Econometria applicata all'intermediazione finanziaria

Regressione con un singolo regressore

Tiziano Razzolini

Università di Siena.

mail: tiziano.razzolini@unisi.it

4 novembre 2020

Regressione con un singolo regressore

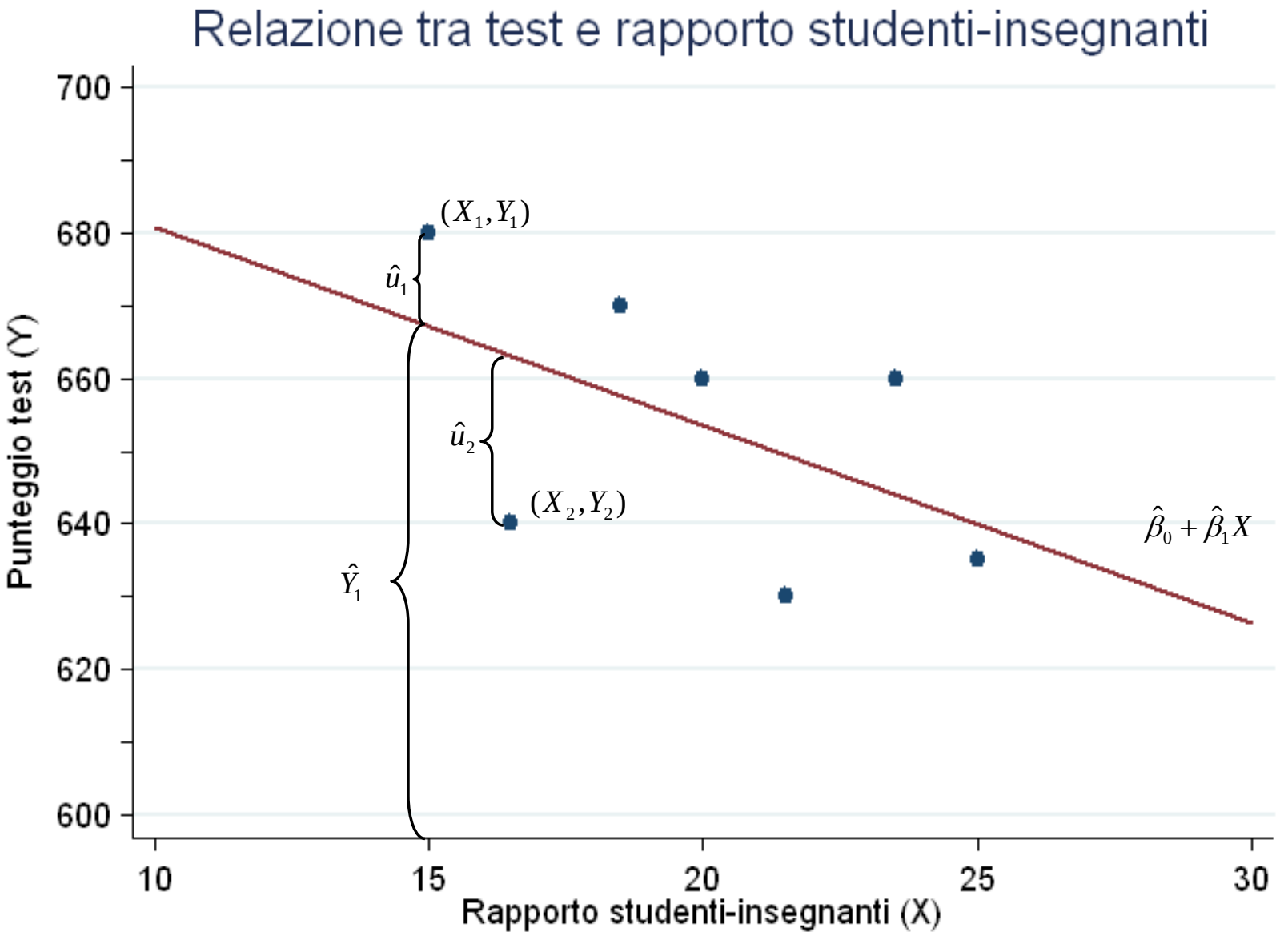
Modello di regressione lineare

Il modello di regressione lineare univariata consiste nella stima di due parametri in una relazione di tipo lineare tra una variabile dipendente Y ed un'unica variabile indipendente X piu' una costante:

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

- Y_i è la variabile dipendente
- X_i è l'unica variabile indipendente
- il pedice $i = 1, ..., n$ indica le n osservazioni
- $\beta_0 + \beta_1 X_1$ è la funzione di regressione della popolazione rapresentabile come una retta.
- β_0 è l'intercetta
- β_1 è la pendenza
- u_i è il termine di disturbo o errore.

