Econometria applicata all'intermediazione finanziaria

Regressione univariata e verifica di ipotesi

Tiziano Razzolini

Università di Siena.

mail: razzolini4@unisi.it

4 novembre 2020

Verifica di ipotesi

La stima del coefficiente $\hat{\beta}_1$ non è sufficiente da sola a guidare scelte di politica economica. In molti casi è necessario testare se il coefficiente stimato è uguale ad un dato valore: nella maggior parte dei casi il test effettuato è $H_0 : \beta_1 = 0$, ad esempio nel caso dell'effetto rapporto studenti-punteggio test.
In termini generali l'ipotesi bilaterale è:

$$H_0 : \beta_1 = \beta_{1,0} \qquad \text{contro} \quad \beta_1 \neq \beta_{1,0}$$

Verifica di ipotesi

La costruzione della statistica t richiede il calcolo di $SE(\hat{\beta}_1)$

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

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

Verifica di ipotesi

Dopo aver calcolato $SE(\hat{\beta}_1)$ ed aver costruito la statistica t, calcoliamo il valore-p che rappresenta la probabilità di osservare un valore lontano dal valore ipotizzato $\beta_{1,0}$ uguale o maggiore al valore stimato $\hat{\beta}_1^{act}$:

$$\text{valore-p:} \quad Pr_{H_0}\left[|\hat{\beta}_1 - \beta_{1,0}| > |\hat{\beta}_1^{act} - \beta_{1,0}|\right] =$$

$$= Pr_{H_0}\left[\left|\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right| > \left|\frac{\hat{\beta}_1^{act} - \beta_{1,0}}{SE(\hat{\beta}_1)}\right|\right] = Pr_{H_0}\left(|t| > |t^{act}|\right) =$$

Nel caso di campioni grandi $\hat{\beta}_1$ si distribuisce come una normale e quindi anche la statistica t sotto l'ipotesi nulla avrà una distribuzione normale standardizzata:

$$= Pr(|Z| > |t^{act}|) = 2\Phi(-|t^{act}|)$$

Verifica di ipotesi

Nel caso di un livello di significatività del 5% osservare un valore-p minore del 5% vuol dire che è improbabile aver ottenuto quel valore di $\hat{\beta}_1$ come risultato della variabilità.

Un modo alternativo per effettuare il test è calcolare il valore critico t associato al livello di significatività scelto: nel caso del 5% di significatività $t = \pm 1,96$. Quindi l'ipotesi nulla viene rifiutata quando $|t^{act}| > 1,96$

Regressione con un singolo regressore

Esposizione dei risultati

Riportiamo l'esempio nel libro di testo: I risultati possono essere riportati nel modo seguente:

$$\widehat{TestScore} = \underset{(10,4)}{698,9} \; \underset{(0,52)}{-2,28} \times STR \quad , \; R^2 = 0,051, \quad SER = 18,6$$
$$\text{n=420}$$

La t statistica è:

$t^{act} = (-2,28 - 0)/(0.52) = -4,38 < -1,96;$

o meglio $\quad |-4,38| = 4,38 > |-1,96| = 1,96$

Pertanto l'ipotesi nulla viene rifiutata. Alternativamente calcoliamo che il valore-p associato a $t^{act} = 4,38$ è circa $0,00001$.

Verifica di ipotesi unilaterale

Diversamente dai test di tipo bilaterale una verifica di ipotesi unilaterale testa:

$$H_0 : \beta_1 = \beta_{1,0} \qquad \text{contro} \quad \beta_1 < \beta_{1,0}$$

Nell'esempio esaminato è rilevante sapere se il rapporto studenti insegnanti ha un effetto negativo sul punteggio dei test. Se il livello di significatività scelto è il 5% nel caso di ipotesi unilaterale è $t = -1,645$.In questo caso si rifiuta l'ipotesi nulla se $\mathrm{t}^{act} < -1,645$

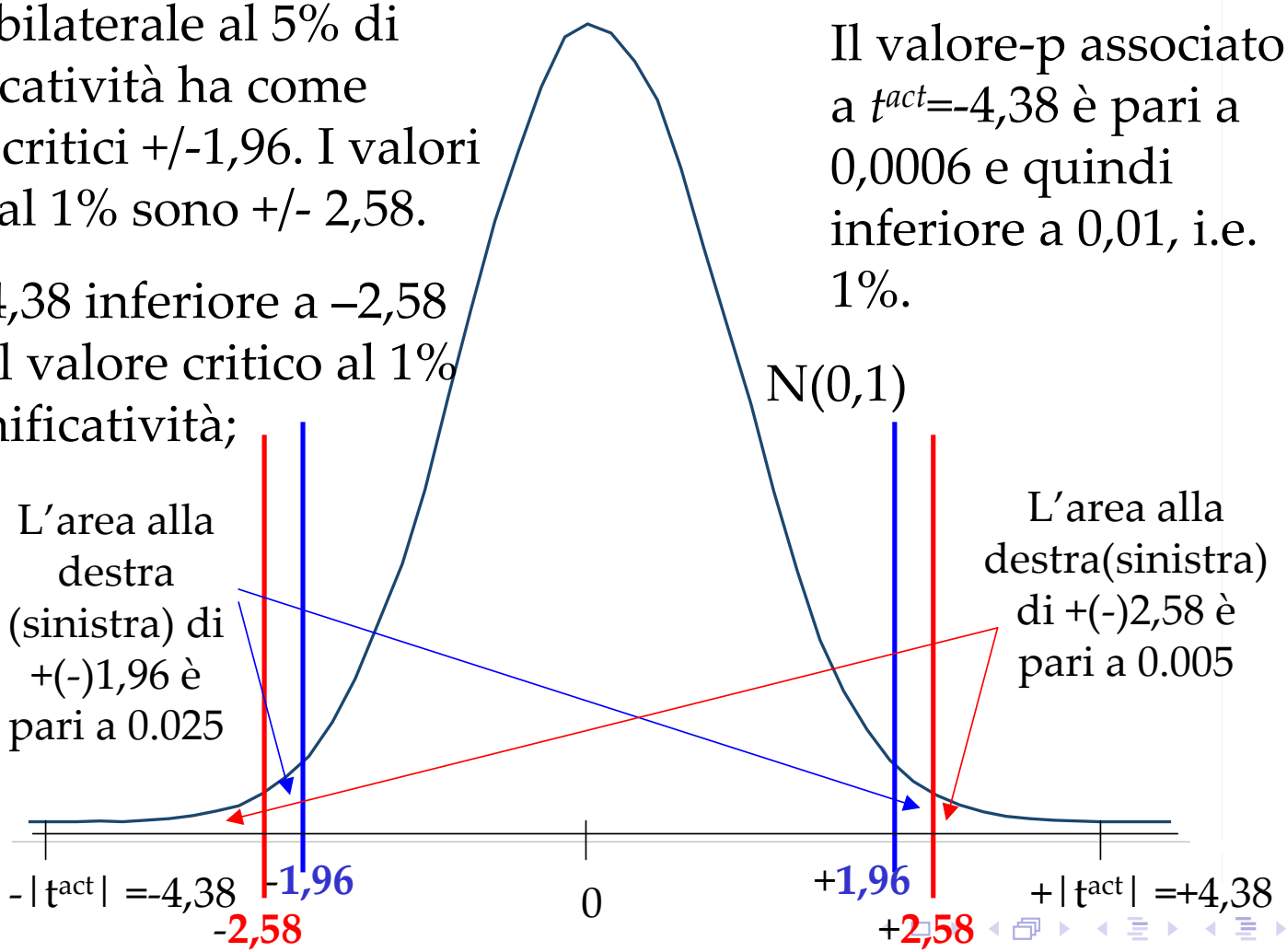
NB: Nel caso bilaterale H_0 è rifiutata al 5% di probabilità se $|t^{act}| > 1,96$

