Tema 6

Multicolinealidad

Contenido

A la hora de estimar un modelo económico, los datos disponibles sobre las variables explicativas o regresores pueden presentar un alto grado de correlación, especialmente en un contexto de series temporales y con series macroeconómicas. Por ejemplo, la población y el PIB en general suelen estar altamente correlacionados. A este fenómeno se le conoce como multicolinealidad. En algún caso puede que los datos de una variable se obtengan como resultado de una identidad contable o de una combinación lineal exacta entre otros regresores. Este último caso se denomina de multicolinealidad exacta o perfecta.

Cuando dos o más variables explicativas en un modelo están altamente correlacionadas en la muestra, es muy difícil separar el efecto parcial de cada una de estas variables sobre la variable dependiente. La información muestral que incorpora una de estas variables es casi la misma que el resto de las correlacionadas con ella. En el caso extremo de multicolinealidad exacta no es posible estimar separadamente estos efectos sino una combinación lineal de ellos. En este tema analizaremos las implicaciones que tiene en la estimación por el método de Mínimos Cuadrados Ordinarios este fenómeno muestral.

6.1. Multicolinealidad perfecta

Dada la especificación del modelo y los datos de las variables, si al menos una de las variables **explicativas** se puede obtener como combinación lineal exacta de alguna o algunas de las restantes, diremos que existe multicolinealidad exacta o perfecta.

Consideremos el siguiente ejemplo. ¿Qué ocurrirá si definimos una nueva variable $F25$ que es una combinación lineal exacta de otra variable explicativa en el modelo, $F25 = 5 \times F2$ y pretendemos estimar los parámetros del siguiente modelo?

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 F25_i + u_i \qquad i = 1, 2, \ldots, N \tag{6.1}$$

Las variables $F25$ y $F2$ son combinación lineal exacta por lo que el rango de la matriz X es $3 = K - 1$, menor que el número de parámetros a estimar, ya que la cuarta columna se obtiene de multiplicar por 5 la segunda columna. El sistema de ecuaciones normales que se obtiene del criterio de estimación del método de Mínimos Cuadrados Ordinarios sería un sistema de cuatro ecuaciones pero solamente tres serán linealmente independientes[1].

$$\begin{aligned}
\sum Y_i &= N\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \hat{\beta}_3 \sum X_{3i} + \hat{\beta}_4 \sum X_{4i} \\
\sum Y_i X_{2i} &= \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{3i} X_{2i} + \hat{\beta}_4 \sum X_{4i} X_{2i} \\
\sum Y_i X_{3i} &= \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2 + \hat{\beta}_4 \sum X_{4i} X_{3i} \\
\sum Y_i X_{4i} &= \hat{\beta}_1 \sum X_{4i} + \hat{\beta}_2 \sum X_{2i} X_{4i} + \hat{\beta}_3 \sum X_{3i} X_{4i} + \hat{\beta}_4 \sum X_{4i}^2
\end{aligned}$$

Si sustituimos en estas ecuaciones la relación lineal exacta $X_{4i} = 5X_{2i}$ y reorganizamos,

[1] La notación utilizada es $Y_i \equiv P_i$, $X_{2i} \equiv F2_i$, $X_{3i} \equiv BEDRMS_i$, $X_{4i} \equiv F25_i$.

obtenemos:

$$\begin{aligned}
\sum Y_i &= N\hat{\beta}_1 + (\hat{\beta}_2 + 5\hat{\beta}_4)\sum X_{2i} + \hat{\beta}_3 \sum X_{3i} \\
\sum Y_i X_{2i} &= \hat{\beta}_1 \sum X_{2i} + (\hat{\beta}_2 + 5\hat{\beta}_4)\sum X_{2i}^2 + \hat{\beta}_3 \sum X_{3i} X_{2i} \\
\sum Y_i X_{3i} &= \hat{\beta}_1 \sum X_{3i} + (\hat{\beta}_2 + 5\hat{\beta}_4)\sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2 \\
5\,[\sum Y_i X_{2i} &= \hat{\beta}_1 \sum X_{2i} + (\hat{\beta}_2 + 5\hat{\beta}_4)\sum X_{2i}^2 + \hat{\beta}_3 \sum X_{3i} X_{2i}]
\end{aligned}$$

Se puede observar que la cuarta ecuación es la misma que la segunda excepto por un factor de escala igual a 5. Por lo tanto, hay cuatro incógnitas $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ y $\hat{\beta}_4$ pero solamente tres ecuaciones linealmente independientes. Consecuentemente, no es posible estimar de forma única todos los coeficientes del modelo. Ahora bien, las tres primeras ecuaciones si podemos resolverlas para $\hat{\beta}_1$, $\hat{\beta}_3$ y la combinación lineal $(\hat{\beta}_2 + 5\hat{\beta}_4)$. Esto mismo se puede comprobar sustituyendo $F25_i = 5 \times F2_i$ en el modelo (6.1).

