Econometria
Tema 4(2): Multicol·linealitat

Ramon Alemany

Grau Estadística UB-UPC

Curs 2017-18

Presentació

Bibliografia

- GREENE, W. (1999)
  Análisis econométrico. 3a Ed.
  Capítol 9

- WOOLDRIDGE, J. (2009)
  Introducción a la Econometría. Un enfoque moderno. 4a Ed.
  Capítol 9

- STOCK, J. & WATSON, M. (2012)
  Introducción a la Econometría. 3a Ed.
  Capítol 6

Multicol·linealitat: Definició i tipologia

1. Definició i tipologia

La multicol·linealitat es defineix com el grau de correlació existent entre les variables explicatives o regressors del MRLM.

Podem distingir entre tres situacions:

- multicol·linealitat perfecta
- absència total de multicol·linealitat
- presència de cert grau de correlació entre variables explicatives

Definició i tipologia

Model de regressió lineal múltiple en desviacions

Sigui el model de regressió següent:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

Per les hipòtesis sobre el terme de pertorbació sabem que:

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{1i} + \beta_2 \bar{x}_{2i}$$

i si restem una de l'altra:

$$y_i - \bar{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i - (\beta_0 + \beta_1 \bar{x}_{1i} + \beta_2 \bar{x}_{2i})$$

$$y_i - \bar{y}_i = \beta_1 (x_{1i} - \bar{x}_{1i}) + \beta_2 (x_{2i} - \bar{x}_{2i}) + u_i$$

El model quedarà doncs:

$$\tilde{y}_i = \beta_1 \tilde{x}_{1i} + \beta_2 \tilde{x}_{2i} + u_i$$

on: $\tilde{y}_i = y_i - \bar{y}_i \quad \tilde{x}_{1i} = x_{1i} - \bar{x}_{1i} \quad \tilde{x}_{2i} = x_{2i} - \bar{x}_{2i}$

Definició i tipologia

El vector d'estimadors dels paràmetres β_1 i β_2 és: $\hat{\beta} = (X'X)^{-1}(X'Y)$

$$\begin{bmatrix} \hat{\beta}_1 \\ \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \sum \tilde{x}_{1i}^2 & \sum \tilde{x}_{1i}\tilde{x}_{2i} \\ \\ \sum \tilde{x}_{1i}\tilde{x}_{2i} & \sum \tilde{x}_{2i}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum \tilde{x}_{1i}\tilde{y}_i \\ \\ \sum \tilde{x}_{2i}\tilde{y}_i \end{bmatrix} =$$

$$= \frac{1}{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2 - (\sum \tilde{x}_{1i}\tilde{x}_{2i})} \begin{bmatrix} \sum \tilde{x}_{2i}^2 & -\sum \tilde{x}_{1i}\tilde{x}_{2i} \\ \\ -\sum \tilde{x}_{1i}\tilde{x}_{2i} & \sum \tilde{x}_{1i}^2 \end{bmatrix} \begin{bmatrix} \sum \tilde{x}_{1i}\tilde{y}_i \\ \\ \sum \tilde{x}_{2i}\tilde{y}_i \end{bmatrix}$$

Per altra banda, el coeficient de correlació lineal entre les dues variables explicatives és:

$$r_{12} = \frac{\sum \tilde{x}_{1i}\tilde{x}_{2i}}{\sqrt{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2}} \qquad r_{12}^2 = \frac{(\sum \tilde{x}_{1i}\tilde{x}_{2i})^2}{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2} \qquad 1 - r_{12}^2 = \frac{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2 - (\sum \tilde{x}_{1i}\tilde{x}_{2i})^2}{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2}$$