Regressione lineare



Regressione con un singolo regressore

Esempio

Esaminare la relazione tra punteggio dei test e numerosità della classe, o meglio rapporto tra numero studenti/insegnanti nel distretto *i* è importante per determinare qual è l'effetto della la numerosità della classe sul punteggio dei test.

La regressione lineare è:

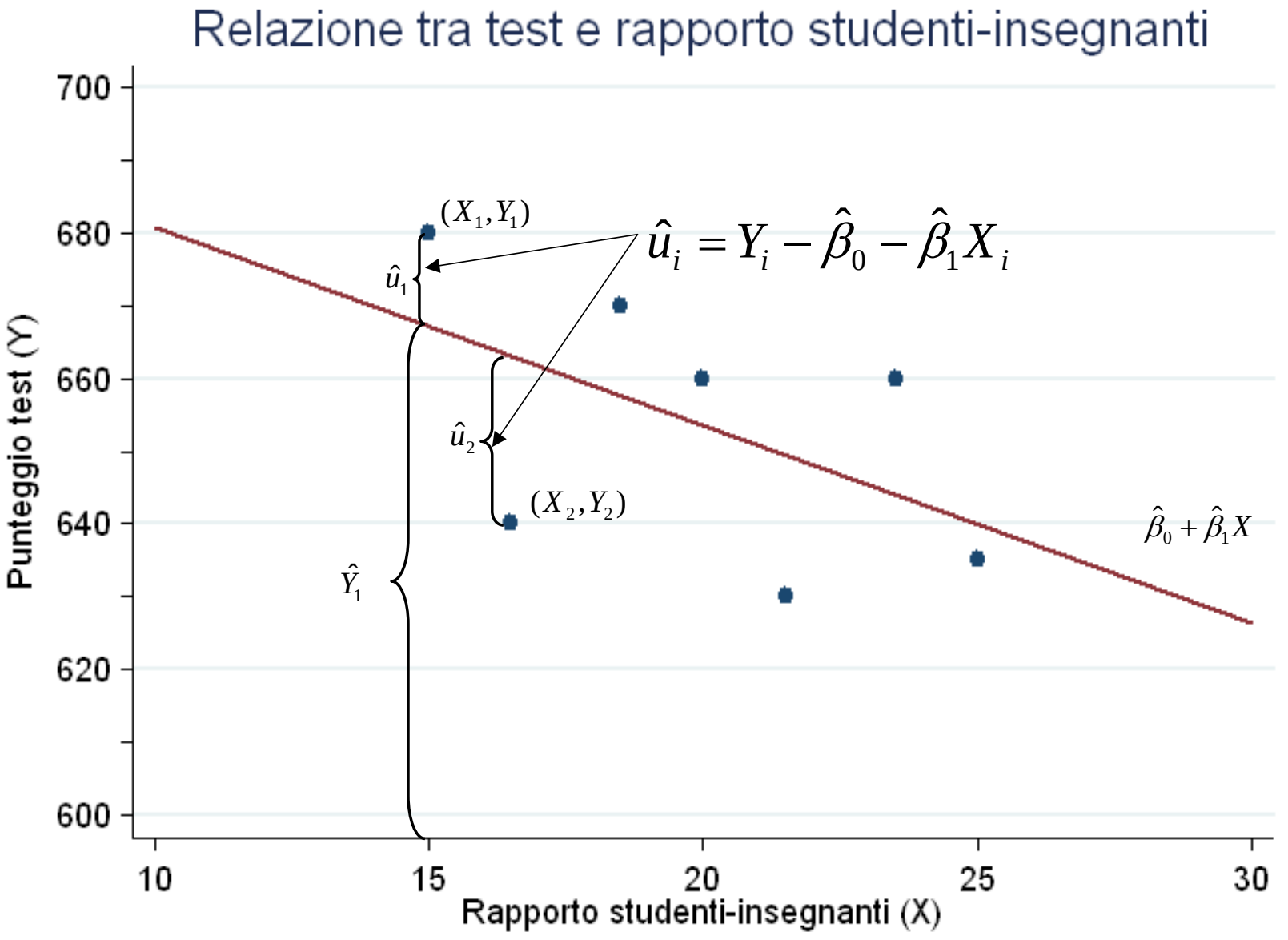
$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize$$

La variabile $\beta_{ClassSize}$ rappresenta la variazione nel punteggio data una variazione unitaria nel rapporto studenti/insegnanti:

$\Delta TestScore = \beta_{ClassSize} \times \Delta ClassSize$ o in altri termini:

$$\beta_{ClassSize} = \frac{\Delta TestScore}{\Delta ClassSize} = \frac{\text{Var. punteggi}}{\text{Var. numerosità}} \qquad \text{Es.: } \beta_{ClassSize} = -0.6$$

