Econometria | 2023/2024

Lezione 10: Nonlinearità - Interazioni

Giuseppe Ragusa

https://gragusa.org

Roma, Marzo 2024

Interazioni fra variabili

Spesso è interessante capire come l'effetto su Y di una variabile indipendente dipende dal valore di un'altra variabile.

Domanda:

è possibile che scuole con molti studenti che apprendono l'inglese `Elpct` traggono più vantaggi da una riduzione delle dimensioni delle classi rispetto a quelle con un numero inferiore di studenti non-madrelingua?

Per rispondere si può usare una regressione multipla con un termine di interazione.

Interazioni

Prendiamo in considerazione tre casi:

1. Interazioni tra due variabili binarie.
2. Interazioni tra una variabile binaria e una continua.
3. Interazioni tra due variabili continue.

Interazioni tra due variabili binarie

Prendiamo due variabili binarie D_1 e D_2 e il modello di regressione della popolazione:

$$Y_i = \beta_0 + \beta_1 \times D_{1i} + \beta_2 \times D_{2i} + u_i,$$

dove

- $Y_i = ln(salari_i)$,
- $D_{1i} =$ {1 se la persona i ha una laurea, 0 altrimenti}.
- $D_{2i} =$ {1 se la persona i è donna, 0 se la persona i-esima è uomo}.

Sappiamo che β_1 misura la differenza media tra individui con e senza laurea e β_2 è la differenza, ceteris paribus, di genere in $log(salari)$.

Questo modello non ci permette di determinare se esiste un effetto specifico di genere nel possedere una laurea e, in caso affermativo, quanto sia forte questo effetto.

Interazioni tra due variabili binarie, ctd.

Consideriamo invece il seguente modello:

$$Y_i = \beta_0 + \beta_1 \times D_{1i} + \beta_2 \times D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$$

$D_{1i} \times D_{2i}$ è il termine di interazione e β_3 misura la differenza nell'effetto (sui salari) di avere una laurea tra donne e uomini

Per interpretare i coefficienti, calcoliamo i valori attesi di Y per ogni possibile combinazione delle variabili binarie:

- $D_1 = 0 \text{ e } D_2 = 0: \quad E(Y|D_1 = 0, D_2 = 0) = \beta_0$
- $D_1 = 1 \text{ e } D_2 = 0: \quad E(Y|D_1 = 1, D_2 = 0) = \beta_0 + \beta_1$
- $D_1 = 0 \text{ e } D_2 = 1: \quad E(Y|D_1 = 0, D_2 = 1) = \beta_0 + \beta_2$
- $D_1 = 1 \text{ e } D_2 = 1: \quad E(Y|D_1 = 1, D_2 = 1) = \beta_0 + \beta_1 + \beta_2 + beta_3$

Ora possiamo confrontare questi valori attesi e interpretare i coefficienti.

Interazioni tra due variabili binarie, ctd.

- β_1 è la differenza tra i valori attesi di Y quando D_1 passa da 0 a 1, mentre $D_2 = 0$:

$$E(Y|D_1 = 1, D_2 = 0) - E(Y|D_1 = 0, D_2 = 0) = \beta_1$$

- β_2 è la differenza tra i valori attesi di Y quando D_2 passa da 0 a 1, mentre $D_1 = 0$:

$$E(Y|D_1 = 0, D_2 = 1) - E(Y|D_1 = 0, D_2 = 0) = \beta_2$$

- $\beta_1 + \beta_3$ è la differenza tra i valori attesi di Y quando D_1 passa da 0 a 1, mentre $D_2 = 1$:

$$E(Y|D_1 = 1, D_2 = 1) - E(Y_i|D_1 = 0, D_2 = 1) = \beta_1 + \beta_3$$

- $\beta_2 + \beta_3$ è la differenza tra i valori attesi di Y quando D_2 passa da 0 a 1, mentre $D_1 = 1$:

$$E(Y|D_1 = 1, D_2 = 1) - E(Y_i|D_1 = 1, D_2 = 0) = \beta_2 + \beta_3$$

Esempio: HiSTR e HiEL

$$HiSTR = \begin{cases} 1, & \text{se } STR \geq 20 \\ 0, & \text{altrimenti}, \end{cases}, \quad HiEL = \begin{cases} 1, & \text{se } PctEL \geq 10 \\ 0, & \text{altrimenti}. \end{cases}$$

Possiamo costruire le variabili sopra utilizzando R nel seguente modo:

```r
library(Ecdat)
library(dplyr)
data(Caschool)
# Aggiungiamo HiSTR e HiEl CASchool
Caschool <- Caschool |> mutate(HiSTR = ifelse(str >= 20, 1, 0),
                               HiEL  = ifelse(elpct >= 10, 1, 0))
```

Procediamo quindi stimando il modello

$$TestScore = \beta_0 + \beta_1 \times HiSTR + \beta_2 \times HiEL + \beta_3 \times (HiSTR \times HiEL) + u_i.$$

Esistono diversi modi per aggiungere il termine di interazione alla formula quando si utilizza `lm()` o `feols()` — ma il modo più intuitivo è utilizzare `HiEL*HiSTR`.[1]

1. Aggiungendo `HiEL*HiSTR` alla formula, verranno aggiunti `HiEL`, `HiSTR` e la loro interazione come regressori, mentre `HiEL:HiSTR` aggiunge solo il termine di interazione.

