Lezione 14 (a cura di Ludovica Peccia)

MULTICOLLINEARITA'

La multicollinearità sorge quando c'è un'elevata correlazione tra due o più variabili esplicative.

In un modello di regressione $Y= X_1, X_2, X_3$

se X_2 è una trasformazione lineare di X_1 e quindi sussiste una relazione del tipo

$X_2 = a + bX_1$, le due variabili sono perfettamente correlate.

In un caso del genere $\hat{\beta}$**=** $(X' X)^{-1} X'$**Y** , che è lo stimatore dei minimi quadrati, non dà risultati attendibili. Infatti se c'è una correlazione perfetta tra le due variabili, la matrice (X'X) diventa singolare $\Rightarrow$ ha determinante uguale a 0 e perciò non esiste la matrice inversa.

I problemi sorgono anche se c'è una forte correlazione: in questo caso il determinante della matrice sarà molto piccolo e le stime di $\hat{\beta}$ non saranno attendibili.

Se le variabili sono fortemente correlate vuol dire che danno la stessa informazione e il modello di regressione non riesce più ad attribuire un significato a ciascuna di esse.

Ad esempio : Y = spesa
X_1 = reddito lordo
X_2= reddito netto
Con X1 e X2 fortemente correlate la regressione di Y su X_1 e X_2 darebbe risultati strani.

Esistono degli indici che evidenziano questo problema di multicollinearità:

$VIF_j = \frac{1}{1 \quad r^2 j}$, con j=1,…,k

$\mathbf{r^2}$ è l'indice di determinazione di X_j contro le altre covariate X_1, …, X_{j-1}, X_{j+1}, X_k

Se $\mathbf{r^2}$ è molto elevato la covariata sarà molto correlata alle altre.

Se ad esempio ho 3 covariate devo calcolare 3 indici:

$VIF_1 = \frac{1}{1-r_1^2}$ dove r^2_1= indice di determinazione di X_1 contro X_2 e X_3

$VIF_2 = \frac{1}{1-r_2^2}$ dove r^2_2= indice di determinazione di X_2 contro X_1 e X_3

$VIF_3 = \frac{1}{1-r_3^2}$ dove r^2_3= indice di determinazione di X_3 contro X_1 e X_2

Se dall'esempio risulta $VIF_2 \geq 5$, posso eliminare la prima covariata (reddito lordo) e considerare solo il reddito netto in quanto più significativo per il modello.

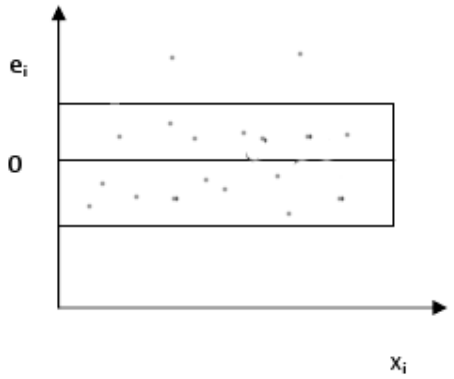
Spesso viene calcolato $\overline{VIF} = \sum_j \frac{VIFj}{k}$

Se $\overline{VIF}$>5 il problema è nel complesso rilevante.

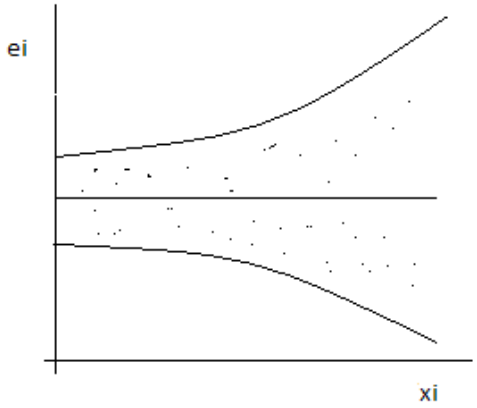
RIMEDI RISPETTO A VIOLAZIONI DELLE ASSUNZIONI DEL MODELLO

Abbiamo detto che la diagnostica è importante perché attraverso i grafici riusciamo a capire se le assunzioni sono verificate. I problemi più frequenti sono:

- OUTLIER: verifichiamo da cosa è dovuta l'osservazione anomala.



  Se le osservazioni anomale sono < 3 possiamo eliminarle
  Se le osservazioni anomale sono > 3 c'è qualche errore nel modello
  Una volta eliminato l'outlier vediamo che le stime dei parametri cambiano fortemente
- ETEROSCHEDASTICITA', ossia varianza dei termini di errore non costante.
  In questo caso l'assunzione $V(\varepsilon_i)=\sigma^2$, $i=1,\ldots,n$ non è valida e ce lo dimostra il grafico dei residui che assumerà una forma a megafono.



