Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Programa Académico de Estadística
Estadística Aplicada II
Cali, mayo de 2018

Tarea 4: Multicolinealidad

Kevin Steven García[a], Alejandro Vargas[b]

1. Introducción

En este trabajo evaluaremos la presencia de multicolinealidad en el modelo lineal ajustado para los 500 datos seleccionados de la base de datos 'cadata'. Se dará una presentación formal de la multicolinealidad y se mostraran algunas consecuencias que tiene la presencia de este, además, se describirán y se aplicarán distintos métodos para verificar la presencia o ausencia de esta en nuestro modelo lineal ajustado, posteriormente se tratara de solucionar esto, llevando a cabo estimación por PCR(regresión por componentes principales) y finalmente, se concluirá acerca del modelo teniendo en cuenta los resultados de los métodos aplicados y de la estimación por PCR comparada con los estimadores por MCO(mínimos cuadrados ordinarios) .

2. Modelo planteado

El modelo sobre el cuál se va a evaluar la presencia de la multicolinealidad, es el modelo que se planteó en la tarea 1, que pretendía explicar la variable 'Valor mediano de las viviendas' con las variables explicativas 'Ingreso mediano','Edad mediana de la vivienda','Total de habitaciones','Total de dormitorios','Población' y 'Hogares'. El modelo planteado para estas variables, sin hacer transformación ni selección de estas es:

$$ValorMediano = \beta_0 + \beta_1(IngresoMediano) + \beta_2(EdadMediana) + \beta_3(TotalDeHabitaciones)$$

$$+\beta_4(TotalDeDormitorios) + \beta_5(Poblacion) + \beta_6(Hogares)$$

3. Multicolinealidad

La multicolinealidad implica una dependencia casi lineal entre los regresores, los cuales son las columnas de la matriz X, por lo que es claro que una dependencia lineal exacta causaría una matriz X^TX singular. La presencia de dependencias casi lineales puede influir en forma dramática sobre la capacidad de estimar coeficientes de regresión afectando la precisión de las estimaciones. Cuando existe multicolinealidad en nuestro modelo, el análisis por mínimos cuadrados puede ser totalmente inadecuado.
La multicolinealidad se da por cuatro fuentes principales:

1. El método de recolección de datos que se empleó:Puede originar problemas de multicolinealidad cuando el analista sólo muestrea un subespacio de la región de los regresores definidos.

[a]Código: 1533173. E-mail: kevin.chica@correounivalle.edu.co
[b]Código: 1525953. E-mail: jose.alejandro.vargas@correounivalle.edu.co

2. Restricciones en el modelo o en la población:Las restricciones en el modelo o en la población que se muestrea pueden causar multicolinealidad; por ejemplo, supóngase que una empresa eléctrica está investigando el efecto del ingreso familiar (X_l) y el tamaño de la vivienda (X_2) sobre el consumo eléctrico residencial.En este ejemplo, una restricción física en la población fue lo que causó este fenómeno: las familias que tienen ingresos mayores en general tienen casas mayores que las familias de menores ingresos, cuando hay restricciones físicas como ésta, habrá multicolinealidad independientemente del método de muestreo que se emplee.

3. Especificación del modelo:También se puede inducir la multicolinealidad por la elección del modelo. Por ejemplo,al agregar términos polinomiales a un modelo de regresión se produce un deterioramiento en X^TX, además, si el rango de X es pequeño, al agregar un término en X^2 puede producirse una multicolinealidad importante.

4. Un modelo sobredefinido:Cuando se tienen mas variables regresoras que observaciones se puede generar un problema de multicolinealidad.

Una multicolinealidad fuerte entre los regresores del modelo,da como resultado grandes varianzas y covarianzas de los estimadores de coeficientes de regresión por mínimos cuadrados.Esto implica que distintas muestras tomadas con los mismos valores de X podrían ocasionar estimaciones muy diferentes de los parámetros del modelo.

