# CAPÍTULO 6

# MULTICOLINEALIDAD

**Luis Quintana Romero y Miguel Ángel Mendoza**

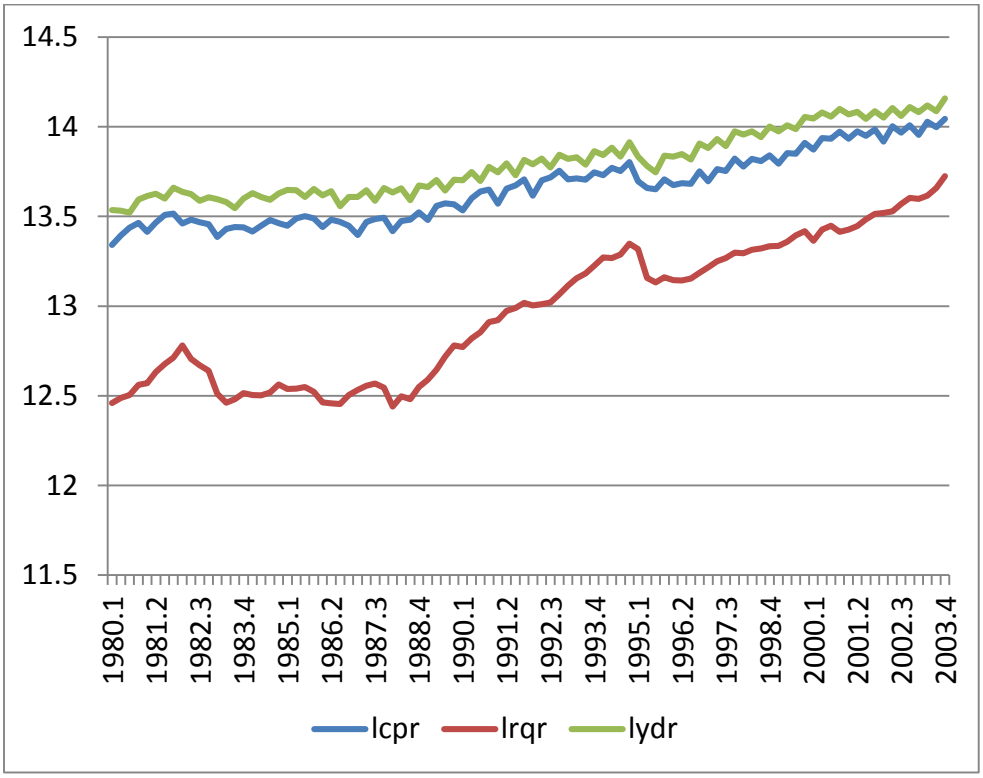
## 1. LA MULTICOLINEALIDAD UN PROBLEMA DE GRADO

La multicolinealidad debe considerarse como un problema de grado que se presenta de manera cotidiana en los modelos econométricos. Esto significa que el comportamiento de buena parte de las variables económicas guarda algún tipo de relación unas con otras y esa relación puede ser de menor o mayor grado. Solamente cuando dicha relación es de mayor grado podría ser un problema dentro de la modelación econométrica tal y como veremos a continuación.

Para ilustrar la relación que existe entre las variables económicas en la gráfica siguiente se muestran los valores logarítmicos trimestrales del consumo privado real, el ingreso nacional disponible real y la riqueza real para la economía mexicana de 1980 a 2003. En la gráfica se observa que el consumo y el ingreso prácticamente tiene el mismo comportamiento, mientras que la riqueza tiene la misma tendencia que las otras dos variables; la gráfica muestra una clara asociación positiva entre las tres variables, lo cual implica que debe de existir algún grado de asociación lineal entre las variables que hemos seleccionado.

**Gráfica 1**

**Consumo, ingreso y riqueza por trimestre en México 1980-2003**




Si bien las variables muestran trayectorias similares existen diferencias entre ellas, por ende están relacionadas de forma aproximada pero no exacta, esto nos permite plantear que la multicolinealidad es la relación perfecta o no, que se da entre variables económicas.

La relación exacta entre las variables se denomina multicolinealidad perfecta, lo cual significa que alguna o algunas de las variables que forman las columnas de la matriz de regresores sería una combinación lineal exacta del resto de columnas. Por ejemplo, si suponemos que la matriz de regresores se compone de tres columnas con las variables $x_1, x_2, x_3$ se obtendría la siguiente relación lineal:

$$\lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 = 0 \qquad (1)$$

Siendo las constantes $\lambda_i$ simultáneamente diferentes de cero, esto es; $\lambda_i \neq 0 \; \forall i$. Lo cual permitiría expresar una variable en términos de las demás, por ejemplo al despejar $x_1$:

$$x_1 = \frac{-\lambda_2 x_2 - \lambda_3 x_3}{\lambda_1} \qquad (2)$$

Si los coeficientes fueran nulos no habría forma de obtener combinación lineal alguna y las columnas de la matriz de regresores serían linealmente independientes y dicha matriz sería no singular.