Definició i tipologia

i per tant el vector d'estimacions dels paràmetres serà:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \frac{1}{1-r_{12}^2} \begin{bmatrix} \frac{1}{\sum \tilde{x}_{1i}^2} & -\frac{r_{12}}{\sqrt{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2}} \\ -\frac{r_{12}}{\sqrt{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2}} & \frac{1}{\sum \tilde{x}_{1i}^2} \end{bmatrix} \begin{bmatrix} \sum \tilde{x}_{1i}\tilde{y}_i \\ \sum \tilde{x}_{2i}\tilde{y}_i \end{bmatrix}$$

i l'estimador de la variància dels paràmetres estimats és:

$$\widehat{\text{Var}(\hat{\beta})} = \frac{\hat{\sigma}_u^2}{1-r_{12}^2} \begin{bmatrix} \frac{1}{\sum \tilde{x}_{1i}^2} & -\frac{r_{12}}{\sqrt{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2}} \\ -\frac{r_{12}}{\sqrt{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2}} & \frac{1}{\sum \tilde{x}_{1i}^2} \end{bmatrix}$$

Definició i tipologia

Multicol·linealitat Perfecta

Presència de Multicol·linealitat Perfecta

Es produeix quan una de les variables explicatives del model és una combinació lineal exacta d'una altra o d'altres variables del MRLM. Difícilment es presenta a la pràctica.

En aquest cas, es comprova que el rang de la matriu de variables explicatives és inferior a k:

$$\rho(X) < k$$

i no es verifica una de les hipòtesis bàsiques del MRLM.

Definició i tipologia

Multicol·linealitat Perfecta

En el model en desviacions, si existeix una relació lineal exacta entre les dues explicatives del model aleshores:

$$r_{12} = 1 \qquad 1 - r_{12} = 0 \qquad \frac{1}{1 - r_{12}} = \frac{1}{0}$$

$$\mid X'X \mid = 0 \qquad \nexists (X'X)^{-1}$$

En conseqüència, no existirà una solució única pel vector d'estimadors MQO.

Definició i tipologia

Absència total de Multicol·linealitat

Absència total de Multicol·linealitat

Implica que les variables explicatives del MRLM són ortogonals entre sí, és a dir, totalment independents entre elles:

$$\text{Cov}(X_i, X_j) = 0 \quad \forall\, i \neq j$$

Difícilment es presenta a la pràctica.

Si les variables són ortogonals, en tota regressió en que s'intenti explicar el comportament d'una d'elles a partir de les altres els coeficients estimats seran zero.

Definició i tipologia

Absència total de Multicol·linealitat

En el cas que els dos regressors siguin ortogonals aleshores $r_{12} = 0$.

El vector d'estimadors prendrà l'expressió:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} \frac{1}{\sum \tilde{x}_{1i}^2} & -\frac{r_{12}}{\sqrt{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2}} \\ -\frac{r_{12}}{\sqrt{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2}} & \frac{1}{\sum \tilde{x}_{1i}^2} \end{bmatrix} \begin{bmatrix} \sum \tilde{x}_{1i}\tilde{y}_i \\ \sum \tilde{x}_{2i}\tilde{y}_i \end{bmatrix} = \begin{bmatrix} \frac{\sum \tilde{x}_{1i}\tilde{y}_i}{\sum \tilde{x}_{1i}^2} \\ \frac{\sum \tilde{x}_{2i}\tilde{y}_i}{\sum \tilde{x}_{2i}^2} \end{bmatrix}$$

és a dir, les estimacions del MRL Simple per cadascuna de les variables explicatives per separat.

Definició i tipologia

Absència total de Multicol·linealitat

I les variàncies dels paràmetres serien:

$$\widehat{\text{Var}(\hat{\beta})} = \frac{\hat{\sigma}_u^2}{1-r_{12}^2}\begin{bmatrix} \frac{1}{\sum \tilde{x}_{1i}^2} & -\frac{r_{12}}{\sqrt{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2}} \\ -\frac{r_{12}}{\sqrt{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2}} & \frac{1}{\sum \tilde{x}_{1i}^2} \end{bmatrix} = \hat{\sigma}_u^2 \begin{bmatrix} \frac{1}{\sum \tilde{x}_{1i}^2} & 0 \\ 0 & \frac{1}{\sum \tilde{x}_{1i}^2} \end{bmatrix}$$

Si les variables són ortogonals, els estimadors i les seves variàncies són les mateixes tant si s'estima el model conjunt com si es fa amb cada una de les variables per separat.