4. Métodos para detectar la multicolinealidad

Para este trabajo se utilizaran los métodos de diagnostico expuestos en el libro (Douglas C. Montgomery 2002), que son los siguientes:

1. Examen de la matriz de correlación:Una medida muy sencilla de la multicolinealidad es la inspección de los elementos r_{ij} no diagonales en la matriz de correlaciones. Si los regresores x_i y x_j son casi linealmente dependientes $|r_{ij}|$ será próximo a 1. Para nuestros datos, la matriz de correlaciones es la siguiente:

$$R = \begin{pmatrix} 1 & 0.01725 & 0.07903 & -0.18136 & -0.16257 & -0.16049 \\ & 1 & -0.31169 & -0.17024 & -0.26517 & -0.14456 \\ & & 1 & 0.86207 & 0.85768 & 0.86527 \\ & & & 1 & 0.83406 & 0.98870 \\ & & & & 1 & 0.85391 \\ & & & & & 1 \end{pmatrix}$$

   La matriz de correlaciones R, nos muestra perfectamente la alta correlación que existe entre las últimas 4 variables regresoras o explicativas (Total de habitaciones, total de dormitorios, población y hogares) , por lo cuál es casi evidente que una o varías de estas sobran en el modelo y pueden ocasionar multicolinealidad.

2. Factores de inflación de varianza:Los elementos diagonales de la matriz $C = (X^TX)^{-1}$ son muy útiles para detectar la multicolinealidad. C_{jj}, el j-ésimo elemento diagonal de C se puede escribir en la forma $C_{jj} = (1 - R_j^2)^-1$ siendo R_j^2 el coeficiente de determinación obtenido cuando se hace la regresión de x_j respecto a los demás $p-1$ regresores. Si x_j es casi ortogonal a los regresores restantes, R_j^2 es pequeño y C_{jj} es cercano a 1, mientras que si x_j es casi linealmente dependiente en algún subconjunto de los regresores restantes, R_j^2 es casi lineal y C_{jj} es grande. Como la varianza de los j-ésimos coeficientes de regresión es $C_{jj}\sigma^2$ se puede considerar que C_{jj} es el factor en el que aumenta la varianza de $\hat{\beta}_j$ debido a dependencias casi lineales entre los regresores.
   Entonces se define el factor de inflación de varianza como:

$$VIF_j = C_{jj} = (1 - R_j^2)^{-1}$$

El factor VIF (de variance inflation factor) para cada término del modelo mide el efecto combinado que tienen las dependencias entre los regresores sobre la varianza de ese término. Si hay uno o más VIF grandes, hay multicolinealidad. El criterio dice que si cualquiera de los VIF es mayor que 5 o 10, es indicio de que los coeficientes asociados de regresión están mal estimados debido a la multicolinealidad.
Los VIF_i para este modelo son:

VARIABLE	Ing.Mediano	Edad Mediana	T.Habitaciones	T.Dormitorios	Población	Hogares
VIF	1.565795	1.271698	8.203393	52.2694	5.675371	56.12648

Podemos ver que hay 2 VIF mayores que 5 y 2 VIF mayores que 10, entonces podemos afirmar que los coeficientes de regresión asociados a estas variables están mal estimados debido a la multicolinealidad.

3. Análisis del eigensistema de X^TX:
Las raíces características, o eigenvalores o valores propios de X^TX, por ejemplo $\lambda_1, \lambda_2, ..., \lambda_p$ se pueden usar para medir el grado de multicolinealidad en los datos. Si hay una o más dependencias casi lineales en los datos, una o más de las raíces características será pequeña. Uno o más eigenvalores pequeños implican que hay dependencias casi lineales entre las columnas de X. Algunos analistas prefieren examinar el número de condición de X^TX, que se define como

$$k = \frac{\lambda_{max}}{\lambda_{min}}$$

En general, si el número de condición es menor que 100, no hay problema grave de multicolinealidad. Los números de condición de 100 a 1 000 implican multicolinealidad de moderada a fuerte, y si k es mayor que 1 000, es indicio de una fuerte multicolinealidad. Los Índices de condición de la matriz X^TX son