Verifica di ipotesi

Il test bilaterale al 5% di significatività ha come valori critici +/-1,96. I valori critici al 1% sono +/- 2,58.

t^{act}=-4,38 inferiore a –2,58 che è il valore critico al 1% di significatività;

Il valore-p associato a t^{act}=-4,38 è pari a 0,0006 e quindi inferiore a 0,01, i.e. 1%.



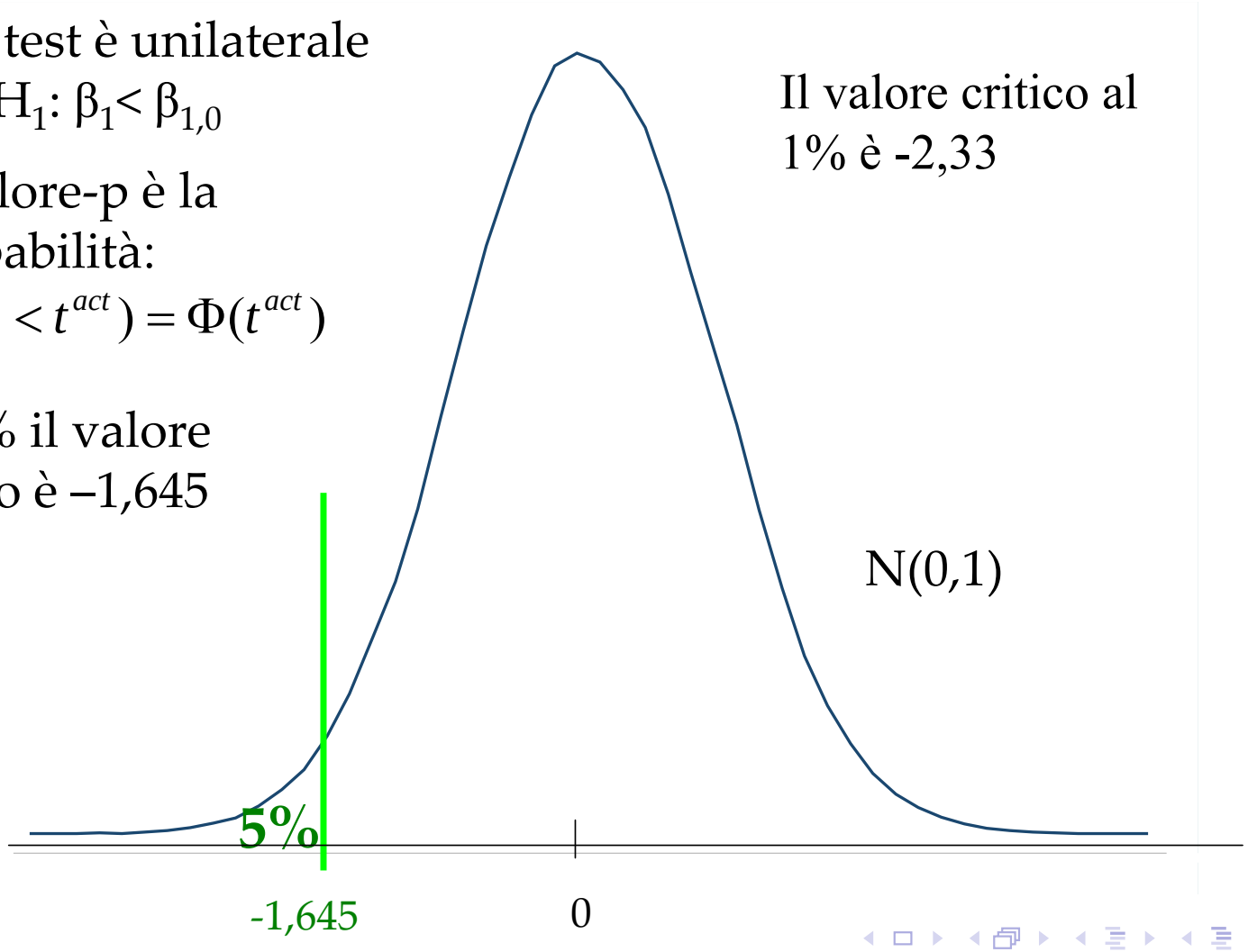
Verifica di ipotesi

Se il test è unilaterale con H_1: $\beta_1 < \beta_{1,0}$

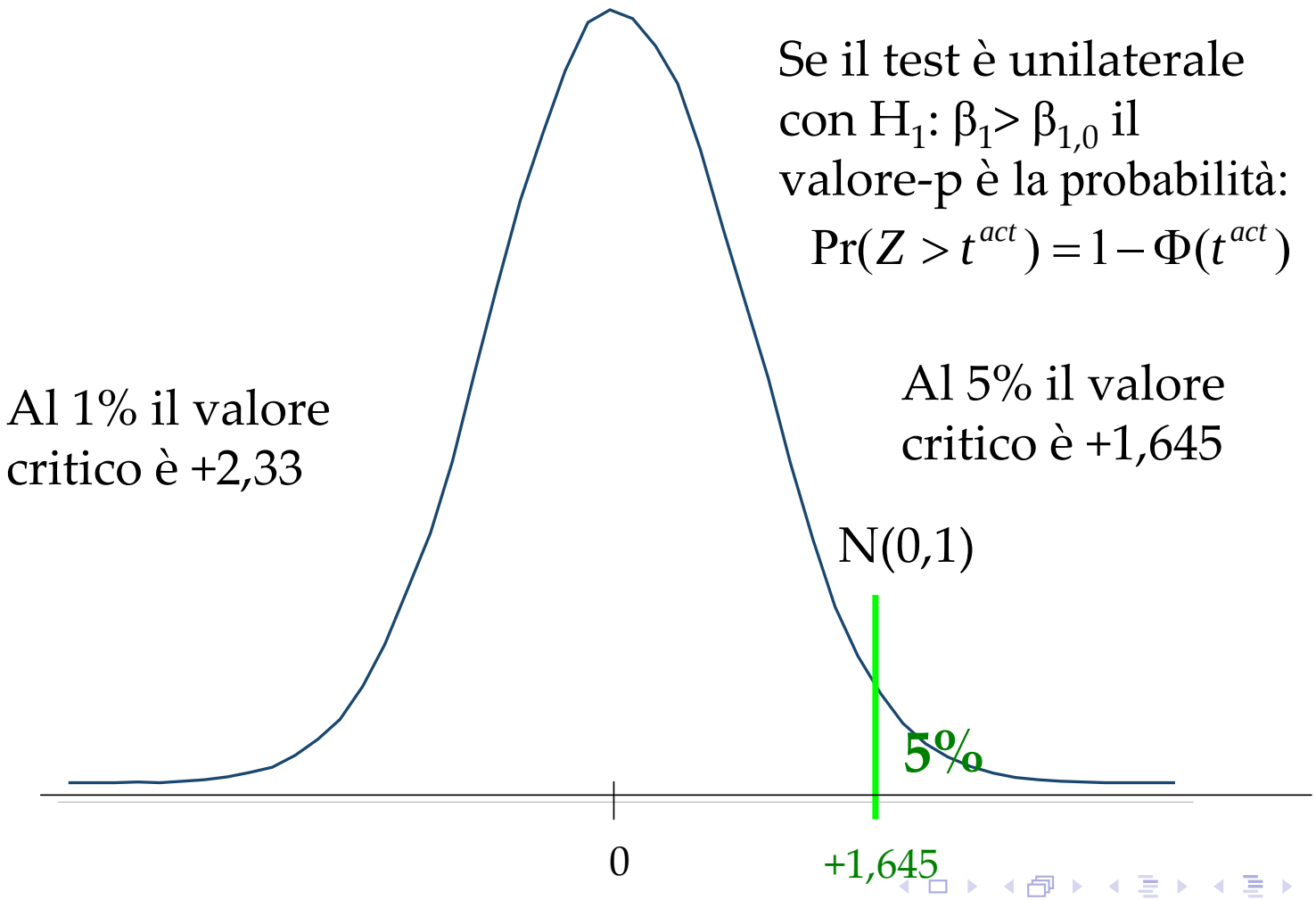
Il valore-p è la probabilità:
$\Pr(Z < t^{act}) = \Phi(t^{act})$

Al 5% il valore critico è –1,645

Il valore critico al 1% è -2,33



Verifica di ipotesi



Verifica di ipotesi sull'intercetta

Il test di ipotesi piu' comune sull'intercetta è di tipo bilaterale:

$$H_0 : \beta_0 = \beta_{0,0} \quad \text{contro} \quad \beta_0 \neq \beta_{0,0}$$

Vedremo successivamente un esempio di tale test

Intervalli di confidenza per coefficienti

L'intervallo di confidenza per β_1 o β_0 è un intervallo che con probabilita pari al livello di confidenza (es: 95%) contiene il vero valore del parametro.

Nel caso di β_1 l'intervallo di confidenza al 95% è pari a:

$$\left[\hat{\beta}_1 - 1,96SE(\hat{\beta}_1), \quad \hat{\beta}_1 + 1,96SE(\hat{\beta}_1)\right]$$

Nel caso di β_0 l'intervallo di confidenza al 95% è pari a:

$$\left[\hat{\beta}_0 - 1,96SE(\hat{\beta}_0), \quad \hat{\beta}_0 + 1,96SE(\hat{\beta}_0)\right]$$

Tali intervalli di confidenza al 95% includono l'insieme di valori che sottoposti a test non rigettano l'ipotesi nulla al livello di significatività al 5%

Regressione con un singolo regressore

Intervallo di confidenza per coefficienti

Nella regressione del punteggio dei test sul rapporto-studenti l'intervallo di confidenza bilaterale al livello di confidenza del 95% è:

$$\left[\hat{\beta}_1 - 1,96SE(\hat{\beta}_1), \hat{\beta}_1 + 1,96SE(\hat{\beta}_1)\right] =$$

$$= [-2,28 - 1,96 \times 0,52; -2,28 + 1,96 \times 0,52] = [-3,30; -1,26]$$

Intervallo di confidenza per variazione dei coefficienti

E' possibile determinare un intervallo di confidenza per l'effetto predetto di una variazione Δx. Ad esempio l'effetto di una riduzione di due studenti

$$\left[\hat{\beta}_1\Delta x - 1,96 \times SE(\hat{\beta}_1) \times \Delta x, \quad \hat{\beta}_1\Delta x + 1,96SE(\hat{\beta}_1) \times \Delta x\right] =$$

$$= [-3,30 \times (-2); -1,26 \times (-2)] = [6,60; 2,52]$$

Regressione con X binaria

La variabile indipendente X puo' essere di tipo binario: ad esempio una variabile "gender" pari a uno se l'individuo è una femmina, pari a 0 se è un maschio. Tale variabile è detta variabile **dummy** Il calcolo del coefficiente $\hat{\beta}_1$ è si ottiene nella stessa maniera.

