Estadística para la Economía y la Gestión
IN 3401
Clase 6
Problemas con los Datos

27 de octubre de 2009

Multicolinealidad

Es prácticamente imposible encontrar dos variables económicas cuyo coeficiente de correlación es una determinada muestra sea numéricamente cero, dicho coeficiente puede tomar valores pequeños pero nunca llegar a ser cero.

La *Multicolinealidad* aparece cuando las variables explicativas en modelo econométrico están correlacionadas entre si, esto tiene efectos negativas cuando se quire estimar los parámetros del modelo por MCO.

Existen diversas fuentes de la multicolinealidad:

- El método de recolección de información empleado, obtención de muestras en un intervalo limitado de valores de los regresores en la población.
- Restricción en el modelo o en la población objeto de muestreo.
- Especificación del modelo.

Consideremos el siguiente modelo:

$$y_i = \beta_1 + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \epsilon_i$$

Si existe la inversa de $X'X$, el estimador MCO de este modelo viene dado por $\hat{\beta}_{MCO} = (X'X)^{-1}X'Y$ y su matriz de covarianzas es $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$.

Supongamos que x_{ji} tiene un alto grado de correlación con las demás variables explicativas de modelo, es decir, que la regresión lineal:

$$x_{ji} = \delta_1 + \delta_2 x_{2i} + ... + \delta_{j-1} x_{j-1i} + \delta_{j+1} x_{j+1i} + ... + \delta_k x_{ki} + \nu_i$$

tiene un coeficiente de determinación alto.

En estas condiciones la variable x_{ji} puede escribirse aproximadamente como una combinación lineal del resto de las variables explicativas del modelo, lo que se puede apreciar en la ecuación.

Como consecuencia una de las columnas de la matriz X, la correspondiente a x_{ji}, puede escribirse como una combinación lineal aproximada de las demás columnas de X, y de esta forma $X'X$ será aproximadamente singular.

En la medida que el determinante de $X'X$ sea distinto de cero, existirá $(X'X)^{-1}$, y por lo tanto también existirá el estimador MCO, el cual sigue cumpliendo con la propiedad de MELI, pero se tienen consecuencias.

Algunas consecuencias...

1. La solución del sistema de ecuaciones normales está mal definido.
2. Pequeñas variaciones muestrales por incorporar o sustraer un número reducido de observaciones muestrales podrían generar importantes cambios en la solución $\hat{\beta}$ del sistema de ecuaciones normales.
3. Al ser la matriz $X'X$ casi singular, es muy pequeña. Como consecuencia la matriz de covarianzas será muy grande, por lo tanto el estimador MCO es poco preciso en este caso.

La presencia de multicolinealidad en un modelo de regresión lineal puede ser de dos formas:

- **Multicolinealidad Exacta**: una de las variables explicativas es una combinación lineal determinística de todas las demás (o algunas de ellas).
- **Multicolinealidad Aproximada**: ocurre cuando una de las variables es aproximadamente igual a una combinación lineal de las restantes, como en la ecuación anterior.

En la práctica, contrario a lo que se pudiera esperara es más complicado la multicolinealidad aproximada que la exacta.

Puesto que la multicolinealidad es un problema de naturaleza muestral, que surge principalmente por el carácter no experimental de la mayoría de la información recopilada en las Ciencias Sociales, no tiene una manera única de ser detectada. Lo que se tiene son algunas reglas prácticas detalladas a continuación:

1. El R^2 es alto, pero los parámetros no resultan ser individualmente significativos.
Por ejemplo: Considere los siguientes datos:

Período	y_i	x_{i2}	x_{i3}	x_{i4}
1	20	5	10	10
2	12	2	8	6
3	28	7	12	16
4	26	6	4	12
5	14	4	16	8
6	24	8	14	14
7	16	3	6	4

