Econometria
Tema 4(3): Observacions atípiques i influents

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

Definició d'observacions atípiques i influents

1.Definició d'observacions atípiques i influents

Sempre és convenient revisar la mostra, no fos cas que existeixi alguna observació "estranya", atípica o massa allunyada de la resta d'observacions, que faci difícil pensar que hagi estat generada pel mateix procés que la resta.

La inclusió d'una observació "estranya" pot arribar a afectar de forma notable els nostres resultats ja que, entre d'altres, poden:

- afectar els coeficients estimats
- a l'ajust del model
- a la inferència

Detecció d'observacions potencialment influents

2. Detecció d'observacions potencialment influents

Direm que una observació presenta palanquejament o que té influència potencial, si és "força" diferent de les altres observacions en termes del valor de les variables explicatives.

El **leverage** o **lever** (h_{ii}) serveix per conèixer el grau de palanquejament d'una observació i-éssima.

El lever h_{ii} s'obté com l'element i-éssim de la diagonal principal de la matriu H ("hat matrix"):

$$H = X(X'X)^{-1}X'$$

Detecció d'observacions potencialment influents

on $H = X(X'X)^{-1}X'$ és:

- una matriu quadrada de dimensió $(N \times N)$ i simètrica
- idempotent $(HH = H)$. Donat que és simètrica es compleix també que $HH' = H'H = H'H' = H$
- traça $H = k$

Es diu "Hat matrix" perquè:

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

A més recordem que:

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

$$= \left[I - X(X'X)^{-1}X'\right] Y = MY$$

$$M = I - H$$

Detecció d'observacions potencialment influents

També podem veure que:

$$\text{Var}(\hat{Y}) = \text{E}\left[(\hat{Y} - \text{E}(\hat{Y}))(\hat{Y} - \text{E}(\hat{Y}))'\right] = \text{E}\left[(X\hat{\beta} - X\beta)(X\hat{\beta} - X\beta)'\right] =$$

$$= \text{E}\left[(X(\hat{\beta} - \beta)(\hat{\beta} - \beta)'X'\right] = (X\text{E}\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right]X' =$$

$$= X\sigma_u^2(X'X)^{-1}X' = \sigma_u^2 X(X'X)^{-1}X' = \sigma_u^2 H$$

D'aquesta manera, es pot concloure que quan més gran sigui el valor del leverage, menys precisa serà l'estimació del valor ajustat de la variable endògena...

Detecció d'observacions potencialment influents

Pel que fa a les propietats del **leverage**, cal dir que és possible comprovar com h_{ii} es pot calcular també a partir de la següent expressió:

$$h_{ii} = \frac{1}{N}[1 + d_i] = \frac{1}{N}\left[1 + (x_{ji} - \bar{x}_j)\mathrm{S}^{-1}(x_{ji} - \bar{x}_j)'\right]$$

- d_i és la **distància de Mahalanobis** (que mesura la distancia d'una observació en relació a una població o conjunt d'observacions)
- $(x_{ji} - \bar{x}_j)$ és un vector fila de dimensió $(1 \times k)$ que conté els valors centrats de l'observació i-éssima respecte de cadascuna de les k variables explicatives
- S és la matriu de variàncies i covariàncies mostrals entre les variables explicatives: $\mathrm{S} = (1/N)X'X$

Detecció d'observacions potencialment influents

De l'expressió anterior podem observar com, a mesura que augmenta d_i, augmenta també el leverage h_{ii}, i que:

$$\frac{1}{N} \leq h_{ii} \leq 1$$

$h_{ii} \geq 2\bar{h}$ l'observació i-éssima presenta palanquejament o influència potencial

$h_{ii} < 2\bar{h}$ l'observació i-éssima NO presenta palanquejament ni influència potencial

amb $\bar{h} = \frac{k}{N}$.

Detecció d'observacions atípiques

3. Detecció d'observacions atípiques

Els **outliers** o observacions atípiques són aquelles observacions que presenten un comportament molt diferent a la resta. Són estranyes en el sentit que el seu procés generador és diferent a la de la resta d'observacions.

Instruments per a la seva detecció:

a) Residu
b) Residu estandarditzat
c) Residu estudentitzat
d) Residu estudentitzat amb omissió

Detecció d'observacions atípiques

Residu associat a cada observació

a) Residu associat a cada observació.

Si una observació és atípica és lògic pensar que tindrà un residu en l'ajust superior al de la resta. Però si s'utilitzen els residus per a detectar outliers hi ha el problema que depenen de les unitats de mesura.

No obstant això, es pot aplicar el criteri de considerar com a outlier aquella observació que té un residu més gran o igual que el doble de la mitjana de tots els residus.

Detecció d'observacions atípiques

Residu estandarditzat

b) Residu estandarditzat:

$$\frac{e_i}{\hat{\sigma}_u} = \frac{e_i}{\sqrt{\frac{e'e}{N-k}}}$$

Utilitzar el residu estandarditzat soluciona els problemes de mesura però tampoc és adient del tot ja que si l'observació té major influència que la resta aleshores l'ajust estarà influït per ella i també l'estimació de la variància del terme de pertorbació.