$$k_j = \frac{\lambda_{max}}{\lambda_j}; j = 1, 2, ..., p$$

La cantidad de índices de condición que son grandes (digamos, ≥ 1000) es una medida útil de la cantidad de dependencias casi lineales en X^TX.
Los valores propios para la matriz $X^TX = R$ (en forma de correlación) correspondiente a nuestros datos son:
$\lambda_1 = 3.72181333, \lambda_2 = 1.05255886, \lambda_3 = 0.93404888, \lambda_4 = 0.19394997, \lambda_5 = 0.08822446, \lambda_6 = 0.00940450$
De aquí, tenemos que $\lambda_{max} = 3.72181333$ y $\lambda_{min} = 0.00940450$, por lo tanto el número de condición es:

$$k = \frac{\lambda_{max}}{\lambda_{min}} = \frac{3.72181333}{0.00940450} = 395.7481344$$

El cuál es demasiado grande, lo que nos indica una multicolinealidad fuerte en nuestro modelo con todas las variables.
Y los indices de condición son:

$$k_1 = \frac{3.72181333}{3.72181333} = 1, k_2 = \frac{3.72181333}{1.05255886} = 3.5359669, k_3 = \frac{3.72181333}{0.93404888} = 3.9846023$$

$$k_4 = \frac{3.72181333}{0.19394997} = 19.189553, k_5 = \frac{3.72181333}{0.08822446} = 42.18573092, k_6 = \frac{3.72181333}{0.00940450} = 395.7481344$$

Podemos ver que 1 indices de condición es mayor que 100, lo que nos dice que al menos hay 1 dependencias fuerte casi lineales en los datos de nuestro modelo (columnas de X^TX).

4. Determinante de X^TX en forma de correlación: Se puede usar el determinante de X^TX como índice de colinealidad; ya que la matriz X^TX está en forma de correlación, el intervalo de posibles valores del determinante es $0 \leq |X^TX| \leq 1$. Si $|X^TX| = 1$, los regresores son

ortogonales, mientras que si $|X^TX| = 0$, hay una dependencia lineal exacta entre ellos. El grado de multicolinealidad se agrava a medida que $|X^TX|$ tiende a cero. Si bien esta medida de multicolinealidad es fácil de aplicar, no proporciona información alguna sobre el origen de la multicolinealidad.
Para nuestro caso el determinante de la matriz de correlaciones X^TX es:

$$|X^TX| = |R| = 0.0005888233$$

Por lo cual podemos afirmar que hay una multicolinealidad grave en nuestro modelo.

5. Regresión por componentes principales (PCR)

Los coeficientes de regresión se pueden obtener por un procedimiento denominado regresión por componentes principales. Sea la forma canónica del modelo:

$$y = Z\alpha + \epsilon$$

Donde $Z = XT$, $\alpha = T'\beta$, $T'X'XT = Z'Z = A$
Y $A = diag(\lambda_1, \lambda_2, ..., \lambda_p)$ es una matriz diagonal de tamaño p x p de los valores propios de $X'X$ y T es una matriz ortogonal cuyas columnas son los vectores propios asociados a los valores propios $\lambda_1, \lambda_2, ..., \lambda_p$. Las columnas de Z, definen un nuevo conjunto de regresores ortogonales, como:

$$Z = [Z_1, Z_2, ..., Z_p]$$

Se llaman componentes principales.
Entonces, el estimador de $\hat{\alpha}$ por minimos cuadrados es

$$\hat{\alpha} = (Z'Z)^{-1}Z'y = A^{-1}Z'y$$

Y la matriz de covarianzas de $\hat{\alpha}$ es

$$Var(\hat{\alpha}) = \sigma^2(Z'Z)^{-1} = \sigma^2A^{-1}$$

El método de regresión por componentes principales combate a la multicolinealidad, al usar en el modelo menos componentes que el conjunto completo de componentes principales. Para obtener el estimador de componentes principales, se supone que los regresores están ordenados por valores propios decrecientes, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$. Supóngase que los últimos s eigenvalores de éstos son aproximadamente iguales a cero. En la regresión por componentes principales, se eliminan los componentes principales que correspondan a eigenvalores cercanos a cero, y se aplican los mínimos cuadrados a los componentes restantes. Esto es,

