Graphical Interpretation of Variance Inflation Factors

Robert A. Stine

Department of Statistics, Wharton School

University of Pennsylvania, Philadelphia PA 19104-6302

Abstract

The variance inflation factor for a regression coefficient is the square of the ratio of t-statistics associated with partial residual and partial regression plots. This alternative characterization, particularly when presented with a dynamic graphical display, can help students interpret the size of the variance inflation factor, relate it to regression diagnostic plots, and understand the impact of multicollinearty. Examples using two small data sets illustrate this approach.

1 Introduction

This note focusses on the connection between the variance inflation factor (VIF) and two added-variable plots for least squares regression, partial regression plots and partial residual plots (also known as component-plus-residual plots). To help students master regression diagnostics, I have found it useful to point out explicitly the connections among them. Introductions to regression diagnostics at the level of Chatterjee and Price (1991) or Fox (1991) offer the student a variety of numerical and graphical diagnostics for judging the adequacy of a regression model. There are diagnostics for specification error, outliers, multicollinearity, nonlinearity, heteroscedasticity, and other faults. Rather than present each diagnostic individually, I find it useful to describe the connections among them, much as one needs to do in presenting the various types of random variables in an introductory course.

The presentation offered here is relatively elementary. The level is appropriate for students who do not know linear algebra, and I have found it useful in more advanced courses as well. The presentation relies upon imbedding the three diagnostics in a single dynamic plot. At one extreme of a slider control, this plot is the partial residual plot which shows none of the effects of collinearity. As the control moves to the other extreme, it becomes the partial regression plot which conveys the effects of multicollinearity. The plot dynamically updates its coordinates to show the effects of intermediate levels of multicollinearity.

2 The Diagnostics

The VIF measures how much multicollinearity has increased the variance of a slope estimate. Suppose that we write the full-rank regression model for n independent observations as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $\mathrm{Var}(\epsilon_i) = \sigma^2$. In vector form, the model is $Y = X\beta + \epsilon$ where X is the $n \times (k+1)$ matrix with columns $X_0, X_1, \ldots, X_k$ and X_0 is a column vector of 1's. The name of this diagnostic arises from writing the variance of the least squares estimator $\hat{\beta}_j$ ($j =$

$1, \ldots, k$) as (e.g. Belsley, Section 2.3)

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_j) &= \sigma^2 (X'X)^{-1}_{jj} \\ &= \frac{\sigma^2}{SS_j} VIF_j, \end{aligned}$$

where $SS_j = \sum_i (x_{ij} - \bar{x}_j)^2$ and

$$VIF_j = \frac{1}{1 - R_j^2}. \tag{1}$$

R_j^2 is the R^2 statistic from the regression of X_j on the other covariates. Unfortunately, there is no well-defined upon critical value for what it is needed in order to have a "large" VIF. Some authors, such as Chatterjee and Price (1991), suggest 10 as being large enough to indicate a problem.

The variance inflation factor is closely tied to the difference between two added variable plots for a regression. The partial regression plot for the jth variable shows two sets of residuals, those from regressing Y and X_j on the other covariates. The associated simple regression has slope $\hat{\beta}_j$ and the same residuals $\hat{\epsilon} = Y - X\hat{\beta}$ as the multiple regression. Indeed, with an adjustment for degrees of freedom, the variance of the slope estimate based on the partial regression plot is the same as that for $\hat{\beta}_j$ in the multiple regression,

$$var_j^{regr} = \frac{n-k-1}{n-2} \hat{\sigma}^2 (X'X)^{-1}_{jj} = \frac{n-k-1}{n-2} \frac{\hat{\sigma}^2}{SS_j(1 - R_j^2)}, \tag{2}$$

where $\hat{\sigma}^2 = \sum_i \hat{\epsilon}_i^2 / (n-k-1)$. While seldom useful for detecting nonlinearity (neither axis shows an observed variable), these plots reveal the presence of multiple outliers (masking) and show the effects of multicollinearity.

