

El paquete “multiColl” para detectar multicolinealidad preocupante

I Congreso & XII Jornadas de Usuarios de R

Román Salmerón (romansg@ugr.es)
Catalina García (cbgarcia@ugr.es)



Dpto. Métodos Cuantitativos para
la Economía y la Empresa
Universidad de Granada



I Congreso & XII Jornadas de usuarios de R
Universidad de Córdoba
Comunidad R Hispano

Córdoba, 23-25 noviembre 2022

Regresión lineal múltiple y multicolinealidad

Breve introducción

Econometría es una rama de la Economía que proporciona una base para refinar o refutar conocimiento teórico y conseguir signos y magnitudes de las relaciones de variables que se desean analizar.

Dado el modelo lineal general con n observaciones y p variables independientes:

$$Y = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_i \cdot X_i + \cdots + \beta_p \cdot X_p + u = X\beta + u,$$

el objetivo anterior se consigue a partir de la estimación numérica de los coeficientes de dicho modelo y su inferencia asociada.

Aplicando Mínimos Cuadrados Ordinarios (MCO), se obtiene la estimación (signo y magnitud) de cada coeficiente mediante $\widehat{\beta} = (X^t X)^{-1} X^t Y$.

Si las variables que forman X son linealmente dependientes, entonces no existe $(X^t X)^{-1}$ (multicolinealidad perfecta) y, en consecuencia, no se puede obtener $\widehat{\beta}$.

En el caso de que las relaciones lineales sean altas aunque no perfectas, existe $(X^t X)^{-1}$, pero puede verse afectado el análisis numérico (estimación de $\widehat{\beta}$) y estadístico (inferencia) del modelo.

Paquete "multiColl": *CNs, CVs, ki, RdetR, SLM, VIF, perturb.n, multiColl.M*.

Para más detalles ver los trabajos de Salmerón et al. [2], [3], [4], [5].

Función *multiCollM* del paquete “multiColl”

Código

Ejemplo con los datos de Longley, J. W. (1967) [1] sobre datos macroeconómicos en los que el número de empleados se explica a partir del PIB, PIB deflactado, número de desempleados, número de personas en las fuerzas armadas, población mayor a los 14 años y el año.

```
attach(longley)
longley.y = Employed
dummy = ifelse(Armed.Forces>200, 1, 0)
longley.X = cbind(array(1,length(Employed)), longley[,-7], dummy)
multiCollM(longley.y, longley.X, dummy=TRUE, pos1=8, n=5000, mu=5, dv=5, tol=0.01,
pos2 = 1:6)
```

- *pos1* posición de las variables binarias/dummies
- *n* número iteraciones en las que se realiza perturbación
- *mu* y *dv* cualquier número real para generar aleatoriamente vector que perturba
- *tol* porcentaje perturbación a incluir en las variables
- *pos2* posición de las variables a perturbar una vez excluida la constante

Función *multiCollM* del paquete "multiColl"

Resultados: análisis del modelo lineal general

\$‘Linear Model’

Residuals:

Min	1Q	Median	3Q	Max
-0.12623	-0.07790	-0.04562	0.04243	0.30007

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-3.710e+03	5.134e+02	-7.227	9.01e-05	***
xGNP.deflator	-3.864e-02	5.021e-02	-0.769	0.46372	
xGNP	-6.434e-02	2.028e-02	-3.173	0.01313	*
xUnemployed	-2.344e-02	2.896e-03	-8.092	4.02e-05	***
xArmed.Forces	-2.229e-02	2.983e-03	-7.472	7.11e-05	***
xPopulation	2.962e-01	1.518e-01	1.951	0.08683	.
xYear	1.934e+00	2.623e-01	7.372	7.82e-05	***
xdummy	3.107e+00	7.061e-01	4.399	0.00229	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1749 on 8 degrees of freedom

Multiple R-squared: 0.9987, Adjusted R-squared: 0.9975

F-statistic: 863.2 on 7 and 8 DF, p-value: 7.151e-11

Función *multiCollM* del paquete "multiColl"

Resultados: perturbación, coeficientes de variación correlaciones simples

\$Perturbation

2.5%	97.5%
100.0553	103.0528

\$multiCol

\$multiCol\$'Coeficients of Variation'