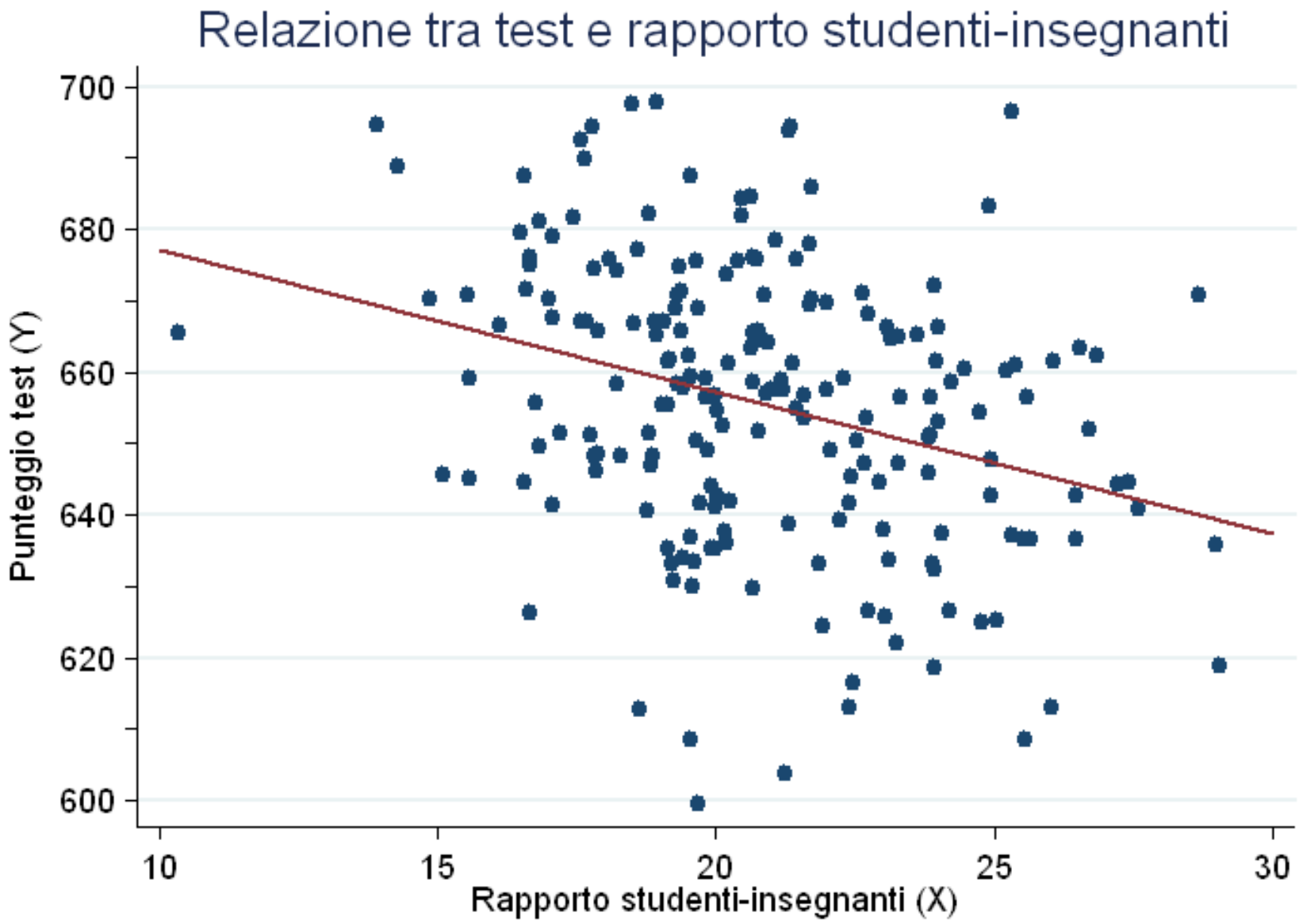
Regressione lineare



Termine di disturbo

Il termine di disturbo o errore u_i raccoglie tutti gli altri fattori, escluso X che contribuiscono a determinare Y_i. Ci potrebbero essere altri fattori rilevanti per la determinazione del punteggio dei test: es: tipo di distretto, qualità insegnanti e testi etc..

Regressione lineare



Come stimare i parametri beta?

Un metodo per stimare i parametri è usare lo stimatore dei minimi quadrati. Avevamo visto che la media campionaria $\bar{Y}_i$ minimizza la somme degli scostamenti al quadrato dello differenze dello stimatore stesso con le osservazioni campionarie. Lo **stimatore dei minimi quadrati ordinari** minimizza la somma per tutte le n osservazioni delle differenza al quadrato tra Y_i ed il valore predetto dalla retta di regressione $\beta_0 + \beta_1 X_i$.