$$P_i = \beta_1 + (\beta_2 + 5\beta_4)\, F2_i + \beta_3\, BEDRMS_i + u_i \qquad i = 1, 2, \ldots, N \tag{6.2}$$

Vemos que en esta regresión son estimables de forma separada y única los coeficientes β_1 y β_3 pero no β_2 y β_4. El coeficiente que acompaña a $F2_i$ recogería la combinación lineal $\beta_2 + 5\beta_4$.

¿Qué hace el programa GRETL si hay multicolinealidad perfecta? Elimina una variable cualquiera de las que forman parte de esa relación exacta, mostrando el siguiente resultado.

```
Modelo 8: estimaciones MCO utilizando las 14 observaciones 1-14
Variable dependiente: P

Omitidas debido a colinealidad exacta: F25

     VARIABLE       COEFICIENTE       DESV.TÍP.      ESTAD T      VALOR P
  const               121,179           80,1778        1,511   0,15888
  F2                  0,148314         0,0212080     6,993   0,00002 ***
  BEDRMS              -23,9106           24,6419       -0,970   0,35274

  Media de la var. dependiente = 317,493
  Desviación típica de la var. dependiente. = 88,4982
  Suma de cuadrados de los residuos = 16832,8
  Desviación típica de los residuos = 39,1185
  R-cuadrado = 0,834673
  R-cuadrado corregido = 0,804613
  Estadístico F (2, 11) = 27,7674 (valor p = 5,02e-005)
  Log-verosimilitud = -69,5093
  Criterio de información de Akaike (AIC) = 145,019
  Criterio de información Bayesiano de Schwarz (BIC) = 146,936
  Criterio de Hannan-Quinn (HQC) = 144,841
```

Por lo tanto, **avisa** de que ha eliminado una variable explicativa de la regresión, en este caso $F25$, y muestra los resultados de la regresión excluyendo esa variable. De hecho, el coeficiente que acompaña a F2 podría considerarse como $\widehat{(\beta_2 + 5\beta_4)}$. Este ha sido un ejemplo ilustrativo de las implicaciones que tiene el problema de multicolinealidad perfecta.

6.2. Multicolinealidad de grado alto

En general es difícil tener en un modelo de regresión variables explicativas o regresores que no presenten cierta correlación muestral. La multicolinealidad, de no ser perfecta, se puede considerar un problema cuando la correlación entre los regresores es tan alto que se hace casi imposible estimar con precisión los efectos individuales de cada uno de ellos.

Si la correlación entre la variables explicativas es alta, es común tener los siguientes síntomas:

- Pequeños cambios en los datos o en la especificación provocan grandes cambios en las estimaciones de los coeficientes.
- La estimaciones de los coeficientes suelen presentar signos distintos a los esperados y magnitudes poco razonables.
- El efecto más pernicioso de la existencia de un alto grado de multicolinealidad es el de incrementar las varianzas de los coeficientes estimados por MCO. Es decir, es difícil estimar separadamente los efectos marginales o individuales de cada variable explicativa por lo que estos se estiman con poca precisión.[2] Como consecuencia, el valor del estadístico para realizar contrastes de significatividad individual tiende a ser pequeño y aumenta la probabilidad de no rechazar la hipótesis nula, por lo que se tiende a concluir que las variables no son significativas individualmente. El problema no reside en que los contrastes no sean correctos estadísticamente, sino en que no estimamos con suficiente precisión estos efectos individuales.
- Se obtienen valores altos del R^2 aún cuando los valores de los estadísticos t de significatividad individual son bajos. El problema reside en la identificación del efecto individual de cada variable explicativa, no tanto en su conjunto. Por eso, si se realiza un contraste de significatividad conjunta de las variables explicativas, el resultado normalmente será rechazar la hipótesis nula por lo que conjuntamente son significativas aunque individualmente cada una de ellas no lo sea.

Si se presentan estos síntomas se puede sospechar que el problema de multicolinealidad esté afectando a nuestros resultados, especialmente a la inferencia sobre los efectos individuales de cada variable explicativa. De todas formas es importante analizar e interpretar adecuadamente los resultados obtenidos sin tomar conclusiones precipitadamente.

¿Cómo podemos analizar si existe un problema de multicolinealidad?

