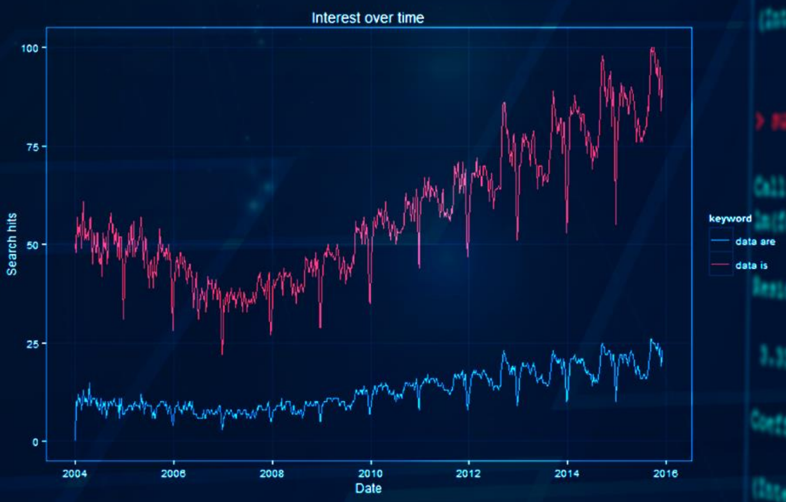
Econometría Aplicada con R



INSTITUTO CIENTÍFICO DEL PACÍFICO

SESIÓN 8: Multicolinealidad

Contenido

Introducción

La multicolinealidad es un problema que suele presentarse en la estimación de modelos econométricos. Este problema implica que cierta parte de las variables del modelo guarden relación entre ellas, y esta relación variará en intensidad.

En la siguiente sesión se verán las implicancias teóricas de la multicolinealidad, así como la identificación de dicho problema y la solución del mismo haciendo uso de R como herramienta principal.

Multicolinealidad

La multicolinealidad es un problema que se presenta de forma común en los modelos de regresión lineal en donde se hacer referencia a la asociación lineal entre las variables explicativas del modelo.

Se dice que es un fenómeno muestral porque en la población no se presenta, por ejemplo, se sabe que el consumo puede ser explicado por la riqueza a nivel poblacional, pero cuando se toman muestras ambas variables resultan ser no significativas, ya el valor de sus respectivos estadísticos t son no significativos.

Por lo general se tiende a sospechar de cierto grado de multicolinealidad cuando el valor de coeficiente de determinación R-cuadrado estimado en el modelo es muy alto.

Multicolinealidad perfecta

Suponiendo que se tiene un modelo de regresión lineal con k variables explicativas $x_1, x_2, \ldots, x_k$ (donde $x_1 = 1$ para todas las observaciones de forma que den cabida al intercepto), se dice que existirá una relación lineal exacta si se satisface la siguiente condición:

$$\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k = 0$$

Donde $\lambda_1, \lambda_2, \ldots, \lambda_k$, son constantes que son simultáneamente iguales a cero.

Consecuencias

Los parámetros estimados son indeterminados debido a que no es posible separar las influencias de las distintas variables explicativas debido a que están relacionadas linealmente.

Multicolinealidad imperfecta

En este caso a la relación lineal entre las variables explicativas se le suma un término denominado error estocástico.

Tiene la siguiente definición:

$$\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k + v_i = 0$$

Donde v_i es el término de error estocástico.

Consecuencias

Podemos estimar los parámetros por MCO pero los valores estimados no son muy confiables. Cuanto más grande es la correlación, más próximo a cero será el determinante de la matriz $X'X$ lo cuál incrementará las varianzas y covarianzas de los parámetros estimados.

Consecuencias de la multicolinealidad

Aunque los estimadores MCO son MELI (mejor estimador lineal insesgado), presentan varianzas y covarianzas grandes, lo que dificulta su estimación precisa.

Debido a la primera consecuencia, los intervalos de confianza tienden a ser mucho más amplios, por lo que se hace propicia una aceptación más fácil de la hipótesis nula (el verdadero coeficiente poblacional).

El valor del estadístico t de uno o más coeficientes tiende a 0.

El coeficiente de determinación R-cuadrado al igual que la prueba F como medidas de asociación conjunta mostrarán valores altos.

Los estimadores MCO y sus errores estándar se hacen sensibles a pequeños cambios en los datos.

Detección de la multicolinealidad

Como la multicolinealidad es en esencia un fenómeno de tipo muestral que surge de información sobre todo no experimental recopilada en la mayoría de las ciencias sociales, no hay un método único para detectarla o medir su fuerza.