La multicolinealidad perfecta en realidad debe considerarse un caso poco frecuente en los modelos econométricos, que de ocurrir tendría como consecuencia la violación del supuesto de rango completo de la matriz de regresores $[\mathbf{X}]$ y en consecuencia tampoco se cumpliría para la matriz $[\mathbf{X'X}]$, siendo singulares ambas matrices y sus determinantes iguales a cero, lo que daría lugar a la indeterminación de los estimadores de mínimos cuadrados ordinarios para los parámetros del modelo. Esta situación se explica debido a que no estaría definida la matriz inversa $[\mathbf{X'X}]^{-1}$, que como sabemos es necesaria para obtener los estimadores de mínimos cuadrados ordinarios:$\widehat{\boldsymbol{\beta}} = [\mathbf{X'X}]^{-1}[\mathbf{X'Y}]$ .

En realidad el problema de la multicolinealidad debe ser visto como un problema de identificación, ya que alternativamente diferentes valores de los parámetros en el modelo generan el mismo valor estimado de la variable dependiente, lo que impide identificar el efecto individual de cada variable.

Resulta más usual que se presente multicolinealidad imperfecta, lo cual intuitivamente implica que los regresores de la regresión se encuentran altamente correlacionadas, pero sin ser esos coeficientes del cien por ciento. En términos de la matriz de regresores, significa que el determinante de la matriz $[\mathbf{X}]$ es cercano a cero, sin embargo ello no impide la obtención de los estimadores de mínimos cuadrados ordinarios, pero se mantiene el problema de identificación debido a que

la variación de alguna de las X's además de afectar a **Y** afectan a las demás variables impidiendo distinguir su efecto individual.

Si suponemos nuevamente que la matriz de regresores se compone de tres columnas con las variables $x_1$, $x_2$, $x_3$ se obtendría la siguiente relación lineal imperfecta entre ellas:

$$\lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + v = 0 \quad (3)$$

Siendo las constantes $\lambda_i$ simultáneamente diferentes de cero, como en el caso previo, pero ahora existe un término de error $v$. Debido a esto último, al despejar y expresar una variable en términos de las demás, por ejemplo al despejar $x_1$, la combinación lineal que se obtiene ya no es exacta y, por ende, la multicolinealidad ya no es perfecta:

$$x_1 = \frac{-\lambda_2 x_2 - \lambda_3 x_3}{\lambda_1} + \frac{v}{\lambda_1} = combinación\ lineal + error \quad (4)$$

Para tener una idea más precisa de lo que ocurre cuando la colinealidad entre las columnas de la matriz de regresores se incrementa, en el cuadro siguiente se muestra los que sucede con el determinante de la matriz y con los errores estándar de los estimadores de mínimos cuadrados ordinarios al irse incrementando el grado de correlación entre las variables. Para simplificar el asunto se supondrá que la varianza residual es una constante iguala la unidad, por ello $\sigma^2 = 1$. Claramente se observa que al ir aumentando la colinealidad entre las columnas de la matriz X, el determinante disminuye y las varianzas de los estimadores se van incrementando. En el caso limite, cuando las columnas de la matriz son iguales, se tiene multicolinealidad perfecta y el determinante se hace cero por lo que es imposible calcular la matriz inversa necesaria para la obtención de los estimadores de mínimos cuadrados ordinarios y las varianzas de los estimadores tienden a infinito.

**Cuadro 1**

**Ejemplo matricial de la multicolinealidad**

| Matriz X | Determinante | Varianza: $\sigma^2[X'X]^{-1}$ |
|---|---|---|
| $\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$ | 1 | $\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$ |
| $\begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}$ | 0.75 | $\begin{bmatrix}1.333 & -0.666\\-0.666 & 1.333\end{bmatrix}$ |
| $\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$ | 0.36 | $\begin{bmatrix}2.777 & -2.222\\-2.222 & 2.777\end{bmatrix}$ |
| $\begin{bmatrix}1 & 0.99\\0.99 & 1\end{bmatrix}$ | 0.02 | $\begin{bmatrix}50.251 & -49.749\\-49.749 & 50.251\end{bmatrix}$ |
| ... | ... | ... |
| $\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$ | 0.00 | No definida |

## 2. Pruebas para la detección de multicolinealidad

Algunas de las pruebas más usuales para detectar multicolinealidad son las siguientes (Quintana y Mendoza, 2008):

### a) Coeficientes t's no significativos y $R^2$ elevada

Una elevada $R^2$ junto con uno u algunos coeficientes t poco significativos es una de las pruebas más tradicionales para evaluar multicolinealidad. Del cuadro 1 es fácil comprender que los estadísticos t tenderán a disminuir debido a que su denominador se va incrementando paulatinamente al elevarse la colinealidad entre las variables.