Las variables x_3 y x_4 tienen las mismas observaciones numéricas solo que en distinto orden, de forma tal que la correlación entre x_2 y estas dos variables son: $\rho_{23} = 0,32$ y $\rho_{24} = 0,93$, altamente diferentes entre sí. Una regresión de y_i sobre x_{2i}, x_{3i} y una constante generó las siguientes estimaciones MCO:

```
. reg y x2 x3

      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  2,     4) =   24.49
       Model |  214.486486     2  107.243243           Prob > F      =  0.0057
    Residual |  17.5135135     4  4.37837838           R-squared     =  0.9245
-------------+------------------------------           Adj R-squared =  0.8868
       Total |         232     6  38.6666667           Root MSE      =  2.0925

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |    2.918919   .4175976     6.99   0.002     1.759482    4.078356
          x3 |   -.5405405   .2087988    -2.59   0.061    -1.120259    .0391779
       _cons |    10.81081   2.557774     4.23   0.013      3.70929    17.91233
------------------------------------------------------------------------------
```

Una regresión de y_i contra una constante, x_{2i} y x_{4i}, produjo las siguientes estimaciones:

```
. reg y x2 x4

      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  2,     4) =    9.60
       Model |         192         2          96       Prob > F      =  0.0297
    Residual |          40         4          10       R-squared     =  0.8276
-------------+------------------------------           Adj R-squared =  0.7414
       Total |         232         6  38.6666667       Root MSE      =  3.1623

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |   1.333333   1.610153     0.83   0.454    -3.137168    5.803835
          x4 |   .6666667   .8050765     0.83   0.454    -1.568584    2.901917
       _cons |   6.666667   3.269225     2.04   0.111    -2.410156    15.74349
------------------------------------------------------------------------------
```

Ambas regresiones no incluyen las mismas variables explicativas y por lo tanto, no son comparables. Sin embargo, en el segundo modelo donde el grado de correlación entre las variables explicativas es alto, podemos apreciar que a pesar de que el R^2 es alto, los parámetros resultan ser insignificativos individualmente ($t_4 = 2{,}78$).

2. Pequeños cambios en los datos, produce importantes variaciones en las estimaciones mínimo cuadráticas.

3. Los coeficientes pueden tener signos opuestos a los esperados o una magnitud poco creíble.

(a) Métodos basados en la correlación entre variables explicativas

Si descomponemos la matriz X de la siguiente forma:

$$X = [x_j; X_j]$$

donde x_j es un vector columna correspondiente a la j-ésima variable explicativa y X_j una matriz de $n \times (k-1)$ con las observaciones de las restantes variables. Entonces, $X'X$ puede escribirse como:

$$X = \begin{pmatrix} x_j'x_j & x_j'X_j \\ X_j'x_j & X_j'X_j \end{pmatrix}$$

De esta forma, el elemento $(1,1)$ de $(X'X)^{-1}$ es (Demostrar):

$$[(x_j'x_j) - x_j'X_j(X_j'X_j)^{-1}(X_j'x_j)]^{-1} = (x_jM_jx_j)^{-1}$$

donde $M_j = I_n - X_j(X_j'X_j)^{-1}X_j'$ y donde $x_j'M_jx_j$ corresponde a la suma de los residuos al cuadrado de una regresión de x_j sobre X_j.

De esta forma se tiene que:

$$V(\hat{\beta}_j) = \frac{\sigma_\epsilon^2}{x_j' M_j x_j}$$

Lo que tiene la siguiente expresión:

$$V(\hat{\beta}_j) = \frac{\sigma_\epsilon^2}{ST_j(1 - R_j^2)}$$

donde ST_j es la suma total de la regresión entre x_j y X_j y R_j^2 es el coeficiente de determinación de esta misma regresión.
La varianza de $\hat{\beta}_j$ depende de tres cosas:

- La varianza del término de error, que es independiente del grado de correlación entre las $x's$.
- La suma total propia de la variable x_j, la que depende solo de esta variable.
- El coeficiente de determinación R_j^2 , el que si depende del grado del grado de correlación entre la variable x_j y las restantes, es decir, depende del grado de multicolinealidad.