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

dove b_0 e b_1 sono gli stimatori di β_0 e β_1. Questi stimatori sono detti **stimatori dei minimi quadrati ordinari** o anche **stimatori OLS** dove OLS è l'acronimo di Ordinary Least Squares. Gli stimatori OLS sono indicati con $\hat{\beta}_0$ e $\hat{\beta}_1$.

Stimatori OLS

La retta di regressione è ottenuta con gli stimatori OLS $\hat{\beta}_0$ e $\hat{\beta}_1$, cioè $\hat{\beta}_0 + \hat{\beta}_1 X_i$

Il **valore predetto** è $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Il **residuo** è $\hat{u}_i = Y_i - \hat{Y}_i$

Calcolo degli stimatori OLS

Minimizziamo la somma degli errori al quadrato calcolando le derivate parziali rispetto a b_0 e b_1 e ponendole pari a zero.

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

$$\frac{\partial}{\partial b_0} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 = -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = 0$$

$$\frac{\partial}{\partial b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 = -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) X_i = 0$$

Regressione con un singolo regressore

Calcolo degli stimatori OLS

1) $-2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0 \Leftrightarrow n\bar{Y} - nb_0 - b_1 n\bar{X} = 0 \Leftrightarrow$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

2) $-2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) X_i = 0 \Leftrightarrow$

$$\sum_{i=1}^{n} X_i Y_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 =$$
$$= \sum_{i=1}^{n} X_i Y_i - (\bar{Y} - b_1 \bar{X}) \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 =$$
$$= \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} + nb_1 \bar{X}^2 - b_1 \sum_{i=1}^{n} X_i^2 = 0$$

Notate che: $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$
e che $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$

Calcolo degli stimatori OLS

Quindi si ha:

$$2) -2\sum_{i=1}^{n}\left(Y_i - b_0 - b_1 X_i\right) X_i = 0 \Leftrightarrow$$

$$= \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} + nb_1\bar{X}^2 - b_1\sum_{i=1}^{n} X_i^2 =$$

$$= \sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) - b_1\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = 0$$

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

Se dividiamo entrambi i membri per $(n-1)$ si ha:

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$$

Proprietà degli OLS

- $\sum_{i=1}^{n} \hat{u}_i = 0$. Si ha infatti: $\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$. Sostituendo $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ si ha che:

  $\hat{u}_i = Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i$ facendo la sommatoria si ha:

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

  Entrambi i termini sono pari zero.

- $\frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i = \bar{Y}$ si ha infatti che: $Y_i = \hat{Y}_i + \hat{u}_i$ e quindi

  $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i + \sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} \hat{Y}_i$

Regressione con un singolo regressore

Proprietà degli OLS

- $\sum_{i=1}^{n} \hat{u}_i X_i = 0$ e $s_{\hat{u}X} = 0$

Vale infatti: $\sum_{i=1}^{n} \hat{u}_i X_i = \sum_{i=1}^{n} \hat{u}_i \left(X_i - \bar{X}\right) =$

$= \sum_{i=1}^{n} \hat{u}_i X_i - \bar{X} \sum_{i=1}^{n} \hat{u}_i$

$\sum_{i=1}^{n} \hat{u}_i X_i = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right) \left(X_i - \bar{X}\right) =$

$= \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) \left(X_i - \bar{X}\right) =$

$= \sum_{i=1}^{n} \left(Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i\right) \left(X_i - \bar{X}\right) =$

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

Dato che: $\hat{\beta}_1 = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right) \left(X_i - \bar{X}\right) / \sum_{i=1}^{n} (X_i - \bar{X})^2$

L'R^2 della regressione

L'R^2 della regressione è un indice della qualità della regressione.

L'R^2 è costruito come la frazione della varianza campionaria di Y_i spiegata da X_i . Ricordando che:

$$Y_i = \hat{Y}_i + \hat{u}_i$$

Si ha che l'R^2 è il rapporto tra la varianza campionaria di $\hat{Y}_i$ e la varianza campionaria di Y_i

L'R^2 della regressione

Definiamo come somma dei quadrati totale o **TSS** (**Total Sum of Squares**):

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

La somma dei quadrati spiegata o **ESS** (**Explained Sum of Squares**):

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

L'R^2 è pari a:

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

L'R^2 della regressione

L'R^2 puo' essere scritto anche come:

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

dove $SSR = \sum_{i=1}^{n} \hat{u}_i^2$

L'R^2 è compreso tra 0 ed 1.