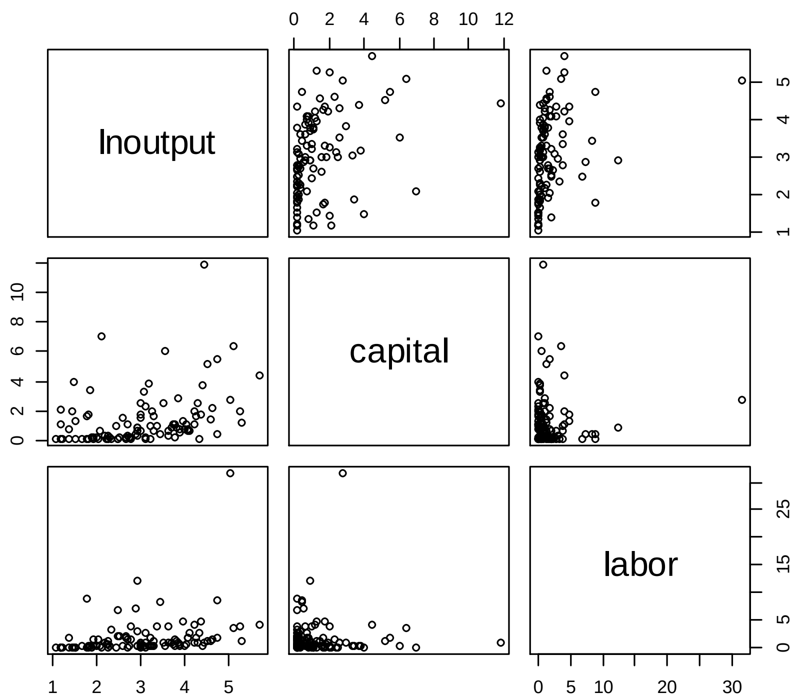
R-cuadrado elevado con coeficientes no significativos

Como ya se ha mencionado, es un síntoma "clásico" de multicolinealidad. Si R^2 es alta, es decir, está por encima de 0.8, la prueba F, en la mayoría de los casos, rechazará la hipótesis de que los coeficientes parciales de pendiente son simultáneamente iguales a cero, pero las pruebas t individuales mostrarán que ningún coeficiente parcial de pendiente, o muy pocos, son estadísticamente diferentes de cero.

Coeficiente de correlación y diagramas de dispersión

Una forma de detectar la alta correlación que puedan existir en cada uno de los predictores es haciendo una matriz de los coeficientes de correlación y observar el valor de cada uno de estos coeficientes.

Otra medida para detectar multicolinealidad es hacer dispersiones entre las variables explicativas, y las que tengan una relación lineal muy alta a nivel gráfico, van a generar sospecha de multicolinealidad.



Análisis del Factor Inflador de Varianzas (VIF)

Este factor representa la velocidad con la que crecen las varianzas en el modelo de regresión.

El VIF (o FIV en español) muestra la forma como la varianza de un estimador se infla por la presencia de la multicolinealidad. A medida que el grado de colinealidad entre los regresores se acerca a 1, el VIF se acerca a infinito. Es decir, a medida que el grado de colinealidad aumenta, la varianza de un estimador también y, en el límite, se vuelve infinita.

Este factor viene dado por la siguiente expresión:

$$VIF(\beta_i) = \frac{1}{1 - R_{xi}^2}$$

En el análisis de multicolinealidad se considera que un VIF cercano o mayor a 10 indica la existencia de alta multicolinealidad.

Se puede utilizar la siguiente regla para interpretar el factor inflador de varianzas:

VIF	**ESTADO DE LOS PREDICTORES**
VIF = 1	No correlacionados
1 < VIF < 5	Moderadamente correlacionados
VIF > 5 A 10	Altamente correlacionados

Detección de multicolinealidad en R

De acuerdo a las diferentes formas de multicolinealidad vistas en el punto anterior, se verá en esta parte la aplicación práctica de cada una de estas haciendo uso de R.

Matriz de correlaciones

Los coeficientes de correlación entre las variables se calculan con la función **cor()** de R.

```
cor(production[c("lnoutput", "capital", "labor")], use="complete")

          lnoutput     capital       labor
lnoutput 1.0000000  0.35652230  0.30362168
capital  0.3565223  1.00000000 -0.01379399
labor    0.3036217 -0.01379399  1.00000000
```

De acuerdo a los resultados de la matriz, se podrá intuir cuáles son los términos regresores que generan multicolinealidad en el modelo.