La cota inferior para la varianza de $\hat{\beta}_j$, cuando $R^2 = 0$, es:

$$V(\hat{\beta}_j^0) = \frac{\sigma_\epsilon^2}{ST_j}$$

Por lo que la relación entre las varianzas de la estimación de β_j en un caso de correlación entre variables explicativas y el caso de independencia lineal es:

$$\frac{V(\hat{\beta}_j)}{V(\hat{\beta}_j^0)} = \frac{1}{1 - R^2}$$

```
. reg  x2 x3

      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  1,     5) =    0.58
       Model |  2.89285714     1  2.89285714            Prob > F      =  0.4821
    Residual |  25.1071429     5  5.02142857            R-squared     =  0.1033
-------------+------------------------------           Adj R-squared =  -0.0760
       Total |          28     6  4.66666667            Root MSE      =  2.2409

------------------------------------------------------------------------------
          x2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x3 |   .1607143   .2117408     0.76   0.482    -.3835829    .7050114
       _cons |    3.392857   2.280519     1.49   0.197    -2.469403    9.255117
------------------------------------------------------------------------------

. reg  x2 x4

      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  1,     5) =   31.30
       Model |  24.1428571     1  24.1428571            Prob > F      =  0.0025
    Residual |  3.85714286     5  .771428571            R-squared     =  0.8622
-------------+------------------------------           Adj R-squared =  0.8347
       Total |          28     6  4.66666667            Root MSE      =  .87831

------------------------------------------------------------------------------
          x2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x4 |   .4642857   .0829925     5.59   0.003     .2509467    .6776247
       _cons |   .3571429   .8938566     0.40   0.706    -1.940589    2.654874
------------------------------------------------------------------------------
```

De acuerdo con este análisis, los coeficientes de determinación obtenidos en las regresiones de cada variable explicativa con el resto son un buen indicador de una posible situación de multicolinealidad.

(b) Métodos basados en el tamaño de la matriz $X'X$:
Cuando tenemos multicolinealidad la matriz $X'X$ es casi singular, de esta manera una medida de tamaño de esta matriz nos permite detectar la presencia de multicolinealidad.

El determinante no es una medida buena, ya que tiene problemas de sensibilidad a los cambios de unidades.

Pero sabemos que el determinante de una matriz simétrica es igual al producto de sus valores propios, y por lo tanto el examen de estos valores nos da una idea del tamaño de la matriz.

De esta forma, Belsley propone la siguiente medida para ver el grado de multicolinealidad:

$$\gamma = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

Esta medida se denomina *número de condición de la matriz X*, y números de este indicador mayores 25 suelen considerarse problemáticos.
Los $\lambda's$ corresponden a los valores propios de la matriz $B = S(X'X)S$, donde S es la siguiente matriz diagonal:

$$S = \begin{pmatrix} \frac{1}{\sqrt{x_2'x_2}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{x_3'x_3}} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{\sqrt{x_k'x_k}} \end{pmatrix}$$

Esta matriz nos permite librarnos del problema de unidad en el tamaño de los valores propios, ya que normaliza cada una de las variables al dividir todas las observaciones por su desviación estándar.

El número de condición de la matriz X (γ), implica que mientras mayor es este valor, el valor de λ_{min} es realmente pequeño al compararlo con λ_{max}, indicando el potencial problema de multicolinealidad.

Se han propuesto varios métodos para hacer frente a la multicolinealidad. La solución más sencilla es eliminar de la regresión las variables que se sospeche son la causa del problema. Obviamente de este método surgen problemas de especificación.

Las soluciones propuestas en la literatura (estimador de ridge o estimador cresta y estimador de componentes principales) tienen como característica buscar un estimador ligeramente sesgado pero cuya varianza sea mucho menor, es decir, un estimador con menor error cuadrático medio.

No existe una metodología que permita eliminar el problema de alta multicolinealidad sin alterar las propiedades y la interpretación de los parámetros. Estas metodologías tienen poco respaldo intuitivo, por lo tanto la interpretación de los parámetros es desconocida.

Error de Medición

Una dificultad en todo trabajo empírico en Economía es la imposibilidad de disponer de las observaciones muestrales de las variables de interés.

Por ejemplo, las variables de contabilidad nacional como el PIB, stock de capital o consumo, son sólo estimaciones de conceptos teóricos que no se observan en la realidad.

En otros casos, como la Renta Permanente, inteligencia o habilidad de un trabajador, no disponemos ni siquiera estimaciones, y debemos utilizar variables Proxies, que aproximan los conceptos que se quieren utilizar. Así por ejemplo se utilizan años de experiencia del trabajador para aproximar su habilidad.