$$\hat{\alpha_{PC}} = B\hat{\alpha}$$

donde $b_1 = b_2 = ... = b_{p-s} = 1$, y $b_{p-s+1} = b_{p-s+2} = ... = b_p = 0$.Así, el estimador de componentes principales es

$$\hat{\alpha_{PC}} = \begin{bmatrix} \hat{\alpha_1} & \hat{\alpha_2} & \cdots & \hat{\alpha_{p-s}} & | & 0 & 0 & \cdots & 0 \end{bmatrix}$$

o, en términos de los regresores normalizados,

$$\hat{\beta}_{PC} = T\hat{\alpha}_{PC} = \sum_{j=1}^{p-1} \lambda_j^{-1} t_j' X'y t_j$$

5.1. Estimación de nuestro modelo por PCR:

Para hacer uso de la regresión por componentes principales en nuestro modelo, se comienza con la transformación lineal $Z = XT$ que transforma los regresores estandarizados originales en un conjunto ortogonal de variables (los componentes principales). Los valores propios λ_j y la matriz T para los 500 datos seleccionados de la base de datos 'cadata' se muestran en la siguiente tabla.La matriz T indica la relación entre Z y los regresores estandarizados. Por ejemplo, la relación entre Z_1 y los regresores estandarizados es

$$z_1 = -0.0785x_1 - 0.1609x_2 + 0.4854x_3 + 0.4987x_4 + 0.4831x_5 + 0.5002x_6$$

VECTORES PROPIOS						VALORES PROPIOS λ_j
-0.07851439	0.88436486	-0.399610859	0.03866445	0.2232029	-0.02703696	3.72181333
-0.16091266	-0.38609154	-0.891249689	-0.14164674	-0.1006961	-0.02260359	1.05255886
0.48537035	0.22689357	-0.072328309	-0.13160716	-0.8298092	0.04242423	0.93404888
0.49876670	-0.09577328	-0.124700897	0.47098126	0.1664129	-0.69064631	0.19394997
0.48308811	-0.02353949	0.003937038	-0.77674426	0.3947270	-0.08315767	0.08822446
0.50021584	-0.08732010	-0.158707136	0.36873545	0.2606814	0.71627560	0.00940450

Y, el estimador de α es:

$$\hat{\alpha} = A^{-1}Z'y = \begin{bmatrix} -0.02697939 \\ 0.16597000 \\ -0.64338425 \\ 0.75560789 \\ -0.28640901 \\ 0.81285787 \end{bmatrix}$$

Podemos ver que hay 3 valores propios muy cercanos a cero, por lo que esos no se tendrán en cuenta y nos quedaremos con las componentes principales correspondientes a los otros tres valores propios, entonces:

$$\alpha_{\hat{PC}} = \begin{bmatrix} -0.02697939 \\ 0.16597000 \\ -0.64338425 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Y, los betas estimados con tres componentes principales son:

$$\hat{\beta}_{PC} = T\hat{\alpha}_{PC} = \begin{bmatrix} 0.40599964 \\ 0.51367772 \\ 0.07109743 \\ 0.05087868 \\ -0.01947330 \\ 0.07412164 \end{bmatrix}$$

6. Comparación resultados

en la siguiente tabla, se van a ver las comparaciones en las estimaciones de los coeficientes de regresión por el método de PCR y por MCO, se debe tener en cuenta que las estimaciones estan-

darizadas se hicieron paso a paso y las estimaciones originales se hicieron con las funciones de R.

