El Factor de Inflación de la Varianza en R

A. Rodríguez y C. García
Doctorado en Ciencias Económicas y Empresariales. Universidad de Granada, España
R. Salmerón y C. B. García
Dept. de Métodos Cuantitativos para la Economía y la Empresa. Universidad de Granada, España

1. Introduction

Cuando en un modelo de regresión lineal existe una fuerte relación lineal entre sus variables independientes se dice que existe multicolinealidad aproximada. En esta situación, el estimador por Mínimos Cuadrados Ordinarios (MCO) puede ofrecer resultados inestables, por lo que no se recomienda su uso y se hace necesario disponer de herramientas que permitan detectar este problema de forma adecuada.
Entre estas herramientas las más usadas son el Factor de Inflación de la Varianza (FIV) y el Número de Condición (NC). Destacar que no se tratan de test estadísticos que contrasten si la existencia de multicolinealidad es grave, sino de reglas de decisión que tratan de establecer umbrales a partir de los cuales el problema planteado se considera preocupante.
El presente trabajo se centra en el cálculo de la primera de ellas en el entorno de programación **R**.

2. Factor de Inflación de la Varianza

El modelo de regresión lineal es una técnica estadística ampliamente utilizada para estudiar la relación entre una variable denominada dependiente o explicada, $\mathbf{Y}$, y un conjunto de variables independientes o explicativas, $\mathbf{X_1}, \ldots, \mathbf{X_p}$, $\mathbf{p \geq 1}$. Dicha relación se define para n observaciones y p variables independientes como sigue a continuación:

$$\mathbf{y}_t = \beta_1 + \beta_2 \mathbf{X}_{2t} + \cdots + \beta_p \mathbf{X}_{pt} + \mathbf{u}_t, \quad t = 1, \ldots, n. \tag{1}$$

donde $\mathbf{u}$ representa la perturbación aleatoria (que se presupone esférica) y se supone que $\mathbf{X}_1$ es un vector de unos.
En este contexto, el FIV se define como sigue a continuación:

$$FIV(i) = \frac{var(\hat{\beta}_i)}{var(\hat{\beta}_i^0)} = \frac{1}{1 - R_i^2}, \quad i = 2, \ldots, p, \tag{2}$$

siendo $\hat{\beta}$ el estimador por MCO del modelo (1), $\hat{\beta}^{\mathbf{0}}$ el estimador por MCO del modelo (1) suponiendo que las variables explicativas son ortogonales, y R_i^2 el coeficiente de determinación de la regresión auxiliar que tiene como variable dependiente a $\mathbf{X_i}$ y como independientes al resto de regresores, para $i = 2, \ldots, p$.
Tradicionalmente, valores del FIV superiores a 10 indicarían que el modelo de regresión lineal presenta un grado de multicolinealidad preocupante.

3. Multicolinealidad en el modelo de regresión lineal simple

El modelo de regresión lineal simple es aquel en el que $p = 2$, es decir, el modelo (1) quedaría:

$$\mathbf{y}_t = \beta_1 + \beta_2 \mathbf{X}_{2t} + \mathbf{u}_t, \quad t = 1, \ldots, n. \tag{3}$$

Si se considera que $\mathbf{X}_1$ no es una variable independiente del modelo sino una simple constante, se podría pensar que en este modelo no puede existir un problema de multicolinealidad grave.
Si por el contrario, sí se la considera como tal, la multicolinealidad grave podría aparecer si la variable $\mathbf{X}_2$ es prácticamente constante. Es decir, cuando su varianza es muy pequeña.
La realidad es que considerando, por ejemplo, los siguientes valores:

$$\mathbf{y_t} = \begin{pmatrix} 3 \\ 2 \\ -1 \\ 5 \end{pmatrix}, \quad \mathbf{X_{2t}} = \begin{pmatrix} 3,1 \\ 2,9 \\ \mathbf{3} \\ 3,1 \end{pmatrix}, \quad \mathbf{X^*_{2t}} = \begin{pmatrix} 3,1 \\ 2,9 \\ \mathbf{3,12} \\ 3,1 \end{pmatrix},$$

se obtienen estimadores muy distintos tanto en signo como en magnitud, $\hat{\beta} = (-39, 13{,}6364)$ y $\hat{\beta}^* = (3{,}66873, -0{,}4644)$, siendo éste uno de los síntomas de la multicolinealidad (inestabilidad en las estimaciones realizadas ante leves cambios en la muestra).