Podemos adelantar que el error de medición o el uso de variables proxies generará sesgos en las estimaciones por MCO, el que será menor:

- cuanto más se aproxime la verdadera variable que debería incluirse en el modelo con que que incluyo efectivamente.
- cuanto más independiente sea el error de medida de las restantes variables del modelo.

Consideremos el siguiente modelo lineal simple:

$$y_i = \beta x_i + \epsilon_i \qquad i = 1, ..., n$$

en el que la variable dependiente y_i está medida con error, es decir, solo observamos:

$$y_i^* = y_i + \nu_i \qquad i = 1, ..., n$$

donde asumimos que $\nu_i \sim N(0, \sigma_\nu^2)$ y es independiente de x_i y ϵ_i. Reemplazando

$$y_i^* = \beta x_i + (\epsilon_i + \nu_i) = \beta x_i + \varepsilon_i$$

Bajo los supuestos mencionados es fácil darse cuenta que el estimador de β será el mismo que si observáramos el verdadero valor de y_i. En consecuencia, los errores de medida en la variable endógena no producen ningún problema importante al estimar por MCO.

Ahora supongamos que la variable x_i esta medida con error, es decir:

$$x_i^* = x_i + \omega_i, \qquad i = 1, ..., n$$

donde $\omega_i \sim N(0, \sigma_\omega^2)$ y es independiente de ϵ_i, x_i y de y_i.

El modelo en términos de las variables observables es:

$$y_i = \beta x_i^* + (\epsilon_i - \beta\omega_i) = \beta x_i^* + \varepsilon_i$$

en este caso tenemos dificultad al estimar por MCO, ya que el término de error ε_i esta relacionado con x_i, lo que va en contra del supuesto 6, veamos:

$$\begin{aligned} Cov(\varepsilon_i, x_i^*) &= Cov(\epsilon_i - \beta\omega_i, x_i + \omega_i) \\ &= Cov(\epsilon_i, x_i) - \beta Cov(\omega_i, x_i) + Cov(\epsilon_i, \omega_i) - \beta Cov(\omega_i, \omega_i) \\ &= 0 - \beta \cdot 0 + 0 - \beta\sigma_\omega^2 \end{aligned}$$

Esto hace que el estimador MCO de β sea sesgado.

$$
\begin{aligned}
\hat{\beta} &= \frac{\sum x_i^* y_i}{\sum (x_i^*)^2} \quad / \cdot \frac{1/n}{1/n} / plim \\
plim\hat{\beta} &= \frac{plim \frac{1}{n} \sum x_i^* y_i}{plim \frac{1}{n} \sum (x_i^*)^2} \\
&= \frac{plim \frac{1}{n} \sum (x_i + \omega_i)(\beta x_i + \epsilon_i)}{plim \frac{1}{n} \sum (x_i + \omega_i)^2} \\
&= \frac{plim \frac{1}{n} \sum (x_i + \omega_i)(\beta x_i + \epsilon_i + \beta\omega_i - \beta\omega_i)}{plim \frac{1}{n} \sum (x_i + \omega_i)^2} \\
&= \beta + \frac{plim \frac{1}{n} \sum (x_i + \omega_i)(\epsilon_i - \beta\omega_i)}{plim \frac{1}{n} \sum (x_i + \omega_i)^2} \\
plim\hat{\beta} &= \beta + \frac{-\beta\sigma_\omega^2}{S_x^2 + \sigma_\omega^2} \\
plim\hat{\beta} &= \frac{\beta}{1 + \frac{\sigma_\omega^2}{S_x^2}}, \quad \text{donde} \;\; S_x^2 = plim \frac{1}{n} \sum x_i \quad \text{existe}
\end{aligned}
$$

El resultado en términos generales es que el estimador MCO en presencia de *error de medición* estará sesgado hacia en origen.

En el caso del modelo de regresión múltiple:

$$\begin{aligned} Y &= X\beta + \epsilon \\ X^* &= X + \omega \end{aligned}$$

donde todas las variables pueden estar medidas con error. Extendiendo lo desarrollado anteriormente:

$$plim\hat{\beta}_{MCO} = \beta - [\Sigma_{xx} + \Sigma_{\omega\omega}]^{-1}\Sigma_{\omega\omega}\beta$$

donde $\Sigma_{xx} = plim\frac{X'X}{n}$ y $\Sigma_{\omega\omega} = plim\frac{\omega'\omega}{n}$.