### b) Coeficientes de correlación

Elevados coeficientes de correlación entre pares de variables son un síntoma a favor de la multicolinealidad. Es usual considerar que coeficientes de correlación

entre las variables por encima de 0.8 u 80% son evidencia de correlación seria, sin embargo también existen modelos con multicolinealidad grave y bajos coeficientes de correlación debido a que dicho coeficiente es sensible a transformaciones de las variables.

**c) Regresiones auxiliares y efecto $R^2$ de Theil**

Se corren regresiones auxiliares de la variable dependiente contra los k regresores menos uno de ellos, al coeficiente de determinación de esas regresiones se le denomina $R_i^2$. El efecto $R^2$de Theil (1971) se obtiene con la siguiente expresión:

$$R^2Theil = R^2 - [\sum_{i=1}^{n}(R^2 - R_i^2)] \quad (5)$$

donde $R^2$ es el coeficiente de determinación de la regresión original con todos los regresores y $R_i^2$ es el coeficiente de determinación de la regresión auxiliar i. Si el efecto de Theil fuera nulo no existiría multicolinealidad, entre mayor sea el efecto más grave es el problema.

**d) Regresiones auxiliares y regla de Klein**

La regla de Klein (1967) es un principio práctico, propuesto por el premio Nobel Lawrence Klein. De acuerdo a dicho principio, la multicolinealidad es un problema a considerar sí la $R_i^2$ de alguna regresión auxiliar es mayor que el coeficiente de determinación $R^2$ de la regresión original.

En este caso, las regresiones auxiliares son diferentes a las de Theil, ya que se efectúan tomando cada uno de los regresores y corriendo regresiones con los regresores restantes. Por ejemplo, si se tuvieran tres regresores $x_1$, $x_2$, $x_3$ en el modelo, las regresiones auxiliares serían las siguientes:

$$x_{1i} = \propto_1 + \propto_2 x_{2i} + \propto_3 x_{3i} + \varepsilon_{1i} \quad (6)$$

$$x_{2i} = \propto_1 + \propto_2 x_{1i} + \propto_3 x_{3i} + \varepsilon_{2i} \quad (6a)$$

$$x_{3i} = \propto_1 + \propto_2 x_{2i} + \propto_3 x_{1i} + \varepsilon_{3i} \quad (6b)$$

siendo i=1,2,...,n y $\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i}$ los usuales términos de perturbación aleatoria.

En este caso tendremos tres coeficientes de determinación de las regresiones auxiliares $R_1^2, R_2^2, R_3^2$ si alguno de ellos es mayor a $R^2$ el problema de multicolinealidad se puede considerar grave.

**f) Índice de la condición de número**

Este método hace uso de las propiedades de los valores característicos de una matriz, como sabemos el número de valores característicos diferentes de cero es igual al rango de la matriz y el producto de los valores característicos es su determinate.

Para calcular el índice de la condición de número (ICN) se deben obtener los valores característicos de la matriz [$\mathbf{X'X}$], a los cuales denominaremos $\lambda_i$ y se divide el máximo valor característico entre el menor valor característico:

$$ICN = \frac{\sqrt{\lambda_{máximo}}}{\sqrt{\lambda_{mínimo}}} \quad (7)$$

Como los valores característicos dependen de las unidades de medida de los datos, es mejor normalizar primero las variables de la matriz **X** para después calcular los valores característicos. Si las columnas de **X** son ortogonale la condición de número será igual a la unidad. En la práctica una condición de número superior a 20 se considera síntoma de multicolinealidad problemática.

**g. Factor de inflación varianza**

El factor de inflación varianza (VIF) se utiliza como una medida del grado en que la varianza del estimador de mínimos cuadrados es incrementada por la colinealidad entre las variables. El VIF se define de la manera siguiente:

$$VIF = \frac{1}{1-R_i^2} \quad (8)$$

En donde $R_i^2$ es el coeficiente de determinación de la regresión auxiliar i, tal y como se mostró en el caso previo. Por ejemplo, ante perfecta multicolinealidad $R_i^2 = 1$, lo cual hace que el VIF tienda a infinito, si la multicolinealidad es imperfecta y elevada, por ejemplo un $R_i^2 = 0.9$, el VIF será igual a 10. Es usual en la práctica que si el VIF resulta mayor a 10 o incluso 5 sea considerado como evidencia de fuerte multicolinealidad.

## 3. UN EJEMPLO PRÁCTICO EN LA DETECCIÓN DE MULTICOLINEALIDAD EN R CON LA FUNCIÓN CONSUMO PARA MÉXICO

Para tener una idea intuitiva de las implicaciones de la multicolinealidad, en esta sección se realiza primero una simulación con datos artificiales y después se procede a abordar un caso real para México,

Para realizar la simulación se deben generar dos variables, en donde una de ellas es independiente y la otra es una combinación lineal de aquella.

