Regresión cresta generalizada en **R** para el modelo de regresión lineal múltiple

https://github.com/rnoremlas/documentos

Román Salmerón (**romansg@ugr.es**)
Catalina García (**cbgarciag@ugr.es**)
Guillermo Hortal Reina (**ghorrei@correo.ugr.es**)

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

III Congreso & XIV Jornadas de Usuarios de R
Instituto de Matemáticas de la
Universidad de Sevilla
Grupo Local de Usuarios de R en Sevilla

III Congreso & XIV Jornadas de Usuarios de **R**
Sevilla, 6-8 noviembre 2024

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:

$$\mathbf{Y} = \beta_1 \cdot \mathbf{X}_1 + \beta_2 \cdot \mathbf{X}_2 + \cdots + \beta_i \cdot \mathbf{X}_i + \cdots + \beta_p \cdot \mathbf{X}_p + \mathbf{u} = \mathbf{X}\beta + \mathbf{u},$$

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 $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1} \mathbf{X}^t\mathbf{Y}$.

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

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

Generalized Ridge Regression

Regresión Cresta Generalizada: menor ECM que MCO

Ya que $\mathbf{X}^t\mathbf{X}$ es una matriz definida positiva, existe una matriz ortogonal $\boldsymbol{\Gamma}$ (esto es $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^t = \mathbf{I} = \boldsymbol{\Gamma}^t\boldsymbol{\Gamma}$) y una matriz diagonal $\boldsymbol{\Lambda}$ (ambas de dimensiones $p \times p$) tales que $\mathbf{X}^t\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^t$. Además, la matriz $\boldsymbol{\Gamma}$ contiene los autovectores de $\mathbf{X}^t\mathbf{X}$ y $\boldsymbol{\Lambda}$ sus autovalores. A partir de esta descomposición...

Hoerl y Kennard [1] (page 70) proporcionan la siguiente expresión:

$$\widehat{\beta}(\mathbf{K}) = \left(\mathbf{X}^t\mathbf{X} + \mathbf{K}\right)^{-1} \mathbf{X}^t\mathbf{Y}, \text{ siendo } \mathbf{K} = \begin{pmatrix} k_1 & 0 & \dots & 0 \\ 0 & k_2 & \dots & 0 \\ \vdots & \vdots & & 0 \\ 0 & 0 & \dots & k_p \end{pmatrix}, \tag{1}$$

cuando realmente es:

$$\widehat{\beta}(\mathbf{K}) = \left(\mathbf{X}^t\mathbf{X} + \boldsymbol{\Gamma}\mathbf{K}\boldsymbol{\Gamma}^t\right)^{-1} \mathbf{X}^t\mathbf{Y}. \tag{2}$$

Es claro que para el caso regular (que es el tradicionalmente usado), $\mathbf{K} = k \cdot \mathbf{I}$, las dos expresiones anteriores coinciden. Sin embargo, para el caso generalizado donde los k_i, $i = 1, \dots, p$, son distintos, estamos ante dos expresiones distintas.

Generalized Ridge Regression

Regresión Cresta Generalizada

Generalized ridge regression: biased estimation for multiple linear regression models [2]

Se obtiene una expresión para su estimador, norma, error cuadrático medio (ECM), bondad de ajuste (BdA) y se propone hacer inferencia *bootstrap*.

Objetivo: puesto que los estimadores propuestos son sesgados, se busca que tengan un menor error cuadrático medio.

Ejemplo de Klein y Goldberger [3]

El consumo, **C**, es analizado en función del ingreso salarial, **IS**, ingreso no salarial, **InS**, y el ingreso agrícola, **IA**.

¿Sin pérdida de generalidad? vamos a considerar datos estandarizados: restamos la media y dividimos entre la raíz cuadrada de la varianza por el número de observaciones.