Valori prossimi ad uno indicano che X_i spiega molto bene la variazione di Y_i; valori prossimi a zero indicano che X_i spiega poco o nulla della variazione di Y_i

L'R^2 della regressione

Si ha che : $TSS = ESS + SSR$. Infatti:

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}\right)^2 =$$

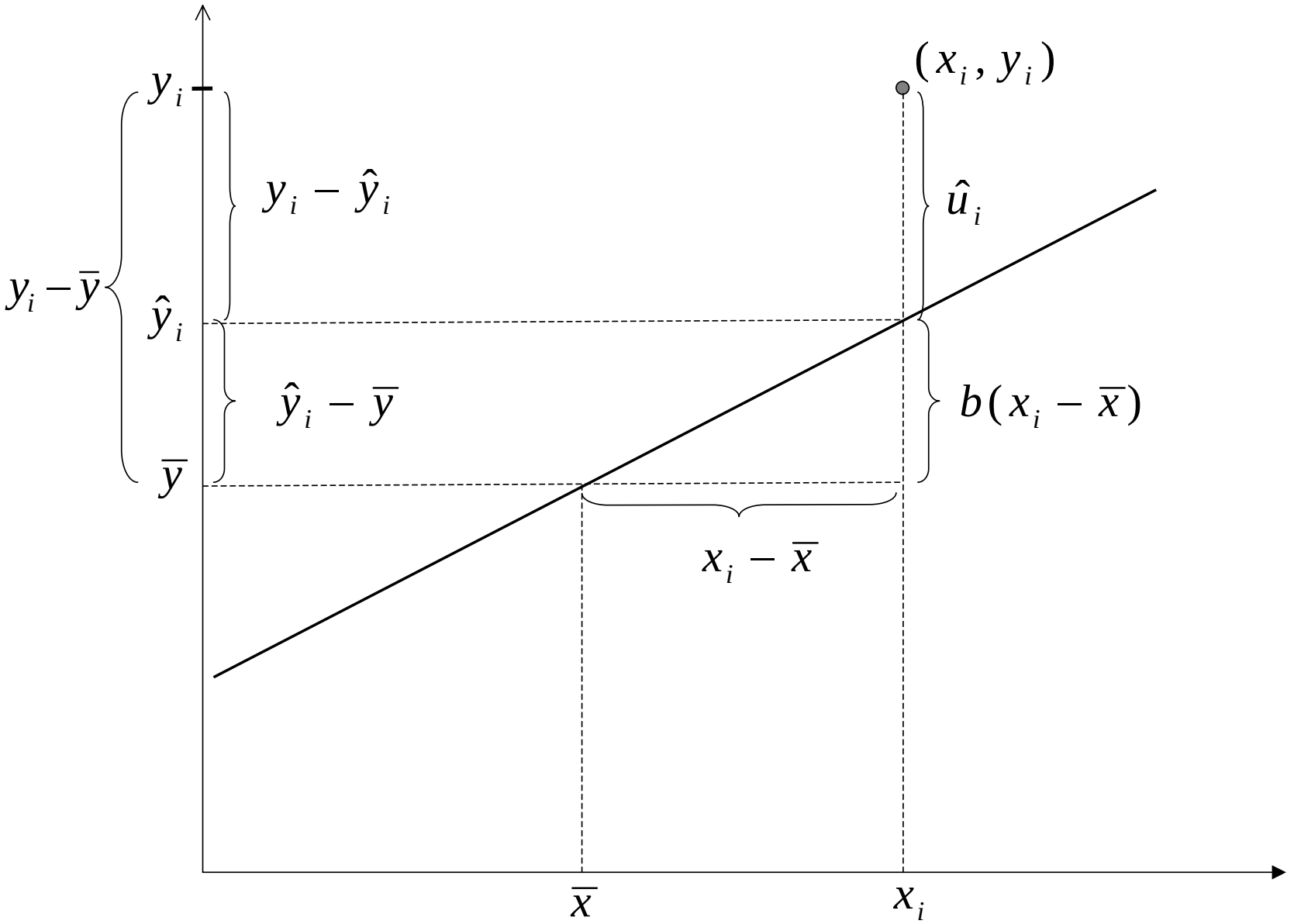
$$= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) =$$

$$= SSR + ESS$$

Dato che : $2\sum_{i=1}^{n} \hat{u}_i \left(\hat{Y}_i - \bar{Y}\right) = 2\sum_{i=1}^{n} \hat{u}_i \hat{Y}_i - 2\bar{Y}\sum_{i=1}^{n} \hat{u}_i =$

$$2\sum_{i=1}^{n} \hat{u}_i(\hat{\beta}_0 + \hat{\beta}_1 X_i) = 2\hat{\beta}_1 \sum_{i=1}^{n} \hat{u}_i X_i = 0$$