El proceso generador de los datos PGD se puede formular como:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \qquad (9)$$

siendo:

$$x_{3i} = \gamma_i + 5x_{2i}$$

$$y_i = 2 + 0.5x_{2i} + 0.1x_{3i} + u_i \qquad (10)$$

donde:

$x_{2i}\ y\quad x_{3i}$ son series de 1000 variables seudo aleatorias generadas artificialmente con distribución normal, media 0 y varianza unitaria.

$\gamma_i$ es una variable aleatoria normalmente distribuida

$u_i$ es un término de perturbación aleatoria con media cero y varianza constante 0.4

Para construir nuestras variables utilizaremos el generador de números seudoaleatorios de R, por lo cual lo primero que debemos hacer es fijar el valor semilla con el que se generarán los números, en este caso lo fijamos en 50:

```
set.seed(50)
```

Ahora generamos nuestras variables aleatorias con rnorm y corremos la regresión con lm:

```
X2=rnorm(100,0,1)
X3=rnorm(100,0,1)+5*X2
Y=2+0.5*X2+0.1*X3+rnorm(100,0,4)
summary(lm(Y~X2+X3))
```

Los resultados de la regresión se muestran a continuación, en ellos se puede observar que el coeficiente de $X_3$ no es estadísticamente significativo y la $R^2$ ajustada es relativamente elevada. Esto significa que debido a la colinealidad entre $X_2$ y $X_3$ no es posible separar el efecto de cada una de las variables en la variable dependiente, además de que la varianza del coeficiente de $X_3$ es muy alta por lo cual el estadístico t es muy bajo.

```
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  1.96761    0.03868      50.867    < 2e-16 ***
X2           0.69746    0.20364      3.425     0.000903 ***
X3           0.05881    0.03994      1.472     0.144144
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3863 on 97 degrees of freedom
Multiple R-squared:  0.8772,    Adjusted R-squared:  0.8747
F-statistic: 346.5 on 2 and 97 DF,  p-value: < 2.2e-16
```

Si la colinealidad fuera perfecta entre $X_2$ y $X_3$, $X_3$ sería una combinación lineal perfecta de $X_2$ y el proceso generador podría ser:

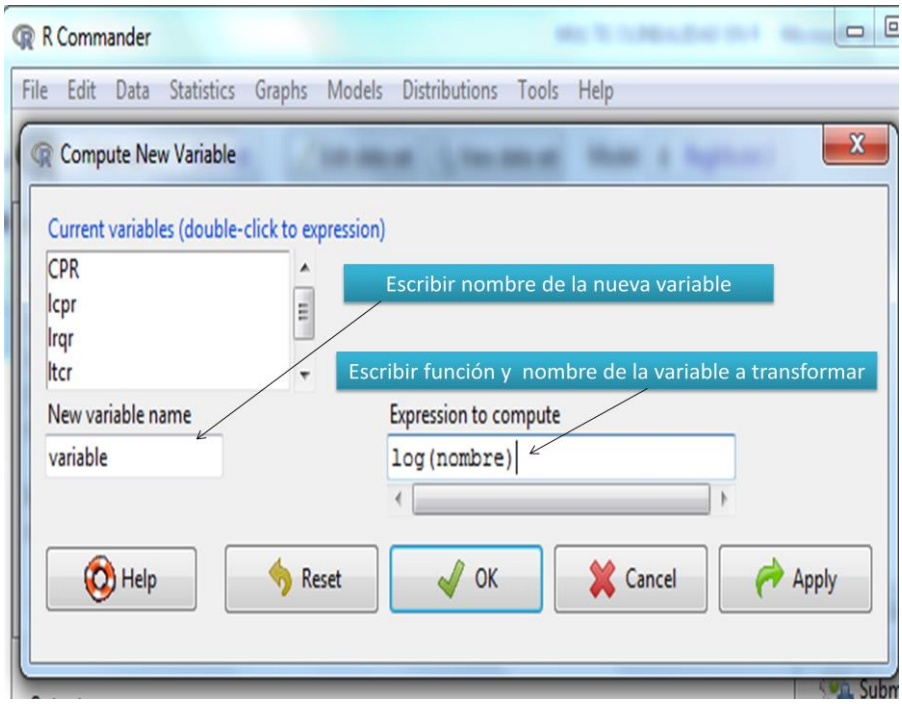
$$x_{3i} = 5x_{2i} \quad (11)$$

Sí incorpora este nuevo proceso en nuestra simulación, el R automáticamente elimina una de las variables y envía una alerta de que uno de los coeficientes no está definido debido a un problema de singularidad en la matriz de regresores, tal y como se observa en el recuadro siguiente:

```
lm(formula = Y ~ X2 + X3)
Residuals:
    Min      1Q  Median      3Q     Max
-0.80422 -0.19019  0.01836  0.17085  0.81986
Coefficients: (1 not defined because of singularities)
             Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)  1.94072    0.03649      53.19    <2e-16 ***
X2            1.04002   0.03741     27.80     <2e-16 ***
X3           NA         NA          NA         NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3643 on 98 degrees of freedom
Multiple R-squared:  0.8875,    Adjusted R-squared:  0.8863
F-statistic: 772.8 on 1 and 98 DF,  p-value: < 2.2e-16
```

En los datos del archivo consumo_fun.txt se presenta información trimestral para la economía mexicana del consumo privado (CPR), la riqueza real (RQR), y el ingreso disponible real (YPD).