1) Una primera aproximación consiste en obtener los coeficientes de correlación muestral simples para cada par de variables explicativas y ver si el grado de correlación entre estas variables es alto.

   Utilizando el ejemplo de los precios de los pisos (Fichero de muestra del Ramanathan *data4-1*) con las variables que ya analizamos en temas anteriores,

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i$$

[2] Los estimadores MCO siguen siendo los de menor varianza dentro de la clase de lineales e insesgados si las hipótesis básicas se satisfacen. Luego no es un problema de pérdida de eficiencia relativamente a otro estimador lineal e insesgado.

obtenemos los siguientes valores de los coeficientes de correlación:

```
Coeficientes de correlación, usando las observaciones 1 - 14
          valor crítico al 5% (a dos colas) = 0,5324 para n = 14

           P            F2          BEDRMS         BATHS
      1,0000        0,9058          0,3156        0,6696  P
                    1,0000          0,4647        0,7873  F2
                                    1,0000        0,5323  BEDRMS
                                                  1,0000  BATHS
```

Como podemos observar, todas las variables explicativas presentan cierto grado de correlación dos a dos, siendo la correlación mayor entre F2 y BATH con un coeficiente igual a 0,7873. Excepto por este valor, no parece que los coeficientes de correlación simple sean demasiado grandes para sospechar que haya un problema de multicolinealidad. De todas formas, aunque es condición suficiente para que exista este problema que todos estos coeficientes fueran altos, lo contrario no necesariamente es cierto. Se puede dar el caso de tener una relación lineal casi perfecta entre tres o más variables y sin embargo las correlaciones simples entre pares de variables no ser mayores que $0,5$.

2) Otra forma de **detectar la multicolinealidad** consiste en realizar la regresión de cada una de las variables explicativas sobre el resto[3] y analizar los coeficientes de determinación de cada regresión. Si alguno o algunos de estos coeficientes de determinación (R_j^2) son altos, estaría señalando la posible existencia de un problema de multicolinealidad.

   Siguiendo con el ejemplo sobre el modelo del precio de la vivienda, esto consistiría en realizar las siguientes regresiones:

```
Modelo 1: estimaciones MCO utilizando las 14 observaciones 1-14
Variable dependiente: F2

  VARIABLE     COEFICIENTE       DESV.TÍP.      ESTAD T  VALOR P
  const        -657,612          809,640        -0,812   0,43389
  BEDRMS         73,9671         254,175         0,291   0,77646
  BATHS         975,371          283,195         3,444   0,00548 ***

  R-cuadrado = 0,622773

Modelo 2: estimaciones MCO utilizando las 14 observaciones 1-14
Variable dependiente: BEDRMS

  VARIABLE     COEFICIENTE       DESV.TÍP.     ESTAD T  VALOR P
  const         2,29560          0,700852       3,275   0,00739 ***
  F2            0,000103288      0,000354931    0,291   0,77646
  BATHS         0,487828         0,459485       1,062   0,31113
```

[3] En cada regresión se incluye el término constante como regresor pero no como variable dependiente.

```
  R-cuadrado = 0,288847

Modelo 3: estimaciones MCO utilizando las 14 observaciones 1-14
Variable dependiente: BATHS

  VARIABLE    COEFICIENTE       DESV.TÍP.    ESTAD T  VALOR P
  const        0,646527         0,583914      1,107   0,29182
  F2           0,000531961      0,000154452   3,444   0,00548 ***
  BEDRMS       0,190531         0,179461      1,062   0,31113

  R-cuadrado = 0,655201
```

Los resultados parecen mostrar que las variaciones muestrales de las variables $F2$ y $BATHS$ son las más explicadas por el resto de variables explicativas, aunque los coeficientes de determinación de esas dos regresiones no son excesivamente altos; alrededor de un 60 % de la variación de $F2$ y de $BATHS$ vienen explicadas por variaciones en el resto de variables explicativas. Si recordamos los resultados obtenidos en el Tema 3, donde al estimar el modelo 3 una vez que incluíamos $F2$ en la regresión, obteníamos que las variables $BATH$ y $BEDRMS$ no eran significativas. ¿Puede ser este hecho consecuencia de un problema de multicolinealidad? ¿Podríamos tener problemas de multicolinealidad entre las variables $F2, BATHS$ y $BEDRMS$? Vamos a utilizar algún procedimiento más formal para detectar si existe este problema.

3) Neter, Wasserman & Kutner (1990) consideran una serie de indicadores para analizar el grado de multicolinealidad entre los regresores de un modelo, como por ejemplo los llamados **Tolerancia** (TOL) y **Factor de Inflación de la Varianza** (VIF) que se definen:

$$VIF_j = \frac{1}{\left(1 - R_j^2\right)} \qquad TOL_j = \frac{1}{VIF_j}$$

siendo R_j^2 el coeficiente de determinación de la regresión auxiliar de la variable X_j sobre el resto de las variables explicativas y $1 \leq VIF_j \leq \infty$.