Regressione lineare



Errore standard della regressione

L'errore standard della regressione (**SER: Standard Error Regression**) è lo stimatore della deviazione standard dell'errore di regressione.

$$SER = s_{\hat{u}} \qquad s_{\hat{u}}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2 = \frac{SSR}{n-2}$$

Si tratta di una deviazione dato che $\sum_{i=1}^{n} \hat{u}_i = 0$. I gradi di libertà sono $n-2$ poichè sono stati stimati i due parametri β_0 e β_1. Poichè u_i e Y_i hanno la stessa unità di misura la SER misura la dispersione delle osservazioni attorno alla retta di regressione.

Ipotesi dei minimi quadrati

Assunzione 1:

La distribuzione condizionata di u_i data X_i è nulla Il valore atteso di u_i condizionata a X_i è pari a zero.

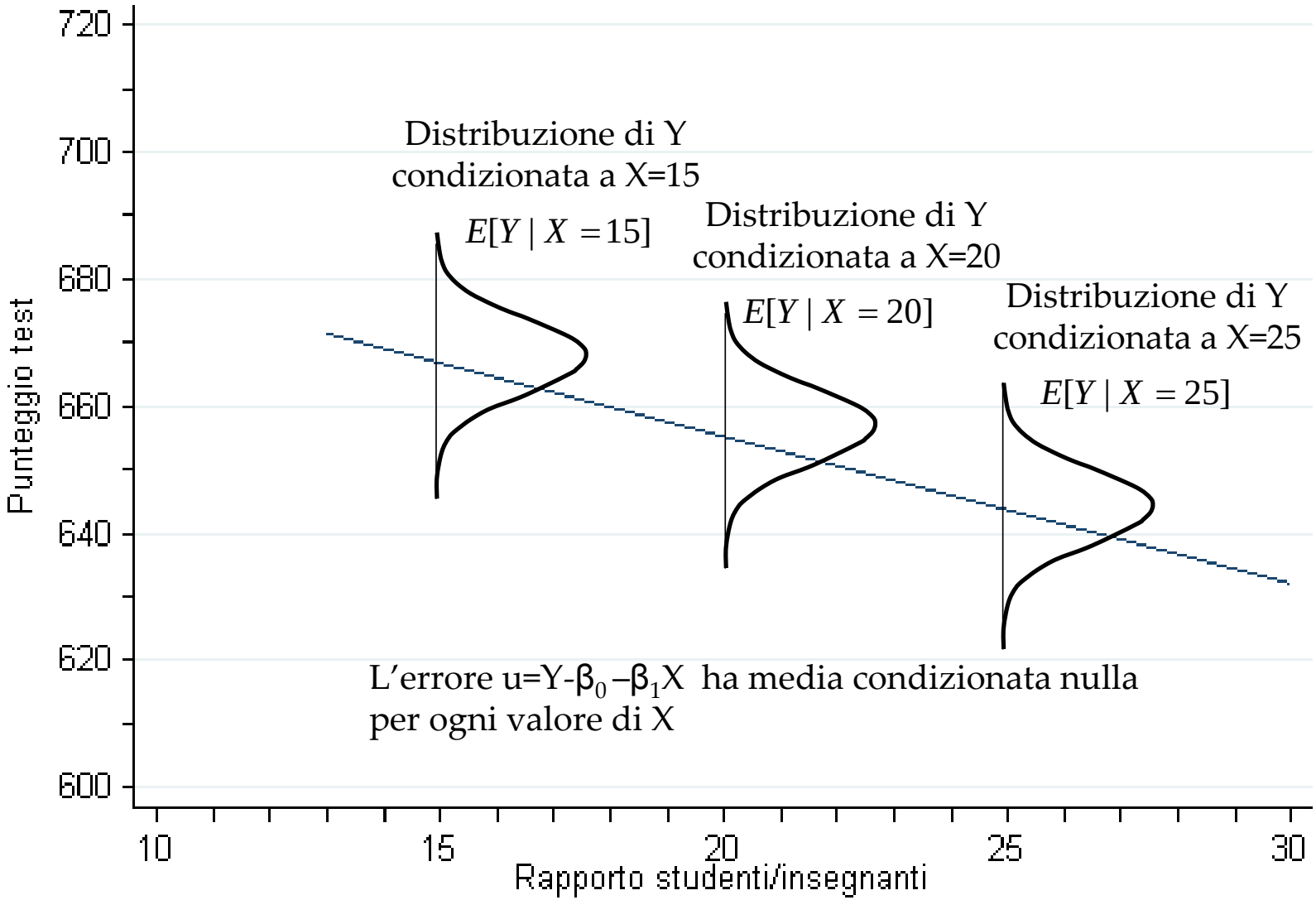
$$E[u_i|X_i] = 0 \qquad \text{oppure} \qquad E[u_i|X_i = x] = 0$$

Cio' significa che gli altri fattori inclusi in u_i che contribuiscono a determinare Y_i hanno una distribuzione condizionata a X con media nulla.