Para utilizar los datos en R los importamos a través del RCommander y una vez cargados en el DATASET realizamos una transformación logarítmica de las variables seleccionando en el menú principal DATA/Manage variables in active dataset/Compute a new variable. Se abrirá una ventana en la cual simplemente en el espacio de New variable name se anota el nuevo nombre de la variable y en el espacio Expression to compute se escribe la función, en este caso log, y en paréntesis el nombre de la variable a transformar, tal y como se muestra en la imagen siguiente.




Con las variables transformadas en logaritmos se estima la siguiente ecuación:

$$lcpr_t = \beta_1 + \beta_2 lrqr_t + \beta_3 lypdr_t + \beta_3 ltcr_t + u_t \qquad (12)$$

donde:

lcprt es el logaritmo del consumo privado real en miles de millones de pesos de 1993

lrqrt es el logaritmo de la riqueza real calculada como el cociente del agregado monetario M4 dividido entre el índice de precios al consumidor.

lyndrt es el logaritmo del ingreso nacional disponible real en miles de millones de pesos de 1993

ltcrt es el logaritmo del tipo de cambio real

Los resultados de la regresión se muestran a continuación, de ellos se desprende que un incremento del diez por ciento en la riqueza da lugar a un aumento del 15.4% en el consumo, mientras que una variación de la misma magnitud en el

ingreso eleva en 71% al consumo. De los resultados también se observa que el tipo de cambio tiene un efecto negativo, pero éste no resulta estadísticamente significativo.

```
Call:
lm(formula = lcpr ~ lrqr + ltcr + lypdr, data = Dataset)
Residuals:
      Min        1Q    Median        3Q       Max
-0.061536 -0.017314 -0.001635  0.020202  0.072171
Coefficients:
             Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)  1.90203     0.54239    3.507     0.000703 ***
lrqr          0.15401   0.03161     4.873     4.57e-06 ***
ltcr         -0.03185    0.02053   -1.551     0.124223
lypdr         0.71042    0.06637   10.704     < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.03154 on 92 degrees of freedom
Multiple R-squared:  0.9744,      Adjusted R-squared:  0.9735
F-statistic:  1165 on 3 and 92 DF,  p-value: < 2.2e-16
```

En los resultados previos es relevante examinar la posible existencia de multicolinealidad en virtud de la fuerte relación que puede existir entre las tres variables explicativas; la riqueza de los individuos se forma a través de su ingreso y estas dos variables son afectadas sensiblemente por lo que ocurre con los precios de los bienes importados, cuyo efecto es tomado en cuenta por el tipo de cambio.

Una primer evidencia de posible elevada colinealidad entre las variables se deriva de la alta $R^2$ ajustada de 0.97 y la nula significancia de una de las variables. Para intentar confirmar esta evidencia es preciso realizar algunas exploraciones adicionales.

**a) Coeficientes de correlación**

Los coeficientes de correlación entre las variables se calculan con la función col del R;

```
cor(Dataset[,c("lydr","lrqr","ltcr")], use="complete")
```

En RCommander basta seleccionar en el menú principal STATISTICS/Summaries/Correlation matrix. En la ventana que se abre basta seleccionar las variables a correlacionar y el tipo de correlación que en este caso es el Pearson. Los resultados son los siguientes:

```
> cor(Dataset[,c("lrqr","ltcr","lypdr")], use="complete")
          lrqr                ltcr      lypdr
lrqr   1.0000000      -0.528662     0.9632604
ltcr  -0.5286620       1.000000     -0.4918170
lypdr  0.9632604      -0.491817     1.0000000
```

En los resultados es posible observar que las correlaciones son muy altas entre el ingreso y la riqueza, 96%, mientras que con el tipo de cambio las correlaciones son relativamente bajas. Por ello, de existir algún problema de colinealidad se deriva de las primeras dos variables.