Lo que implica que un sólo error basta para generar inconsistencias en todos los coeficientes del modelo.

La estimación consistente de los parámetros en presencia de errores de medida es posible si se dispone de *instrumentos*.

Definición: Un instrumento es una variable no incluida en el modelo, que cumple con:

- No estar correlacionada con el término de error.
- Esta correlacionada con la variable explicativa para la cual actúa como instrumento (en este caso la variable medida con error).

Volviendo al modelo anterior, el sesgo del estimador MCO de β surge por la correlación entre la variable x_i^* y ε_i. Supongamos ahora que se dispone de la variable z_i, tal que:

$$E(z_i\varepsilon_i) = 0 \qquad E(z_i x_i^*) \neq 0$$

Entonces el estimador de variables instrumentales para este caso es:

$$\hat{\beta}_{IV} = \frac{\sum z_i y_i}{\sum z_i x_i^*}$$

En un modelo de regresión múltiple, tenemos que encontrar una matriz Z que contenga los instrumentos de las variables medidas con error.

El estimador de Variables Instrumentales se obtiene de una regresión MCO en dos etapas:

1. En la primera etapa, se hace una regresión entre X^* y la matriz de instrumentos Z, para obtener el valor estimado de X^*:

$$\begin{aligned} X^* &= Z\phi + \epsilon \\ \hat{\phi} &= (Z'Z)^{-1}Z'X^* \\ \hat{X}^* &= Z(Z'Z)^{-1}Z'X^* \end{aligned}$$

2. En la segunda etapa se reemplaza el valor estimado de X^* en el modelo de regresión original:

$$\begin{aligned} Y &= X^*\beta + \varepsilon \\ Y &= \hat{X}^*\beta + \varepsilon \end{aligned}$$

y obtengo el estimador de β mediante MCO:

$$\begin{aligned} \hat{\beta}_{IV} &= (\hat{X}^{*'}\hat{X}^*)^{-1}\hat{X}^{*'}Y \\ &= [X^{*'}Z(Z'Z)^{-1}Z'X^*]^{-1}X^{*'}Z(Z'Z)^{-1}Z'Y \end{aligned}$$

Si todas las variables explicativas están medidas con error cada una de ellas se necesita un instrumento, entonces Z tiene dimensión $n \times k$ al igual que X^*, en este caso se puede demostrar (Hacerlo) que:

$$\hat{\beta}_{IV} = (Z'X^*)^{-1}Z'Y$$

con matriz de varianzas y covarianzas (también demostrar):

$$V(\hat{\beta}_{IV}) = \sigma_\epsilon^2(Z'\hat{X}^*)^{-1}(Z'Z)(\hat{X}^{*'}Z)^{-1}$$

Notemos que:

- Bajo errores de medida, el estimador MCO es inconsistente, mientras que el estimador de variables instrumentales es consistente.
- Si en realidad no hubiese errores de medida, ambos estimadores serán consistentes, y MCO es además eficiente, lo que no ocurre con cualquier estimador de variables instrumentales (es un estimador en dos etapas, lo que hace perder eficiencia).

Para contrastar la existencia de errores de medida Hausman plantea realizar un test estadístico comparando $(\hat{\beta}_{MCO} - \hat{\beta}_{VI})$ con su matriz de varianzas y covarianzas.

La hipótesis nula es que no existe error de medida, es decir:

$$H_0 : \hat{\beta}_{MCO} - \hat{\beta}_{VI} = 0$$

Hausman demuestra que la matriz de varianzas y covarianzas de $(\hat{\beta}_{MCO} - \hat{\beta}_{VI})$ es igual a $V(\hat{\beta}_{VI}) - V(\hat{\beta}_{MCO})$. De esta forma, se puede construir el siguiente estadístico de Wald para la hipótesis nula anterior:

$$W = (\hat{\beta}_{MCO} - \hat{\beta}_{VI})'(V(\hat{\beta}_{VI}) - V(\hat{\beta}_{MCO}))^{-1}(\hat{\beta}_{MCO} - \hat{\beta}_{VI}) \sim \chi^2_k$$