Abbiamo visto che $E[u_i|X_i] = 0$ implica $Corr(u_i, X_i) = 0$; il contrario non è necessariamente vero. Pero' se X_i e u_i fossero correlati la media condizionata $E[u_i|X_i]$ sarebbe necessariamente non nulla.

Regressione con un singolo regressore

Retta di regressione e distribuzioni di probabilità condizionate



Ipotesi dei minimi quadrati

Assunzione 1:

La distribuzione condizionata di u_i data X_i è nulla

Con dati non sperimentali tale ipotesi equivale ad affermare che è come se la variabile X fosse assegnata casualmente. E' difficile immaginare situazioni reali in cui tale ipotesi si verifichi.

Ipotesi dei minimi quadrati

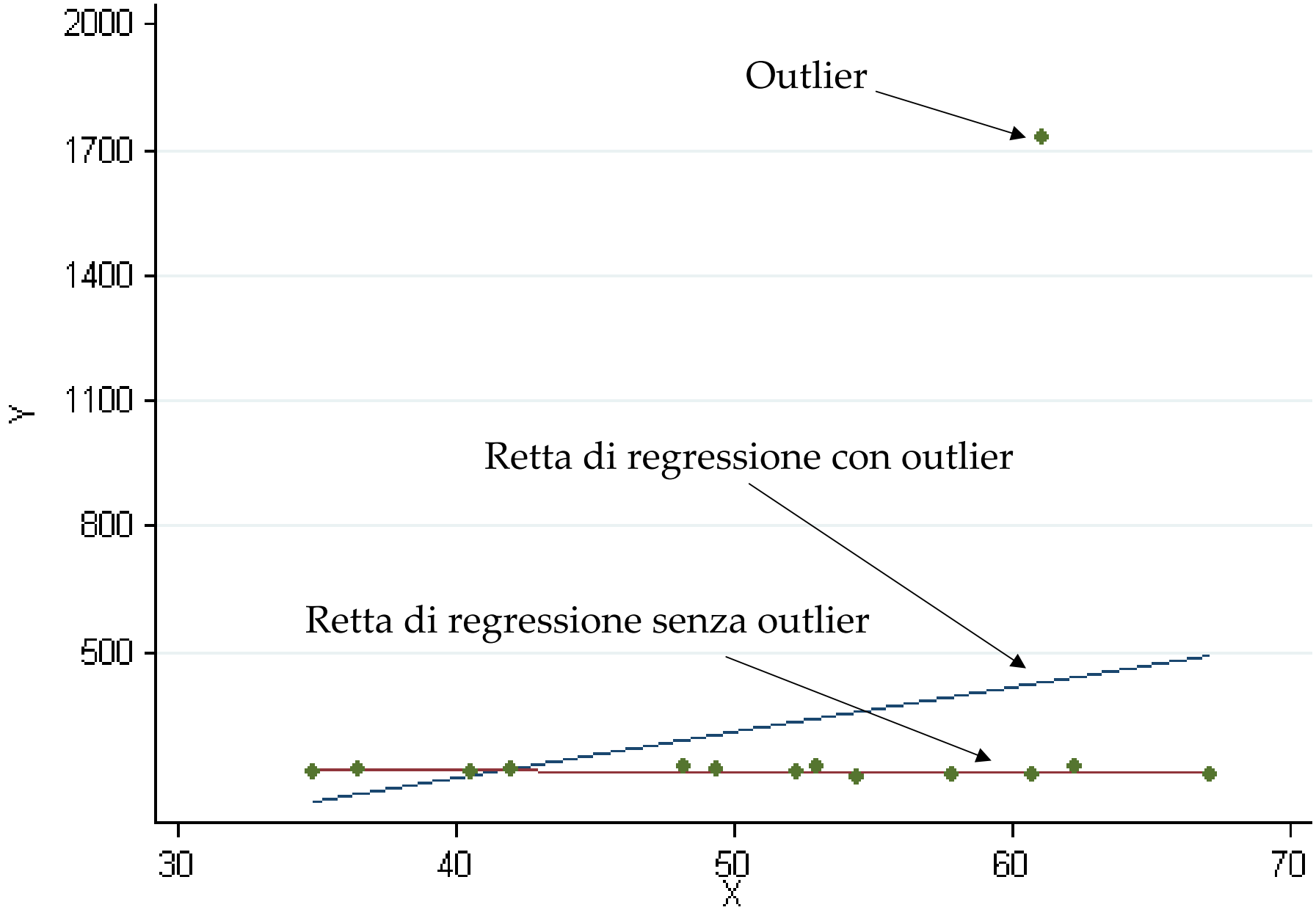
Assunzione 2:

$(X_i, Y_i), i = 1, \ldots, n$ sono indipendentemente e identicamente distribuite.