Definició i tipologia

Absència total de Multicol·linealitat

Per tant, cadascuna de les variables explicatives explica una part de la variabilitat total de la variable endògena. Es pot repartir la SQR entre les diferents variables:

$$SQR = SQR_2 + \ldots + SQR_k$$

El problema de fer dues estimacions separades dels corresponents MRLS seria que obtindríem una estimació esbiaixada de la variància del terme de pertorbació.

Definició i tipologia

Presència d'un cert grau de Multicol·linealitat

Presència d'un cert grau de Multicol·linealitat

És la situació més habitual en la pràctica i suposa que existeix un cert grau de correlació entre les variables explicatives.

Les causes són diverses: pot ser que les variables explicatives presentin una tendència comuna; que tinguem dues variables que captin un concepte semblant; que les observacions recollides siguin molt homogènies, o que presentin poca variabilitat, etc...

Definició i tipologia

Presència d'un cert grau de Multicol·linealitat

Donat que cap variable explicativa és una combinació lineal de la resta, es compleix que $\rho(X) = k$ i, per tant, podem obtenir les estimacions MQO úniques dels paràmetres del MRLM, les quals seguiran sent no esbiaixades, consistents i eficients.

A més, la multicol·linealitat no afecta a l'estimació de la variància del terme de perturbació.

Malgrat això, les conseqüències sobre les estimacions del model poden arribar a ser importants, tant més com més gran sigui el grau de multicol·linelitat.

Conseqüències de l'elevada multicol·linealitat

2. Conseqüències de l'elevada multicol·linealitat

Entre les principals conseqüències de la presència d'un cert grau de multicol·linealitat es troben les següents:

- Augment de la variància dels paràmetres estimats.
- Disminució de la precisió de les prediccions i limitacions importants en l'anàlisi estructural del model.
- Impossibilitat d'aïllar l'efecte de cadascuna de les variables explicatives sobre la variable endògena.

Conseqüències de l'elevada multicol·linealitat

La variància dels paràmetres estimats és:

$$\widehat{\text{Var}(\hat{\beta})} = \frac{\hat{\sigma}_u^2}{1-r_{12}^2} \begin{bmatrix} \frac{1}{\sum \tilde{x}_{1i}^2} & -\frac{r_{12}}{\sqrt{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2}} \\ -\frac{r_{12}}{\sqrt{\sum \tilde{x}_{1i}^2 \sum \tilde{x}_{2i}^2}} & \frac{1}{\sum \tilde{x}_{1i}^2} \end{bmatrix}$$

$$\widehat{\text{Var}(\hat{\beta}_1)} = \frac{\hat{\sigma}_u^2}{\sum \tilde{x}_{1i}^2 (1-r_{12}^2)} \qquad \widehat{\text{Var}(\hat{\beta}_2)} = \frac{\hat{\sigma}_u^2}{\sum \tilde{x}_{2i}^2 (1-r_{12}^2)}$$

Per tant, a mesura que augmenti r_{12} cap a la unitat, el denominador es farà més petit i el quocient més gran, de manera que la variància de l'estimador serà més gran.

Conseqüències de l'elevada multicol·linealitat

Menor precisió i menor estabilitat de les estimacions

Una elevada multicol·linealitat provoca una **menor precisió i una menor estabilitat de les estimacions per MQO.**

A mesura que augmenta el grau de correlació existent entre les variables explicatives, és a dir, la multicol·linealitat, la variància de l'estimador MQO de β serà més elevada i, per tant, la seva precisió i estabilitat serà menor.

Els paràmetres seran poc estables (canviaran substancialment de valor quan afegim o treiem observacions a la mostra) i ens podem trobar amb signes no esperats per la T^{a} Econòmica.

Conseqüències de l'elevada multicol·linealitat

Falsa no significació individual

La **hipòtesi nul·la de no significació individual** dels paràmetres **serà no rebutjada més freqüentment**.

$$\begin{cases} H_0: & \beta_j = 0 \\ H_A: & \beta_j \neq 0 \end{cases} \qquad \frac{\hat{\beta}_j}{\text{e.s.}(\hat{\beta}_j)} \sim t_{N-k;\alpha/2}$$

Si augmenta la variància de l'estimador, disminueix el valor de l'estadístic de contrast i tindrem una tendència més gran a quedar per sota del valor crític, i per tant, a acceptar la hipòtesi nul·la.