Esempio: HiSTR e HiEL, ctd

$$TestScore = \beta_0 + \beta_1 \times HiSTR + \beta_2 \times HiEL + \beta_3 \times (HiSTR \times HiEL) + u_i.$$

```r
library(fixest)
# stimiamo il modello con un termine di interazione binaria
bi_model <- feols(testscr ~ HiSTR * HiEL, data = Caschool, vcov = "hetero")
summary(bi_model)
```

```
OLS estimation, Dep. Var.: testscr
Observations: 420
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error    t value   Pr(>|t|)
(Intercept) 664.14329    1.38809 478.458830  < 2.2e-16 ***
HiSTR        -1.90784    1.93221  -0.987386 3.2403e-01
HiEL        -18.16295    2.34595  -7.742249 7.5024e-14 ***
HiSTR:HiEL   -3.49434    3.12123  -1.119539 2.6356e-01
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 16.0   Adj. R2: 0.290475
```

Esempio: `HiSTR` e `HiEL`, ctd.

Il modello di regressione stimato è

$$\widehat{TestScore} = \underset{(1.39)}{664.1} - \underset{(1.93)}{1.9} \times HiSTR - \underset{(2.33)}{18.3} \times HiEL - \underset{(3.12)}{3.3} \times (HiSTR \times HiEL)$$

L'effetto del passaggio da un distretto scolastico con un basso rapporto studenti-insegnanti ($HiStr = 0$) a uno con un alto rapporto studenti-insegnanti ($HiStr = 1$), a seconda dell'alto o basso percentuale di studenti che parlano inglese, sia di $-1.9 - 3.3 \times HiEL$.

- per i distretti con una bassa percentuale di studenti che parlano inglese ($HiEL = 0$), l'effetto previsto è una diminuzione di 1.9 punti
- per i distretti con una alta percentuale di studenti che parlano inglese ($HiEL = 1$), l'effetto previsto è una diminuzione di $1.9 + 3.3 = 5.2$ punti

Interazioni tra variabile binaria e continua

Consideriamo adesso il caso in cui X_i è il numero di anni di esperienza lavorativa della persona i, una variabile continua. Abbiamo

$$\begin{aligned} Y_i &=, \ln(salari_i), \\ X_i &= \text{esperienza lavorativa della persona } i, \end{aligned}$$

$$D_i = \begin{cases} 1, & \text{se la persona } i\text{-esima ha una laurea universitaria} \\ 0, & \text{altrimenti.} \end{cases}$$

Il modello è quindi

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i,$$

Questo modello ci consente di stimare il beneficio medio di avere una laurea universitaria (tenendo costante l'esperienza lavorativa) e l'effetto medio dell'esperienza lavorativa sui salari (tenendo costante la laurea universitaria)

Interazioni tra variabile binaria e continua, ctd.

Aggiungendo il termine di interazione $X_i \times D_i$, consentiamo all'effetto di un anno aggiuntivo di esperienza lavorativa di differire tra individui con e senza laurea universitaria,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i.$$

Qui, β_3 rappresenta la differenza attesa nell'effetto di un anno aggiuntivo di esperienza lavorativa tra laureati e non laureati.

Un'altra possibile specificazione è

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i.$$

Secondo questo modello l'impatto atteso di un anno aggiuntivo di esperienza lavorativa sui guadagni differisce tra laureati e non laureati ma la laurea di per sé non aumenta i guadagni.

Interazioni tra variabile binaria e continua

Un termine di interazione come $X_i \times D_i$ (dove X_i è continuo e D_i binario) consente alla pendenza di dipendere dalla variabile binaria D_i.

Ci sono tre possibilità:

1. Intercetta diversa e stessa pendenza:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$$

2. Intercetta diversa e pendenza diversa:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 \times (X_i \times D_i) + u_i$$

3. Stessa Intercetta e pendenza diversa:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i$$

Esempio: `str` e `HiEL`

Possiamo rispondere alla domanda se l'effetto sulla punteggio dei test di ridurre il rapporto studenti-insegnanti dipenda dal fatto che ci siano molti o pochi studenti non-madrelingua usando `str` invcece di `HiSTR`.