In contrast, partial residual plots offer as a means for identifying nonlinearity. The partial residual plot corresponding to X_j shows $\hat{\epsilon} + \hat{\beta}_j X_j$ versus X_j. These plots ignore the effects of multicollinearity and convey a misleading impression of the significance of the fit, as noted by various authors. Although the associated simple regression again has slope $\hat{\beta}_j$ and residuals $\hat{\epsilon}$, the estimated variance of the fitted slope is

$$var_j^{res} = \frac{n-k-1}{n-2} \frac{\hat{\sigma}^2}{SS_j}. \tag{3}$$

The variance equations (2) and (3) are well-known; see, for example, Atkinson (1985) or Cook and Weisberg (1982). Noting the form of the VIF (1), it is immediate (though not explicitly given in the literature) that

$$VIF_j = \frac{var_j^{regr}}{var_j^{res}}. \tag{4}$$

In other words, VIF_j is the square of the ratio of the t-statistics from fits in the partial residual plot and partial regression plot.

3 The Dynamic Plot

A single dynamic plot ties these diagnostics together. Let $P_{(-j)}$ denote the projection matrix associated with all of the covariates but X_j. Following Cook and Weisberg (1982), define

$$\hat{\epsilon}_j(\lambda) \quad = \quad (I - \lambda P_{(-j)})(X_j - \overline{X}_j),$$

and

$$\begin{aligned} \hat{\epsilon}_Y(\lambda) &= (I - \lambda P_{(-j)})(Y - \overline{Y}) \\ &= \hat{\epsilon} + (I - \lambda P_{(-j)})\hat{\beta}_j(X_j - \overline{X}_j) \\ &= \hat{\epsilon} + \hat{\beta}_j \hat{\epsilon}_j(\lambda). \end{aligned}$$

The dynamic plot of $\hat{\epsilon}_Y(\lambda)$ on $\hat{\epsilon}_j(\lambda)$ allows the viewer to manipulate $0 \leq \lambda \leq 1$ using slider tools like those in *Lisp-Stat* (Tierney 1990).

The animation opens with $\lambda = 0$ which offers the greatest variation in the x-axis and is the (centered) partial residual plot. Intuitively, this is the relationship beteen Y and X_j were X_j uncorrelated with the other covariates. As λ varies from zero to one, the animation shows how the points move in response to the changing amounts of collinearity. As λ approaches one, it becomes the partial regression plot and shows the full impact of the multicollinearity present in the data. The simplicity of the calculations makes real-time animations possible on personal computers. The display also gives the effective variance inflation factor associated with the plotted data,

$$VIF_j(\lambda) = \frac{1}{1 - \lambda R_j^2}.$$

The following examples using two small data sets illustrate the use of this plot. Cook and Weisberg (1989) present other dynamic regression diagnostics. As in their examples, I give a sequence of several frames which attempt to represent the animated display.

An implementation of this dynamic graphic is available from the author via e-mail. It requires that the user have *Lisp-Stat*. The code consists of several methods that enhance the standard regression model object in this package.

4 Two Examples

The first example considers a time-series regression which has substantial collinearity. The regression considers the dependence of domestic U.S. crude oil production (*OUTPUT*) upon gross national product (*GNP*), price, a time trend (*YEAR*), and level of wildcat drilling activity during the 31 years 1948-1978. The data appear in problem 7.17 of Gujarati (1988). The OLS fit including the VIF's for this model appear in Table 1. The VIF for GNP is 62.1 — clearly a "large" VIF.