Casos considerados:

1. Estimación cresta regular: $\mathbf{K} = k \cdot \mathbf{I}$ con $k = 0$ (MCO), $k = p \cdot \frac{\sigma^2}{\beta^t \beta} = 0{,}0500635667908736$ (Hoerl, Kennard y Baldwin [4]) y $k = \frac{\sigma^2}{\xi^2_{max}} = 0{,}0221644584150512$ (minimiza el ECM según Hoerl y Kennard [1]).
2. Estimación cresta generalizada: $\mathbf{K} = diag(0{,}02216446, 0{,}06954418, 2{,}34086947)$ (minimiza el ECM según [1]) y $\mathbf{K} = diag(0, 0, 2{,}34086947)$ (minimiza el ECM según [2]).

Generalized Ridge Regression
Regresión Cresta Generalizada

Generalized ridge regression: biased estimation for multiple linear regression models [2]

Se obtiene una expresión para su estimador, norma, error cuadrático medio (ECM), bondad de ajuste (BdA) y se propone hacer inferencia *bootstrap*.
Objetivo: puesto que los estimadores propuestos son sesgados, se busca que tengan un menor error cuadrático medio.

Ejemplo de Klein y Goldberger [3]

El consumo, **C**, es analizado en función del ingreso salarial, **IS**, ingreso no salarial, **InS**, y el ingreso agrícola, **IA**.

¿Sin pérdida de generalidad? vamos a considerar datos estandarizados: restamos la media y dividimos entre la raíz cuadrada de la varianza por el número de observaciones.
Casos considerados:

1. Estimación cresta regular: $\mathbf{K} = k \cdot \mathbf{I}$ con $k = 0$ (MCO), $k = p \cdot \frac{\sigma^2}{\beta^t \beta} = 0{,}0500635667908736$ (Hoerl, Kennard y Baldwin [4]) y $k = \frac{\sigma^2}{\xi^2_{max}} = 0{,}0221644584150512$ (minimiza el ECM según Hoerl y Kennard [1]).
2. Estimación cresta generalizada: $\mathbf{K} = diag(0{,}02216446, 0{,}06954418, 2{,}34086947)$ (minimiza el ECM según [1]) y $\mathbf{K} = diag(0, 0, 2{,}34086947)$ (minimiza el ECM según [2]).

Generalized Ridge Regression

Ejemplo de Klein y Goldberger [3]

$$\mathbf{C} = \beta_1 \mathbf{IS} + \beta_2 \mathbf{InS} + \beta_3 \mathbf{IA} + \mathbf{u}$$

	$k = 0$	$k = k_{HKB}$	$k = k_{HK}$	$k_i = \frac{\sigma^2}{\xi_i^2}$	$\mathbf{K} = diag(0, 0, k_3)$
IS	0.3852 (-0.589, 1.0535)	0.3883 (-0.0137, 0.6861)	0.3898 (-0.1841, 0.799)	0.4081 (0.3059, 0.4987)	0.427 (0.3156, 0.5343)
InS	0.5394 (-0.0444, 1.2903)	0.4904 (0.1931, 0.8293)	0.5135 (0.1189, 0.9859)	0.4694 (0.364, 0.6037)	0.5047 (0.3934, 0.7112)
IA	0.0592 (-0.2309, 0.3481)	0.0884 (-0.1114, 0.2321)	0.0729 (-0.1623, 0.249)	0.0998 (-0.1067, 0.2558)	0.0504 (-0.247, 0.2362)
ECM	0.1811	0.0614	0.0975	0.0265	0.0315
BdA	0.9187 (0.8201, 0.9944)	0.9177 (0.8128, 0.9897)	0.9185 (0.8182, 0.9924)	0.9174 (0.7923, 0.9853)	0.9186 (0.7963, 0.986)

- Inicialmente ($k = 0$, MCO), ningún coeficiente es significativamente distinto de cero (inferencia tradicional).
- En el caso regular, el coeficiente de **InS** es significativamente distinto de cero (inferencia *bootstrap*).
- En el caso generalizado, los coeficientes de **IS** y **InS** son significativamente distintos de cero (inferencia *bootstrap*).

