Econometria | 2023/2024

Lezione 8: Inferenza nel modello multivariato

Giuseppe Ragusa

https://gragusa.org

Roma, marzo 2024

Sommario

- Verifica di ipotesi e intervalli di confidenza per un singolo coefficiente
- Verifica di ipotesi congiunte su più coefficienti
- Altri tipi di ipotesi che implicano più coefficienti

Verifica di ipotesi e CI per un singolo β

- Per verifica di ipotesi e intervalli di confidenza nella regressione multipla si segue la stessa logica utilizzata per la pendenza in un modello a singolo regressore.
- La distribuzione di

$$\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\underbrace{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}}_{=SE(\hat{\beta}_1)}} \text{ è approssimata da una} N(0,1)$$

- Perciò le ipotesi su β_1 possono essere verificate mediante la consueta statistica-t e gli intervalli di confidenza costruiti come

$$\{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\}$$

- Stesso discorso vale per $\hat{\beta}_2, \ldots, \hat{\beta}_K$.

Esempio: dati California

```r
library(Ecdat)
data("Caschool")

lm1 <- feols(testscr~str, data = Caschool, vcov = "hetero")
confint(lm1)
```

```
            2.5 % 97.5 %
(Intercept) 678.6 719.31
str          -3.3  -1.26
```

```r
lm2 <- feols(testscr~str+elpct, data = Caschool, vcov = "hetero")
confint(lm2)
```

```
              2.5 %  97.5 %
(Intercept) 668.875 703.189
str          -1.952  -0.250
elpct        -0.711  -0.589
```

Esempio: dati California

$$\widehat{testscr} = \underset{(10.461)}{698.933} - \underset{(0.524)}{2.280} \times str$$

$$\widehat{testscr} = \underset{(8.812)}{686.032} - \underset{(0.437)}{1.101} \times str - \underset{(0.031)}{0.650} \times elpct$$

- Il coefficiente di str\$ in (2) è l'effetto su $testscr$ dell'aumento di una unità in str, mantenendo costante la percentuale di studenti non di madrelingua nel distretto
- Il coefficiente di str si dimezza
- L'intervallo di confidenza al 95% per il coefficiente di str in (2) è

$$\{-1.101 \pm 1.96 \times 0.437\} = (-1.958, -0.244)$$

- Il test della statistica-t per $\beta_1 = 0$ è $t = -1.101/0.437 = -2.52$, perciò rifiutiamo l'ipotesi al livello di significatività del 5%

Verifica di ipotesi congiunte(Paragrafo 7.2)

Consideriamo il modello di regressione:

$$testscr_i = \beta_0 + \beta_1 str_i + \beta_2 expnstu_i + \beta_3 elpct_i + u_i$$

L'ipotesi nulla — le risorse scolastiche non contano — e l'alternativa sono:

$$H_0 : \beta_1 = 0 \text{ e } \beta_2 = 0 \text{ vs } H_1 : \textbf{almeno uno dei due coefficienti è diverso da zero}$$

- ipotesi congiunta: specifica un valore per due o più coefficienti, ossia impone una restrizione su due o più coefficienti.
- Un'ipotesi congiunta impone q restrizioni sui parametri. Nell'esempio precedente, $q = 2$ e le due restrizioni sono $\beta_1 = 0 \text{ e } \beta_2 = 0$.
- Un'idea di "buon senso" è quella di rifiutare se l'una o l'altra delle statistiche-t supera 1.96 in valore assoluto.
- Ma questa verifica "coefficiente per coefficiente" non è valida: il test ha un tasso di rifiuto troppo elevato sotto l'ipotesi nulla (più di α)!

Perché non possiamo verificare coefficiente per coefficiente?

- Proviamo a calcolare la probabilità di rifiutare in modo non corretto l'ipotesi nulla le due statistiche-t singole ($\alpha = 0.05$)
- Per semplificare, supponiamo che siano distribuite in modo indipendente (non è vero in generale). Siano t_1 e t_2 le statistiche-t:

$$t_1 = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \text{ e } t_2 = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)}$$

- La verifica "coeff. per coeff." è:

$$\text{rifiuta } H_0 = \beta_1 = \beta_2 = 0 \text{ se } |t_1| > 1.96 \text{ e/o } |t_2| > 1.96$$

- Qual è la probabilità che questa verifica "coeff. per coeff." rifiuti H_0, quando H_0 è effettivamente vera (errore di I tipo)? Dovrebbe essere 5%.

Supponiamo che t_1 e t_2 siano indipendenti (per questo esempio).