Es pot arribar a concloure que una variable no és rellevant quant en realitat sí que ho és.

Conseqüències de l'elevada multicol·linealitat

Prediccions imprecises i anàlisi estructura poc fiable

La **predicció per interval** de Y serà **menys precisa**:

$$\hat{y}_{N+h} \pm t_{N-k;\alpha/2}\sqrt{\hat{\sigma}_u^2\left(1 + x'_{N+h}(x'x)^{-1}x_{N+h}\right)}$$

Es redueix la possibilitat de fer anàlisi estructural basada en el valor dels paràmetres estimats.

Com a conseqüència de la imprecisió de les estimacions no és recomanable fer una anàlisi estructural del model o, si més no, cal relativitzar les conclusions.

Conseqüències de l'elevada multicol·linealitat

Per tant, en la majoria dels casos, valors elevats dels coeficients de correlació dos a dos entre les variables explicatives poden provocar problemes importants de multicol·linealitat.

Aquest efecte pot ser contrarestat si la variància del terme de pertorbació és prou petita o bé si la variància de la variable explicativa considerada és prou gran.

Detecció i valoració de la multicol·linealitat

3. Detecció i valoració de la multicol·linealitat

Hi ha diversos instruments de detecció de problemes de multicol·linealitat al MRLM.

- Anàlisi dels paràmetres estimats
- Contradicció entre els contrastos de significació individual i conjunta
- Coeficient de correlació entre parells de variables
- Determinant de la matriu de correlacions
- Anàlisi del R^2 de les regressions auxiliars
- Anàlisi del R^2 de les regressions on s'eliminen variables explicatives
- El Factor d'Increment de la Variància (FIV_j)

Detecció i valoració de la multicol·linealitat

Anàlisi dels paràmetres estimats

Anàlisi dels paràmetres estimats

Si els resultats de les estimacions MQO són molt sensibles a variacions en la mida de la mostra, és a dir, afegint o traient unes poques observacions,...

O si s'obtenen estimacions dels paràmetres del model de regressió força allunyades d'aquelles que preveu la teoria econòmica (o, fins i tot, amb signe contrari), aleshores,...

Podem sospitar que hi ha un problema important o greu d'elevada multicol·linealitat.

Detecció i valoració de la multicol·linealitat

Contradicció entre els contrastos de significació

Contradicció entre els contrastos de significació individual i conjunta.

La presència de multicol·linealitat elevada pot afectar al test de significació individual dels paràmetres estimats portant a concloure que una variable no és rellevant quan en realitat sí que ho és. Aquest problema però no el té el contrast de la F-Snedecor de significació conjunta.

Per tant, podem sospitar que existeix un problema de multicol·linealitat elevada quan els paràmetres no resultin ser significatius a nivell individual però el model sí resulti ser significatiu globalment.

Detecció i valoració de la multicol·linealitat

Coeficient de correlació entre parells de variables

Coeficient de correlació entre parells de variables

Com ja hem vist anteriorment, un indicador per valorar la possible existència d'un problema de multicol·linealitat és el càlcul dels coeficients de correlació simple entre parells de variables explicatives del MRLM (r_{ij}).

Així, coeficients de correlació simple alts entre parells de variables explicatives ens indicaran una multicol·linealitat elevada.

