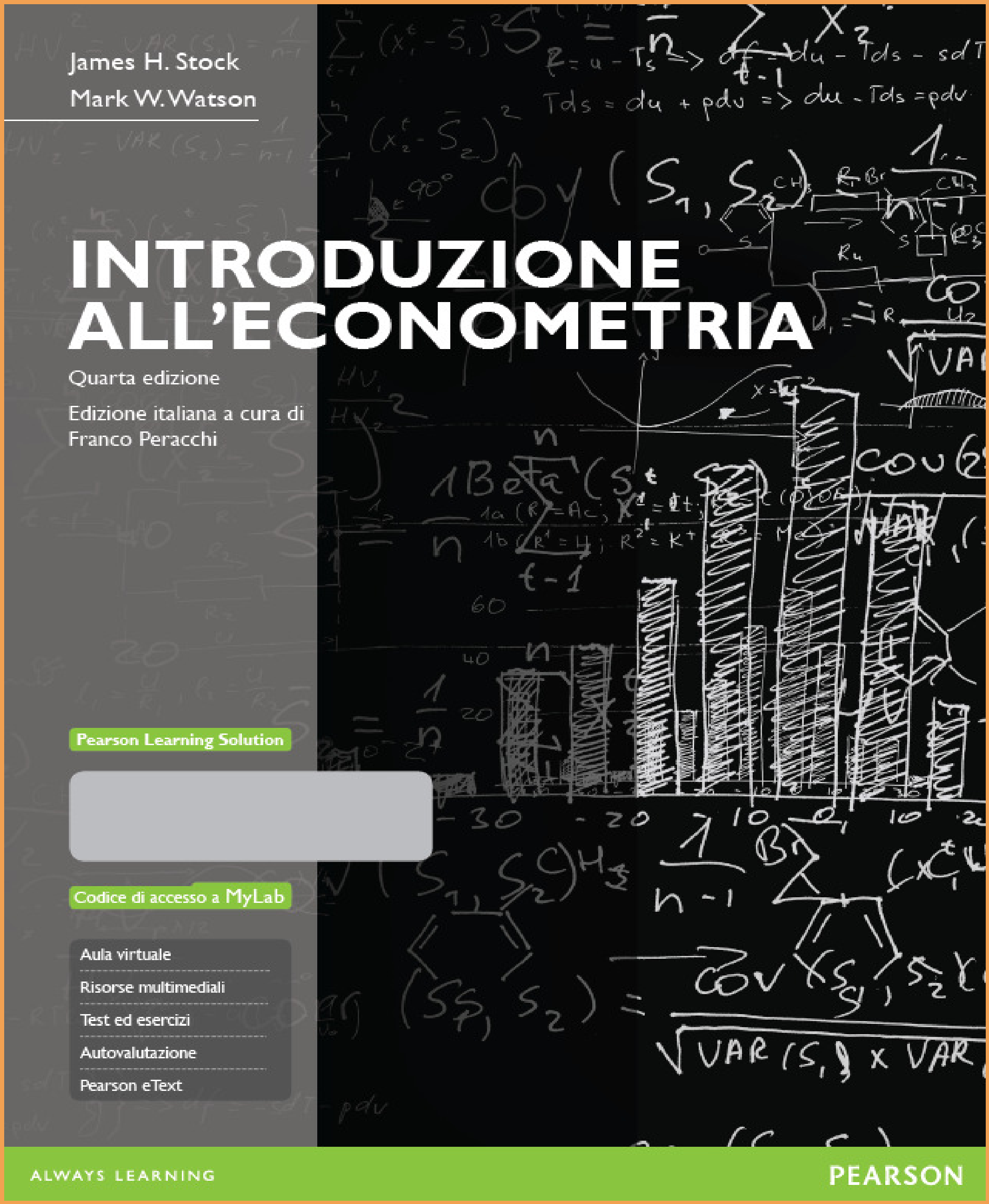
Capitolo 5

Regressione con un singolo regressore: verifica di ipotesi e intervalli di confidenza



Sommario

1. L'errore standard di $\hat{\beta}_1$
2. Verifiche di ipotesi concernenti β_1
3. Intervalli di confidenza per β_1
4. La regressione quando X è variabile binaria
5. Eteroschedasticità e omoschedasticità
6. Efficienza OLS e distribuzione t di Student

Panoramica sul percorso che ci attende…

Vogliamo conoscere la pendenza della retta di regressione. Disponiamo dei dati di un campione, perciò sussiste l'incertezza della campionatura. Per raggiungere l'obiettivo si procede in cinque passaggi:

1. Definire la popolazione oggetto di interesse
2. Fornire uno stimatore di questa popolazione
3. Derivare la distribuzione campionaria dello stimatore (ciò richiede alcune assunzioni). In grandi campioni questa distribuzione campionaria sarà normale per il TLC.
4. La radice quadrata della varianza stimata della distribuzione campionaria è l'errore standard (SE) dello stimatore
5. Utilizzare *SE* per costruire statistiche-*t* (per le verifiche di ipotesi) e intervalli di confidenza.

Oggetto di interesse: β_1 in

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\dots, n$$

$\beta_1 = \Delta E[Y \mid X] / \Delta X$, per una variazione autonoma in X (***effetto causale***)

Stimatore: lo stimatore OLS $\hat{\beta}_1$.

La distribuzione campionaria di $\hat{\beta}_1$:

Per derivare la distribuzione in grandi campioni di $\hat{\beta}_1$, effettuiamo le seguenti assunzioni:

Le assunzioni dei minimi quadrati:

1. $E(u|X = x) = 0$.
2. (X_i, Y_i), $i = 1,\dots,n$, sono i.i.d.
3. Gli outlier sono rari ($E(X^4) < \infty$, $E(Y^4) < \infty$.

La distribuzione campionaria di $\hat{\beta}_1$ (continua)

Sotto le assunzioni dei minimi quadrati, per *n* grande, la distribuzione di $\hat{\beta}_1$ è approssimata da

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2}{n(\sigma_X^2)^2}\right), \text{ dove } v_i = (X_i - \mu_X)u_i$$

Verifica di ipotesi ed errore standard di $\hat{\beta}_1$ (Paragrafo 5.1)

L'obiettivo è di verificare un'ipotesi, quale $\beta_1 = 0$, utilizzando i dati – per determinare sperimentalmente che l'ipotesi (nulla) sia corretta o meno.