- La probabilità di rifiutare in modo non corretto l'ipotesi nulla mediante la verifica "coeff. per coeff." è:

$$
\begin{aligned}
&= Pr_{H_0}(|t_1| > 1.96 \text{ e/o } |t_2| > 1.96) \\
&= 1 - Pr_{H_0}(|t_1| \leq 1.96 \textbf{ e } |t_2| \leq 1.96) \\
&= 1 - Pr_{H_0}(|t_1| \leq 1.96) \times Pr_{H_0}(|t_2| \leq 1.96) \\
&= 1 - (0.95)^2 \\
&= 0.0975 = 9.75\%
\end{aligned}
$$

- Quindi la probabilità di commettere l'errore di tipo I sarebbe 9.75% che non è il 5% desiderato!

La dimensione di una verifica è l'effettivo tasso di rifiuto sotto l'ipotesi nulla.

- La dimensione della verifica del "buon senso" non è 5% (in generale diverso da α)!
- In effetti, la sua dimensione dipende dalla correlazione tra t_1 e t_2 (e quindi dalla correlazione tra $\hat{\beta}_1$ e $\hat{\beta}_2$).

Due soluzioni:

- Utilizzare un valore critico diverso in questa procedura – non 1.96 (questo è il "metodo Bonferroni") (in ogni caso, questo metodo è utilizzato raramente nella pratica)
- Utilizzare una statistica di test diversa studiata per verificare subito sia $\hat{\beta}_1$ che $\hat{\beta}_2$: la statistica F (questa è pratica comune)

Statistica di Wald

$$H_0 : \underbrace{\beta_1 = \beta_{1,0},\, \beta_2 = \beta_{2,0}, \ldots}_{q \text{ ipotesi}}$$

$$H_1 : \text{almeno una delle ipotesi è falsa}$$

Nel caso di $q = 2$

$$W = \begin{pmatrix} \hat{\beta}_1 - \beta_{1,0} \\ \hat{\beta}_2 - \beta_{2,0} \end{pmatrix}' \begin{pmatrix} \hat{\sigma}^2_{\hat{\beta}_1} & \hat{\sigma}_{\hat{\beta}_1, \hat{\beta}_2} \\ \hat{\sigma}_{\hat{\beta}_2, \hat{\beta}_1} & \hat{\sigma}^2_{\hat{\beta}_2} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\beta}_1 - \beta_{1,0} \\ \hat{\beta}_2 - \beta_{2,0} \end{pmatrix}$$

Sotto l'ipotesi nulla, la statistica di Wald ha una distribuzione χ^2_q, ($q = 2$ nel caso di due coefficienti)

Quindi rigettiamo con un livello di significatività α se W è maggiore del valore critico $w_{1-\alpha}$ che soddisfa

$$\Pr(\chi^2_2 \leq w_{1-\alpha}) = 1 - \alpha.$$

χ_2^2 valori critici

$$\Pr(\chi_2^2 > w_{1-\alpha}) = 1 - \alpha.$$

q	$w_{.95}$
1	3.84
2	6.00
3	7.80
4	9.48
5	11.10

La statistica di Wald in R

```r
library(car)
lm2 <- feols(testscr~str+expnstu+elpct, data = Caschool, vcov = "hetero")
linearHypothesis(lm2, c("str=0", "expnstu=0"))
```

```
Linear hypothesis test

Hypothesis:
str = 0
expnstu = 0

Model 1: restricted model
Model 2: testscr ~ str + expnstu + elpct

  Df Chisq Pr(>Chisq)
1
2  2  10.9     0.0044 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La statistica F

Formula per il caso speciale dell'ipotesi congiunta $\hat{\beta}_1 = \beta_{10}$ e $\hat{\beta}_2 = \beta_{20}$ in una regressione con due regressori:

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}}\right) = W/q$$

- La statistica F è grande quando t_1 e/o t_2 è grande.
- Difficile dare la formula per più di due β, a meno che non si utilizzi l'algebra matriciale.
- In grandi campioni, F è distribuita come χ_q^2/q.

Valori critici di χ_q^2/q

$$\Pr(F \leqslant w_\alpha) = 1 - \alpha.$$

q	$w_{.95}$
1	3.84
2	3.00
3	2.60
4	2.37

Esempio: dati sulle dimensioni delle classi in California

```r
library(car)
lm2 <- feols(testscr~str+expnstu+elpct, data = Caschool, vcov = "hetero")
linearHypothesis(lm2, c("str=0", "expnstu=0"), test="F")
```

```
Linear hypothesis test

Hypothesis:
str = 0
expnstu = 0

Model 1: restricted model
Model 2: testscr ~ str + expnstu + elpct

  Df Chisq Pr(>Chisq)
1
2  2  10.9     0.0044 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $F = 5.37$ Il 5% del valore critico per $q = 2$ è 3.00
- $Pr(> F) = 0.005$ (R calcola il p-value)

F classica

Esiste una formula semplice per la statistica F, valida solo in condizini di omoschedasticità (perciò non molto utile), che tuttavia può aiutare a comprendere che cosa fa la statistica F.