Stimiamo il modello di regressione

$$\widehat{TestScore}_i = \beta_0 + \beta_1 \times str_i + \beta_2 \times HiEL_i + \beta_2(str_i \times HiEL_i) + u_i.$$

```r
bci_model <- feols(testscr ~ str + HiEL + str * HiEL, data = Caschool, vcov = "hetero")
summary(bci_model)
```

```
OLS estimation, Dep. Var.: testscr
Observations: 420
Standard-errors: Heteroskedasticity-robust
               Estimate Std. Error   t value  Pr(>|t|)
(Intercept) 682.245839  11.867814 57.487067 < 2.2e-16 ***
str          -0.968460   0.589102 -1.643961   0.10094
HiEL          5.639141  19.514556  0.288971   0.77275
str:HiEL     -1.276613   0.966919 -1.320289   0.18746
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 15.8   Adj. R2: 0.305368
```

Esempio: `str` e `HiEL`, ctd.

The estimated regression model is

$$\widehat{TestScore} = \underset{(11.87)}{682.2} - \underset{(0.59)}{0.97} \times size + \underset{(19.51)}{5.6} \times HiEL - \underset{(0.97)}{1.28} \times (size \times HiEL).$$

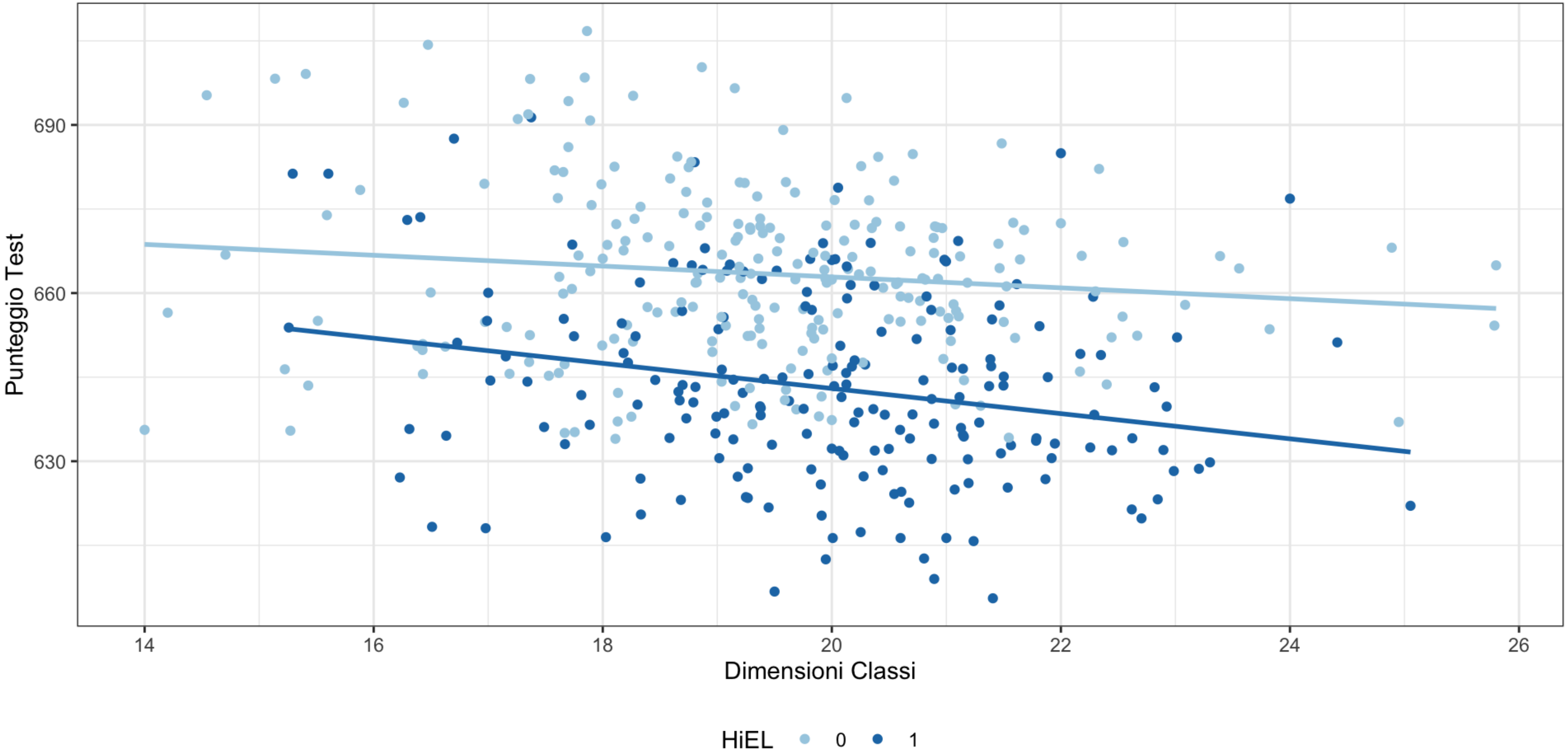
La retta di regressione stimata per i distretti con una bassa frazione di studenti non-madrelingua ($HiEL_i = 0$) è:

$$\widehat{TestScore} = 682.2 - 0.97 \times str_i.$$

Per i distretti con una grande frazione di studenti non-madrelingua abbiamo invece:

$$\begin{aligned}\widehat{TestScore} &= 682.2 + 5.6 - 0.97 \times str_i - 1.28 \times str_i \\ &= 687.8 - 2.25 \times str_i.\end{aligned}$$

Esempio: `str` e `HiEL`, ctd.