Impostazione generale

Ipotesi nulla e alternativa **bilaterale**:

$$H_0:\ \beta_1 = \beta_{1,0} \text{ vs. } H_1:\ \beta_1 \neq \beta_{1,0}$$

dove $\beta_{1,0}$ è il valore ipotizzato nella situazione nulla.

Ipotesi nulla e alternativa **unilaterale**:

$$H_0:\ \beta_1 = \beta_{1,0} \text{ vs. } H_1:\ \beta_1 < \beta_{1,0}$$

***Soluzione generale*: costruire la statistica-*t* e calcolare il valore-*p* (o confrontare il valore critico di *N* (0,1))**

- ***In generale***:

$$t = \frac{\text{stimatore} - \text{valore ipotizzato}}{\text{errore standard dello stimatore}}$$

dove l'*SE* dello stimatore è la radice quadrata di uno stimatore della varianza dello stimatore.

- ***Per verificare la media di Y***: $t = \dfrac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}}$

- ***Per verificare β_1***, $t = \dfrac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$,

dove $SE(\hat{\beta}_1)$ = la radice quadrata di uno stimatore della varianza della distribuzione campionaria di $\hat{\beta}_1$

Formula per SE($\hat{\beta}_1$)

Si ricordi l'espressione per la varianza di (*n* grande):

var($\hat{\beta}_1$) = $\dfrac{\text{var}[(X_i - \mu_x)u_i]}{n(\sigma_X^2)^2} = \dfrac{\sigma_v^2}{n(\sigma_X^2)^2}$, dove $v_i = (X_i - \mu_X)u_i$.

Lo stimatore della varianza di $\hat{\beta}_1$ sostituisce i valori di popolazione ignoti di σ_v^2 e σ_X^2 con gli stimatori ricavati dai dati:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\text{stimatore di } \sigma_v^2}{(\text{stimatore di } \sigma_X^2)^2} = \frac{1}{n} \times \frac{\dfrac{1}{n-2}\sum_{i=1}^{n} \hat{v}_i^2}{\left[\dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

dove $\hat{v}_i = (X_i - \bar{X})\hat{u}_i$.

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{v}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}, \text{ dove } \hat{v}_i = (X_i - \bar{X})\hat{u}_i.$$

$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$ = errore standard di $\hat{\beta}_1$

È leggermente complicato, tuttavia:

- lo è meno di quanto sembri. La varianza var(v) è stimata dal numeratore, mentre [var(X)]^2 è stimata dal denominatore.
- Perché la correzione dei gradi di libertà n – 2? Perché sono stati stimati due coefficienti (β_0 e β_1).
- $SE(\hat{\beta}_1)$ viene calcolato dal software di regressione
- Il software di regressione ha questa formula memorizzata, perciò non dovete eseguire voi il calcolo.

Riepilogo:
Per verificare H_0: $\beta_1 = \beta_{1,0}$ v. H_1: $\beta_1 \neq \beta_{1,0}$,

- Costruire la statistica-*t*

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}}$$

- Si rifiuta al livello di significatività del 5% se |t| > 1,96
- Il valore *p* è $p = \Pr[|t| > |t^{act}|]$ = probabilità nell'area delle code normale $|t^{act}|$; si rifiuta al livello di significatività del 5% se il valore *p* è < 5%.
- Questa procedura si affida all'approssimazione di *n* grande che $\hat{\beta}_1$ sia distribuito normalmente; in generale $n = 50$ è grande abbastanza per un'approssimazione eccellente.

Esempio: *Punteggi nei test* (*TestScore*) e *STR*, dati della California

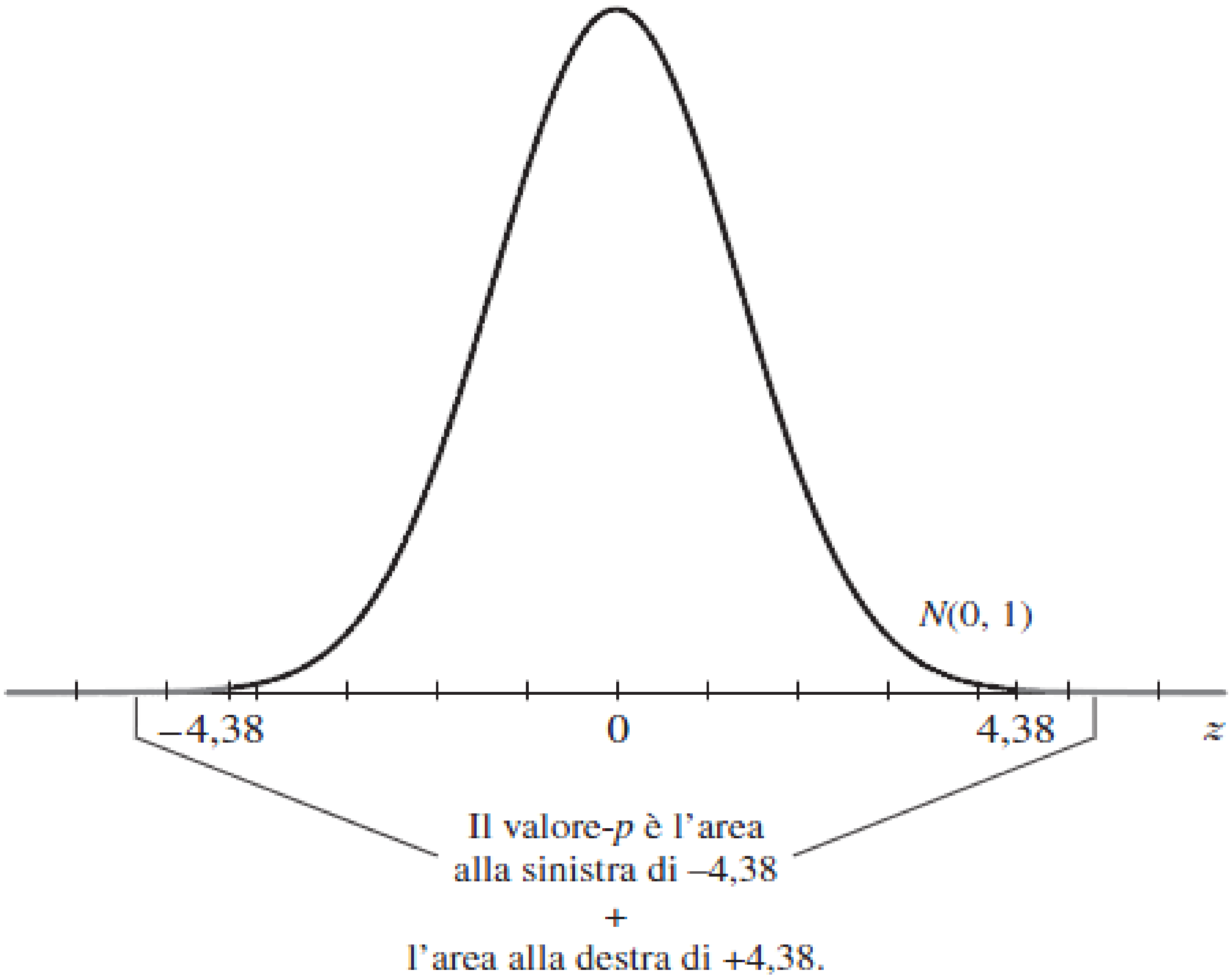
Regressione lineare stimata: $\widehat{TestScore} = 698{,}9 - 2{,}28 \times STR$