Seran preocupants quan $r_{ij} > R^2_{y,x_1x_2...x_k}$

Detecció i valoració de la multicol·linealitat

Determinant de la Matriu de Correlacions

Determinant de la Matriu de Correlacions

Un altre instrument de detecció és el càlcul del determinant de la matriu de correlacions entre les variables explicatives, R_x.

$$|R_x| = \begin{vmatrix} 1 & r_{23} & \dots & r_{2k} \\ r_{32} & 1 & \dots & r_{3k} \\ \vdots & \vdots & \dots & \vdots \\ r_{k2} & r_{k3}1 & \dots & 1 \end{vmatrix} \qquad |R_x| \in [0,1]$$

$|R_x| = 0$ Multicol·linealitat Perfecta

$|R_x| = 1$ Absència Total de Multicol·linealitat

$|R_x| \approx 0$ Elevada Multicol·linealitat

Detecció i valoració de la multicol·linealitat

Anàlisi del R^2 de regressions auxiliars

Anàlisi del R^2 de regressions auxiliars

Un altre mètode consisteix en analitzar els coeficients de determinació de les regressions auxiliars en les quals figuri com a variable endògena successivament cada variable explicativa i com a exògenes, la resta de variables explicatives.

En cas que el coeficient de determinació d'alguna de les regressions auxiliars sigui elevat (R_j^2), voldrà dir que la variable que actua com a endògena es troba altament correlacionada amb la resta de variables explicatives, havent-hi doncs multicol·linealitat elevada en el model original.

Detecció i valoració de la multicol·linealitat

Anàlisi del R^2 de regressions auxiliars

Per exemple, per un MRLM amb k=4:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i$$

trobaríem les tres regressions auxiliars següents:

$$x_{2i} = \gamma_0 + \gamma_1 x_{3i} + \gamma_2 x_{4i} + v_i$$

$$x_{3i} = \gamma_0 + \gamma_1 x_{2i} + \gamma_2 x_{4i} + v_i$$

$$x_{4i} = \gamma_0 + \gamma_1 x_{2i} + \gamma_2 x_{3i} + v_i$$

calculant per cadascuna d'elles el seu coeficient de determinació $R_j^2,\ j = 2,3,4.$

Si algun dels coeficients de determinació de les regressions auxiliars és major o igual que 0.8 aleshores tindrem indicis de multicol·linealitat elevada en el model.

Detecció i valoració de la multicol·linealitat

Anàlisi del R^2 de regressions successives

Anàlisi del R^2 de regressions on cada vegada s'elimini una variable explicativa

S'estimen successives regressions en les quals figuri com a variable endògena la mateixa que la del model original i on en cada una d'elles s'elimina una variable explicativa.

Si el R^2 no es modifica gaire respecte a l'obtingut amb totes les variables explicatives significa que la variable explicativa no inclosa és poc rellevant o bé que el que explica ja queda explicat per la resta de variables, és a dir, que hi ha multicol·linealitat elevada associada a aquesta variable.

Detecció i valoració de la multicol·linealitat

Anàlisi del R^2 de regressions successives

Per exemple, per un MRLM amb k=4:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i$$

trobaríem les tres regressions auxiliars següents:

$$y_i = \gamma_0 + \gamma_1 x_{3i} + \gamma_2 x_{4i} + v_i$$

$$y_i = \gamma_0 + \gamma_1 x_{2i} + \gamma_2 x_{4i} + v_i$$

$$y_i = \gamma_0 + \gamma_1 x_{2i} + \gamma_2 x_{3i} + v_i$$

Si algun dels coeficients de determinació d'aquestes regressions és molt semblant al del model original aleshores la variable que s'ha eliminat no és rellevant o bé el model original presenta problemes d'elevada multicol·linealitat.

Detecció i valoració de la multicol·linealitat

El Factor d'Increment de la Variància (FIV)

El Factor d'Increment de la Variància (FIV)

El Factor d'increment de la variància per la variable x_j (FIV_j) compara la variància real del paràmetre β_j amb la que s'obtindria en cas que hi hagués absència total de multicol·linealitat (ortogonalitat entre les variables explicatives del model).