La statistica F in condizioni di omoschedasticità pura

- In presenza di “omoschedasticità pura”:
  - Eseguire due regressioni, una sotto l’ipotesi nulla (regressione “vincolata”) e una sotto l’ipotesi alternativa (regressione senza vincolo).
  - Confrontare gli adattamenti delle regressioni – gli R^2 – se il modello “non vincolato” si adatta sufficientemente meglio, rifiutare l’ipotesi nulla

Regressione "vincolata" e "non vincolata"

Esempio: $H_0 : \beta_1 = \beta_2 = 0 \text{ vs } H_1 : \text{almeno uno è} \neq 0$

- Regressione "senza vincolo" (sotto H_1)

$$testscr_i = \beta_0 + \beta_1 str_i + \beta_2 expnstu_i + \beta_3 elpct_i + u_i$$

- Regressione "vincolata" (sotto H_0)

$$testscr_i = \beta_0 + \beta_3 elpct_i + u_i$$

  - Il numero di vincoli sotto H_0 è $q = 2$ (perché?).
  - L'adattamento risulterà migliore (R^2 sarà maggiore) nella regressione non vincolata (perché?)

Formula semplice per la statistica F classica

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n - k_{ur} - 1)}$$

dove:

- R^2_r : R^2 della regressione vincolata
- R^2_{ur}= R^2 della regressione non vincolata
- $q =$ numero di restrizioni sotto l'ipotesi nulla
- $k_{ur} =$ numero di regressori nella regressione non vincolata.

```r
lm_ur <- lm(testscr~str+expnstu+elpct, data = Caschool)
cat("R.square unrestricted: ", summary(lm_ur)$r.square)
```

```
R.square unrestricted:  0.437
```

```r
lm_r <- lm(testscr~elpct, data = Caschool)
cat("R.square restricted: ", summary(lm_r)$r.square)
```

```
R.square restricted:  0.415
```