Il software di regressione segnala gli errori standard:

$$SE(\hat{\beta}_0) = 10{,}4 \qquad\qquad SE(\hat{\beta}_1) = 0{,}52$$

verifica statistica-t: $\beta_{1,0} = 0$; tact $= \dfrac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \dfrac{-2{,}28 - 0}{0{,}52} = -4{,}38$

- Il livello di significatività bilaterale dell'1% è 2,58, perciò rifiutiamo l'ipotesi nulla al livello di significatività dell'1%.
- In alternativa, possiamo calcolare il valore *p*…



Il valore p basato sull'approssimazione normale standard con n grande alla statistica t è 0,00001 (10^{-5})

Intervalli di confidenza per β_1 (Paragrafo 5.2)

Si ricordi che un intervallo di confidenza al 95% equivale a:

- la serie di punti che non può essere rifiutata al livello di significatività del 5%;
- una funzione polidroma (un intervallo funzione dei dati) che contiene il reale valore del parametro il 95% delle volte nei campioni ripetuti.

Poiché la statistica *t* per β_1 è N(0,1) in grandi campioni, la costruzione di un intervallo di confidenza al 95% per β_1 equivale al caso della media campionaria:

$$\text{intervallo di confidenza al 95\% per } \beta_1 = \{\hat{\beta}_1 \pm 1{,}96 \times SE(\hat{\beta}_1)\}$$

Esempio di intervallo di confidenza: *TestScore* e *STR*
Retta di regressione stimata: $\widehat{TestScore}$ **= 698,9 – 2,28** $\times$ ***STR***

$SE(\hat{\beta}_0) = 10{,}4$

$SE(\hat{\beta}_1) = 0{,}52$

Intervallo di confidenza al 95% per $\hat{\beta}_1$:

$$\{\hat{\beta}_1 \pm 1{,}96 \times SE(\hat{\beta}_1)\} = \{-2{,}28 \pm 1{,}96 \times 0{,}52\}$$

$$= (-3{,}30,\ -1{,}26)$$

Le due affermazioni seguenti sono equivalenti *(perché?)*

- L'intervallo di confidenza al 95% non include lo zero;
- L'ipotesi $\beta_1 = 0$ è rifiutata al livello del 5%

Un modo conciso (e tradizionale) per la notazione delle regressioni:

Porre gli errori standard tra parentesi sotto i coefficienti stimati ai quali sono pertinenti.

$$\underset{}{\widehat{TestScore}} = \underset{(10,4)}{698,9} - \underset{(0,52)}{2,28} \times STR,\ R^2 = 0,05,\ SER = 18,6$$

<u>Questa espressione fornisce molte informazioni</u>

- La retta di regressione stimata è

$$\widehat{TestScore} = 698,9 - 2,28 \times STR$$

- L'errore standard di $\hat{\beta}_0$ è 10,4
- L'errore standard di $\hat{\beta}_1$ è 0,52
- R^2 è 0,05; l'errore standard della regressione è 18,6

Regressione OLS: lettura output STATA

```
• regress testscr str, robust
•
• Regression with robust standard errors               Number of obs =     420
•                                                      F(  1,   418) =   19.26
•                                                      Prob > F      =  0.0000
•                                                      R-squared     =  0.0512
•                                                      Root MSE      =  18.581
• -------------------------------------------------------------------------
•          |               Robust
•  testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
• ---------+---------------------------------------------------------------
•      str |  -2.279808   .5194892    -4.38   0.000    -3.300945   -1.258671
•    _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
• -------------------------------------------------------------------------
```

quindi:

$$\widehat{TestScore} = 698{,}9 - 2{,}28 \times STR, \; , \; R^2 = \mathbf{0{,}05}, \; SER = \mathbf{18{,}6}$$

$$(10{,}4) \quad (0{,}52)$$

t (β_1 = 0) = **–4,38**, valore-p = **0,000** (bilaterale)

L'intervallo conf. bilaterale al 95% per β_1 è (–3,30, –1,26)

Riepilogo di inferenza statistica in merito a β_0 e β_1

Stima:

- Gli stimatori OLS $\hat{\beta}_0$ e $\hat{\beta}_1$
- $\hat{\beta}_0$ e $\hat{\beta}_1$ hanno approssimativamente distribuzioni campionarie normali in grandi campioni

Verifica:

- H_0: $\beta_1 = \beta_{1,0}$ v. $\beta_1 \neq \beta_{1,0}$ ($\beta_{1,0}$ è il valore di β_1 sotto H_0)
- $t = (\hat{\beta}_1 - \beta_{1,0})/SE(\hat{\beta}_1)$
- valore-p = area sotto la normale standard al di fuori di t^{act} (n grande)

Intervalli di confidenza:

- l'intervallo di confidenza al 95% per β_1 è $\{\hat{\beta}_1 \pm 1{,}96 \times SE(\hat{\beta}_1)\}$
- Questo è l'insieme di β_1 che non è rifiutato al livello del 5%
- L'IC al 95% contiene il β_1 reale nel 95% di tutti i campioni.

