, that is,

$$SS_{xd} = n_0\left[\frac{-n_1\bar{x}_1}{n}\right]^2 + \sum_{j=1}^{n_1}\left[x_{1j} - (\frac{n_1\bar{x}_1}{n})\right]^2 = SS_{x_1} + \frac{n_0 n_1 \bar{x}_1^2}{n},$$

with the mean of column four being

$$\bar{x}_{xd} = \frac{\sum_{j=1}^{n_1} x_{1j}}{n} = \frac{n_1\bar{x}_1}{n} \quad .$$

With similarities to both VIF_x and VIF_{dummy}, the $VIF_{x\times dummy}$ is "large" if the square of the group 1 x-mean is large relative to the two within-group sums of squares, but is also large if n_0 and n_1 are close to $n/2$,as for VIF_{dummy} (3.3) and if the sum of squares for group 1, is large, relative to that in group 0, as for VIF_x (3.2). Hence $VIF_{x\times dummy}$, like VIF_x, is not invariant with respect to the coding of the dummy variable, but is equal to 2 if $\bar{x}_1 = 0$ and $SS_{x_0} = SS_{x_1}$.

Similarly to VIF_x, $VIF_{x\times dummy} = 1$ if and only if $\bar{x}_1 = 0$ and $SS_{x_1} = 0$, which again will not happen in "real life".

Finally, the closest that the correlation matrix can be to "no correlation" in "real life" is when $\bar{x}_1 = \bar{x}_0 = 0$:

$$R = \begin{bmatrix} 1 & 0 & \sqrt{SS_{x_1}/SS_x} \\ & 1 & 0 \\ symmetric & & 1 \end{bmatrix}.$$

As with Model 1, should we take large VIFs in Model 2 to indicate a "collinearity problem"?

3.4 Regression Model 3: Two Numeric Predictors

The model with two numeric predictors x and w is included to provide a comparison for the model with two numeric predictors and a dummy variable which is discussed in the following section.

The $n\times 3$ model matrix is

$$X_{n\times 3}=\begin{bmatrix} J_{n_0} & X_0 & W_0 \\ J_{n_1} & X_1 & W_1 \end{bmatrix},$$

and the 2×2 correlation matrix is

$$R=\begin{bmatrix} 1 & \dfrac{SS_{xw}}{\sqrt{SS_x SS_w}} \\ symmetric & 1 \end{bmatrix}=\begin{bmatrix} 1 & r_{xw} \\ symmetric & 1 \end{bmatrix}.$$

Therefore the VIF for *x* and *w* is

$$VIF_x = VIF_w = \frac{1}{1-r_{xw}^2}=\frac{SS_x SS_w}{SS_x SS_w - SS_{xw}^2}. \qquad (3.5)$$

This VIF has the classic interpretation, based on the R^2 of the regression between the two numeric predictors *x* and *w*. Clearly, if $VIF_x = VIF_w = 1$, then $R = \mathrm{I}_2$. Of interest is how this VIF changes when a dummy variable is included.

3.5 Regression Model 4: Two Numeric Predictors and a Dummy Variable

The model with two numeric predictors *x* and *w* and a dummy variable has $n\times 4$ model matrix:

$$X_{nx4}=\begin{bmatrix} J_{n_0} & X_0 & 0 & W_0 \\ J_{n_1} & X_1 & J_{n_1} & W_1 \end{bmatrix},$$

with columns corresponding to intercept, *x*, dummy variable, and *w*, in that order. The 3×3 correlation matrix is

$$R=\begin{bmatrix} 1 & \dfrac{\sqrt{n_0 n_1 / n}\ (\bar{x}_1-\bar{x}_0)}{\sqrt{SS_x}} & \dfrac{SS_{xw}}{\sqrt{SS_x\ SS_w}} \\ & 1 & \dfrac{\sqrt{n_0 n_1 / n}\ (\bar{w}_1-\bar{w}_0)}{\sqrt{SS_w}} \\ symmetric & & 1 \end{bmatrix}.$$

