A new function for the VIF in Ridge Regression

A reviewed function to obtain the Variance Inflation Factor in Ridge estimation

Claudia García - Ainara Rodríguez
Phd students. Universidad de Granada
Catalina García - Román Salmerón
Quantitative methods for economics and business. Universidad de Granada

Corresponding mail: claugar@correo.ugr.es

Abstract

Marquardt [3] and McDonald [4] presented an expression for the Variance Inflation Factor (VIF) to be applied in Ridge Regression (RR) that leads to values of VIF lesser than 1, contrarily to its theoretical concept. However, these expressions have been widely applied and developed to be applied in R software. This work presents an alternative expression that that satisfies the afore mentioned condition, and also presents other interesting properties. The function R is also presented and applied in an empirical application.

Introduction

The ridge estimation was presented by Hoerl and Kennard [2] as a mechanical method to solve collinearity. It defines a class of estimators which depend on the non-negative scalar parameter k:

$$\widehat{\boldsymbol{\beta}}(k) = \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{Y}. \tag{1}$$

Its covariance matrix is:

$$\text{var}\left(\widehat{\beta}(k)\right) = \sigma^2\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}. \tag{2}$$

- The estimator given in (1) is a biased estimator when $k > 0$ and when $k = 0$ it coincides with the least square estimator.
- After selecting the value of the parameter k we can apply the ridge estimation but then we will need to know if the collinearity has been solved.
- Thus, it is necessary to extend the different collinearity diagnostic measures to be applied in Ridge Regression (RR).

A widely applied measure to analyze the problem of collinearity is the Variance Inflator Factor (VIF) which is defined in an standardized model where the exogenous variables are orthogonal to the variable $\mathbf{X}_i$ (then $R_i^2 = 0$) and then it is possible to obtain the generally accepted definition of VIF due to [5]:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}. \tag{3}$$

The VIFs presented by Marquardt [3] and McDonald [4] for the RR are incorrectly calculated in the matrix $\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}$ leading to a wrong definition of VIF that have been widely applied in scientific literature and in R software.

Origin of the incorrect expression

By taking the following standardized model:

$$y_j = \beta_1 x_{1j} + \beta_2 x_{2j} + v_j, \quad j = 1, \ldots, n, \tag{4}$$

where the following conditions are assumed: $\sum_{j=1}^n x_{1j} = 0$, $\sum_{j=1}^n x_{1j}^2 = 1$, $\sum_{j=1}^n x_{2j} = 0$, $\sum_{j=1}^n x_{2j}^2 = 1$, and $\sum_{j=1}^n x_{1j}x_{2j} = \rho$, the variance inflator factor could be defined, when $p = 2$, as the corresponding element of the main diagonal of the matrix $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$. That is to say:

$$\text{VIF} = \frac{1}{1 - \rho^2}. \tag{5}$$

By using this last definition of VIF in the ridge estimator with expression (1) and with the matrix $\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}$, Marquardt [3] and McDonald [4] presented the following expressions, respectively:

$$\text{VIF}_M(k) = \frac{(1+k)^2 - 2(1+k)\rho^2 + \rho^2}{\left[(1+k)^2 - \rho^2\right]^2}, \tag{6}$$

$$\text{VIF}_{McD}(k) = \frac{\lambda_1\left(\lambda_1 + k\right)^{-2} + \lambda_2\left(\lambda_2 + k\right)^{-2}}{2}, \tag{7}$$

where $\lambda_1 = 1 + \rho$ and $\lambda_2 = 1 - \rho$ are the latent roots of $\mathbf{X}'\mathbf{X}$.

Note that these authors considered the elements of the main diagonal of the matrix:

$$\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}$$

as the variance inflation factors leading to the following consequences:

- Expression (6) can take values lower than 1 while expression (5) will be always equal to or greater than 1 since $-1 \leq \rho \leq 1$.
- It is also evident that the VIF should increases as the correlation coefficient ρ increases. However, the $\text{VIF}_M(k)$ does not verify this condition. Note that the $\text{VIF}_M(k)$ decreases for values of ρ higher than 0.9 (when collinearity is serious), even taking values less than one.

Our contribution

García *et al.* [1] presented an alternative expression of the VIF obtained from the matrix $\mathbf{X}_A$ calculated to obtain the ridge estimator by using Ordinary Least Squares (OLS) regression. Marquardt [3] and, more explicitly, Zhang and Ibrahim [6] pointed out that the ridge estimator can be calculated by OLS regression from the matrix $\mathbf{X}_A$ as:

$$\widehat{\boldsymbol{\beta}}_R(k) = \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{Y} = \left(\mathbf{X}_A'\mathbf{X}_A\right)^{-1}\mathbf{X}_A'\mathbf{Y}_A, \quad k \geq 0, \tag{8}$$

where $\mathbf{X}_A = \begin{pmatrix}\mathbf{X} \\ \sqrt{k}\mathbf{I}\end{pmatrix}$ and $\mathbf{Y}_A = \begin{pmatrix}\mathbf{Y} \\ \mathbf{0}\end{pmatrix}$ being $\mathbf{I}$ the identity matrix and $\mathbf{0}$ the null vector both of order p.