Generalized Ridge Regression
Ejemplo de Klein y Goldberger [3]

$$\mathbf{C} = \beta_1 \mathbf{IS} + \beta_2 \mathbf{InS} + \beta_3 \mathbf{IA} + \mathbf{u}$$

	$k = 0$	$k = k_{HKB}$	$k = k_{HK}$	$k_i = \frac{\sigma^2}{\xi_i^2}$	$\mathbf{K} = diag(0, 0, k_3)$
IS	0.3852	0.3883	0.3898	0.4081	0.427
	(-0.589, 1.0535)	(-0.0137, 0.6861)	(-0.1841, 0.799)	(0.3059, 0.4987)	(0.3156, 0.5343)
InS	0.5394	0.4904	0.5135	0.4694	0.5047
	(-0.0444, 1.2903)	(0.1931, 0.8293)	(0.1189, 0.9859)	(0.364, 0.6037)	(0.3934, 0.7112)
IA	0.0592	0.0884	0.0729	0.0998	0.0504
	(-0.2309, 0.3481)	(-0.1114, 0.2321)	(-0.1623, 0.249)	(-0.1067, 0.2558)	(-0.247, 0.2362)
ECM	0.1811	0.0614	0.0975	0.0265	0.0315
BdA	0.9187	0.9177	0.9185	0.9174	0.9186
	(0.8201, 0.9944)	(0.8128, 0.9897)	(0.8182, 0.9924)	(0.7923, 0.9853)	(0.7963, 0.986)

- En el caso generalizado se obtienen los menores errores cuadráticos medios.
- La bondad de ajuste empeora con respecto a MCO (aunque no mucho) y, en el caso regular, decrece conforme aumenta el valor de k.

Generalized Ridge Regression

Ejemplo de Klein y Goldberger [3]: comparación con paquetes existentes en R

- El paquete *lrmest* [6] no permite transformar los datos: hay que introducirlos ya transformados.
- Los paquetes *lmridge* [5] y *ridge* [7] sí tienen argumentos para transformar los datos: para estandarizar los datos opciones "sc" y "corrform"; para tipificarlos las opciones "scaled" y "scale".

$k = 0$ (MCO)

	GRR	*lrmest*	*lmridge*			*ridge*	
			sc	scaled	centered	corrform	scale
IS	0.3852	0.3852	0.3803	0.3803	0.3803	0.3803	0.3803
InS	0.5394	0.5394	1.4186	1.4186	1.4186	1.4186	1.4186
IA	0.0592	0.0592	0.5331	0.5331	0.5331	0.5331	0.5331
ECM	0.1811	0.1811	818.0679	62.9283	2.3416		
BdA	0.9187		0.9187	0.9187	0.9187		

- Las estimaciones porporcionadas por GRR y *lrmest* son iguales entre sí y distintas a las proporcionadas por *lmridge* y *ridge* que, a su vez, coinciden entre sí (aunque se observan diferencias en cuanto a la significación individual).
- El ECM proporcionado por GRR y *lrmest* son iguales entre sí y distintos a los proporcionados por *lmridge* (que proporciona un valor distinto en cada transformación considerada).
- La bondad de ajuste de GRR coincide con la de *lmridge*.

Generalized Ridge Regression

Ejemplo de Klein y Goldberger [3]: comparación con paquetes existentes en R

$k = k_{HKB} = 0{,}05006357$

	GRR	*lrmest*	*lmridge* sc	*lmridge* scaled	*lmridge* centered	*ridge* corrform	*ridge* scale
IS	0.3883	0.3883	0.3834	0.3818	0.3808	0.3834	0.3818
InS	0.4904	0.4904	1.2897	1.4046	1.4176	1.2897	1.4046
IA	0.0884	0.0884	0.7954	0.5547	0.5315	0.7954	0.5547
ECM	0.0614	0.0614	274.2663	55.1524	2.3309		
BdA	0.9177		0.8778	0.9153	0.9187		

- Al igual que antes, las estimaciones porporcionadas por GRR y *lrmest* son iguales entre sí y distintas a las proporcionadas por *lmridge* y *ridge* que, a su vez, coinciden entre sí cuando la transformación coincide (aunque hay diferencias en cuanto a la significación individual).
- El ECM proporcionado por GRR y *lrmest* son iguales entre sí y distintos a los proporcionados por *lmridge* (que proporciona un valor distinto en cada transformación considerada).
- La bondad de ajuste de GRR no coincide con la de *lmridge* (que, además, proporciona un valor distinto para cada transformación usada).