La regressione quando *X* è una variabile binaria (Paragrafo 5.3)

A volte un regressore è binario:

- $X = 1$ se classe piccola, $= 0$ altrimenti
- $X = 1$ se femmina, $= 0$ se maschio
- $X = 1$ se trattato (farmaco sperimentale), $= 0$ altrimenti

I regressori binari sono a volte chiamati variabili "dummy".

Fin qui β_1 è stato chiamato "pendenza" ma ciò non ha senso se la variabile X è binaria.

Come interpretiamo la regressione con un regressore binario?

Interpretazione delle regressioni con un regressore binario

$Y_i = \beta_0 + \beta_1 X_i + u_i$, dove X è binaria (X_i = 0 o 1):

Quando $X_i = 0$, $Y_i = \beta_0 + u_i$

- la media di Y_i è β_0
- cioè $E(Y_i|X_i=0) = \beta_0$

Quando $X_i = 1$, $Y_i = \beta_0 + \beta_1 + u_i$

- la media di Y_i è $\beta_0 + \beta_1$
- cioè $E(Y_i|X_i=1) = \beta_0 + \beta_1$

quindi:

$$\beta_1 = E(Y_i|X_i=1) - E(Y_i|X_i=0)$$
$$= \text{differenza tra medie}$$

Esempio: Sia $D_i = \begin{cases} 1 \text{ se } STR_i \leq 20 \\ 0 \text{ se } STR_i > 20 \end{cases}$

regressione OLS: $\widehat{TestScore} = \mathbf{650{,}0} + \mathbf{7{,}4} \times D$

$(1{,}3) \quad (\mathbf{1{,}8})$

Tabulazione delle medie:

Dimensione classe	**Punteggio medio ($\bar{Y}$)**	**Dev. std. (s_Y)**	***N***
Piccola ($STR > 20$)	657,4	19,4	238
Grande ($STR \geq 20$)	**650,0**	17,9	182

Differenza tra medie: $\bar{Y}_{\text{small}} - \bar{Y}_{\text{large}} = 657{,}4 - 650{,}0 = \mathbf{7{,}4}$

Errore standard $SE = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}} = \sqrt{\frac{19{,}4^2}{238} + \frac{17{,}9^2}{182}} = \mathbf{1{,}8}$

Riepilogo: regressione quando la variabile X_i è binaria (0/1)

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

- β_0 = media di Y quando X = 0
- $\beta_0 + \beta_1$ = media di Y quando X = 1
- β_1 = differenza tra medie, X =1 meno X = 0
- SE($\hat{\beta}_1$) ha l'interpretazione consueta
- statistica-*t*, intervalli di confidenza costruiti come di consueto
- Questo è un altro modo (facile) per eseguire l'analisi della differenza tra medie
- La formulazione della regressione è particolarmente utile quando abbiamo regressori supplementari (*e li avremo molto presto*)

Eteroschedasticità e omoschedasticità, ed errori standard per l'omoschedasticità pura (Paragrafo 5.4)

1. Di che cosa si tratta...?
2. Conseguenze dell'omoschedasticità
3. Implicazioni per il calcolo degli errori standard

Che cosa significano questi due termini?

Se var(u|X=x) è costante – ossia se la varianza della distribuzione di u condizionata a X non dipende da X – allora u è detto ***omoschedastico***. In caso contrario, u è ***eteroschedastico***.

Esempio: etero/omoschedasticità nel caso di un regressore binario (ossia confronto tra medie)

- Errore standard quando le varianze sono ***ineguali***:

$$SE = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}$$

- Errore standard quando le varianze sono ***uguali***:

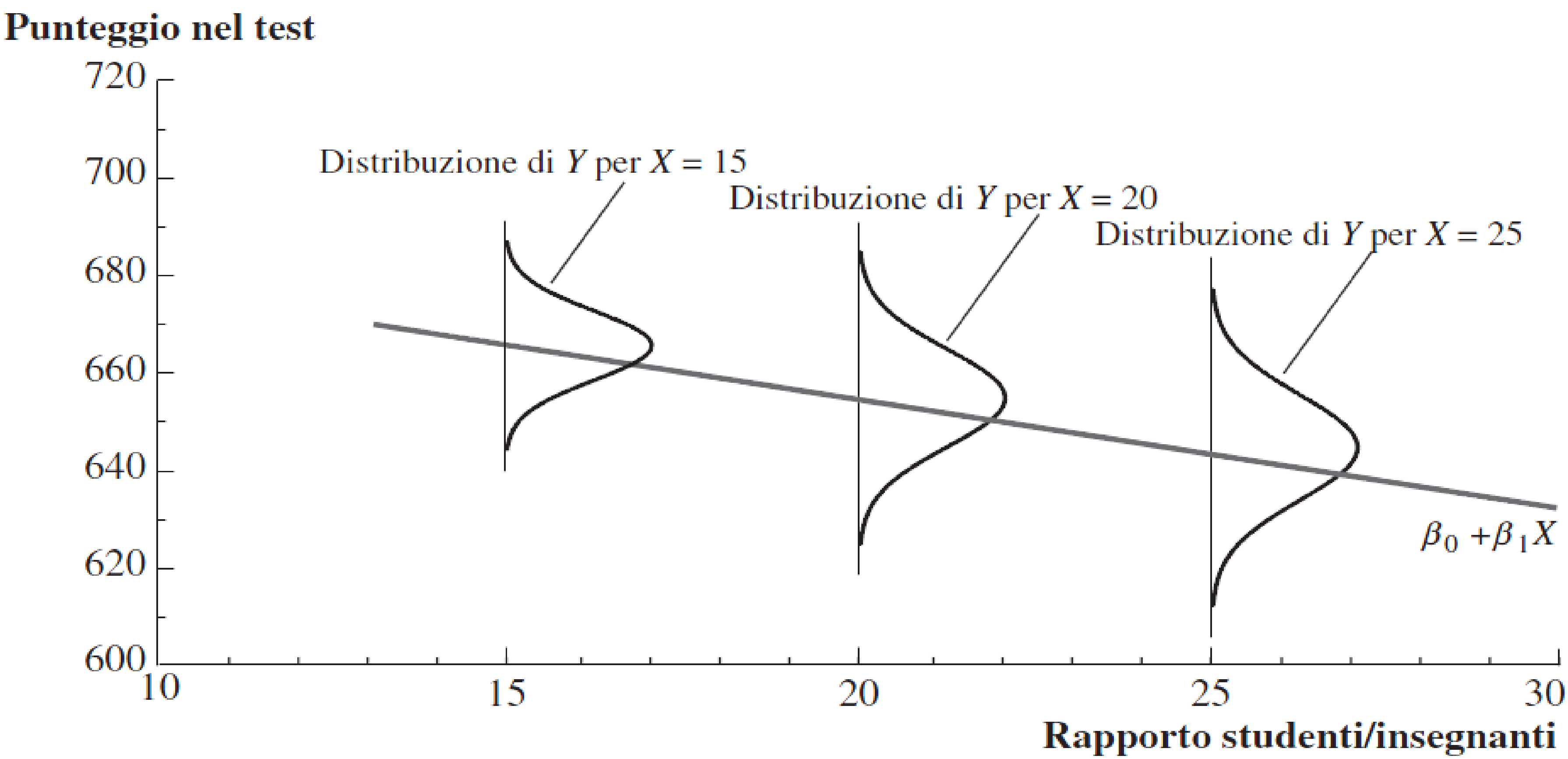
$$SE = s_p \sqrt{\frac{1}{n_s} + \frac{1}{n_l}}$$

Dove $s_p^2 = \dfrac{(n_s - 1)s_s^2 + (n_l - 1)s_l^2}{n_s + n_l - 2}$ (SW, Sez. 3.6)

s_p = "stimatore aggregato di σ^2" quando $\sigma_l^2 = \sigma_s^2$

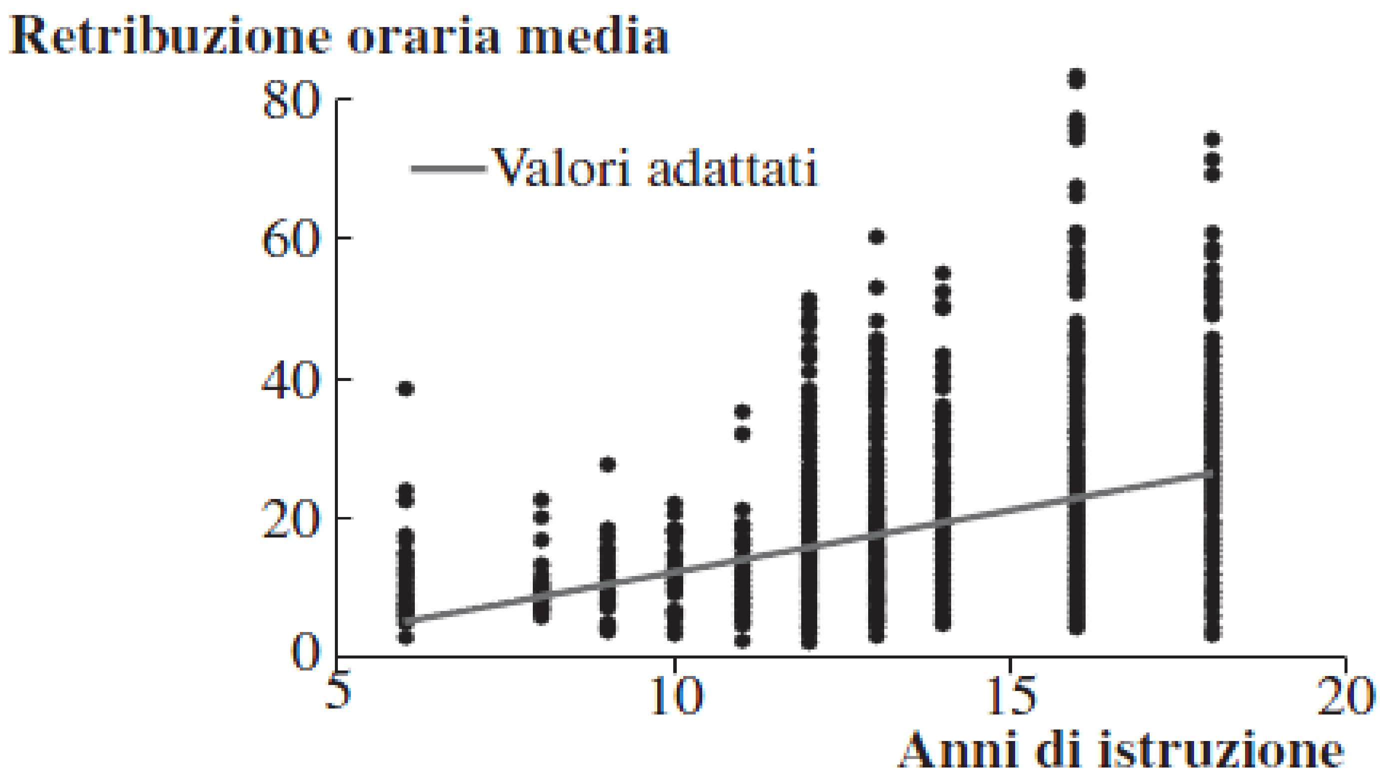
- Varianze **uguali** = **omo**schedasticità
- Varianze **ineguali** = **etero**schedasticità