	3 Componentes	principales	Minimos cuadrados	Ordinarios
Parámetro	Estimador Est.	Estimador Ori.	Estimador Est.	Estimador Ori.
β_0	1.130918e-17	-5495.492	-1.047e-17	57720.52
β_1	0.4060	28198.53521	0.3493	24261.20
β_2	0.5136	4241.22091	0.4171	3443.94
β_3	0.0711	5.56604	0.2438	19.09
β_4	0.0508	17.0307	-0.2023	-67.72
β_5	-0.0195	-3.01025	-0.7870	-121.66
β_6	0.0741	27.21851	0.8603	315.92

	β_1	β_2	β_3	β_4	β_5	β_6
$V_{MCO}(\hat{\beta})$	0.0014	0.0012	0.0076	0.0485	0.0052	0.0521
$V_{PCR}(\hat{\beta})$	0.0018	0.0015	0.0096	0.0616	0.0067	0.0662

Podemos ver que en todas las estimaciones la varianza aumenta un poco, esto nos indica que las estimaciones por MCO son un poco mas precisas que la estimaciones por PCR con tres componentes.
Con respecto al R^2 del modelo, también son mejores las estimaciones por MCO, ya que, $R^2_{MCO} = 0.5425316$ y $R^2_{PCR} = 0.4183462$

7. Implementación en R

Para llevar a cabo todo el análisis estadístico se utilizó el software R Core Team (2017).Para leer y cargar la base de datos 'cadata' se utilizó el paquete Wickham et al. (2017),para todo lo relacionado con el álgebra matricial implícito (valores y vectores propios, multiplicación y manejo de vectores) se realizó con las operaciones base que trae R, para las estimaciones de los coeficientes o de los $\beta's$ por MCO se utilizó la función Zeileis & Hothorn (2002) y para la estimación de los coeficientes por componentes principales se usó el paquete Mevik et al. (2016), más específicamente la función pcr, que nos arroja los componentes principales y los coeficientes estimados por el número de componentes que se desee. Se utilizó un paquete adicional para calcular los VIF, cabe decir que también se hicieron a mano con la formula dada, y los resultados fueron idénticos, ese paquete fue Nakazawa (2018).

8. Conclusiones

La regresión PCR resuelve el problema de multicolinealidad reduciendo el numero de componentes principales que el modelo completo o regresión por mínimos cuadrados utiliza, sin embargo la varianza de este método aumentó, también la de sus estimaciones, lo cual es un punto en contra, aun así es una buena alternativa para solucionar este problema. También apreciamos que el R^2 disminuyó considerablemente, esto tal vez es debido a que se quitan componentes que explican poco sobre la variabilidad del modelo y esto hace que el R^2 disminuya en proporción a cuantos componentes se eliminan del modelo.

Referencias

Douglas C. Montgomery, Elizabeth A. Peck, G. G. V. (2002), *Introducción al análisis de regresión lineal*, tercera edn, Compañía Editorial Continental.

Faraway, J. J. (2005), *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Vol. 1 edn., Chapman y Hall / CRC.

Mevik, B.-H., Wehrens, R. & Liland, K. H. (2016), *pls: Partial Least Squares and Principal Component Regression*. R package version 2.6-0.
*https://CRAN.R-project.org/package=pls

Nakazawa, M. (2018), *fmsb: Functions for Medical Statistics Book with some Demographic Data*. R package version 0.6.3.
*https://CRAN.R-project.org/package=fmsb

Norman R. Draper, H. S. (1998), *Applied Regression Analysis, Third Edition (Wiley Series in Probability and Statistics)*, third edn.
*http://gen.lib.rus.ec/book/index.php?md5=903983CE6C8FBCBB1F81265002FE576B

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

Wickham, H., Hester, J. & Francois, R. (2017), *readr: Read Rectangular Text Data*. R package version 1.1.1.
*https://CRAN.R-project.org/package=readr

Zeileis, A. & Hothorn, T. (2002), 'Diagnostic checking in regression relationships', *R News* **2**(3), 7–10.
*https://CRAN.R-project.org/doc/Rnews/