**b) Factor de inflación-varianza (VIF)**

Para calcular el VIF en RCommander seleccionamos del menú principal MODELS/Numeric diagnostics/Variance inflation factors. En los resultados siguientes es posible establecer la existencia de problemas de colinealidad graves en virtud de que las variables de riqueza y de ingreso presentan un VIF muy por arriba de diez unidades.

```
> vif(RegModel.3)
    lrqr          ltcr          lypdr
14.673133  1.396047        13.945404
```

**c) Regresiones auxiliares: La regla de Klein.**

Al correr una regresión auxiliar tomando al ingreso como variable dependiente y a la riqueza y el tipo de cambio como explicatorias obtenemos una $R^2$ ajustada de 0.9267 la cual es inferior a la de 0.9765 del modelo original, tal y como se observa en los resultados del recuadro siguiente. Esto implica que el problema de multicolinealidad no es muy grave.

```
Call:
lm(formula = lypdr ~ lrqr + ltcr, data = Dataset)
Residuals:
      Min        1Q       Median        3Q       Max
-0.125362 -0.036361  0.004442  0.034763  0.108628
Coefficients:
                Estimate      Std. Error     t value        Pr(>|t|)
(Intercept)     7.88543       0.22250        35.439         <2e-16 ***
lrqr             0.45315      0.01519        29.837         <2e-16 ***
ltcr            0.02364       0.03198        0.739          0.462
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.04928 on 93 degrees of freedom
Multiple R-squared:  0.9283,        Adjusted R-squared:  0.9267
F-statistic:   602 on 2 and 93 DF,  p-value: < 2.2e-16
```

**d) Regresiones auxiliares: El efecto de Theil.**

Con base en los resultados de la regresión auxiliar previa y los de las regresiones auxiliares excluyendo a uno de los regresores se puede calcular el efecto de Theil.

```
Call:
lm(formula = lcpr ~ lrqr + ltcr, data = Dataset)
Residuals:
```

```
     Min       1Q   Median       3Q      Max
-0.120729 -0.035090  0.002992  0.037276  0.102336
Coefficients:
             Estimate     Std. Error   t value      Pr(>|t|)
(Intercept)  7.50399      0.21225      35.354       <2e-16 ***
lrqr         0.47593      0.01449      32.851       <2e-16 ***
ltcr         -0.01506     0.03051      -0.494       0.623
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.04701 on 93 degrees of freedom
Multiple R-squared:  0.9424,      Adjusted R-squared:  0.9412
F-statistic: 760.9 on 2 and 93 DF,  p-value: < 2.2e-16

Call:
lm(formula = lcpr ~ lrqr + lypdr, data = Dataset)
Residuals:
     Min       1Q       Median       3Q      Max
-0.063740 -0.020311  0.000018  0.019144  0.069434
Coefficients:
             Estimate     Std. Error   t value      Pr(>|t|)
(Intercept)  1.81882      0.54380      3.345        0.00119 **
lrqr         0.16552      0.03096      5.347        6.36e-07 ***
lypdr        0.70255      0.06667      10.537       < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.03178 on 93 degrees of freedom
Multiple R-squared:  0.9737,      Adjusted R-squared:  0.9731
F-statistic:  1720 on 2 and 93 DF,  p-value: < 2.2e-16

Call:
lm(formula = lcpr ~ ltcr + lypdr, data = Dataset)
Residuals:
     Min       1Q       Median       3Q      Max
-0.074269 -0.020489 -0.001975  0.018901  0.082560
Coefficients:
             Estimate     Std. Error   t value      Pr(>|t|)
(Intercept)   -0.31390     0.32972      -0.952       0.3436
ltcr         -0.05534     0.02226      -2.486       0.0147 *
lypdr        1.01813      0.02277      44.711       <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.03519 on 93 degrees of freedom
Multiple R-squared:  0.9677,      Adjusted R-squared:  0.967
F-statistic:  1395 on 2 and 93 DF,  p-value: < 2.2e-16
```

Con los datos del recuadro previo es posible calcular el efecto de Theil utilizando la $R^2$ original de 0.9744 y las $R^2$ de las ecuaciones auxiliares de la manera siguiente:

0.9744-(0.9744-0.9424)-( 0.9744-0.9737)-( 0.9744-0.9677)=0.935

El resultado es la reducción en el efecto individual de la suma de las variables explicatorias debido a la multicolinealidad en relación con el que hubieran tenido de ser independientes las variables.

**e) La condición de número**

En RCommander se pueden calcular los valores característicos para la matriz de regresores del modelo de la ecuación (12). Para ello se debe seleccionar en el menú principal la secuencia de opciones: STATISTICS/Dimensional analysis/Principal componente analysis. A continuación se abre una ventana en la que se deben seleccionar las variables que componen la matriz de regresores, que en este caso son lrqr, lypdr y ltcr. También se deben establecer las opciones las cuales permiten analizar la matriz de correlaciones, generar una gráfica de las componentes y sus varianzas, además de permitir añadir las componentes a la tabla de datos.

En el caso de los regresores de la función consumo el R nos presenta las tres raíces características ordenadas de mayor a menor tal y como se observa en el recuadro siguiente.

```
> .PC$sd^2  # component variances
   Comp.1    Comp.2    Comp.3
2.34925314 0.61498514 0.03576172
```

Al sustituir estos resultados en la fórmula del ICN obtenemos:

$$\text{ICN} = \frac{\sqrt{\lambda_{\text{máximo}}}}{\sqrt{\lambda_{\text{mínimo}}}} = \frac{\sqrt{2.34925314}}{\sqrt{0.03576172}} = 8.105$$

El valor del ICN es inferior al umbral de 20 que se ha definido en la literatura para establecer un grado de multicolinealidad grave, por consiguiente no habría que preocuparse de este problema en el modelo.