We now know the matrix $\mathbf{X}_A$ that has generated the matrix $\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)$ and we can calculate the determination coefficient between the independent variables and the VIF from its general definition. By developing the matrix $\mathbf{X}_A$ with $p = 2$ we have:

$$\mathbf{X}_A = \begin{pmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \\ \sqrt{k} & 0 \\ 0 & \sqrt{k} \end{pmatrix} = \begin{pmatrix} z_{11} & z_{21} \\ z_{12} & z_{22} \\ \vdots & \vdots \\ z_{1n+2} & z_{2n+2} \end{pmatrix}, \tag{9}$$

from which we can estimate the model $z_{1j} = \beta_1 + \beta_2 z_{2j} + w_j$, with $j = 1, \ldots, n+2$ (without standardized variables) where $\sum_{j=1}^{n+2} z_{1j} = \sqrt{k}$ and $\sum_{j=1}^{n+2} z_{2j} = \sqrt{k}$, and obtain the coefficient of determination. And the VIF for the ridge regression will be obtained as:

$$\text{VIF}_R(k, n) = \frac{1}{1 - R_i^2} = \frac{\left[(n+2)(1+k) - k\right]^2}{(n+2)^2\left[(1+k)^2 - \rho^2\right] - 2(n+2)k(1+k-\rho)}. \tag{10}$$

R function

```
VIF <- function(independientes, salto=0.1, tope=1, graf=F)

        observaciones = dim(independientes)[1]
        variables = dim(independientes)[2]

        discretizacion = seq(0, tope, salto)
        identidad = diag(variables)
        ceros = matrix(0, variables, 1)

# Estandardizing
                X = matrix( , observaciones, variables)
                for (i in 1:variables) {
                        media = mean(independientes[,i])
                        varianza = ((observaciones-1)/observaciones)*var(independientes[,i])
                        for (j in 1:observaciones) {
                                X[j,i] = (independientes[j,i] - media)/sqrt(observaciones*varianza)
                        }
                }

                XX = crossprod(X)
# Apply Theil definition to standardized data
theil2 = matrix(, length(discretizacion), variables+1)
                determinante = array( , c(length(discretizacion),2))
                j = 1
                for (k in discretizacion)
                {
                        Ik = sqrt(k)*identidad
                        Xa2 = rbind(X,Ik)
                        determinante[j,1] = k
                        determinante[j,2] = det(cor(Xa2))
                        theil2[j,1] = k
                        for (i in 1:variables)
                        {
                                reg4 = lm(Xa2[,i] ~ Xa2[,-i] - 1)
                                R24 = as.numeric(summary(reg4)[8])
                                theil2[j,i+1] = 1/(1-R24)
                        }
                        j = j+1
                }
# Output
                filas = c()
                for(k in discretizacion)
                {
                        filas = c("",filas)
                }
                rownames(theil2) = filas
                columnas1 = c()
                columnas2 = c("k")
                columnas3 = c("␣")
                for (i in 1:variables)
                {
                        columnas1 = c(columnas1, paste("X",i))
                        columnas2 = c(columnas2, paste("X",i))
                }
                colnames(theil2) = columnas2
                rownames(XX) = columnas1
                colnames(XX) = columnas1
                resultado = list(XX, determinante, theil2)
                names(resultado) = c("Matriz␣de␣correlaciones", "Determinante␣matriz␣de␣correlaciones", "VIF␣in␣Ridge␣Regression")
                resultado
```

Example

In this example the total mortality rate, Y, is related to the nitrogen oxide pollution potential, X_1, and the hydrocarbon pollution potential, X_2, for 60 cities. From this information, we can obtain the value of $n = 60$ and $\rho = 0.984$.

k	$\text{VIF}(k, 60)$	$\text{VIF}_M(k)$	$\text{VIF}(k, 60) - \text{VIF}_M(k)$
0	31.5020	31.5020	0.0000
0.01	19.6732	12.0838	7.5894
0.02	14.4167	6.4199	7.9968
0.03	11.4461	4.0253	7.4208
0.04	9.5369	2.7932	6.7437
0.05	8.2066	2.0763	6.1302
0.06	7.2266	1.6225	5.6041
0.07	6.4748	1.3168	5.1580
0.08	5.8799	1.1009	4.7790
0.09	5.3975	0.9426	4.4548
0.1	4.9984	0.8229	4.1754
0.11	4.6628	0.7301	3.9327
0.12	4.3767	0.6566	3.7201
0.13	4.1300	0.5973	3.5327
0.14	3.9150	0.5486	3.3664
0.15	3.7261	0.5082	3.2179
0.16	3.5587	0.4741	3.0846
0.17	3.4094	0.4450	2.9644
0.18	3.2755	0.4201	2.8554
0.19	3.1547	0.3984	2.7563
0.2	3.0451	0.3794	2.6657
0.21	2.9453	0.3627	2.5826
0.22	2.8541	0.3479	2.5062
0.23	2.7703	0.3346	2.4358
0.24	2.6932	0.3226	2.3706
0.25	2.6219	0.3118	2.3101
0.26	2.5559	0.3020	2.2539
0.27	2.4945	0.2931	2.2015
0.28	2.4373	0.2848	2.1525
0.29	2.3840	0.2773	2.1067
0.3	2.3340	0.2703	2.0638

References

[1] C. B. García, J. García, M. M. López Martín, and R. Salmerón. Collinearity: Revisiting the variance inflation factor in ridge regression. *Journal of Applied Statistics*, 42(3):648–661, 2015.

[2] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[3] D. W. Marquardt. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics*, 12(3):591–612, 1970.

[4] G. C. McDonald. Tracing ridge regression coefficients. *Wiley Interdiscip. Rev.: Comput. Statist.*, 2:695–703, 2010.

[5] H. Theil. *Principles of econometrics*. Wiley, New York, 1971.

[6] J. Zhang and M. Ibrahim. A simulation study on spss ridge regression and ordinary least squares regression procedures for multicollinearity data. *J. App. Stat.*, 32(6):571–588, 2005.