En cas de regressors ortogonals: $\text{Var}(\hat{\beta}_j^*) = \dfrac{\sigma_u^2}{\sum \tilde{x}_j^2}$

En cas d'alguna correlació lineal entre les variables explicatives:

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma_u^2}{(1 - R_j^2) \sum \tilde{x}_j^2}$$

on R_j^2 és el de la regressió auxiliar de la variable x_j i la resta d'explicatives

Detecció i valoració de la multicol·linealitat

El Factor d'Increment de la Variància (FIV)

Per tant, el factor d'increment de la variància per la variable x_j (FIV_j) serà:

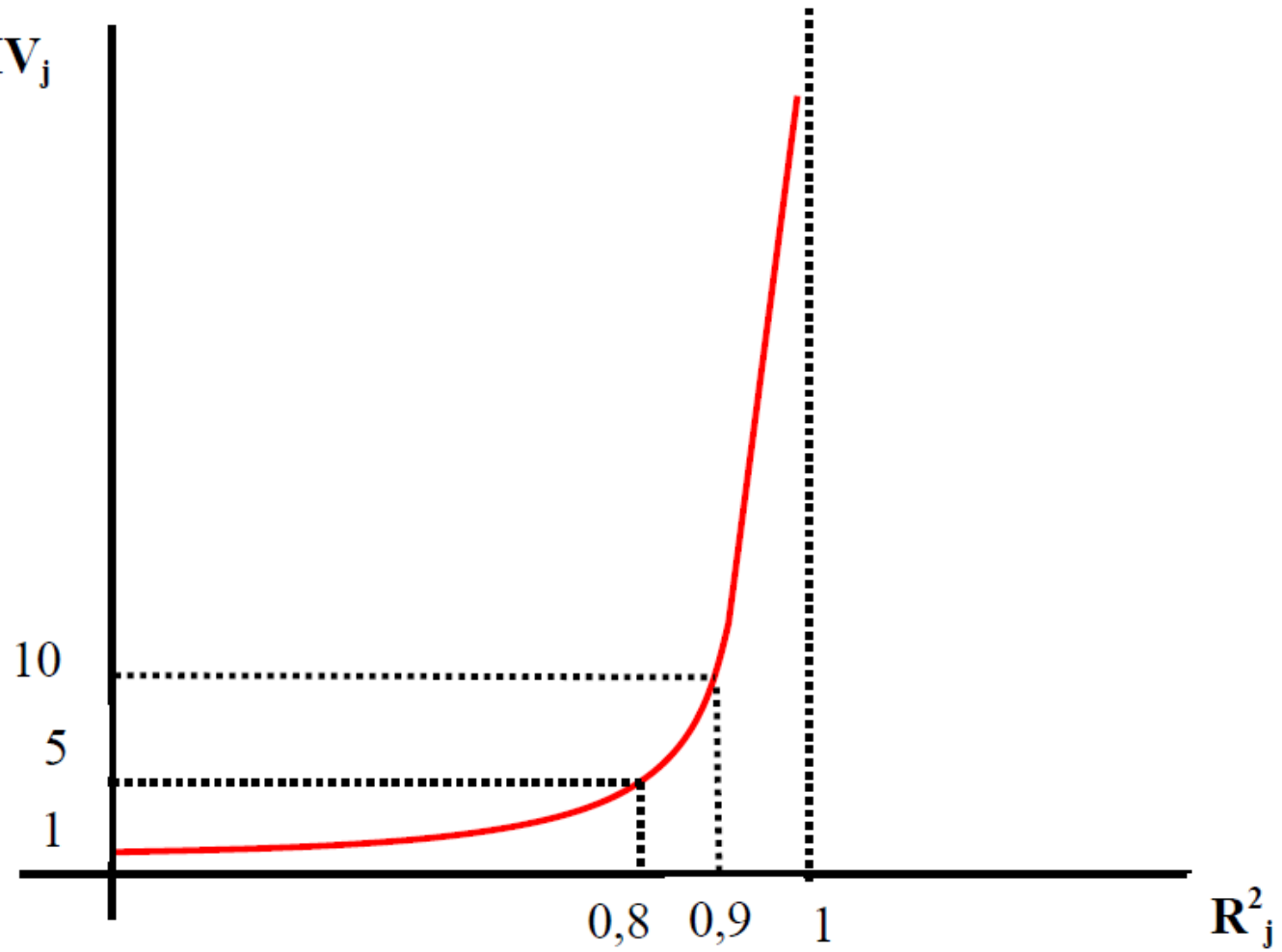
$$\text{FIV}_j = \frac{\text{Var}(\hat{\beta}_j)}{\text{Var}(\hat{\beta}_j^*)} = \frac{1}{(1 - R_j^2)}$$

i ens mostra l'augment que experimenta la variància de l'estimador pel fet de passar de l'ortogonalitat a una determinada correlació amb la resta de variables explicatives.