## 4. Soluciones al problema de la multicolinealidad

Una vez que se ha detectado que el grado de multicolinealidad del modelo es grave, se puede optar por una serie de métodos de corrección. Debe señalarse que si el problema de multicolinealidad no es severo más vale no hacer nada, ya que los remediales generalmente pueden implicar problemas más fuertes que el que se buscaba corregir. Debe considerarse que frente a un problema de multicolinealidad los estimadores de mínimos cuadrados ordinarios siguen siendo insesgados, de modo que si el problema no es grave el modelo puede utilizarse sin que afecte en gran medida a la inferencia estadística. Incluso si el objetivo de la modelación no fuera el análisis estructural sino el mero pronóstico, la multicolinealidad no tendría mayor efecto dado que la relación entre las variables se mantiene tanto en el horizonte histórico como en el futuro de las variables.

De cualquier forma, si se quiere hacer algo para resolver el problema los remedios usuales son los siguientes:

### a) Imponer restricciones al modelo

Se deben restringir los parámetros de aquellas variables altamente colineales. Por ejemplo, si las variables $x_2$ $y$ $x_3$son altamente colineales es posible restringir el modelo utilizando información a priori o bien por estimaciones de corte transversal.

Suponga que nuestro modelo es:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \quad (13)$$

con i=1,2,..n

Al aplicar pruebas de detección de multicolinealidad se encontró que esta era grave y se debía a una elevada colinealidad entre $x_2$ $y$ $x_3$. Si, en publicaciones acerca de modelos similares al que se está estimando, existiera evidencia sobre los coeficientes se podría usar esa información para corregir. Por ejemplo, suponga que la evidencia encontrada es que el coeficiente $\beta_3$ es un medio del coeficiente $\beta_2$. Esto nos permite aplicar la siguiente restricción:

$$\beta_3 = 0.5\beta_2 \quad (14)$$

Sustituyendo en el modelo original obtenemos la ecuación restringida:

$$y_i = \beta_1 + \beta_2 x_{2i} + 0.5\beta_2 x_{3i} + u_i \quad (15)$$

$$y_i = \beta_1 + \beta_2 (x_{2i} - 0.5x_{3i}) + u_i \quad (15a)$$

$$y_i = \beta_1 + \beta_2 x_{2i}^* + u_i \quad (15b)$$

Donde: $x_{2i}^* = x_{2i} - 0.5x_{3i}$

Una vez restringido el modelo la multicolinealidad se ha eliminado y al obtener el estimador de MCO $\hat{\beta}_2$ es posible obtener $\hat{\beta}_3$ si se sustituye el primero en la restricción (14).

La principal limitante de este método es la carencia de antecedentes empíricos acerca de los coeficientes de interés en los modelos econométricos.

Otra alternativa que implica restringir el modelo original es la estimación de un modelo en corte transversal. Por ejemplo, para el caso que nos ocupa se podría estimar $\beta_3$ en un modelo de corte transversal y sustituir su valor estimado en el modelo de series de tiempo. Suponga que en la estimación de corte transversal se obtiene que:

$$\hat{\beta}_3 = 0.5 \quad (16)$$

Se restringe el modelo sustituyendo ese valor en el modelo original:

$$y_i = \beta_1 + \beta_2 x_{2i} + 0.5x_{3i} + u_i \quad (17)$$

$$y_i - 0.5x_{3i} = \beta_1 + \beta_2 x_{2i} + u_i \quad (17a)$$

$$y_i^* = \beta_1 + \beta_2 x_{2i} + u_i \quad (17b)$$

Donde: $y_i^* = y_i - 0.5x_{3i}$

La limitante de este procedimiento es que la interpretación de los parámetros de corte transversal y series de tiempo puede diferir ampliamente al calcularse sobre conjuntos de datos diferentes.

**b) Componentes principales**

El método de componentes principales busca eliminar el problema de multicolinealidad a través de la obtención de un conjunto de variables a partir de las originales y sin implicar grandes pérdidas de información (Everitt y Hothorn, 2006). Las nuevas variables o componentes cumplen con la condición de ser ortogonales entre sí.

El método parte de una forma cuadrática **x'Ax** que se minimiza sujeta a la condición de normalidad **x'x**=1:

$$\mathbf{x'Ax} - \boldsymbol{\lambda}(\mathbf{x'x} - \mathbf{1}) \quad (18)$$

Donde **A** es una matriz simétrica.

Al derivar con respecto a x:

$$\mathbf{2Ax} - \mathbf{2\lambda x} = \mathbf{0} \quad (19)$$

Al factorizar encontramos la ecuación característica:

$$(\mathbf{A} - \boldsymbol{\lambda}\mathbf{I})\mathbf{x} = \mathbf{0} \quad (20)$$

Al obtener el determinante de la ecuación característica se genera un polinomio característico y al encontrar sus raíces nos permite obtener los valores característicos $\boldsymbol{\lambda_i}$.

