Variance inflation factor

From Wikipedia, the free encyclopedia

In statistics, the **variance inflation factor (VIF)** quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

Contents

Definition

Consider the following linear model with k independent variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon.$$

The standard error of the estimate of β_j is the square root of the $j+1, j+1$ element of $s^2(X'X)^{-1}$, where s is the root mean squared error (RMSE) (note that RMSE^2 is an unbiased estimator of the true variance of the error term, σ^2); X is the regression design matrix — a matrix such that $X_{i,j+1}$ is the value of the j^{th} independent variable for the i^{th} case or observation, and such that $X_{i,1}$ equals 1 for all i. It turns out that the square of this standard error, the estimated variance of the estimate of β_j, can be equivalently expressed as

$$\widehat{\operatorname{var}}(\hat{\beta}_j) = \frac{s^2}{(n-1)\widehat{\operatorname{var}}(X_j)} \cdot \frac{1}{1-R_j^2},$$

where R_j^2 is the multiple R^2 for the regression of X_j on the other covariates (a regression that does not involve the response variable Y). This identity separates the influences of several distinct factors on the variance of the coefficient estimate:

- s^2: greater scatter in the data around the regression surface leads to proportionately more variance in the coefficient estimates
- n: greater sample size results in proportionately less variance in the coefficient estimates
- $\widehat{\operatorname{var}}(X_j)$: greater variability in a particular covariate leads to proportionately less variance in the corresponding coefficient estimate

The remaining term, $1 / (1 - R_j^2)$ is the VIF. It reflects all other factors that influence the uncertainty in the coefficient estimates. The VIF equals 1 when the vector X_j is orthogonal to each column of the design matrix for the regression of X_j on the other covariates. By contrast, the VIF is greater than 1 when the vector X_j is not orthogonal to all columns of the design matrix for the regression of X_j on the other covariates. Finally, note that the VIF is invariant to the scaling of the variables (that is, we could scale each variable X_j by a constant c_j without changing the VIF).

Calculation and Analysis

We can calculate k different VIFs (one for each X_i) in three steps:

Step one

First we run an ordinary least square regression that has X_i as a function of all the other explanatory variables in the first equation.
If $i = 1$, for example, the equation would be

$$X_1 = \alpha_2 X_2 + \alpha_3 X_3 + \cdots + \alpha_k X_k + c_0 + e$$

where c_0 is a constant and e is the error term.

Step two

Then, calculate the VIF factor for $\hat{\beta}_i$ with the following formula:

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

where R^2_i is the coefficient of determination of the regression equation in step one, with X_i on the left hand side, and all other predictor variables (all the other X variables) on the right hand side.

Step three

Analyze the magnitude of multicollinearity by considering the size of the $\mathrm{VIF}(\hat{\beta}_i)$. A rule of thumb is that if $\mathrm{VIF}(\hat{\beta}_i) > 10$ then multicollinearity is high.[1]

Some software instead calculates the tolerance which is just the reciprocal of the VIF. The choice of which to use is a matter of personal preference.

Interpretation

The square root of the variance inflation factor tells you how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other predictor variables in the model.

Example
If the variance inflation factor of a predictor variable were 5.27 ($\sqrt{5.27} = 2.3$) this means that the standard error for the coefficient of that predictor variable is 2.3 times as large as it would be if that predictor variable were uncorrelated with the other predictor variables.

References

1. Kutner, M. H.; Nachtsheim, C. J.; Neter, J. (2004). *Applied Linear Regression Models* (4th ed.). McGraw-Hill Irwin.

- Allison, P. D. (1999). *Multiple Regression: A Primer*. Thousand Oaks, CA: Pine Forge Press. p. 142.
- Hair, J. F.; Anderson, R.; Tatham, R. L.; Black, W. C. (2006). *Multivariate Data Analysis*. Upper Saddle River, NJ.
- Kutner, M. H.; Nachtsheim, C. J.; Neter, J. (2004). *Applied Linear Regression Models* (4th ed.). McGraw-Hill Irwin.
- Longnecker, M. T.; Ott, R. L. (2004). *A First Course in Statistical Methods*. Thomson Brooks/Cole. p. 615.
- Marquardt, D. W. (1970). "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation". *Technometrics* **12** (3): 591–612 [pp. 605–7]. doi:10.1080/00401706.1970.10488699.
- Studenmund, A. H. (2006). *Using Econometrics: A Practical Guide* (5th ed.). Pearson International. pp. 258–259.