Confronto con `linearHypothesis`

$$F = \frac{(R^2_{ur} - R^2_{r})/q}{(1 - R^2_{ur})/(n - k_{ur} - 1)} = \frac{(0.4366 - 0.4149)/2}{(1 - 0.4366)/(420 - 3 - 1)} = 8.01$$

```r
lm_ur <- lm(testscr~str+expnstu+elpct, data = Caschool)
linearHypothesis(lm_ur, c("str=0", "expnstu=0"))
```

```
Linear hypothesis test

Hypothesis:
str = 0
expnstu = 0

Model 1: restricted model
Model 2: testscr ~ str + expnstu + elpct

  Res.Df   RSS Df Sum of Sq    F  Pr(>F)
1    418 89000
2    416 85700  2      3300 8.01 0.00039 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La statistica F classica – riepilogo

Se le quattro assunzioni dei minimi quadrati per la regressione multipla valgono e, in aggiunta:

1. u_i è omoschedastico, ossia $var(u|X_1, \ldots, X_k)$ è costante;
2. $u_1, \ldots, u_n$ sono normalmente distribuiti

allora:

- la statistica F classica ha la distribuzione $F_{q,n-k-1}$, dove: q = numero delle restrizioni e k = numero dei regressori sotto l'alternativa (modello non vincolato).
- Se gli errori sono omoschedastici, la statistica W ha una distribuzione in grandi campioni che è χ_q^2.
- Se gli errori sono omoschedastici, la statistica F classica ha una distribuzione in grandi campioni che è χ_q^2/q.
- Se gli errori sono eteroschedastici, la statistica F classica non ha una distribuzione χ_q^2/q e non è valida

Riepilogo: la statistica F classica e la distribuzione F

- La statistica F classica è giustificata solo sotto condizioni molto forti – troppo forti per essere realistiche.
- Dovreste utilizzare la statistica F robusta all'eteroschedasticità robusta ossia $F_{q,\infty}$.
- Per $n \geq 100$, la distribuzione F è essenzialmente la distribuzione χ^2_q/q.
- Per n piccolo, a volte i ricercatori utilizzano la distribuzione F perché ha valori critici più grandi e in tal senso è più prudente.

Riepilogo: verifica di ipotesi congiunte

- L'approccio "coefficiente per coefficient" che prevede il rifiuto se l'una o l'altra statistica t supera 1.96 rifiuta più del 5% delle volte sotto l'ipotesi nulla (la dimensione supera il livello di significatività desiderato)
- La statistica F robusta all'eteroschedasticità (=Wald stat/q) è integrata in R (comando " linearHypothesis"); questa verifica tutte le restrizioni q allo stesso tempo.
- Per n grande, la statistica F ha distribuzione χ_q^2/q (=$F_{q,\infty}$).
- La statistica F classica è storicamente importante (e così anche nella pratica) e può aiutare l'intuizione, ma non è valida in presenza di eteroschedasticità.

Verifica di restrizioni singole su coefficienti multipli (Paragrafo 7.3)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\ i = 1, \ldots, n$$

Considerate di voler testare:

$$H_0 : \beta_1 = \beta_2 \text{ vs } H_1 : \beta_1 \neq \beta_2$$

Questa ipotesi nulla impone una singola restrizione ($q = 1$) su coefficienti multipli – non si tratta di ipotesi congiunte con restrizioni multiple (confrontate con $H_0 : \beta_1 = \beta_2 = 0$).

Due metodi per la verifica di restrizioni singole su coefficienti multipli

1. Trasformare la regressione

Trasformare i regressori in modo che la restrizione diventi una restrizione su un singolo coefficiente

2. Eseguire la verifica direttamente

Alcuni software, tra cui R, consentono di verificare le restrizioni utilizzando direttamente coefficienti multipli

Metodo 1: Riorganizzare ("trasformare") la regressione

$$
\begin{aligned}
&Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\ i = 1, \ldots, n \\
&H_0 : \beta_1 = \beta_2 \text{ vs } H_1 : \beta_1 \neq \beta_2
\end{aligned}
$$

Sottrarre e sommare $\beta_2 X_{1i}$:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \\
Y_i &= \beta_0 + \beta_1 X_{1i} - \beta_2 X_{1i} + \beta_2 X_{1i} + \beta_2 X_{2i} + u_i \\
Y_i &= \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + u_i
\end{aligned}
$$

Alternativamente definiamo:

$$
Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i
$$

dove:

$$
\gamma_1 = (\beta_1 - \beta_2) \text{ e } W_i = (X_{1i} + X_{2i}).
$$

Riorganizzare la regressione (continua)

a. Equazione originale

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$H_0 : \beta_1 = \beta_2 \text{ vs } H_1 : \beta_1 \neq \beta_2$$

b. Equazione riorganizzata (“trasformata”):

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i, \quad \gamma_1 = (\beta_1 - \beta_2), W_i = (X_{1i} + X_{2i})$$

$$H_0 : \gamma_1 = 0 \text{ vs } H_1 : \gamma_1 \neq 0$$

- Queste due regressioni (a. e b.) hanno lo stesso R^2, gli stessi valori previsti e gli stessi residui.
- Il problema di verifica è ora semplice: verificare se $\gamma_1 = 0$ nella regressione b.

Riorganizzare la regressione (continua)

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i, \quad \gamma_1 = (\beta_1 - \beta_2), W_i = (X_{1i} + X_{2i})$$

$$H_0 : \gamma_1 = 0 \text{ vs } H_1 : \gamma_1 \neq 0$$

```r
Caschool <- Caschool |> mutate(W=str+expnstu)
lm2 <- feols(testscr~str+W+elpct, data = Caschool, vcov = "hetero")
lm2
```

```
OLS estimation, Dep. Var.: testscr
Observations: 420
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 649.57795   15.45834  42.021 < 2.2e-16 ***
str          -0.29027    0.48124  -0.603  0.546731
W             0.00387    0.00158   2.447  0.014821 *
elpct        -0.65602    0.03178 -20.640 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 14.3   Adj. R2: 0.432529
```

Metodo 2: Eseguire la verifica direttamente

$$testscr_i = \beta_0 + \beta_1 str_i + \beta_2 expnstu_i + \beta_3 elpct_i + u_i$$
$$H_0 : \beta_1 = \beta_2 \text{ vs } H_1 : \beta_1 \neq \beta_2$$

```r
lm2 <- feols(testscr~str+expnstu+elpct, data = Caschool, vcov = "hetero")
linearHypothesis(lm2, "str=expnstu")
```

```
Linear hypothesis test

Hypothesis:
str - expnstu = 0

Model 1: restricted model
Model 2: testscr ~ str + expnstu + elpct

  Df Chisq Pr(>Chisq)
1
2  1  0.36       0.55
```