Si partimos de la matriz de regresores **X**, el método de componentes principales consiste en encontrar una función lineal de las variables originales, **Z**=**a'x**, que maximice la varianza de **X** sujeta a la condición de normalidad, **a'a**=1. Al resolver el polinomio podemos encontrar la raíz característica máxima y su correspondiente vector característico, el cual es el vector a que necesitamos para encontrar **Z**.

La principal limitante de este método es que las nuevas variables **Z** pueden no tener interpretación económica alguna.

**c) Eliminar variables**

La eliminación de variables sospechosas de colinealidad puede ser otra opción para evitar el problema de multicolinealidad, sin embargo puede llevarnos a un problema más grave como el de variable relevante omitida. En nuestro ejemplo la eliminación de la variable $x_{3i}$ deja el modelo como:

$$y_i = \beta_1 + \beta_2 x_{2i} + u_i \quad (21)$$

Sin embargo, si la variable omitida fuera relevante se genera un problema de sesgo en los estimadores de MCO.

**d) Transformar variables**

La transformación de variables con primeras diferencias o calculando porcentajes es otro remedio que busca diferenciar más las variables entre sí. Sin embargo, su principal limitante es que, por una lado la teoría relevante pudiera estar interesada únicamente en las variables de nivel y no en sus diferencias ni en sus porcentajes, por otro lado la variable dependiente pudiera estar relacionada con las demás en niveles pero no en porcentajes ni en diferencias.

## 5. UN EJEMPLO PRÁCTICO EN R DE SOLUCIÓN A LA MULTICOLINEALIDAD EN LA FUNCIÓN CONSUMO.

Una vía para buscar corregir cualquier síntoma de multicolinealidad en el modelo que hemos estimado para la función consumo podría ser el de componentes principales. Para lo cual podemos seguir el mismo procedimiento que ya hemos aplicado para calcular los valores característicos de la prueba del ICN, Es decir, se debe seleccionar en el menú principal la secuencia de opciones: STATISTICS/Dimensional analysis/Principal componente analysis, pero ahora solamente consideraremos las dos variables que ya hemos confirmado antes guardan una elevada colinealidad entre sí, nos referimos a lrqr y lypdr.

Los resultados que se muestran a continuación indican que la componente primera representa el 98.16% de la varianza total, por lo cual si tomamos esa componente para realizar la combinación lineal de los dos regresores prácticamente no habría perdida de información.

```
> .PC <- princomp(~lrqr+lydr, cor=TRUE, data=Dataset)
> unclass(loadings(.PC))  # component loadings
        Comp.1       Comp.2
lrqr    0.7071068   -0.7071068
lydr    0.7071068    0.7071068
> .PC$sd^2  # component variances
   Comp.1     Comp.2
1.96326036  0.03673964
> summary(.PC) # proportions of variance
Importance of components:
                          Comp.1      Comp.2
Standard deviation        1.4011639   0.19167588
Proportion of Variance    0.9816302   0.01836982
Cumulative Proportion     0.9816302   1.00000000
```

Si ahora se corre la regresión sustituyendo los dos regresores pos la combinación lineal de los mismos en la componente principal primera que ha sido guardada en la tabla de datos con el nombre PC1, es posible replantear el modelo de la ecuación (12) de la siguiente manera:

$$lcpr_t = \beta_1 + \beta_2 PC1_t + \beta_3 ltcr_t + u_t \quad (22)$$

Los resultados de esta regresión se muestran en seguida, de ellos se observa que la variable PC1 es estadísticamente significativa y que representa el efecto combinado de la riqueza y el ingreso en el consumo de los individuos.

```
m(formula = lcpr ~ ltcr + PC1, data = Dataset)
Residuals:
      Min        1Q    Median        3Q       Max
-0.070837 -0.018441 -0.003601  0.020371  0.070356
Coefficients:
             Estimate   Std. Error  t value    Pr(>|t|)
(Intercept)  13.686850  0.028197    485.397     <2e-16 ***
ltcr         -0.022570  0.021034    -1.073      0.286
PC1           0.134105  0.002781    48.221     <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.03273 on 93 degrees of freedom
Multiple R-squared:  0.9721,      Adjusted R-squared:  0.9715
F-statistic:  1619 on 2 and 93 DF,  p-value: < 2.2e-16
```

REFERENCIAS


L. R. Klein, An Introduction to Econometrics , Prentice-Hall, 1962;

Theil, H, Principles of Econometrics, Wiley, 1971.

Everitt,S. Brian y Torsten Hothorn, A handbook of statistical analysis using R, Chapman / Hall/CRC, 2006.

Quintana Romero, Luis y Miguel Ángel Mendoza, Econometría básica, Plaza y Valdés, 2008.


**ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO**

consumo_fun.txt