Tale ipotesi è soddisfatta se le osservazione sono estratte in maniera casuale: (X_i, Y_i) sono indipendentemente e identicamente distribuite.

Nel caso di serie temporali le osservazioni vicine nel tempo non sono indipendentemente distribuite.

Regressione con un singolo regressore



Ipotesi dei minimi quadrati

Assunzione 3:

La presenza di outlier è improbabile

In termini formali cioè equivale a dire che i momenti quarti sono finiti e non nulli:

$0 < E(X^4) < \infty$ e $0 < E(Y^4) < \infty$

o che X ed Y hanno curtosi finita.

In genere la presenza di outlier è dovuta ad errori nella raccolta dati. Un'analisi accurata dei dati e delle osservazioni estreme permette di identificare e correggere tali osservazioni.

Distribuzione campionaria degli OLS

Correttezza: La correttezza degli stimatori $\hat{\beta}_0$ e $\hat{\beta}_1$ vale per qualunque valore di n (anche per piccolo campioni).

$$E(\hat{\beta}_0) = \beta_0 \qquad E(\hat{\beta}_1) = \beta_1$$

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

Abbiamo che: $Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + u_i - \bar{u}$

Quindi:

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

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

Regressione con un singolo regressore

Distribuzione campionaria degli OLS

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

$$\beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2} - \underbrace{\frac{\bar{u}\sum_{i=1}^{n}(X_i-\bar{X})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}_{=0}$$

$$E(\hat{\beta}_1) = \beta_1 + E\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}\right] =$$

usando la legge delle aspettative iterate e l'assunzione 1 si ha:

$$E(\hat{\beta}_1) = \beta_1 + E\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})E(u_i|X_i)}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}\right] = \beta_1$$

=

Distribuzione campionaria degli OLS

Lo stimatore $\hat{\beta}_1$ è non distorto

$$E(\hat{\beta}_1) = \beta_1$$

$$E(\hat{\beta}_1 - \beta_1) = 0$$

Anche $\hat{\beta}_0$ è non distorto:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = (\beta_0 + \beta_1 \bar{X} + \bar{u}) - \hat{\beta}_1 \bar{X} = \beta_0 + \bar{X}(\hat{\beta}_1 - \beta_1) + \bar{u}$$

Prendendo il valore atteso si ha:

$$E(\hat{\beta}_0) = \beta_0$$

Distribuzione degli stimatori OLS in grandi campioni

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

$\bar{X}$ è consistente per μ_X. Il numeratore è quindi:
$\bar{v} = \frac{1}{n}\sum_i^n (X_i - \mu_X)u_i = \frac{1}{n}\sum_i^n v_i$. Per la prima ipotesi $\bar{v}$ ha media nulla. La varianza di v_i è $\sigma_v^2 = var[(X_i - \mu_X)u_i]$ che per la terza assunzione è finita.

Vale quindi il teorema del limite centrale per $\bar{v}/\sigma_{\bar{v}} = \bar{v}/(\sqrt{\sigma_v^2/n})$ che per grandi campioni di distribuisce come $N(0, \sigma_{\bar{v}}^2)$. Il denominatore è la varianza campionaria di X divisa n invece di $(n-1)$; al crescere di n tale differenza diminuisce. La varianza campionaria è uno stimatore consistente della varianza della popolazione.

Distribuzione degli stimatori OLS in grandi campioni

Quindi $\hat{\beta}_1$ si distribuisce in grandi campioni come $N(\beta_1, \sigma^2_{\hat{\beta}_1})$ con $\sigma^2_{\hat{\beta}_1} = var(\bar{v})/[var(X_i)]^2 = var[(X_i - \mu_X)u_i]/\{n[var(X_i)]^2\}$

In maniera simile si ottiene che in grandi campioni di β_0 è $N(\beta_0, \sigma_{\hat{\beta}_0})$ con:

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

La varianze $\sigma^2_{\hat{\beta}_1}$ e $\sigma^2_{\hat{\beta}_0}$ tendono a zero per n grande.

La varianza $\sigma^2_{\hat{\beta}_1}$ è inversamente proporzionale alla varianza di X_i, quindi maggiore è la varianza di X piu' preciso è $\hat{\beta}_1$