Esempio: immaginiamo una regressione del punteggio dei test sulla seguente dummy:

$$D_i = \begin{cases} 1 \text{ se il rapporto studenti insegnanti è } < 20 \\ 0 \text{ se il rapporto studenti insegnanti è } \geq 20 \end{cases}$$

La regressione è:

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

Regressione con un singolo regressore

Regressione con X binaria

Se $D_i = 0$ si ha $Y_i = \beta_0 + u_i$ e quindi $E(Y_i|D_i = 0) = \beta_0$
Se $D_i = 1$ si ha $Y_i = \beta_0 + \beta_1 + u_i$ e quindi
$E(Y_i|D_i = 1) = \beta_0 + \beta_1$

$\beta_1 + \beta_0$ è la media di Y_i per la popolazione con $D_i = 1$.

β_0 è la media per la popolazione con $D_i = 0$

$\beta_1 = \beta_1 + \beta_0 - \beta_0$ è la differenza tra le due medie. β_1 è quindi la differenza tra la media campionaria di Y_i tra i due gruppi: il primo con $D_i = 1$, il secondo con $D_i = 0$

Regressione con X binaria

Si puo' testare l'eguaglianza delle due medie testando l'ipotesi $H_0 : \beta_1 = 0$ contro $H_1 : \beta_1 \neq 0$. Si rifiuta H_0 al 5% se $|t| > 1,96$

$$\widehat{TestScore} = \underset{(1,3)}{650,0} + \underset{(1,8)}{7,4D}, \quad R^2 = 0,035, \quad SER = 18,7 \quad \text{n=420}$$

La statistica t è uguale a $4,04 = 7,4/1,8$ ed è maggiore di $1,96$

Si puo' costruire un intervallo di confidenza al 95% pari a:
$7,4 \pm 1,96 \times 1,8 = (3,9; 10,9)$

Regressione con un singolo regressore

Regressione con X binaria

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{n_1 \bar{Y}_1 - n\bar{Y}(n_1/n)}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

dove n_1 è il numero di individui con $X = 1$ e $\bar{Y}_1$ è la media di Y per questi n_1 individui. $\bar{Y}_0$ è la media di Y per gli $n_0 = n - n_1$ individui con $X = 0$.

$\sum_{i=1}^{n}(X_i - \bar{X})^2$ se diviso per n è la varianza di una variabile binaria e pertanto è uguale a $\frac{n_1}{n}\left(1 - \frac{n_1}{n}\right)$

$$\bar{Y} = \frac{n_1}{n}\bar{Y}_1 + \left(1 - \frac{n_1}{n}\right)\bar{Y}_0$$

Regressione con X binaria

$$\hat{\beta}_1 = \frac{(n_1/n)\bar{Y}_1 - (n_1/n)\bar{Y}}{(n_1/n)(1-n_1/n)} =$$

$$\hat{\beta}_1 = \frac{\bar{Y}_1 - \bar{Y}}{(1-n_1/n)} =$$

$$\hat{\beta}_1 = \frac{\bar{Y}_1 - (n_1/n)\bar{Y}_1 - (1-(n_1/n)\bar{Y}_0)}{(1-n_1/n)} =$$

$$\hat{\beta}_1 = \frac{\bar{Y}_1(1-(n_1/n)) - (1-(n_1/n)\bar{Y}_0)}{(1-n_1/n)} =$$

$$= \bar{Y}_1 - \bar{Y}_0$$

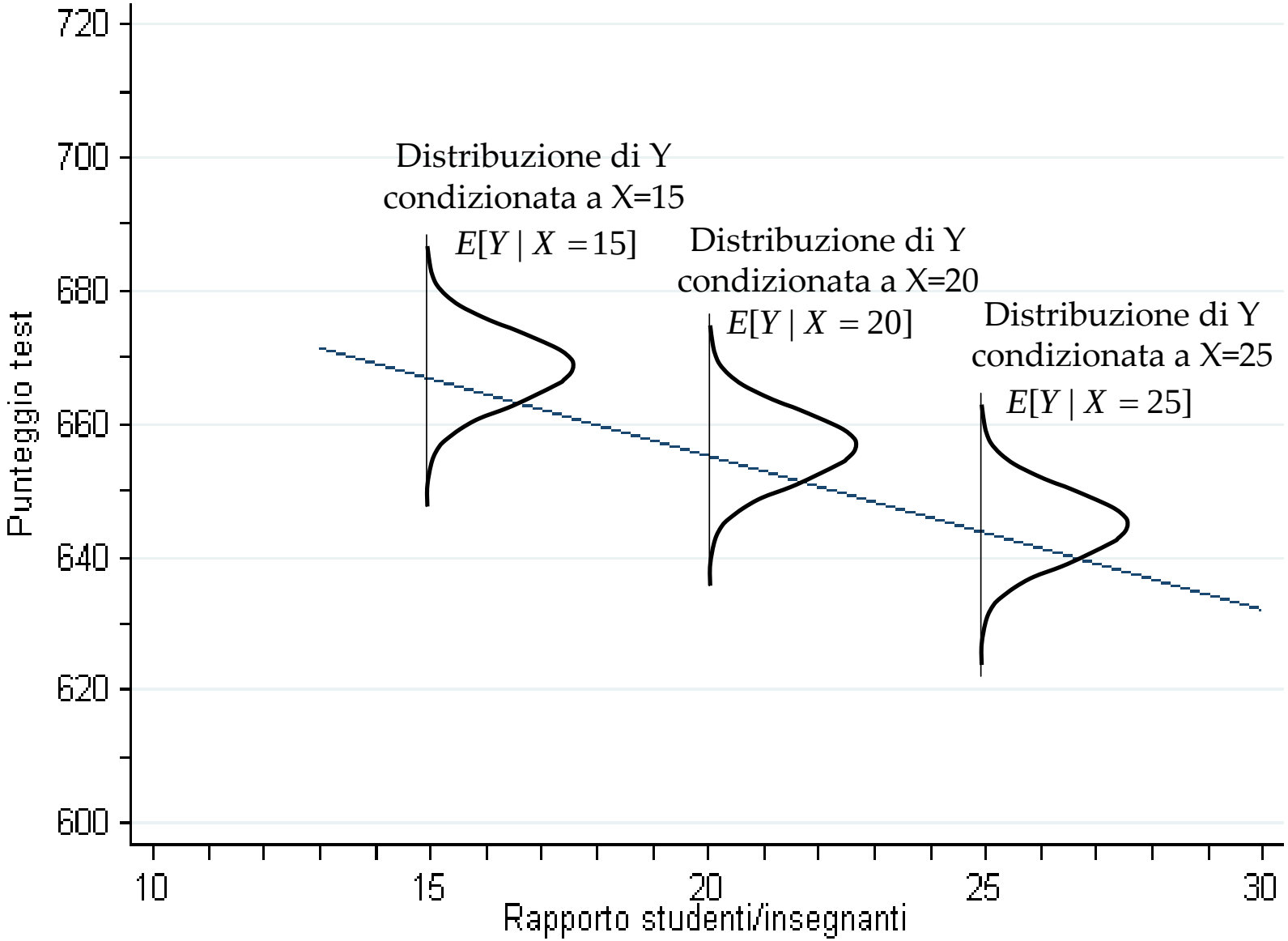
Eteroschedasticità e omoschedasticità

Se la varianza di u_i non dipende da X_1 si puo' dire che gli errori sono **omoschedastici** i.e. se $Var(u_i|X_i = x) = c$ con c costante per $i = 1, \ldots, n$.