Detecció d'observacions atípiques

Residu estudentitzat

c) Residu estudentitzat:

Estandarditzem el residu per la seva desviació estàndard en lloc de per la del terme de pertorbació.

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}_u^2(1-h_{ii})}} = \frac{e_i}{\sqrt{\frac{e'e}{N-k}(1-h_{ii})}}$$

on al numerador trobem el valor del residu e_i de l'estimació associat a l'observació i-éssima i al denominador la desviació estàndard d'e_i.

Detecció d'observacions atípiques

Residu estudentitzat amb omissió

d) Residu estudentitzat amb omissió:

Estandarditzem el residu per la seva desviació estàndard en lloc de per la del terme de pertorbació, però sense tenir en compte el residu en el seu càlcul.

$$r_i^* = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}(1-h_{ii})}} = \frac{e_i}{\sqrt{\frac{e'_{(i)}e_{(i)}}{N-k-1}(1-h_{ii})}}$$

on al numerador trobem el valor del residu e_i de l'estimació associat a l'observació i-éssima i al denominador la desviació estàndard del terme de pertorbació estimada sense el residu de l'observació i-éssima.

Detecció d'observacions atípiques

Els residus estudentitzats es distribueixen segons una t d'Student amb $N-k$ graus de llibertat:

$r_i \geq t_{N-k;\alpha/2}$ l'observació i-éssima es pot considerar **outlier**

$r_i < t_{N-k;\alpha/2}$ l'observació i-éssima NO es pot considerar **outlier**

En el cas de calcular els residus estudentitzats amb omissió aleshores la t d'Student tindrà $N-k-1$ graus de llibertat.

Detecció d'observacions amb influència real

4. Detecció d'observacions amb influència real

Una observació té influència real quan té un major efecte en l'ajust que la resta d'observacions.

La **Distància de Cook** (DC_i) i el **DFFITS** serveixen per detectar si l'observació i-éssima presenta influència real sobre l'ajust.

La influència real suposa que una observació:

- pot provocar canvis en la recta d'ajust, és a dir, en els valors dels paràmetres estimats.
- pot indicar la no verificació d'alguna de les hipòtesis bàsiques (linealitat, normalitat i homoscedasticitat del terme de pertorbació).
- té capacitat per fer que una variable sigui no significativa
- pot deteriorar la capacitat predictiva del model

Detecció d'observacions amb influència real

Distància de Cook

La Distancia de Cook es calcula fent:

$$DC_i = \frac{(\hat{Y} - \hat{Y}_{(i)})'(\hat{Y} - \hat{Y}_{(i)})/k}{e'e/(N-k)} = \frac{(\hat{\beta} - \hat{\beta}_{(i)})'X'X(\hat{\beta} - \hat{\beta}_{(i)})/k}{e'e/(N-k)}$$

- $\hat{Y}_{(i)}$ és l'estimació de la variable endògena sense incloure l'observació i-éssima
- $\hat{\beta}_{(i)}$ és l'estimació dels paràmetres sense incloure l'observació i-éssima
- $e'e$ és la SQE del model estimat per MQO amb totes les observacions

Detecció d'observacions amb influència real

Distància de Cook

La Distància de Cook es pot obtenir també com:

$$DC_i = \frac{r_i^2}{k} \frac{h_{ii}}{1 - h_{ii}}$$

on r_i és el residu estudentitzat de la i-éssima observació i h_{ii} el seu "leverage".

D'aquesta manera, s'observa com la distància de Cook mesura la influència real d'una observació valorant tant el seu grau de palanquejament com la concordança entre el seu procés generador de dades i el de la resta d'observacions.

Detecció d'observacions amb influència real

Distància de Cook

La Distància de Cook es distribueix segons una F de Snedecor amb k graus de llibertat en el numerador i $N-k$ graus de llibertat en el denominador.

Així doncs,

$\text{DC}_i \geq F_{k,N-k;\alpha}$ l'observació i-éssima té influència real en l'ajust del model

$\text{DC}_i < F_{k,N-k;\alpha}$ l'observació i-éssima NO té influència real en l'ajust del model

Per mostres grans direm que una observació te una influència substancial en les estimacions si $\text{DC}_i > 1$ o bé si $\text{DC}_i > \frac{4}{N-k}$.

Detecció d'observacions amb influència real

DFFITS

El DFFITS és una altra mesura que pot indicar si l'observació i-éssima presenta influència real sobre l'ajust.

$$\text{DFFITS}_i = \frac{\hat{Y} - \hat{Y}_{(i)}}{\sqrt{\frac{e'_{(i)} e_{(i)}}{(N-k-1)} h_{ii}}}$$

$\text{DFFITS}_i \geq 2\sqrt{\frac{k}{N}}$ l'observació i-éssima té influència real en l'ajust del model

$\text{DFFITS}_i < 2\sqrt{\frac{k}{N}}$ l'observació i-éssima NO té influència real en l'ajust del model

Econometria
Tema 4(3): Observacions atípiques i influents

Ramon Alemany

Grau Estadística UB-UPC

Curs 2017-18