Eteroschedasticità in un'immagine:



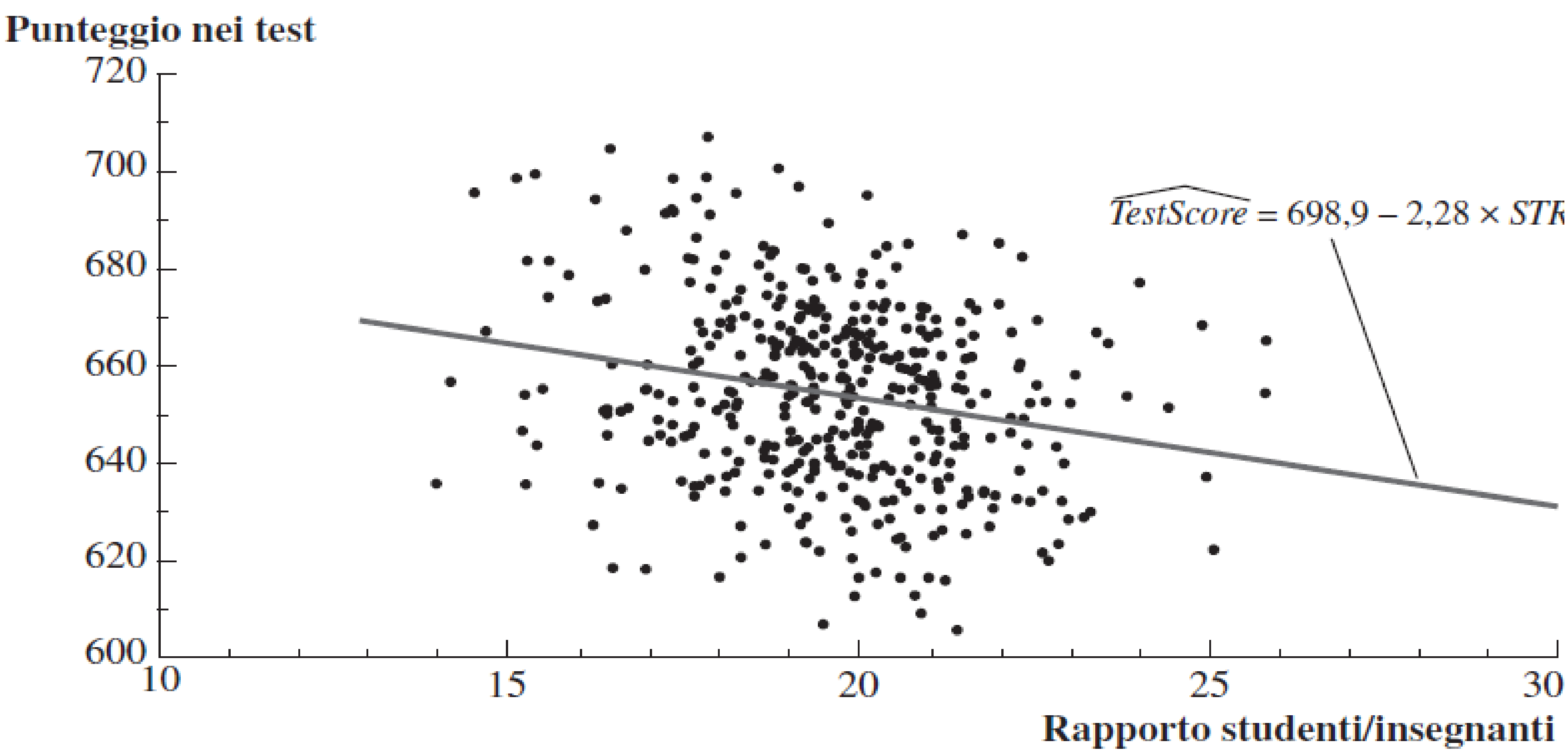
- $E(u|X=x) = 0$ (u soddisfa l'assunzione dei minimi quadrati n. 1)
- La varianza di u dipende da x

Un esempio con dati reali dall'economica del lavoro: retribuzione oraria media rispetto agli anni di istruzione (fonte dati: Current Population Survey):



Eteroschedasticità o omoschedasticità?

Dati dimensione classe:



Eteroschedasticità o omoschedasticità?

Finora abbiamo assunto (senza dirlo) che *u* poteva essere eteroschedastico.

Ricordiamo le tre assunzioni dei minimi quadrati:

1. $E(u|X = x) = 0$
2. (X_i, Y_i), $i = 1,\ldots,n$, sono i.i.d.
3. Gli outlier sono rari

Eteroschedasticità e omoschedasticità concernono var($u|X=x$). Poiché non abbiamo assunto esplicitamente gli errori omoschedastici, abbiamo ammesso implicitamente l'eteroschedasticità.

Che cosa succede se gli errori sono in effetti omoschedastici?

- Si può dimostrare che l'OLS ha la varianza minore tra gli stimatori lineari in *Y*… un risultato chiamato teorema di Gauss-Markov, sul quale ritorneremo a breve.
- La formula per la varianza di $\hat{\beta}_1$ e per l'errore standard OLS si semplifica: se var(u_i|X_i=x) = σ_u^2 , per ogni x, allora

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}[(X_i - \mu_x)u_i]}{n(\sigma_X^2)^2} \quad \text{(formula generale)}$$

$$= \frac{\sigma_u^2}{n\sigma_X^2} \quad \text{(semplificazione se } u \text{ è omoschedastico)}$$

Nota: var($\hat{\beta}_1$) è inversamente proporzionale a var(*X*): maggiore dispersione in *X* significa più informazioni su $\hat{\beta}_1$ – abbiamo discusso di questo in precedenza, ma risulta più chiaro da questa formula.

- Insieme a questa formula di omoschedasticità pura per la varianza di $\hat{\beta}_1$, abbiamo errori standard di omoschedasticità pura:

Formula errore standard omoschedasticità pura:

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Alcune persone (per esempio i programmatori di Excel) trovano più semplice la formula dell'omoschedasticità pura – ma ciò è sbagliato a meno che gli errori non siano realmente omoschedastici.