Se cio' non fosse vero l'errore è **eteroschedastico**

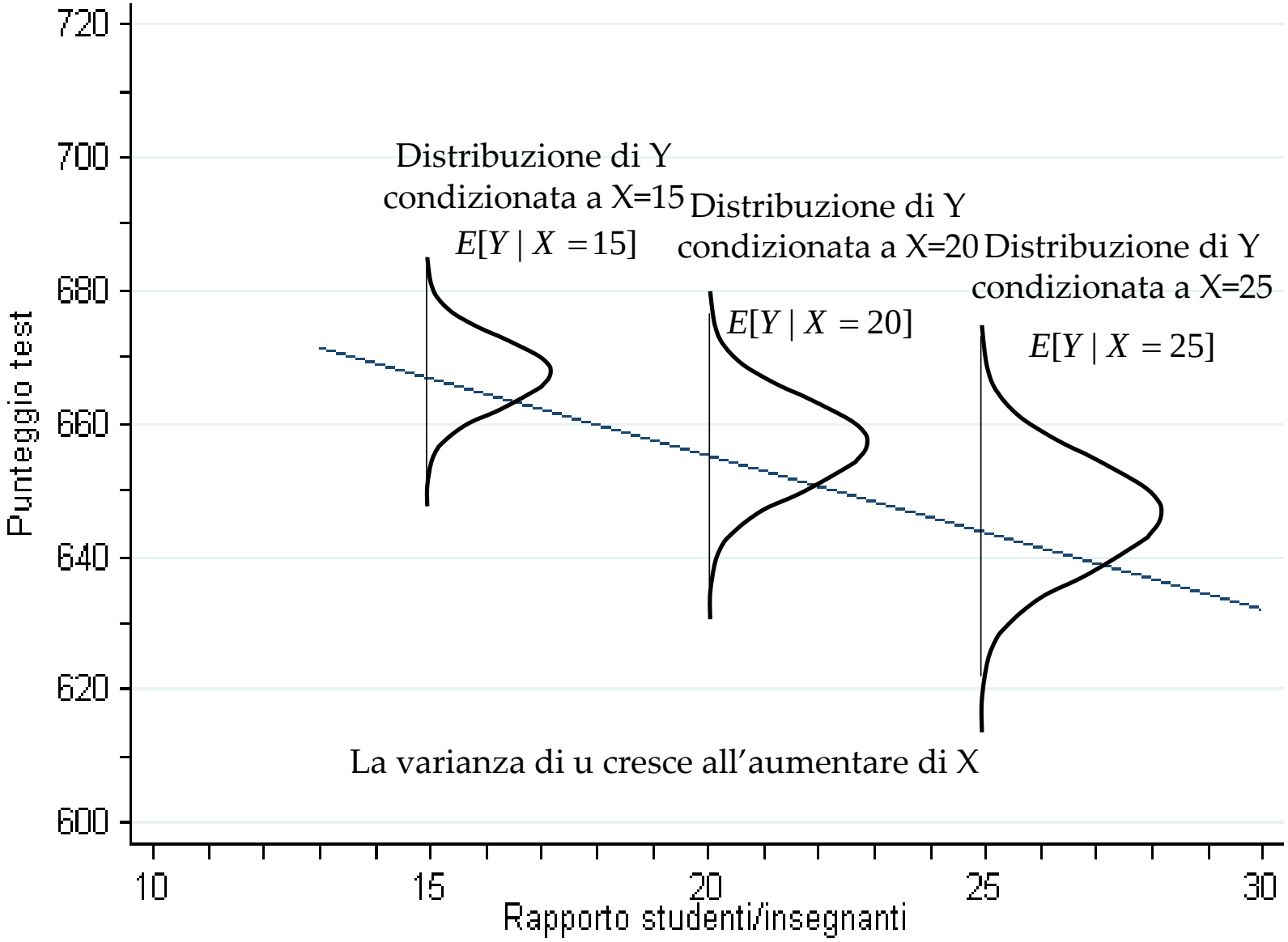
Regressione con un singolo regressore

Omoschedasticità



Regressione con un singolo regressore

Eteroschedasticità



Regressione con un singolo regressore

Eteroschedasticità

Esempio: Immaginiamo di fare una regressione sul reddito da lavoro di uomini e donne.
$Log\_Earnings = \beta_0 + \beta_1 \times MALE_i + u_i$

β_1 indica la differenza delle medie tra uomini e donne.

Se l'errore è omoschedastico la varianza di u_i non dipende dal sesso.

Per testare se la varianza è la stessa si puo' scomporre l'equazione precedente in due parti:
$Log\_Earnings = \beta_0 + u_i$ per le femmine

$Log\_Earnings = \beta_0 + \beta_1 + u_i$ per i maschi

Se la varianza non dipende da MALE allora l'errore è omoschedastico. In caso contrario è eteroschedastico.

Errori standard con omoschedasticità

La varianza di $\hat{\beta}_1$ come già visto è:

$\sigma^2_{\hat{\beta}_1} = \frac{var[(X_i - \mu_X)u_i]}{\{n[var(X_i)]^2\}}$

$var[(X_i - \mu_X)u_i] = E(\{(X_i - \mu_X)u_i - E[(X_i - \mu_X)u_i]\}^2) =$
$E[(X_i - \mu_X)^2 u_i^2] = E[(X_i - \mu_X)^2 var(u_i|X_i)]$

Poichè $E[(X_i - \mu_X)u_i] = 0$ e l'ultima uguaglianza data dalla legge delle aspettative iterate.

Con omoschedasticità $var(u_i|X_i) = \sigma^2_u$. E quindi:

$E[(X_i - \mu_X)^2 \sigma^2_u] = \sigma^2_u \sigma^2_X$.