La varianza de cada uno de los coeficientes de la regresión MCO ($\hat{\beta}_j$) de un modelo de regresión lineal general se puede expresar como:

$$var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{N} \left(X_{ji} - \bar{X}_j\right)^2} \frac{1}{\left(1 - R_j^2\right)} = \frac{\sigma^2}{\sum_{i=1}^{N} \left(X_{ji} - \bar{X}_j\right)^2} VIF_j$$

donde β_j, es el coeficiente que acompaña a la variable X_j y R_j^2 es el coeficiente de determinación de la regresión auxiliar de la variable X_j en función del resto de las variables explicativas. Como vemos existe una relación inmediata entre el valor VIF_j y la varianza del coeficiente estimado. Cuanto más se acerque R_j^2 a la unidad, es decir, cuanto mayor sea la colinealidad de la variable X_j con el resto, mayor es el valor de VIF_j y mayor es la varianza del coeficiente estimado, porque tal y como hemos dicho,

la multicolinealidad "infla" la varianza. Según estos autores, si $VIF_j > 10$, entonces concluiremos que la colinealidad de X_j con las demás variables es alta.

La utilización de los coeficientes TOL y VIF para detectar la presencia de la multicolinealidad ha recibido múltiples críticas, porque la conclusión obtenida con estos valores no siempre recoge adecuadamente la información y problema de los datos. Tal y como hemos visto anteriormente, las varianzas de los estimadores depende del VIF_j, σ^2 y $\sum \left(X_{ji} - \bar{X}_j\right)^2$, por lo que un alto VIF_j no es condición suficiente ni necesaria para que dichas varianzas sean elevadas ya que es posible que σ^2 sea pequeño o $\sum \left(X_{ji} - \bar{X}_j\right)^2$ grande y se compensen.

Los indicadores TOL y VIF se pueden obtener con el programa GRETL de forma muy sencilla. Siguiendo con el ejemplo de los precios de las viviendas, calcularemos la Inflación de la Varianza para analizar la posible presencia de multicolinealidad. Para ello, en la ventana de la estimación por MCO del modelo de interés, elegimos la opción

Contrastes → *Colinealidad*

obteniendo la siguiente información:

```
Factores de inflación de varianza (VIF)

 Mínimo valor posible = 1.0
 Valores mayores que 10.0 pueden indicar un problema de colinealidad

   2)            F2      2,651
   3)          BEDRMS    1,406
   4)           BATHS    2,900

VIF(j) = 1/(1 - R(j)^2), donde R(j) es el coeficiente de
correlación múltiple entre la variable j y las demás variables
independientes
```

Como podemos observar, según los valores del VIF_j, podríamos concluir que no existen problemas de multicolinealidad.

Aunque no es fácil, se pueden considerar las siguientes "soluciones" para intentar resolver el problema:

- Si realmente es un problema muestral, una posibilidad es cambiar de muestra porque puede ser que con nuevos datos el problema se resuelva, aunque esto no siempre ocurre. La idea consiste en conseguir datos menos correlacionados que los anteriores, bien cambiando toda la muestra o simplemente incorporando más datos en la muestra inicial. De todas formas, no siempre resulta fácil obtener mejores datos por lo que muy probablemente debamos convivir con el problema teniendo cuidado con la inferencia realizada y las conclusiones de la misma.

- En ocasiones, si se incorpora información a priori sobre los coeficientes del modelo desaparece el problema. Aún así, sería conveniente tener en cuenta dicha información antes de la detección del problema de multicolinealidad y no posteriormente, ya que así estimaremos el modelo más eficientemente.

- Quitar del modelo alguna de las variables colineales. Es una medida que puede provocar otro tipo de problemas, ya que si la variable que eliminamos del modelo realmente sí es significativa, estaremos omitiendo una variable relevante. Por consiguiente, los estimadores de los coeficientes del modelo y de su varianza serían sesgados por lo que la inferencia realizada no sería válida.

- Existen otros métodos de estimación sugeridos en la literatura econométrica que mejorarían la estimación en términos de eficiencia o precisión, pero los estimadores así obtenidos serían sesgados. Explicar estos métodos no entran dentro de los objetivos de este curso.

Bibliografía

Neter, J., Wasserman, W. y M. H. Kutner (1990), *Applied Linear Statistical Models*, 3ª edn., M.A: Irwin.