The VIFs for *x* and *w* are symmetric versions of each other:

$$VIF_x = \frac{(SS_{w_0} + SS_{w_1})SS_x}{(SS_{x_0} + SS_{x_1})(SS_{w_0} + SS_{w_1}) - (SS_{x_0w_0} + SS_{x_1w_1})^2}$$

$$= \left[\frac{1}{1 - r^2_{x_i w_i}}\right]\left[1 + \frac{t_x^2}{n_0 + n_1 - 2}\right] \tag{3.6}$$

and

$$VIF_w = \frac{(SS_{x_0} + SS_{x_1})SS_w}{(SS_{x_0} + SS_{x_1})(SS_{w_0} + SS_{w_1}) - (SS_{x_0w_0} + SS_{x_1w_1})^2}$$

$$= \left[\frac{1}{1 - r^2_{x_i w_i}}\right]\left[1 + \frac{t_w^2}{n_0 + n_1 - 2}\right], \tag{3.7}$$

where $r_{x_i w_i}$ is the correlation coefficient between *x* and *w* taking group into account (that is, pooling the within-group sums of cross-products and sums of squares)

$$r_{x_i w_i} = \frac{SS_{x_0w_0} + SS_{x_1w_1}}{[(SS_{x_0} + SS_{x_1})(SS_{w_0} + SS_{w_1})]^{1/2}}, \tag{3.8}$$

and t_x and t_w are the usual two independent-sample *t* statistics with equal variances, as in (3.1). Both VIF_x and VIF_w are invariant with respect to the dummy coding.

These two VIFs show two modifications over the VIF for the model with just two numeric predictors (3.5). First, they contain the correlation coefficient between *x* and *w* accounting for dummy groups (3.8), as opposed to the correlation ignoring groups as in Model 3 (3.5). Second, these VIFs also include a factor containing the two-sample *t* (for either *x* or *w*), which reflects how far apart the group means are, relative to the within-group variability, as in the Model 1 VIF (3.1). If there is no within-group correlation between *x* and *w* (i.e., $SS_{x_0w_0} = SS_{x_1w_1} = 0$), then either or both of VIF_x and VIF_w could be large solely because of the magnitudes of the respective t^2's, relative to the sample sizes.

The VIF for the dummy variable is

$$VIF_{dummy} = \frac{SS_x SS_w - SS_{xw}^2}{(SS_{x_0} + SS_{x_1})(SS_{w_0} + SS_{w_1}) - (SS_{x_0w_0} + SS_{x_1w_1})^2}. \tag{3.9}$$

It can be shown that VIF_{dummy} will be equal to one (indicating no collinearity) if

$$\frac{n_0 n_1}{n}[(\bar{x}_1 - \bar{x}_0)^2(SS_{w_0} + SS_{w_1}) + (\bar{w}_1 - \bar{w}_0)^2(SS_{x_0} + SS_{x_1}) - 2(\bar{x}_1 - \bar{x}_0)(\bar{w}_1 - \bar{w}_0)(SS_{x_0w_0} + SS_{x_1w_1})] = 0,$$

that is, if $\bar{x}_1 = \bar{x}_0$ and/or $\bar{w}_1 = \bar{w}_0$ and/or if $SS_{x_0w_0} = -SS_{x_1w_1}$ (which can occur if $SS_{x_0w_0} = SS_{x_1w_1} = 0$). Therefore, the VIF_{dummy} again can be large if the group means are far apart for one or both of the numeric predictors, even if the pooled within-group correlation is small. Again, VIF_{dummy} is increased if n_0 and n_1 are close to $n/2$.

Note that the pooled within-group sum of cross-products ($SS_{x_0w_0} + SS_{x_1w_1}$) appears in all three VIFs for Model 4. If the within-group sum of cross-products correlation between *x* and *w* is equally high in absolute value for each group (indicating within-group collinearity) but opposite in sign, both VIF_x (3.6) and VIF_w (3.7) could be close to zero and would not signal a potential collinearity problem.