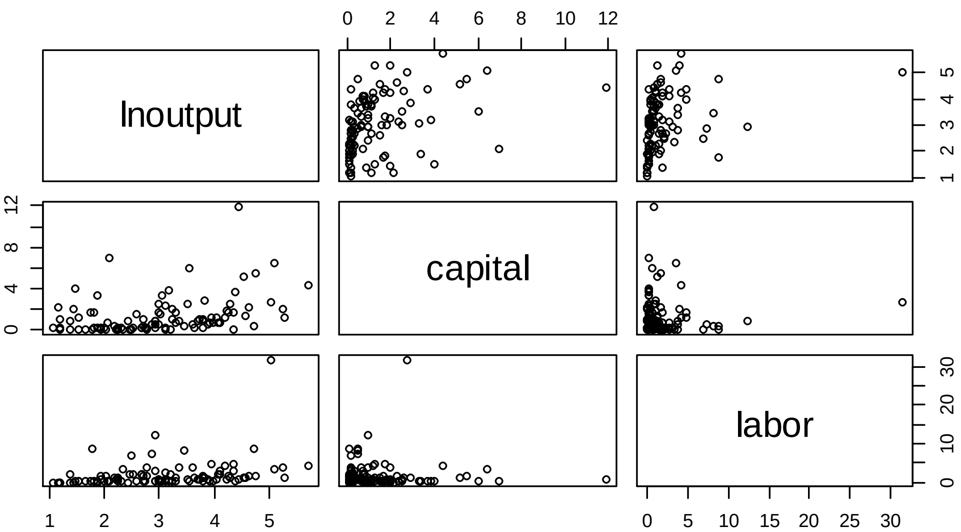
Matriz de dispersiones

Este método también es válido, y que nos permitirá observar la relación visible entre cada una de las variables que se analizarán.

Para hacer matrices de dispersión en R se hace uso del comando **pairs()**, este comando muestra la matriz de correlación para cada una de las variables de las que se quiere hacer la evaluación.

```
pairs(~lnoutput + capital + labor)
```

El resultado mostrado es el siguiente:



Factor Inflador de Varianzas

Para estimar el factor de varianzas haciendo uso de R, se debe usar el comando **vif()**, que pertenece al paquete **car**, con el cuál se pueden hacer diversos tests para modelo de regresión lineal.

```
library(car)
modelo = lm(lnoutput ~ capital + labor, data = production)
vif(modelo)

capital   labor
1.00019 1.00019
```

En primer lugar, se carga el paquete car, luego de esto se genera un objeto llamado modelo, definido por una regresión lineal. Una vez generado el objeto se hace uso del comando **vif()**, indicando que se hará un análisis del factor inflador de varianza del modelo estimado, como los resultados muestran estos factores muy cercanos a 1, se concluye que en este modelo las variable independiente no generan multicolinealidad.

Soluciones a la multicolinealidad

Cuando se ha detectado que el grado de multicolinealidad del modelo estimado 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.

Existen diversas formas de dar correcciones a la multicolinealidad:

Eliminación de variables colineales

Si en el modelo se incluyen dos variables explicativas que cumplen funciones parecidas, no se pierde consistencia teórica si se elimina alguna de estas variables colineales. Igualmente, si la colinealidad se debe a la inclusión de variables irrelevantes, la eliminación de variables no significativas reducirá el grado de colinealidad.

Se debe tener cuidado cuando se eliminan variables en el modelo, esto no es recomendables ya que también se podría incurrir en un problema de variable omitida en el modelo de regresión.

Transformación de variables en el modelo

Si la colinealidad ocurre cuando las variables están expresadas en niveles (datos observados), podría modificarse una expresión en primeras diferencias, se deberá tener cuidado por la eventual generación de heterocedasticidad o autocorrelación.

Por ejemplo, en el siguiente modelo:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \mu_t \qquad (1)$$

$$y_{t-1} = \beta_0 + \beta_1 x_{1t-1} + \beta_2 x_{2t-1} + \mu_{t-1} \qquad (2)$$

(1) - (2):

$$\Delta y_t = \beta_1 \Delta x_{1t} + \beta_2 \Delta x_{2t} + v_t$$

El modelo inicial se transformó en un modelo donde se expresan las primeras diferencias de las variables.

Hay casos donde la transformación de pasa por el concepto de deflactores o en términos relativos. La idea es que si 2 variables son colineales en niveles es poco probable que la colinealidad se mantenga en términos relativos.

Suponiendo que en el siguiente modelo las variables x_1 y x_2 son colineales:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Al dividir el modelo entre x_1, se tendrá:

$$\frac{y}{x_1} = \beta_0 + \beta_1 + \frac{\beta_2 x_2}{x_1} + \frac{\varepsilon}{x_1}$$

Bibliografía

Kleiber, C. & Zeileis, A. (2008). *Applied econometrics with R*. Springer Science & Business Media.

Quintana, L. & Mendoza, A. (2016). *Econometria aplicada usando R*. Universidad Nacional Autónoma de México.