Interazioni tra due variabili continue

Consideriamo adesso un modello di regressione con Y come logaritmo dei guadagni e due regressori continui X_1, gli anni di esperienza lavorativa, e X_2, gli anni di istruzione.

Vogliamo stimare l'effetto sui salari di un anno aggiuntivo di esperienza lavorativa a seconda di un determinato livello di istruzione. Questo effetto può essere valutato includendo il termine di interazione $(X_{1i} \times X_{2i})$ nel modello:

$$\Delta Y_i = \beta_0 + \beta_1 \times X_{1i} + \beta_2 \times X_{2i} + \beta_3 \times (X_{1i} \times X_{2i}) + u_i$$

L'effetto su Y di una variazione di X_1 tenendo costante X_2 is

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2.$$

Un valore positivo di β_3 implica che l'effetto sul logaritmo dei salari di un anno aggiuntivo di esperienza lavorativa cresce linearmente con gli anni di istruzione.

Interazioni tra due variabili continue, ctd.

Abbiamo anche che

$$\frac{\Delta Y}{\Delta X_2} = \beta_2 + \beta_3 X_1$$

come l'effetto su logaritmo dei salari di un anno aggiuntivo di istruzione mantenendo costante l'esperienza lavorativa.

Complessivamente, troviamo che β_3 misura l'effetto di un aumento unitario in X_1 e X_2 oltre gli effetti di aumentare X_1 da solo e X_2 da solo di una unità.

La variazione complessiva in Y è quindi

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2, \quad \frac{\Delta Y}{\Delta X_2} = \beta_2 + \beta_3 X_1, \quad \frac{\Delta Y}{\Delta X_1 \Delta X_2} = \beta_3.$$

Esempio: `str` e `elpct`

Possiamo rispondere alla domanda se l'effetto sulla punteggio dei test di una diminuzione del rapporto studenti-insegnanti dipenda dal numero di studenti che apprendono l'inglese utilizzando le variabili continue `str` e `elpct`.

Stimiamo il modello

$$\widehat{TestScore}_i = \beta_0 + \beta_1 \times str_i + \beta_2 \times elpct_i + \beta_3(str_i \times elpct_i) + u_i.$$

```r
# Stimiamo il modello
bci_model <- feols(testscr ~ str + elpct + str * elpct, data = Caschool, vcov = "hetero")
summary(bci_model)
```

```
OLS estimation, Dep. Var.: testscr
Observations: 420
Standard-errors: Heteroskedasticity-robust
              Estimate Std. Error   t value  Pr(>|t|)
(Intercept) 686.338525  11.759345 58.365370 < 2.2e-16 ***
str          -1.117018   0.587514 -1.901264  0.057958 .
elpct        -0.672911   0.374123 -1.798636  0.072801 .
str:elpct     0.001162   0.018536  0.062676  0.950054
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 14.4   Adj. R2: 0.422299
```

$$\widehat{TestScore} = \underset{(11.76)}{686.3} - \underset{(0.59)}{1.12} \times str - \underset{(0.37)}{0.67} \times elpct + \underset{(0.02)}{0.0012} \times (str \times elpct).$$

Per l'interpretazione, consideriamo i quartili di `elpct`

```r
summary(Caschool$elpct)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.000   1.941   8.778  15.768  22.970  85.540
```

- Se `elpct` è al suo valore mediano di 8.78, stimiamo una pendenza di $-1.12 + 0.0012 \times 8.78 = -1.11$ che significa che aumentare il rapporto studenti-insegnanti di una unità deteriora i punteggi dei test di 1.11 punti.
- Se `elpct` è al 75 quantile, un aumento di una unità di str è $-1.12 + 0.0012 \times 23.0 = -1.09$, quindi la pendenza è leggermente inferiore. L'interpretazione: in un distretto scolastico con una quota di studenti non-madrelingua del 23, una riduzione di STR di una unità è associato ad una diminuzione dei punteggi dei test di circa 1.09 punti.

Tuttavia, la differenza dell'effetto per diversi valori di `elpct` non è statisticamente significativa. Infatti, non è possibile respingere l'$H_0 : \beta_3 = 0$.