Finally, $R = \mathrm{I}_3$ if and only if $\bar{x}_1 = \bar{x}_0$ and $\bar{w}_1 = \bar{w}_0$ and $SS_{xw} = 0$. The last condition is equivalent to the "traditional" condition of Model 3 with two numeric predictors but the other two are the same as in Model 1.

3.6 General Conclusions about the VIF Formulas

In general, the VIFs in the three models with dummy variables can be "large" for three basic reasons that are not related to classical concepts of collinearity:

1) if dummy group means are far apart from each other or a group mean is far from zero, relative to measures of variability (Model 1, $VIF_x = VIF_{dummy}$; Model 2, VIF_x, VIF_{dummy} and $VIF_{x\times dummy}$; and Model 4, VIF_x, VIF_w and VIF_{dummy});

2) if there are very unequal within-group sums of squares (Model 2, VIF_x and $VIF_{x\times dummy}$);

and

3) if sample sizes are small in each dummy group (Model 1, $VIF_x = VIF_{dummy}$; Model 2, VIF_x and VIF_{dummy}; and Model 4, VIF_x and VIF_w); and, to a lesser extent, if sample sizes are approximately equal (making $n_0 n_1 / n$ large) (Model 1, $VIF_x = VIF_{dummy}$; Model 2, $VIF_{x\times dummy}$; Model 4, VIF_{dummy}).

4. A Small Numeric Example: the Economic Example

As indicated previously, the economic data consisted of the response annual per capita consumption and the numeric predictor annual per capita income, both in hundred thousand U.S. Dollars, from 1976 to 2004, as well as a dummy variable. The dummy variable was due to a policy reform in Taiwan in 2000 that disaggregated the income-consumption relationship over 1976 - 2004 into two different stages, the first "pre-reform" stage from 1976 to 1999 with the "stage dummy" designated as 1 and the second "post-reform" stage from 2000 to 2004 with the "stage dummy" designated as 0. The data were retrieved from the website of Statistical Yearbook of the Republic of China (Taiwan R.O.C.) at http://www.stat.gov.tw.

Figures 1 and 2 show the trends of income and consumption over time and the relationship between consumption and income, respectively. Pertinent summary statistics for this dataset are (data divided by \$100,000): $n_0 = 5$, $n_1 = 24$. $\bar{x}_0 = 8.82$, $\bar{x}_1 = 4.74$, $\bar{x} = 5.45$, $SS_{x_0} = .04$, $SS_{x_1} = 159.87$, and $SS_x = 228.55$. All regression models were fitted using the SAS® REG procedure with VIF option (SAS® Institute 2007).

In the initial analysis, Model 2 was fitted to this data, with all three VIFs being greater than 5800. Model 1 was also fitted, with VIFs showing no sign of collinearity according to the "5" rule of thumb (Hocking and Pendleton 1983, Craney and Surles 2002). Nguyen (2008) subsequently obtained VIF formulas for Models 1 (equation 3.1) and 2 (equations 3.2, 3.3, and 3.4), verifying the formulas numerically. For this paper, two other models were fitted for comparison purposes (Table 1): Model 0 (the simple linear regression model with no dummy variable); and Model 5 (model with the numeric predictor and the numeric*dummy cross-product. In addition, Models 1, 2, and 5 were fitted with reverse dummy coding (i.e., 1976-1999 was coded as 0 and 2000-2004 was coded as 1) to see the effect of coding on VIFs (Table 1). Note, however, that when there are only two predictors (other than the intercept) in the model, that dummy coding will have no effect as the two predictors will always have the same numeric VIF. Thus, only Model 2 will show the effect of recoding on values of VIFs.

Table 1 summarizes the adjusted R-square, VIFs and regression coefficient p-values for all four models. Note that Models 0, 1, and 5 do not have "large" VIFs, according to the "5" rule of thumb.