Per risolvere questo problema dobbiamo riformulare il modello di regressione. Assumendo che :

$V(\varepsilon_i)=\sigma^2 i$, i=1,…,n cioè che la varianza sia diversa per ogni i

nel modello di regressione $\mathbf{Y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$

$$\rightarrow \varepsilon \sim N(\mathbf{0}, \sigma^2\Omega)$$

Ω è una matrice diagonale con elementi diversi da 1 che generalizza la matrice identità. La covarianza è sempre uguale a 0.

$$\Omega=\begin{pmatrix} d1 & \cdots & 0 \\ \vdots & d2 & \vdots \\ 0 & \cdots & dn \end{pmatrix}$$

Vediamo ora il comportamento dello stimatore $\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{Y}$

$\rightarrow$ $E(\widehat{\boldsymbol{\beta}})= (X'X)^{-1}X'E(\mathbf{Y}) = \cancel{(X'X)^{-1}}\cancel{X'X}\boldsymbol{\beta}= \boldsymbol{\beta}$

Anche se c'è eteroschedasticità lo stimatore non è distorto perché non viene utilizzata la matrice varianza-covarianza.

$\rightarrow$ $V(\widehat{\boldsymbol{\beta}})= (X'X)^{-1} X'V(\mathbf{Y}) X(X'X)^{-1}$

$V(\mathbf{Y}) = \sigma^2\Omega$

$V(\hat{\boldsymbol{\beta}}) = \sigma^2(X'X)^{-1} X'\Omega X(X'X)^{-1}$ La varianza viene modificata

→ $\sigma^2(X'X)^{-1}$ OLS non è lo stimatore più preciso (efficiente) ed è sbagliato usarlo per errori standard.

Useremo perciò uno stimatore alternativo GLS (stimatore dei minimi quadrati generalizzato).

$$\boldsymbol{\beta}^* = (X'WX)^{-1} X'W\mathbf{Y}$$

dove $W=\Omega^{-1}$

W è la matrice dei pesi, è diagonale ed è l'inversa di Ω.

$$W=\begin{pmatrix} 1/d1 & \dots & 0 \\ \vdots & 1/d2 & \vdots \\ 0 & \dots & 1/dn \end{pmatrix}$$

Si si dà più rilievo alle variabili Y che hanno maggior peso.

→ $E(\hat{\boldsymbol{\beta}}^*) = (X'WX)^{-1} X'WX\boldsymbol{\beta} = \boldsymbol{\beta}$

→ $V(\hat{\boldsymbol{\beta}}^*) = \sigma^2(X'WX)^{-1}$ ~~$X'W$~~ ~~$(\Omega W) X (X'WX)^{-1}$~~

⇒ $V(\hat{\boldsymbol{\beta}}^*) = \sigma^2(X'WX)^{-1}$ che è uno stimatore non distorto

Sulla base di questa espressione otteniamo errori standard e intervalli di confidenza corretti

$\hat{V}(\hat{\boldsymbol{\beta}}^*) = S^{2^*}(X'WX)^{-1}$

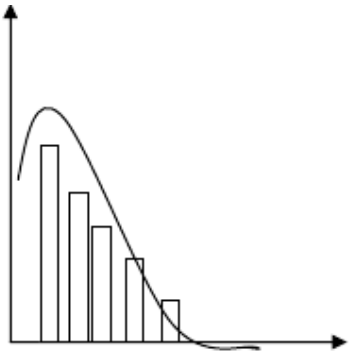
Dove $S^{2^*} = \sum_i \frac{ei^2 wt}{n-(k+1)}$, $w_{i=}1/d_i$

$$= \frac{\boldsymbol{e'We}}{n-(k+1)}$$

Una volta ottenuta la matrice, in diagonale troveremo lo standard error:

→ $s.e\,(\hat{\boldsymbol{\beta}}_j^*) = diag_j\, \hat{V}(\hat{\boldsymbol{\beta}}^*)$

- ε_i NON HA DISTRIBUZIONE NORMALE $\varepsilon_i \sim N(0, \sigma^2)$
  L'istogramma e il qq-plot ci mostravano se l'assunzione di normalità era verificata.
  Se l'assunzione di normalità non è verificata avremo una asimmetria positiva.



Il rimedio tipico, soprattutto se Y è sempre positiva, è la trasformazione logaritmica:

$$\log(Y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

questa trasformazione corregge il problema della normalità.
Per interpretare le stime dei parametri devo riportare la Y sulla scala originaria, ossia:

$$Y_i = e^{\beta 0 + \beta 1 xi1 + \ldots + \beta kxik + \varepsilon i}$$

Mi accorgo che si tratta di un problema moltiplicativo, cioè di quanto si moltiplica la Y se la x aumenta di 1, cioè:

$$Y_i = e^{\beta 0} \times e^{\beta 1 xi1} \times \ldots \times e^{\beta kxik} + e^{\varepsilon i}$$

BOX-COX sono trasformazioni per correggere il problema di non normalità dei termini di errore

$$(y*) = \begin{cases} \log(y), & \lambda = 0 \\ \dfrac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \end{cases}$$

Il caso $\boldsymbol{\lambda}$=0 è quello della asimmetria positiva.