The sequence of five frames shown in Figure 1 convey a sense of how the plot changes as the value of λ ranges over the interval $[0, 1]$. The year 1973, the year of the first oil embargo, is an influential outlier and is highlighted throughout. Initially, the fit looks quite good in Figure 1a, the partial residual plot. As λ increases, collinearity compacts the variation on the x-axis and the fit grows weaker. With $\lambda = 0.5$ in Figure 1c, the points are halfway to the partial regression plot, but $VIF(.5) = 2$ remains small. As λ nears one, the apparent VIF rapidly grows, reflecting the nonlinear definition of $VIF(\lambda)$. The sequence of plots shows that most of the "damage" is done by the time that $VIF(\lambda)$ reaches the range 5 to 10, supporting the intuitive cutoff but allowing students to form their own opinions. Since the points move parallel to the fitted line, the plot also reinforces the notion that multicollinearity does not directly affect the residuals. The outlier year 1973 is just as far from the fitted line in Figure 1a as in Figure 1e. The last frame of Figure 1 repeats the fifth, only with the x-axis expanded to reveal the structure of the partial regression plot.

The second example demonstrates how collinearity affects a model with much less correlation among covariates. This example uses the data for 22 jet fighters reported in Cook and Weisberg (1982, p.47). The dependent variable is the log of the number of months after January, 1940, of the first flight of the particular model of aircraft. The covariates are SPR (power per unit weight), RGF (range), PLF (payload as fraction of total weight), SLF (sustained load factor), and CAR (a dummy which is 1 if the plane can land on an aircraft carrier). Table 2 summarizes the fitted model. In contrast to the first example, all of the VIF's are less than 1.5.

While less dramatic than the first example, the dynamic plot for SLF is still interesting. In their analysis of this data, Cook and Weisberg note that two models are outliers whose effects are disguised in the partial residual plot but evident in the partial regression plot. The sequence of four frames of the dynamic VIF plot in Figure 2 show how collinearity moves these two from being relatively innocent in Figure 2a to being quite influential (attenuating the slope) in Figure 2d.

References

1. Atkinson, A.C. (1985). *Plots, Transformations, and Regression.* Oxford Publications, Oxford.

2. Belsley, D.A. (1991). *Conditioning Diagnostics.* Wiley, New York.

3. Chatterjee, S. and B. Price (1991). *Regression Diagnostics.* Wiley, New York.

4. Cook, R.D. and S. Weisberg (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York.

5. —– (1993). Regression diagnostics with dynamic graphics (with discussion). *Technometrics*, 31, 277-311.

6. Fox, J. (1991). *Regression Diagnostics.* Sage, Newbury Park, CA.

7. Gujarati, D.N. (1988). *Basic Econometrics.* McGraw-Hill, New York.

8. Tierney, L. (1990). *Lisp-Stat.* Wiley, New York.

Table 1. Summary of the least squares regression model fitted to the oil production data for 31 observations 1948-1978. The square of the multiple correlation is $R^2 = 0.94$ and $\hat{\sigma} = 0.29$.

Variable	Estimate.	Standard.Error	t-Statistic	VIF
Constant	2.62	1.52	1.7	*n.a.*
GNP	.0011	.0015	.7	62.1
YEAR	.0960	.044	2.2	58.1
PRICE	-.699	.070	-9.9	1.2
WILDCATS	.094	.029	3.2	1.7

Table 2. Summary of the least squares regression for the jet fighter data. The response is $LOG(FFD)$, the log of the first flight date in months after January, 1940. The $R^2 = 0.83$ and $\hat{\sigma} = .16$.

Variable	Estimate.	Standard.Error	t-Statistic	VIF
Constant	3.72	0.27	14	*n.a.*
SPR	0.085	0.022	3.9	1.45
RGF	0.22	0.062	3.6	1.32
PLF	-0.48	0.47	-1.0	1.15
SLF	0.084	0.046	1.8	1.31
CAR	-0.23	0.088	-2.7	1.27

Figure 1. Frames from the dynamic plot for *GNP* in the model for oil production.