Model 2, the original motivator of this work, is more interesting and we discuss it in some detail. First, the VIF (3.2) for income is high ($VIF_x \approx 5863$) because $SS_x = 228.55$ is very much larger than $SS_{x_0} = .04$. This VIF value can be broken down into the contribution due to the ratio of SS_{x_1} to SS_{x_0} ($159.87/.04 \approx 4101$) and the contribution due to the difference in group means (i.e., $(n_0 n_1 / n)(\bar{x}_1 - \bar{x}_0)^2 / SS_{x_0} \approx 1761$). Note that if the dummy coding is reversed, the VIF for income becomes 228.55/159.88 =1.43 (definitely not a red flag).

The high VIF for the stage dummy variable (3.3) ($VIF_{dummy} \approx 8255$) is due almost entirely to the high ratio between the square of the mean and the sum of squares for dummy group 0 (i.e., $(n_0 n_1 / n)(\bar{x}_0)^2 / SS_{x_0} \approx 8254$). Note that this VIF is invariant with respect to the dummy coding. Also note that if the sample sizes were close to evenly split between the two dummy groups (i.e., 14 and 15) that the multiplier $(n_0 n_1 / n)$ would increase from about 4.14 to about 7.24.

The VIF for the cross-product between income and the stage dummy variable (3.4) is high ($VIF_{x \times dummy} \approx 6493$) because of reasons similar to those of the VIF for income. Again, the ratio of SS_{x_1} to SS_{x_0} contributes about 4101 to this VIF, while the contribution of the part due the term $(n_0 n_1 / n)(\bar{x}_1)^2 / [1/SS_{x_0} + 1/SS_{x_1}]$ is about 2391. Note that if the dummy coding is reversed, this VIF is about 8257 and the contribution of the term $(n_0 n_1 / n)(\bar{x}_1)^2 / [1/SS_{x_0} + 1/SS_{x_1}]$ is about 8050.

With respect to the practicalities of model selection in this specific case, predicted values and their standard errors are virtually identical to 4 decimal places. Adjusted R-squares (Table

1) are all above 0.99. Mean square error is likewise very similar (about 0.012), except that Model 0 (the simple linear regression model has MSe that is almost double that of the other three models (about 0.023). Therefore, any choice among the four models makes little difference, if one's aim is estimation and/or prediction.

5. Summary

None of the conditions for "large" VIFs for Models 1, 2, and 4 (summarized in section 3.6) would seem to measure linear dependencies among columns of the X -matrix, the classic definition of collinearity", and although some might interpret the VIFs as a form of variance inflation, it doesn't seem that such (potentially large) VIFs should be reasons for deleting predictors from models.

Fortunately, in the example presented in Section 4, one can look at all possible reasonable models, and plotting the data (Figures 1 and 2) gives a very clear idea of what is occurring. In many other situations (say 8 numeric predictors and 5 dummy variables), there are too many possible models and examining bivariate plots can never be truly informative.

So consider the situation of 8 numeric predictors and 5 dummy variables with large VIFs for the entire set of predictors. How can one use the results of this paper for decision making in more complex models, given that obtaining explicit VIFs formulas is not an appealing option and bivariate plots are of limited help? First, as indicated in many regression textbooks (e.g., Mendenhall and Sincich 2012) one needs to decide if one is interested in interpreting individual regression coefficients or "only" in estimating and predicting. If the latter, then one is interested in the goodness of the model as a whole, and collinearity may not be of concern. If the former, then collinearity (as measured by VIFs or condition indices) must be addressed, which typically means deleting a predictor that participates in each collinearity. A possible strategy to investigate what predictors might be deleted is to check VIFs in several steps. First, check VIFs in a model with only the numeric predictors, as it would be helpful to know if numerics are "co-linear" or "co-planar" without the influence of the dummies. If there is "enough" data, one might also check VIFs for numerics for separate categories of a dummy, to see if collinearity exists within each dummy category for numerics. Third, if the numerics by themselves don't appear to be causing a problem, check VIFs adding in one dummy at a time.