4. El Factor de Inflación de la Varianza en el modelo de regresión lineal simple

Para calcular el FIV en el modelo (3) se tiene que calcular el coeficiente de determinación de la regresión auxiliar que tiene como variable dependiente a $\mathbf{X}_2$ y como independiente sólo al término independiente. Esto es:

$$\mathbf{X}_2 = \beta_1 + \mathbf{v}_t, \quad t = 1, \ldots, n. \tag{4}$$

Este modelo es usualmente conocido como modelo restringido.
Por otro lado, el coeficiente de determinación del modelo (1) se obtiene como:

$$R^2 = 1 - \frac{SCR}{SCT}, \tag{5}$$

donde *SCR* y *SCT* son, respectivamente, la suma de cuadrados de los residuos y totales del modelo global (1).

Ahora bien, se verifica que la *SCT* coincide con la suma de cuadrados de los residuos del modelo restringido ya que en tal caso la estimación del término independiente coincide con la media de la variable dependiente.
Teniendo en cuenta ésta última interpretación, es claro que el coeficiente de determinación del modelo (4) es siempre cero independientemente de cuáles sean los datos con los que se trabaja. Esto se debe a que en este caso el modelo global y el restringido son el mismo y, por tanto, el cociente de sumas de cuadrados es igual a 1.
En tal caso, el FIV del modelo (3) será siempre igual a 1, su mínimo valor. Por tanto, si el FIV es siempre igual a 1 independientemente de cuáles sean los datos, es claro que no se puede usar en este caso como herramienta que detecta la existencia de multicolinealidad grave.

5. Cálculo del Factor de Inflación de la Varianza en el modelo de regresión lineal simple en el entorno R

Dentro de nuestro conocimiento, en el entorno de programación **R** se puede calcular el FIV a partir de los paquetes *fmsb* y *car*. A continuación se muestra cómo se calcularía el FIV en el ejemplo considerado en el apartado 3 y los resultados obtenidos:

```
> y = c(3,2,-1,5)
> x2 = c(3.1, 2.9, 3, 3.1)
> x2bis = c(3.1, 2.9, 3.12, 3.1)
>
> library(fmsb)
> VIF(lm(y~x2))
[1] 1.375
> VIF(lm(y~x2bis))
[1] 1.000372
>
> library(car)
> vif(lm(y~x2))
Error in vif.default(lm(y ~ x2)): model contains fewer than 2 terms
> vif(lm(y~x2bis))
Error in vif.default(lm(y ~ x2bis)): model contains fewer than 2 terms
>
```

Se puede observar que con el primer paquete, *fmsb*, se obtienen valores distintos de 1, lo cual se contradice totalmente con lo demostrado en el apartado 4.
Mientras que con el segundo, *car*, se obtiene un mensaje informando de que no se puede calcular el FIV.

6. Conclusiones

El FIV es una de las herramientas más usadas para detectar si el grado de multicolinealidad existente en un modelo de regresión lineal es grave. En el presente trabajo se muestra que, para la regresión simple, de las dos librerías usadas para calcular el FIV en **R**:

- la primera, *fmsb*, proporciona valores erróneos (incluso lo hemos comprobado en el caso general), y
- la segunda, *car*, da un mensaje de error diciendo que no puede calcularse, cuando esto es falso.

Por tanto, por un lado, se aconseja no calcular el FIV a partir de la librería *fmsb* y, por otro, mejorar la función de la librería *car* indicando que en el caso del modelo lineal simple el FIV es siempre igual a uno independientemente de los datos. Y que de esta forma esta herramienta queda totalmente inavalidada para detectar si existe multicolinealidad grave en este modelo.

7. Referencias

- John Fox and Sanford Weisberg (2011). An R Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.
- Minato Nakazawa (2017). fmsb: Functions for Medical Statistics Book with some Demographic Data. R package version 0.6.1. URL: https://CRAN.R-project.org/package=fmsb.