Sabent que $R_j^2 \in [0, 1]$aleshores $\text{FIV}_j \in [1, \infty)$.

Detecció i valoració de la multicol·linealitat

Gràficament, la relació entre FIV_j i R_j^2 és:



Detecció i valoració de la multicol·linealitat

D'aquesta manera:

$\text{FIV}_j > 5$ $(R_j^2 > 0.8)$ Problema rellevant de multicol·linealitat
$\text{FIV}_j > 10$ $(R_j^2 > 0.9)$ Problema greu de multicol·linealitat

Per últim, cal dir que els FIV es poden obtenir també a partir de la diagonal principal de la matriu inversa de correlacions:

$$(R_x)^{-1} = \begin{pmatrix} 1 & r_{23} & \dots & r_{2k} \\ r_{32} & 1 & \dots & r_{3k} \\ \vdots & \vdots & \dots & \vdots \\ r_{k2} & r_{k3}1 & \dots & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \text{FIV}_2 & \dots & \dots \\ \dots & \text{FIV}_3 & \dots \\ \vdots & \vdots & \vdots \\ \dots & \dots & \text{FIV}_k \end{pmatrix}$$

Possibles solucions

4. Possibles solucions

Caldrà corregir l'elevada multicol·linealitat si les seves conseqüències són importants per la finalitat per la qual hem construït el model.

Si és exclusivament predictiva podem mantenir el model tot i existir multicol·linealitat, donat que aquesta no impedeix obtenir un bon ajust (malgrat que els intervals de predicció seran grans).

Si és l'anàlisi estructural, caldrà cercar una solució doncs els estimadors són poc estables i imprecisos.

No hi ha cap solució que sigui satisfactòria al 100 % de manera general; dependrà de l'objecte d'anàlisi i de les dades.

Possibles solucions

Incorporar informació addicional

1) Incorporar informació addicional

Augmentar la grandària de la mostra, esperant que es redueixi el problema d'elevada multicol·linealitat.

Es tracta d'una solució no sempre aplicable degut als problemes amb la disponibilitat d'informació estadística. A més, és possible que la introducció de noves dades pugui generar problemes de canvi estructural.

Si les dades del model són temporals habitualment no serà possible disposar de més observacions.

Possibles solucions

Incorporar informació addicional

Incorporar informació extra-mostral provinent d'altres treballs o estudis.

Per exemple, donant algun valor inicial a algun paràmetre, o incorporant restriccions sobre alguns dels mateixos (a partir d'informació "a priori"), que puguin portar a reduir el nombre de paràmetres a estimar, simplificant el model, i reduint el grau de multicol·linealitat.

Possibles solucions

Reespecificació del model

2) Reespecificar el model

Transformar les variables del model prenent les seves diferències o bé dividint el model per una de les variables explicatives.

Diferenciar el model és una solució que té sentit únicament quan les dades són temporals però no si les dades són de tall transversal. Presenta, a més, el problema de que genera autocorrelació en el terme de pertorbació.

Per altra part, **dividir el model per una de les variables explicatives** redueix la multicol·linealitat, però fa que el terme de pertorbació no sigui homoscedàstic.

Possibles solucions

Reespecificació del model

Eliminar alguna/es variable/s del model.

Una de les solucions més habituals al problema de la multicol·linealitat passa per eliminar aquella variable o variables que estiguin més afectades pels problemes de l'elevada multicol·linealitat.

Malgrat això, no sempre és fàcil saber quina variable eliminar i, a més, podem incòrrer en un problema d'omissió de variables rellevants amb els subseqüents problemes que això comporta (estimadors esbiaixats).

Possibles solucions

Reespecificació del model

Atès que eliminar una variable explicativa és el mateix que estimar un model restringit on el paràmetre associat a la variable en qüestió és igual a zero, es determina que l'eliminació d'una variable dóna lloc a estimadors no sempre no esbiaixats però amb una variància inferior a la del model original.