One final comment: Issues with VIFs in the presence of dummy variables identified in this paper as contributing to high VIFs may be thought to produce "variance inflation", but unless the researcher can exercise more control over data than typically happens in regression situations, the "variance inflation" would appear to just be a characteristic of the dummy categories under consideration. In this case, then, if even one is interested in interpreting regression coefficients, perhaps dummy variables should be treated differently from numerics and not deleted just to reduce "collinearity".

Acknowledgements: The authors thank a reviewer for questioning the result that two of the VIFs reduce to 2 under certain conditions, which led to the more interesting result of when they reduce to 1.

References

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

Chatterjee S. and Price B. (1977) *Regression Analysis by Example*, New York: John Wiley & Son.

Craney, T.A., and Surles, J.G. (2002) "Model-Dependent Variance Inflation Factor Cutoff Values," *Quality Engineering*, 14(3), 391-403.

Fox, J. and Monette, G. (1992) "Generalized Collinearity Diagnostics," *Journal of the American Statistician Association*, 87, 178-183.

Graybill, F. A. and Iyer, H. (1994) *Regression Analysis: Concepts and Applications.* Duxbury Press. Belmont, CA.

Hocking, R. R. and Pendelton, O. J. (1983) "The Regression Dilemma," *Communications in Statistics—Theory and Methods*, 12(5), 497-527.

Kendall, M. G. and Stuart, A. (1961) *The Advanced Theory of Statistics: Volume 2 Inference and Relationship.* Hafner Publishing Company, New York.

Kleinbaum, D. G., Kupper, L.L., Muller, K.E., and Nizam, A. (1998) *Applied Regression Analysis and Other Multivariable Methods.* Duxbury Press. Belmont, CA.

Kutner, M. H., Nachtsheim, C. J. and Neter J. (2004) *Applied Linear Regression Models* (4th ed.), Homewood, IL: Irwin.

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

Mendenhall, W. and Sincich, T. (2004) *A Second Course in Statistics: Regression Analysis.* Prentice Hall. Boston, MA.

Nguyen, H. (2008) The Variance Inflation Factor in Linear Regression Models in Presence of a Dummy Variable," Unpublished masters report. New Mexico State University, Las Cruces, NM.

O'Brien, R.M. (2007), "A Caution Regarding Rules of Thumb for Variance Inflation Factors," *Quality & Quantity*, 41, 673-690.

SAS Institute, Inc. (2007) SAS Online Doc® 9.1.3. Cary, NC: SAS Institute, Inc.

Figure 1: Plot of annual per capita income and consumption versus time. Data retrieved from Statistical Yearbook of the Republic of China (Taiwan R.O.C.) at http://www.stat.gov.tw.