Sostituendo e semplificando σ^2_X si ha: $\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_u}{n\sigma^2_X}$

Regressione con un singolo regressore

Errori standard con omoschedasticità

Partendo da $\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_u}{n\sigma^2_X}$

σ^2_u è sostituita dalla varianza campionaria $s^2_{\hat{u}}$; $n\sigma^2_X$ è sostituita da $\sum_{i=1}^n (X_i - \bar{X})^2$

Quindi: $\tilde{\sigma}^2_{\hat{\beta}_1} = \frac{s^2_{\hat{u}}}{\sum_{i=1}^n (X_i - \bar{X})^2}$ con $s^2_{\hat{u}} = \frac{1}{n-2} \sum_{i=1}^n \hat{u}^2_i$

In maniera simile da:

$\sigma^2_{\hat{\beta}_0} = \frac{1}{n} \frac{var(H_i u_i)}{[E(H_i^2)]}$, con $H_i = 1 - \frac{\mu_X}{E(X_i^2)} X_i$

si ottiene: $\hat{\sigma}^2_{\hat{\beta}_0} = \frac{E(X_i^2)}{n\sigma^2_X} s^2_{\hat{u}}$.

Sostituendo con le stime campionarie si ottiene:

$\tilde{\sigma}^2_{\hat{\beta}_0} = \frac{\left(\frac{1}{n} \sum_{i=1}^n X_i^2\right) s^2_{\hat{u}}}{\sum_{i=1}^n (X_i - \bar{X})^2}$ con $s^2_{\hat{u}} = \frac{1}{n-2} \sum_{i=1}^n \hat{u}^2_i$

Errori standard con eteroschedasticità

Lo stimatore di: $\sigma^2_{\hat{\beta}_1} = \frac{var[(X_i - \mu_X)u_i]}{\{n[var(X_i)]^2\}}$

si ottiene sostituendo il numeratore con $\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2\hat{u}_i^2$

e il denominatore con: $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

si ottiene quindi: $\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2\hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$

Lo stimatore di: $\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{var(H_i u_i)}{[E(H_i^2)]^2}$, con $H_i = 1 - \frac{\mu_X}{E(X_i^2)}X_i$

è : $\hat{\sigma}^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{H}_i^2\hat{u}_i^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{H}_i^2\right)^2}$ con $\hat{H}_i^2 = 1 - [\bar{X}/\frac{1}{n}\sum_{i=1}^{n}X_i^2]X_i$

Regressione con un singolo regressore

Implicazioni omoschedasticità

Poichè le ipotesi utilizzate fino ad ora non riguardano la varianza di u_i condizionata ad X , gli stimatori OLS sono non distorti, consistenti e asintoticamente normali sia che nel caso gli errori siano eteroschedastici o omoschedastici.

Se valgono le ipotesi:

1. $E[U_i|X_i] = 0$
2. (X_i, Y_i) sono i.i.d. per $i = 1, \ldots, n$
3. $0 < E(X^4) < \infty$ e $0 < E(Y^4) < \infty$

   Piu' l'ipotesi:
4. $Var(u_i|X_i = x) = c$ con c costante per $i = 1, \ldots, n$

Si ottengono le tre condizioni di Gauss-Markov

Implicazioni omoschedasticità

Da queste 4 ipotesi si hanno le tre condizioni di Gauss-Markov:

1. $E[u_i|X_1,\ldots,X_n] = 0$ poichè

   $$E[u_i|X_1,\ldots,X_n] \overbrace{=}^{i.i.d.Hp.2} \overbrace{E[u_i|X_i] = 0}^{Hp.1}$$

2. $Var(u_i|X_1,\ldots,X_n) = \sigma_u^2 \qquad 0 < \sigma_u^2 < \infty$ poichè

   $$Var(u_i|X_1,\ldots,X_n) \overbrace{=}^{HP.2 i.i.d.} Var(u_i|X_i) \overbrace{=}^{hp:omoschedasticità} \sigma_u^2$$

3. $E(u_i u_j|X_1,\ldots,X_n) \overbrace{=}^{hp.2:i.i.d.}$

   $$=\mathrm{E}(\mathrm{u}_i u_j)|X_i,X_j) \overbrace{=}^{hp.2:i.i.d.} E(u_i|X_i)E(u_j|X_j) \overbrace{=}^{hp.1} 0$$

Allora gli stimatori OLS $\hat{\beta}_0$ e $\hat{\beta}_1$ sono i piu' efficienti tra tutti gli stimatori che sono lineari in $Y_1,\ldots,Y_n$.

Regressione con un singolo regressore

Teorema di Gauss-Markov

Se valgono le tre condizioni di Gauss-Markov lo stimatore OLS $\hat{\beta}_1$ è lo stimatore lineare condizionatamente non distorto piu' efficiente tra la classe degli stimatori lineari. i.e. è **BLUE:** Best Linear Unbiased Estimator.

La classe di stimatori lineari non distorti di β_1 sono tutti quegli stimatori che sono funzioni lineari di $Y_1, \ldots, Y_n$. Un generico stimatore lineare di $\tilde{\beta}_1$ puo' essere scritto come:

$$\tilde{\beta}_1 = \sum_{i=1}^{n} a_i Y_i.$$

Tale stimatore è lineare se i pesi a_1 dipendono da $X_1, \ldots, X_n$ ma non da $Y_1, \ldots, Y_n$.

Lo stimatore $\tilde{\beta}_1$ è condizionatamente non distorto se:

$$E(\tilde{\beta}_1 | X_1, \ldots, X_n) = \beta_1$$

Regressione con un singolo regressore

$\hat{\beta}_1$ stimatore lineare OLS

Abbiamo visto che β_1 è uno stimatore non distorto. β_1 è uno stimatore lineare poichè:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})Y_i - \overbrace{\sum_{i=1}^{n}(X_i-\bar{X})\bar{Y}}^{=0}}{\sum_{i=1}^{n}(X_i-\bar{X})^2} =$$

$$= \underbrace{\frac{\sum_{i=1}^{n}(X_i-\bar{X})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}_{=\hat{a}_i} \times Y_i = \sum_{i=1}^{n}\hat{a}_i Y_i$$

Sotto le condizioni di Gauss-Markov la varianza di $\hat{\beta}_1$ condizionata a $X_1, \ldots, X_n$ è:

$var(u_i|X_1, \ldots, X_n) = \sigma_u^2$

Dimostrazione Gauss-Markov

Immaginiamo esista un altro stimatore lineare condizionatamente non distorto $\tilde{\beta}_1 = \sum_{i=1}^{n} a_i Y_i$. Sostituendo il valore di Y_i si ha:
$\tilde{\beta}_1 = \sum_{i=1}^{n} a_i(\beta_0 + \beta_1 X_1 + u_i)$.
$\tilde{\beta}_1 = \beta_0 \left(\sum_{i=1}^{n} a_i\right) + \beta_1 \left(\sum_{i=1}^{n} a_i X_i\right) + \sum_{i=1}^{n} a_i u_i$

Calcolando le aspettative condizionate per entrambi i lati, e utilizzando il fatto che:
$E(\sum_{i=1}^{n} a_i u_i | X_1, \ldots, X_n) = \sum_{i=1}^{n} a_i E(u_i | X_1, \ldots, X_n) = 0$ si ha:
$E(\tilde{\beta}_1 | X_1, \ldots, X_n) = \beta_0 \left(\sum_{i=1}^{n} a_i\right) + \beta_1 \left(\sum_{i=1}^{n} a_i X_i\right)$.

Affinchè lo stimatore sia non distorto deve valere: $\sum_{i=1}^{n} a_i = 0$ e $\sum_{i=1}^{n} a_i X_i = 1$ (ciò vale anche per lo stimatore OLS $\hat{\beta}_1$)

Regressione con un singolo regressore

Dimostrazione Gauss-Markov

Se vale :$\sum_{i=1}^{n} a_i = 0$ e $\sum_{i=1}^{n} a_i X_i = 1$ si ha:

$\tilde{\beta}_1 = \beta_1 + \sum_{i=1}^{n} a_i u_i \quad \Rightarrow \tilde{\beta}_1 - \beta_1 = \sum_{i=1}^{n} a_i u_i$

Si ha quindi: $Var(\tilde{\beta}_1 | X_1, \ldots, X_n) = Var(\sum_{i=1}^{n} a_i u_i | X_1, \ldots, X_n) =$

$= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j cov(u_i u_j | X_1, \ldots, X_n)$

Utilizzando la 2^a e la 3^a condizione di Gauss-Markov si ha:

$Var(\tilde{\beta}_1 | X_1, \ldots, X_n) = \sigma_u^2 \sum_{i=1}^{n} a_i^2$

Questo vale anche per gli $a_i = \hat{a}_i$ degli OLS

Mostriamo ora che ogni $\tilde{\beta}_1$ non distorto diverso dallo stimatore OLS $\hat{\beta}_1$ ha una varianza maggiore della varianza di $\hat{\beta}_1$

Dimostrazione Gauss-Markov

Sia $\tilde{\beta}_1$ diverso da $\hat{\beta}_1$. Avremo quindi che $\tilde{\beta}_1$ utilizza dei pesi $a_i = \hat{a}_i + d_i$ (i.e. diversi da $\hat{a}_i$ dello stimatore OLS) . Avremo quindi:

$\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} (\hat{a}_i + d_i)^2 = \sum_{i=1}^{n} \hat{a}_i^2 + 2\sum_{i=1}^{n} \hat{a}_i d_i + \sum_{i=1}^{n} d_i^2$

$$\sum_{i=1}^{n} \hat{a}_i d_i = \frac{\sum_{i=1}^{n}(X_i - \bar{X})d_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} == \frac{\sum_{i=1}^{n} X_i d_i - \bar{X}\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} =$$

$$= \frac{[(\sum_{i=1}^{n} a_i X_i - \sum_{i=1}^{n} \hat{a}_i X_i)] - \bar{X}\,(\sum_{i=1}^{n} a_i - \sum_{i=1}^{n} \hat{a}_i)}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = 0$$

Poichè sia per a_i sia per $\hat{a}_i$ vale :$\sum_{i=1}^{n} a_i = 0$ e $\sum_{i=1}^{n} a_i X_i = 1$

Dimostrazione Gauss Markov

Si ha quindi:

$var(\tilde{\beta}_1|X_1,\ldots,X_n) = \sigma_u^2 \sum_{i=1}^n a_i^2 = \sigma_u^2 \sum_{i=1}^n \hat{a}_i^2 + \sigma_u^2 \sum_{i=1}^n d_i^2 =$
$var(\hat{\beta}_1|X_1,\ldots,X_n) + \sigma_u^2 \sum_{i=1}^n d_i^2$

e pertanto: $var(\tilde{\beta}_1|X_1,\ldots,X_n) - var(\hat{\beta}_1|X_1,\ldots,X_n) = \sigma_u^2 \sum_{i=1}^n d_i^2$

Ogni $\tilde{\beta}_1$ condizionatamente non distorto diverso da $\hat{\beta}_1$ ha varianza maggiore.

La varianza minima di uno stimatore lineare condizionatamente non distorto sia ha per $d_i = 0$, cioè per lo stimatore OLS $\hat{\beta}_1$ che puo' quindi definirsi **BLUE**

Regressione con un singolo regressore

Teorema di Gauss Markov sulla media campionaria

Il Teorema di Gauss Markov puo' essere utilizzato su una regressione senza β_1 per provare che $\hat{\beta}_0 = \bar{Y}$ è lo stimatore lineare piu' efficiente di $E(Y)$, i.e. la media campionaria è **BLUE**

Standard errors robusti o no?

Poichè gli stimatori:

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

$$\text{e : } \hat{\sigma}^2_{\hat{\beta}_0} = \frac{1}{n} \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{H}_i^2 \hat{u}_i^2}{\left(\frac{1}{n}\sum_{i=1}^{n} \hat{H}_i^2\right)^2}$$

sono validi anche sotto l'ipotesi di omoschedasticità è sempre conveniente usare gli errori standard robusti all'eteroschedasticità o errori standard di Eicker-Huber-White.

Stimatore dei minimi quadrati ponderati

Lo stimatore OLS non è **BLUE** in presenza di eteroschedasticità

Se la formula della varianza di u_i condizionata a X è nota a meno di un fattore costante di proporzionalità si puo' costruire uno stimatore piu' efficiente di quello OLS.

Lo stimatore dei minimi quadrati ponderati trasforma le variabili della regressione pesando l'osservazione i-esima per l'inverso della radice quadrata della varianza di u_i condizionata a X.