Generalized Ridge Regression

Ejemplo de Klein y Goldberger [3]: comparación con paquetes existentes en R

$k = k_{HK} = 0{,}02216446$

	GRR	*lrmest*	*lmridge*			*ridge*	
			sc	scaled	centered	corrform	scale
IS	0.3898	0.3898	0.3849	0.3810	0.3805	0.3849	0.3810
InS	0.5135	0.5135	1.3506	1.4122	1.4181	1.3506	1.4123
IA	0.0729	0.0729	0.6559	0.5426	0.5324	0.6559	0.5426
ECM	0.0975	0.0975	437.8791	59.2723	2.3369		
BdA	0.9185		0.8998	0.9172	0.9187		

Además de las cuestiones anteriores:

- El paquete *lrmest* no proporciona la bondad de ajuste.
- El paquete *lmridge* porporciona el ECM y bondad de ajuste, pero no coincide con el calculado mediante GRR (excepto la BdA para $k = 0$).
- El paquete *ridge* no porporciona ni el ECM y ni la bondad de ajuste.

Conclusiones
y agradecimientos

Hemos visto que la versión generalizada del estimador cresta:

1. proporciona estimadores con menor error cuadrático medio que la versión regular,
2. no es una estimación proporcionada por los paquetes existentes en R para el estimador cresta,
3. incluso en su versión regular, hemos encontrado diferencias entre los resultados obtenidos (estimación, ECM y bondad de ajuste) y los que proporcionan dichos paquetes.

Por tanto, para un futuro...

a) vamos a intentar crear un paquete en R a partir del código que hemos generado y

b) abordaremos el análisis de su utilidad para mitigar el problema de multicolinealidad (no sólo buscar un menor ECM).

Este trabajo ha sido financiado por:

- el Departamento de Métodos Cuantitativos para la Economía y la Empresa de la Universidad de Granada.

Referencias (por orden de aparición)

- Hoerl, A.E. y Kennard, R.W. (1970). Ridge regression: applications to nonorthogonal problems. Technometrics 12(1), 69–82.
- Salmerón, R., García, C.B. y Hortal, G. (2024). Generalized ridge regression: biased estimation for multiple linear regression models, arXiv: `https://arxiv.org/abs/2407.02583`.
- Klein, L. y Goldberger, A. (1964). An Economic Model of the United States, 1929–1952; North Holland Publishing Company: Amsterdan, The Netherlands.
- Hoerl, A.E., Kennard, R.W. y Baldwin, K.F. (1975). Ridge regression: some simulations. Communications in Statistics - Theory and Methods 4, 105-123.
- Imdad, M.U. y Aslam, M. (2023). *lmridge*: Linear Ridge Regression with Ridge Penalty and Ridge Statistics. R package version 1.2.2, `https://CRAN.R-project.org/package=lmridge`.
- Dissanayake, A. y Wijekoon, P. (2016). *lrmest*: Different Types of Estimators to Deal with Multicollinearity. R package version 3.0, `https://CRAN.R-project.org/package=lrmest`.
- Cule, E., Moritz, S. y Frankowski, D. (2022). *ridge*: Ridge Regression with Automatic Selection of the Penalty Parameter. R package version 3.3, `https://CRAN.R-project.org/package=ridge`.

Regresión cresta generalizada en **R** para el modelo de regresión lineal múltiple

https://github.com/rnoremlas/documentos

Román Salmerón (**romansg@ugr.es**)
Catalina García (**cbgarciag@ugr.es**)
Guillermo Hortal Reina (**ghorrei@correo.ugr.es**)

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

III Congreso & XIV Jornadas de Usuarios de R
Instituto de Matemáticas de la Universidad de Sevilla
Grupo Local de Usuarios de R en Sevilla

III Congreso & XIV Jornadas de Usuarios de **R**
Sevilla, 6-8 noviembre 2024

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