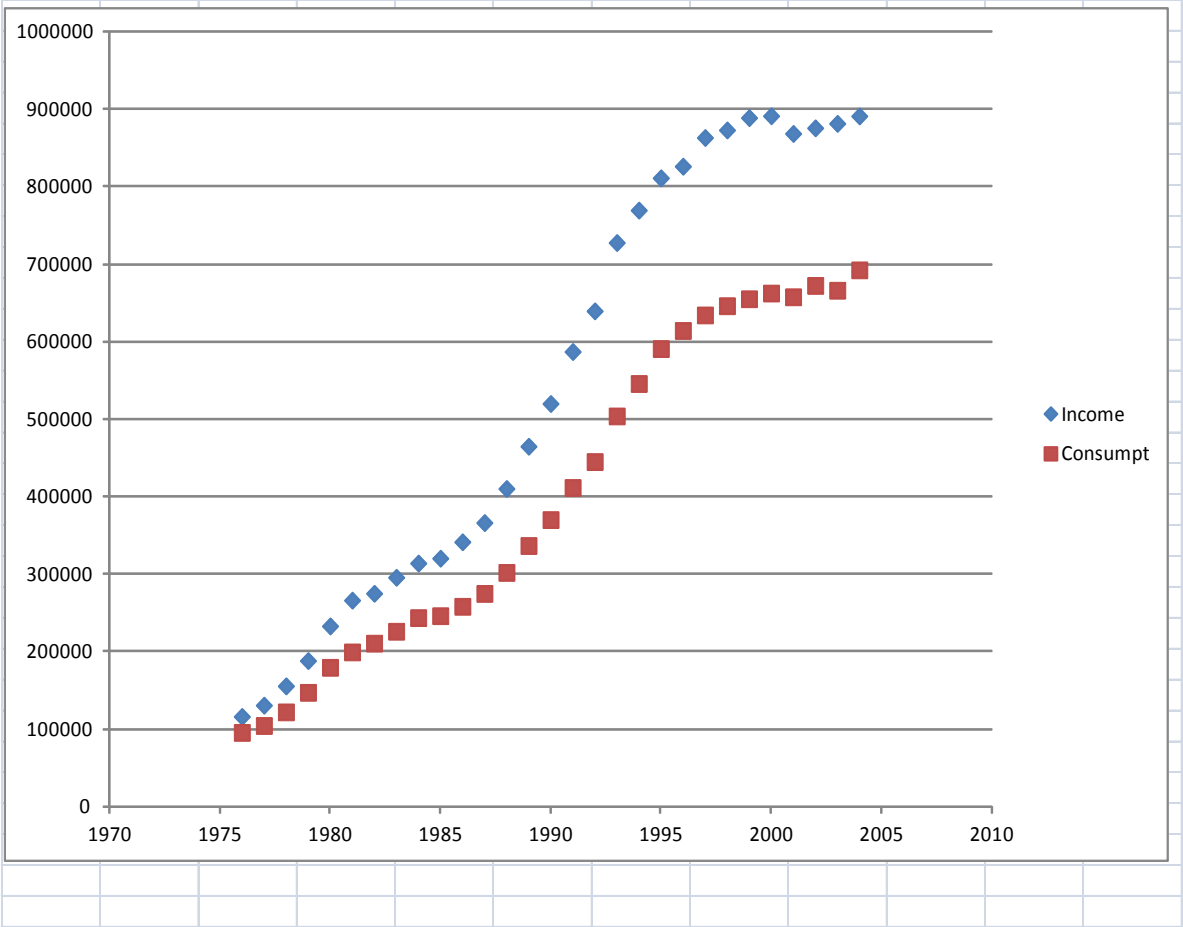

Figure 2. Plot of Annual per capita consumption versus annual per capita income. Data retrieved from Statistical Yearbook of the Republic of China (Taiwan R.O.C.) at http://www.stat.gov.tw.