Stimatore dei minimi quadrati ponderati

Ipotizziamo quindi che valga: $Var(u_i|X) = \lambda h(X_i)$ con λ costante e la funzione h nota. Attuando la seguente trasformazione delle variabili:

$$\tilde{Y}_i = Y_i / \sqrt{h(X_i)} \qquad \tilde{X}_{0i} = 1 / \sqrt{h(X_i)}$$

$$\tilde{X}_{1i} = X_{1i} / \sqrt{h(X_i)} \qquad \tilde{u}_i = u_i / \sqrt{h(X_i)}$$

si ha: $\tilde{Y}_i = \beta_0 \tilde{X}_{0i} + \beta_1 \tilde{X}_{1i} + \tilde{u}_i$

dove β_0 di $\tilde{X}_{0i}$ è la nuova intercetta. La regressione sulle variabile trasformate con OLS da' lo stimatore dei minimi quadrati ponderati o **WLS: Weighted Least Squares**

Stimatore dei minimi quadrati ponderati

Da tale regressione si ottiene infatti:

$$var(\tilde{u}_i|X_i) = var\left(\frac{u_i}{\sqrt{h(X_i)}}|X_i\right) = \frac{var(u_i|X_i)}{h(X_i)} = \frac{\lambda h(X_i)}{h(X_i)} = \lambda$$

La trasformazione rende $\tilde{u}_i$ omoschedastico. torna quindi a valere il teorema di Gauss-Markov e si ha che **WLS** è **BLUE**.

WLS stimati

In genere non si conosce la funzione h. E' possibile pero' stimarla. Se ad esempio: $Var(u_i|X_i) = \theta_0 + \theta_1 X_i^2$ con i parametri θ_0 e $\theta_1 > 0$ si puo' anche in tale caso trasformare le variabili per rendere l'errore omoschedastico. In tal caso:

1. Si stimano i residui $\hat{u}_i$ dalla regressione OLS di Y_i su X_i
2. Si stima la funzione della varianza condizionata $var(u_i|X_i)$. Si stimano quindi θ_0 e θ_1 dalla regressione di $\hat{u}_i^2$ su X_i^2
3. Si utilizzano le stime per calcolare il valore predetto della varianza condizionata $\widehat{var(u_i|X_i)}$
4. Si divide le variabili dipendenti e indipendenti per la radice quadrata di $\widehat{var(u_i|X_i)}$
5. Si effettua un OLS sulle variabili trasformate

Questo metodo è detto dei WLS stimati o dei feasible WLS.

WLS versus errori standard robusti a eteroschedasticità

Se la forma funzionale della varianza di u_i condizionata a X_i fosse nota i WLS sono piu' efficienti del OLS con errori standard robusti all'eteroschedasticità.

Tuttavia un'incorretta specificazione della forma funzione della varianza condizionata porterebbe a degli standard error da WLS non validi. Gli standard errors robusti all'eteroschedasticità sono asintoticamente corretti anche se non si conosce la forma funzionale della varianza condizionata.

Distribuzioni degli stimatori OLS per piccoli campioni

Abbiamo visto che per grandi campioni $\hat{\beta}_1$ si distribuisce come una normale.

In piccoli campioni la distribuzione esatta degi stimatori dipende dalla distribuzione degli errori.

Ipotizziamo che valgano tutte le ipotesi di minimi quadrati, l'ipotesi di omoschedasticità ed assumiamo inoltre che gli errori siano i.i.d e abbiano una distribuzione normale. Se valgono queste ipotesi $\hat{\beta}_1$ condizionatamente a $X_1, \ldots, X_n$ si distribuisce come:

$N(\beta_1, \sigma^2_{\hat{\beta}_1|X})$ con $\sigma^2_{\hat{\beta}_1|X} = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

Distribuzioni degli stimatori OLS per piccoli campioni

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

si ha che $\hat{\beta}_1$ è una media ponderate di variabili che si distribuiscono normalmente (u_i sono i.i.d. come $N(0, \sigma_u^2)$ e u_i e X_i sono indipendenti). Quindi condizionatamente a $X_1, \ldots, X_n$, $\hat{\beta}_1$ si distribuisce normalmente.

La media di tale distribuzione si ottiene dal fatto che: $E(\hat{\beta}_1 | X_1, \ldots, X_n) = \beta_1$. La varianza di $\hat{\beta}_1$ condizionata a $X_1, \ldots, X_n$ si ottiene da:

$$var(\hat{\beta}_1 | X_1, \ldots, X_n) = var\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} | X_1, \ldots, X_n\right) =$$

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 Var(u_i | X_1, \ldots, X_n)}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sigma_u^2}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2} = \frac{\sigma_u^2}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]} =$$

$$= \sigma^2_{\hat{\beta}_1 | X}$$

Distribuzioni della statistica t per piccoli campioni

La statistica $t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$ si distribuisce come una t di Student con $n-2$ gradi di libertà. Sotto l'ipotesi nulla $\hat{\beta}_1$ si distribuisce come una normale $N(\beta_{1,0}, \sigma^2_{\hat{\beta}_1|X})$. Si ha infatti:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{s^2_{\hat{u}}/(\sum_{i=1}^n (X_i - \bar{X})^2}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\sigma^2_u/(\sum_{i=1}^n (X_i - \bar{X})^2}} \sqrt{\frac{\sigma^2_u}{s^2_{\hat{u}}}} = \frac{(\hat{\beta}_1 - \beta_{1,0})/\sigma_{\hat{\beta}_1|x}}{\sqrt{W/(n-2)}}$$

con $s^2_{\hat{u}} = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2$ e $W = \frac{\sum_{i=1}^n \hat{u}_i^2}{\sigma^2_u}$.

Poichè gli errori sono normali, con varianza σ^2_u si ha che W è una chi-quadro con $n-2$ gradi di libertà.

Quindi $E(W)/(n-2) = (n-2)/(n-2) = 1$.

Poichè $E(\frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2) = \sigma^2_u$, si ha che $s_{\hat{u}^2}$ è uno stimatore non distorto di σ^2_u

t di Student per campioni piccoli

Abbiamo quindi una normale al numeratore e una chi-quadro al denominatore divisa per i suoi gradi di libertà. Essendo le due variabili indipendenti, la statistica t si distribuisce come una t di Student con $(n-2)$ gradi di libertà.

Se gli errori sono omoschedastici e sono distribuiti normalmente, in campioni piccoli per testare le ipotesi deve essere utilizzata la t di Student.