Abbiamo ora due formule per gli errori standard per $\hat{\beta}_1$ ***.***

- ***Errori standard per l'omoschedasticità pura*** (**SE classici**)– sono validi solo se gli errori sono omoschedastici.
- Gli errori standard consueti – per differenziare i due, è convenzione chiamarli ***errori standard robusti all'eteroschedasticità***, poiché sono validi (**consistenti**) a prescindere dall'eteroschedasticità o meno degli errori. Sono detti anche **SE di White**, dal nome dell'autore che li ha proposti.
- Il principale vantaggio degli errori standard per l'omoschedasticità pura è che la formula è più semplice. Lo svantaggio, però, è che la formula è corretta solo se gli errori sono omoschedastici.

Implicazioni pratiche…

- La formula dell'omoschedasticità pura per l'errore standard di $\hat{\beta}_1$ e la formula "robusta all'eteroschedasticità" sono diverse – quindi, in generale, *si ottengono errori standard diversi utilizzando formule differenti*.
- Gli errori standard per l'omoschedasticità pura sono l'impostazione predefinita nei software di regressione – a volte l'unica impostazione (per esempio in Excel). Per ottenere gli errori standard "robusti all'eteroschedasticità" generali occorre modificare l'impostazione di default.
- **Se non si modifica l'impostazione di default e vi è eteroschedasticità, gli errori standard (e la statistica-*t* e gli intervalli di confidenza) saranno errati – generalmente, gli *SE* per l'omoschedasticità pura sono troppo piccoli.**

Errori standard robusti all'eteroschedasticità in STATA

```
regress testscr str, robust
Regression with robust standard errors  Number of obs =     420
        F(  1,   418) =   19.26
         Prob > F      =  0.0000
         R-squared     =  0.0512
        Root MSE      =  18.581
-------------------------------------------------------------------------
                        Robust
testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------------------------------------------------------
qqqqstr |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
qq_cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
-------------------------------------------------------------------------
```

- Se si utilizza l'opzione **“, robust"**, STATA calcola gli errori standard robusti all'eteroschedasticità
- In caso contrario, STATA calcola gli errori standard per l'omoschedasticità pura

***Il punto essenziale*:**

- Se gli errori sono omoschedastici o eteroschedastici e si utilizzano errori standard robusti all'eteroschedasticità, va bene
- Se gli errori sono eteroschedastici e si utilizza la formula dell'omoschedasticità pura per gli errori standard, gli errori standard saranno errati (lo stimatore dell'omoschedasticità pura della varianza di $\hat{\beta}_1$ è incoerente (**inconsistente**) in presenza di eteroschedasticità).
- Le due formule coincidono (quando *n* è grande) nel caso speciale di omoschedasticità
- Quindi si dovrebbero sempre utilizzare errori standard robusti all'eteroschedasticità.

Fondamenti teorici dei minimi quadrati ordinari (Paragrafo 5.5)

Abbiamo già appreso molto sugli stimatori dei minimi quadrati ordinari: lo stimatore OLS è non distorto e consistente; abbiamo una formula per gli errori standard robusti all'eteroschedasticità e possiamo costruire intervalli di confidenza e statistiche di test.

Una buona ragione per utilizzare i minimi quadrati ordinari è anche l'impiego universale, perciò gli altri saranno in grado di capire ciò che fate. In effetti, l'OLS è il linguaggio dell'analisi di regressione, e se utilizzate uno stimatore diverso, parlerete un linguaggio differente.

Eppure potreste ancora chiedervi…

- Tutto quanto detto è davvero una buona ragione per utilizzare OLS? Non esistono altri stimatori che potrebbero essere migliori – in particolare che potrebbero avere una varianza inferiore?
- Inoltre, che ne è stato della nostra vecchia cara amica, la distribuzione *t* di Student?

Ora risponderemo a queste domande – ma per farlo abbiamo bisogno di assunzioni più forti delle tre relative ai minimi quadrati che abbiamo già visto.

Le assunzioni dei minimi quadrati estese

Consistono nelle tre assunzioni dei minimi quadrati, più altre due:

1. $E(u|X = x) = 0$.
2. (X_i,Y_i), $i = 1,\ldots,n$, sono i.i.d.
3. Gli outlier sono rari ($E(Y^4) < \infty$, $E(X^4) < \infty$).
4. u è omoschedastico
5. u ha distribuzione $N(0,\sigma^2)$

- Le assunzioni 4 e 5 sono più restrittive – perciò si applicano a un numero inferiori di casi pratici. Tuttavia, facendo queste assunzioni, determinati calcoli matematici si semplificano e si possono dimostrare risultati più robusti –che valgono se tali assunzioni aggiuntive sono vere.
- Iniziamo con una discussione sull'efficienza dello stimatore OLS

Efficienza dello stimatore OLS, parte I: il teorema di Gauss-Markov

Nelle assunzioni dei minimi quadrati ordinari estese 1-4 (le tre di base, più l'omoschedasticità), $\hat{\beta}_1$ ha la varianza minima tra *tutti gli stimatori lineari* (stimatori che sono funzioni lineari di $Y_1,\ldots, Y_n$), condizionatamente a tutte le Xi. Questo è il ***teorema di Gauss-Markov***.

Commenti

- Il teorema di GM è dimostrato nell'Appendice 5.2

Il teorema di Gauss-Markov (continua)

- $\hat{\beta}_1$ è uno stimatore lineare, cioè può essere scritto come funzione lineare di $Y_1,\ldots, Y_n$, condiz. alle Xi:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{1}{n}\sum_{i=1}^{n} w_i Y_i \quad ,$$

dove $w_i = \dfrac{(X_i - \bar{X})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

- Il teorema di G-M dice che, fra tutte le scelte possibili di $\{w_i\}$, i pesi OLS portano alla minima var($\hat{\beta}_1$)

Efficienza dello stimatore OLS, parte II:

- In <u>tutte e cinque</u> le assunzioni dei minimi quadrati estese – compresa la distribuzione normale degli errori – $\hat{\beta}_1$ ha la varianza più piccola di *<u>tutti</u>* gli estimatori consistenti (funzioni lineari *o* non lineari di $Y_1,\ldots,Y_n$), per $n \to \infty$.
- Questo è un risultato assai sorprendente – afferma che, se (in aggiunta alle assunzioni dei minimi quadrati 1-3) gli errori sono omoschedastici e normalmente distribuiti, OLS è la scelta migliore di qualsiasi altro stimatore consistente. E poiché uno stimatore che non sia consistente è una scelta scadente, ciò afferma che l'OLS è davvero la miglior scelta che si possa fare – se valgono tutte e cinque le assunzioni dei minimi quadrati estese. (La dimostrazione di questo risultato va oltre l'ambito di questo corso e non è fornita nel testo – viene svolta generalmente nei corsi post laurea).