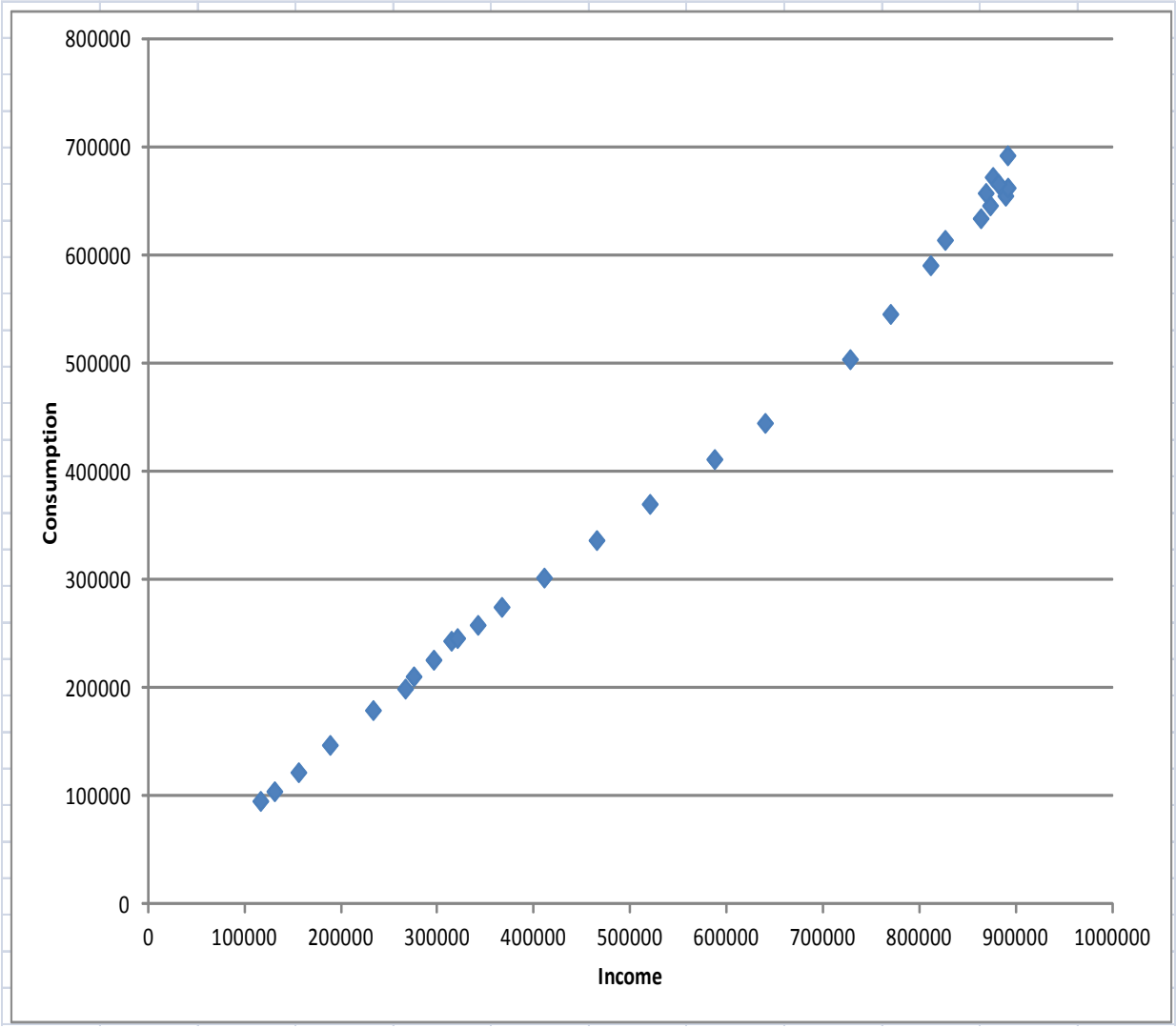

Table 1. Adjusted R-squares, variance inflation factors (VIF) and regression coefficient p-values for four regression models for the Taiwan economic data. Overall model p-value is p<0.0001 for all models. Original dummy coding is 1 for 1976-1999, 0 for 2000-2004.

	Original coding		Reverse coding	
	p-value	**VIF**	**p-value**	**VIF**
Model 0				
Intercept	0.4818	n/a	n/a	n/a
X	<0.0001	1.0000	n/a	n/a
Adjusted Rsquare=0.9948				
Model 1				
Intercept	<0.0001	n/a	<0.0001	n/a
X	<0.0001	1.4290	<0.0001	1.4290
Dummy	<0.0001	1.4290	<0.0001	1.4290
Adjusted Rsquare=0.9973				
Model 2				
Intercept	0.9711	n/a	0.0187	n/a
X	0.1979	5862	<0.0001	1.4296
Dummy	0.9901	8255	0.9901	8255
X*Dummy	0.9575	6493	0.9575	8257
Adjusted Rsquare=0.9972				
Model 5				
Intercept	0.0157	n/a	0.0157	n/a
X	<0.0001	1.1240	<0.0001	1.1240
X*Dummy	<0.0001	1.1240	<0.0001	1.1240
Adjusted Rsquare=0.9973				