D'aquesta manera, usem el criteri de l'EQM (que és igual al $[\text{biaix}^2 + \text{variància}]$) per decidir si és millor eliminar una variable explicativa o no.

En un model amb dues variables explicatives, x_2 i x_3, el quocient entre els EQM de les estimacions de β_2 no incloent x_3 i incloent-la són:

Possibles solucions

Reespecificació del model

$$\frac{\text{EQM}(\hat{\beta}_2^*)}{\text{EQM}(\hat{\beta}_2)} = 1 + R_{23}^2(t^2 - 1)$$

Així, si el valor del estadístic de la t-Student associat a l'estimació del paràmetre β_3 de la variable explicativa x_3 és inferior a 1 en valor absolut, serà millor en termes d'EQM eliminar la variable explicativa x_3 del model inicial.

$$t = \left| \frac{\hat{\beta}_3}{\text{e.s.}(\hat{\beta}_3)} \right| < 1$$

Per contra, si el valor d'aquest estadístic és superior a 1 en valor absolut, resultarà millor en termes d'EQM mantenir la variable x_j en el model.

Possibles solucions

Reespecificació del model

Fer una **anàlisi de Components Principals sobre les variables explicatives** i usar les components com a noves variables explicatives.

L'anàlisi de components principals és una tècnica que permet obtenir unes noves variables que són ortogonals entre elles. Si aquestes noves variables s'usen com a variables explicatives, el model presentarà absència total de multicol·linealitat.

A nivell teòric aquesta solució és òptima però en la pràctica els problemes que presenta impossibiliten la seva aplicació: la difícil interpretació econòmica de les components principals i la dificultat d'usar el model pels seus objectius (anàlisi estructural o predicció).

Possibles solucions

"Ridge Regression"

3) "Ridge Regression"

La regressió Ridge consisteix en introduir un biaix en l'estimació dels paràmetres per MQO però reduïnt la seva variància de forma que tinguin un menor EQM i siguin més estables.

En la "Ridge Regression" hem de treballar amb les variables estandarditzades. Així, l'expressió de l'estimador MQO seria:

$$\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}(\tilde{X}'\tilde{Y})$$

on $\tilde{X}$ i $\tilde{Y}$ representen la matriu de variables explicatives estandarditzades i el vector de la variable dependent estandarditzat.

Possibles solucions

"Ridge Regression"

Sabem que:

$$\mathrm{E}(\hat{\beta}) = \beta$$

i que

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(\tilde{X}'\tilde{X})^{-1} = \sigma^2 R_{\tilde{X}}^{-1} = R_{\tilde{X}}^{-1}$$

atès que $(\tilde{X}'\tilde{X}) = R_X$ (matriu de correlacions entre explicatives)

i que $\sigma^2 = 1$ donada l'estandardització de Y.

Per tant,

$$\mathrm{Var}(\hat{\beta}_j) = \frac{1}{1 - R_j^2}$$

és a dir, en aquesta formulació la variància coincideix amb el FIV.

Possibles solucions

"Ridge Regression"

Ara, la "Ridge Regression" afegeix un petit valor $k, (k > 0)$ als elements de la diagonal principal de la matriu de correlacions, de forma que el determinant de la matriu de correlacions sigui més gran (que no sigui tan pròxim a zero).

L'estimador Ridge és:

$$\tilde{\beta}_{\text{RIDGE}} = (\tilde{X}'\tilde{X} + kI)^{-1}(\tilde{X}'\tilde{Y}) = (R + kI)^{-1}(\tilde{X}'\tilde{Y})$$

Es pot demostrar que existeix un valor de k pel qual l'EQM (biaix^2 + Var) de l'estimador Ridge és menor que el de l'estimador MQO.

La determinació del valor de k es farà per mètodes gràfics o numèrics seleccionant el valor de k més petit que estabilitza els valors de les estimacions dels paràmetres.

Econometria
Tema 4(2): Multicol·linealitat

Ramon Alemany

Grau Estadística UB-UPC

Curs 2017-18