Alcuni aspetti critici di OLS

I risultati precedenti sono impressionanti, tuttavia tali risultati – e lo stimatore OLS – hanno limitazioni importanti.

1.Il teorema di GM non è poi così avvincente:

- La condizione di omoschedasticità spesso non regge (l'omoschedasticità è speciale)
- Il risultato vale solo per gli stimatori lineari – solo un piccolo sottoinsieme di stimatori (ulteriori informazioni a breve)

2.Il risultato di ottimalità più robusto ("parte II" precedente) richiede errori normali omoschedastici – cosa non plausibile nelle applicazioni (si pensi ai dati delle retribuzioni orarie!)

Limitazioni di OLS (continua)

3. L'OLS è più sensibile agli outlier rispetto ad altri stimatori. Nel caso di stima della media della popolazione, se sussistono outlier, la mediana è preferibile alla media poiché la prima è meno sensibile agli outlier – ha varianza minore rispetto all'OLS in presenza di outlier. In modo analogo, nella regressione l'OLS può essere più sensibile agli outlier e, se vi sono outlier, altri stimatori possono essere più efficienti (hanno varianza minore). Uno di essi è lo stimatore delle minime deviazioni assolute (MDA):

$$\min_{b_0, b_1} \sum_{i=1}^{n} \left| Y_i - (b_0 + b_1 X_i) \right|$$

In praticamente tutte le analisi di regressione reali si utilizza l'OLS – ed è quello che faremo anche in questo corso.

Inferenza se *u* è omoschedastico e normalmente distribuito: la distribuzione *t* di Student (Paragrafo 5.6)

Ricordiamo le cinque assunzioni dei minimi quadrati estese:

1. $E(u|X = x) = 0$.
2. (X_i, Y_i), $i = 1,\ldots,n$, sono i.i.d.
3. Gli outlier sono rari ($E(Y^4) < \infty$, $E(X^4) < \infty$).
4. u è omoschedastico
5. u ha distribuzione $N(0,\sigma^2)$

Se tutte le cinque assunzioni valgono, allora:

- $\hat{\beta}_0$ e $\hat{\beta}_1$ sono normalmente distribuiti *per tutti gli n* (!)
- la statistica-*t* ha una distribuzione *t* di Student con $n - 2$ gradi di libertà – questo vale esattamente *per tutti gli n* (!)

Normalità della distribuzione campionaria sotto le assunzioni 1–5:

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \frac{1}{n}\sum_{i=1}^{n} w_i u_i, \quad \text{dove } w_i = \frac{(X_i - \bar{X})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Qual è la distribuzione di una media ponderata di normali?

Sotto le assunzioni 1 – 5, condizionatamente a X1,…,Xn:

$$\hat{\beta}_1 - \beta_1 \sim N\left(0, \frac{1}{n^2}\left(\sum_{i=1}^{n} w_i^2\right)\sigma_u^2\right) \qquad (*)$$

Sostituendo w_i in (*) si produce la formula della varianza dell'omoschedasticità pura.

Inoltre, sotto le assunzioni 1 – 5, nell'ipotesi nulla la statistica *t* ha una distribuzione *t* di Student con *n – 2* gradi di libertà

- Perché $n - 2$? Perché abbiamo stimato 2 parametri, β_0 e β_1
- Per $n < 30$ i valori critici *t* possono essere un po' più grandi dei valori critici $N(0,1)$
- Per $n > 50$ o simile, la differenza nelle distribuzioni t_{n-2} e $N(0,1)$ è trascurabile. Ricordiamo la tabella *t* di Student:

gradi di libertà	*valore critico distribuzione-t al 5%*
10	2,23
20	2,09
30	2,04
60	2,00
∞	1,96

Implicazioni pratiche:

- Se $n < 50$ **e** credete davvero che, per la vostra applicazione, u sia omoschedastico e normalmente distribuito, utilizzate t_{n-2} invece dei valor critici $N(0,1)$ per le verifiche di ipotesi e gli intervalli di confidenza.

- Nella maggior parte delle applicazioni econometriche, non vi è alcun motivo di ritenere che u sia omoschedastico e normale – solitamente vi sono ottime ragioni per credere che né l'una né l'altra assunzione valga.

- Fortunatamente, nelle applicazioni moderne $n > 50$, così possiamo affidarci ai risultati per n grande presentati in precedenza, basati sul teorema limite centrale, per eseguire verifiche di ipotesi e costruire intervalli di confidenza usando l'approssimazione normale per n grande.

Riepilogo e valutazione (Paragrafo 5.7)

- Il problema iniziale:

  Si supponga che vengano assunti nuovi insegnanti cosicché il rapporto studenti/insegnanti si riduca a uno studente per classe. Qual è l'effetto dell'intervento di questa politica ("trattamento") sui punteggi nei test?
- La nostra analisi di regressione utilizzando i dati della California risponde in modo convincente?

  Non proprio – i distretti con *STR* basso tendono a essere quelli con molte altre risorse e famiglie a reddito più elevato, che offrono ai figli maggiori opportunità di apprendimento al di fuori della scuola… ciò suggerisce che corr(u_i, STR_i) > 0, quindi $E(u_i|X_i)\neq 0$.
- Sembra che abbiamo omesso alcuni fattori o variabili dalla nostra analisi e ciò ha distorto i risultati...