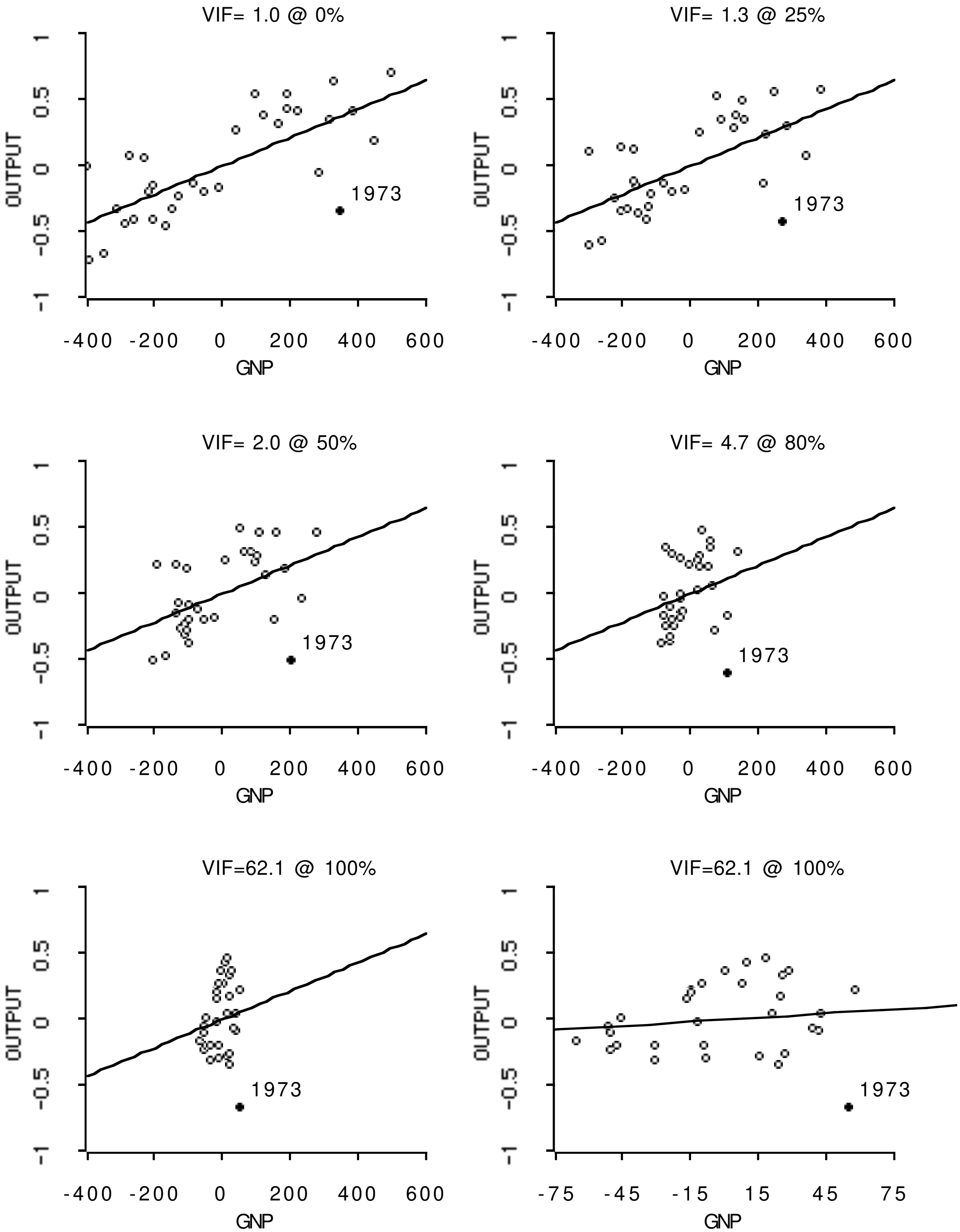

Figure 2. Frames from the dynamic plot for SLF in the model for the log of the first flight date of jet fighters.