[1] 0.102761096 0.248230907 0.283339359 0.258497138 0.057358090 0.002358543

\$multiCol\$'Proportion of ones in the dummys variable'

[1] 75

\$multiCol\$'R and det(R)'

\$multiCol\$'R and det(R) '\$'Correlation matrix'

	GNP.deflator	GNP	Unemployed	Armed.Forces	Population	Year
GNP.deflator	1.0000000	0.9915892	0.6206334	0.4647442	0.9791634	0.9911492
GNP		0.9915892	1.0000000	0.6042609	0.4464368	0.9910901
Unemployed			0.6206334	0.6042609	1.0000000	-0.1774206
Armed.Forces				0.4647442	0.4464368	-0.1774206
Population					1.0000000	0.3644163
Year						0.4172451

\$multiCol\$'R and det(R) '\$'Correlation matrix's determinant'

[1] 1.579615e-08

Función *multiCollM* del paquete "multiColl"

Resultados: factores inflación varianza, número de condición e índice Stewart

```
$multiCol$'Variance Inflation Factors'
```

GNP.deflator	GNP	Unemployed	Armed.Forces	Population	Year
144.04453	1992.20279	35.93856	21.14690	547.10740	765.31143

```
$multiCol$CN
```

```
$multiCol$CN$'Condition Number without intercept'
```

```
[1] 1296.066
```

```
$multiCol$CN$'Condition Number with intercept'
```

```
[1] 46046.61
```

```
$multiCol$CN$'Increase (in percentage)'
```

```
[1] 97.18532
```

```
$multiCol$ki
```

```
$multiCol$ki$'Stewart index'
```

```
[1] 1.377246e+08 1.379075e+04 3.431439e+04 4.835976e+02 3.376216e+02 1.672515e+05  
1.376534e+08
```

```
$multiCol$ki$'Proportion of essential collinearity in i-th independent variable  
(without intercept)'
```

```
[1] 1.04450115 5.80573510 7.43150092 6.26349181 0.32711668 0.00055597
```

```
$multiCol$ki$'Proportion of non-essential collinearity in i-th independent variable  
(without intercept)'
```

```
[1] 98.95550 94.19426 92.56850 93.73651 99.67288 99.99944
```

Conclusiones y agradecimientos

El paquete "multiColl" permite determinar:

- 1 si pequeños cambios en los datos supone importantes cambios en las estimaciones,
- 2 si las varianzas de los estimadores de los coeficientes pueden estar infladas,
- 3 el papel del término constante en las relaciones lineales (esenciales o no esenciales), y
- 4 tratar de manera diferenciada a las variables binarias/dummies.

En definitiva:

- determinar si la multicolinealidad aproximada existente es preocupante y de qué tipo es.

Y, en consecuencia:

- determinar qué método de estimación alternativo a los MCO es el más idóneo: centrar variables, ridge, alzado, LASSO, etc.

Este trabajo ha sido financiado por:

- el proyecto PP2019-El-02 de la Universidad de Granada titulado "Redefinición del factor de inflación de la varianza y de sus umbrales" y
- por el proyecto FEDER A-SEJ-496-UGR20 de la Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades titulado "Disrupción en el problema de multicolinealidad: una nueva visión para su diagnóstico y tratamiento".

Referencias

-  Longley, J. W. (1967). An appraisal of least-squares programs from the point of view of the user. *Journal of the American Statistical Association* 62, 819–841.
-  Salmerón, R., García, C. y García, J. (2019). “multicoll”: An R package to detect multicollinearity. *arXiv preprint:1910.14590*.
-  Salmerón, R., García, C.B. y García, J. (2021). A guide to using the R package “multiColl” for detecting multicollinearity. *Computational Economics*, 57, 529-536. DOI: 10.1007/s10614-019-09967-y.
-  Salmerón, R., García, C.B. y García, J. (2022). The “multiColl” package versus other existing packages in R to detect multicollinearity. *Computational Economics*, 60, 439-450. DOI: 10.1007/s10614-021-10154-1.
-  Salmerón, R., García, C.B. y García, J. (2022). “multiColl”: Collinearity Detection in a Multiple Linear Regression Model. R package version 2.0. URL: <https://CRAN.R-project.org/package=multiColl>.

¡¡¡Muchas gracias por su atención/paciencia!!!