Statistica per l'impresa

6. Misura delle relazioni tra variabili

Correlazione e regressione

Affrontiamo l'analisi delle relazioni tra variabili di interesse da due diversi punti di vista:

- Visualizzare e sintetizzare il legame tra due o più variabili di interesse (analisi della correlazione)
- *Spiegare* l'andamento di una *variabile obiettivo* mediante le informazioni su una o più *variabili esplicative* (analisi di regressione)

Esempi di relazioni "interessanti":

- assenze dal lavoro e qualifiche professionali, e/o anzianità
- incidenti sul lavoro e orario, e/o età del lavoratore
- costo degli input e quantità prodotte
- vendite e spese di promozione
- ...

Deams

Campioni bi- (multi-) variati

Consideriamo dunque (almeno) due variabili con un indice comune:

i	X	Y
1	x_1	y_1
2	x_2	y_2
...	...	...
i	x_i	y_i
...	...	...
n	x_n	y_n

Per esempio, consideriamo il volume totale della produzione (Y) e il corrispondente costo (X) di un'azienda alimentare, misurati negli stabilimenti produttivi di 22 diversi centri (Esempio 6.1)

Analisi grafica della correlazione

La *correlazione* può essere misurata per mezzo di indici sintetici. E' sempre opportuno, tuttavia, affrontare il problema partendo da una visualizzazione dei dati su un *diagramma di dispersione* o *scatterplot*, dove ogni punto rappresenta, nel piano definito dalle due caratteristiche (X, Y), la coppia di osservazioni (x_i, y_i)

Analisi e misura della correlazione

Il momento generalmente usato per misurare l'associazione statistica tra due variabili è la *covarianza*: ovvero la media dei prodotti degli scarti dalle medie individuali.
Distinguiamo la *covarianza della popolazione*

$$\frac{\sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X})}{N}$$

dalla *covarianza campionaria (corretta)*

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1}$$

La seconda è uno stimatore campionario corretto (e consistente) della prima.

Analisi e misura della correlazione

La covarianza dipende dall(e) unità di misura delle variabili. Essa può essere *standardizzata* dividendola per il prodotto dei rispettivi errori standard: denotando questi ultimi $\sigma_x = \sqrt{Var(x)}$ e $\sigma_y = \sqrt{Var(y)}$, il *coefficiente di correlazione di Pearson*

$$\rho_{xy} = \frac{Cov(x, y)}{\sigma_x \cdot \sigma y}$$

è un numero puro (indipendente dall'unità di misura) compreso tra -1 e 1. Nella popolazione, è quindi:

$$\rho_{xy} = \frac{\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n}}{\sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}} \sqrt{\frac{\sum_i (y_i - \bar{y})^2}{n}}}$$

La correlazione campionaria: stima e inferenza

La correlazione nella popolazione può essere stimata con lo stimatore campionario (corretto)

$$r_{xy} = \frac{\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}} \sqrt{\frac{\sum_i (y_i - \bar{y})^2}{n-1}}}$$

La correlazione campionaria è una variabile aleatoria r_{xy}, funzione del campione bivariato $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$. Come tale essa ha un'errore standard che – *solo se* $\rho = 0$ – è dato da:

$$ES_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n-2}}$$

e che, sotto opportune ipotesi di normalità congiunta sulla distribuzione di X, Y, può essere usato per verificare ipotesi su ρ.

Proprietà utili di (medie) varianze e covarianze

Per definizione,

$$Cov(X, X) = Var(X)$$

Trasformazioni lineari: se $Z = a + bX$ è

$$E(Z) = a + b \cdot E(X)$$

$$Var(Z) = b^2 Var(X)$$

Inoltre,

$$Cov(X, Y) = E(X \cdot Y) - E(X) \cdot E(Y)$$

e, caso particolare,

$$Var(X) = E(X^2) - [E(X)]^2$$

Scienza induttiva e falsificazionismo

Secondo Karl Popper (1902-1994):

- La mente umana sovrappone alle osservazioni i propri schemi mentali (*teorie*). I *fatti* sono indistinguibili dalle *opinioni*, cosicché un processo puramente induttivo è impossibile.
- Le teorie scientifiche non sono suscettibili di *verifica* ma soltanto di *falsificazione*. Ogni teoria scientifica è pertanto un'approssimazione alla realtà frutto di un processo di prova ed errore, e verrà mantenuta finché non venga smentita dall'osservazione empirica.
- La *falsificabilità* è il criterio che definisce la *scienza* e la distingue dalle teorie non scientifiche.

In particolare, ogni teoria economica con pretesa di scientificità non può prescindere dalla verifica empirica, che assumerà la veste di *non falsificazione*. La statistica fornirà lo strumento per trarre dai fenomeni collettivi eventuali smentite alle ipotesi teoriche.

La verifica di ipotesi - 1

La verifica (*test*) di ipotesi statistiche consiste nel

- formulare un'ipotesi sulla popolazione di interesse
- tradurla in termini di uno o più parametri (incogniti) della popolazione
- estratto un campione, valutare se tale ipotesi è supportata dai dati

Il fenomeno studiato deve essere rappresentabile con una distribuzione di probabilità definita da *parametri*. A questo punto,

- si specificano:
  - l'ipotesi di interesse (detta *ipotesi nulla*, o H_0)
  - e l'ipotesi *alternativa*, o H_A

  in termini del parametro, o dei parametri, di interesse
- si considera una *statistica test*, la cui distribuzione è nota sotto H_0
- si estrae un campione, si calcola il valore assunto dalla statistica test e se ne valuta la coerenza con l'ipotesi di partenza. Come?

La verifica di ipotesi - 2

La procedura di verifica si basa sulla distribuzione di probabilità che *assumerebbe* la statistica test τ *se* H_0 *fosse vera*.
Data questa,

- si fissa il *livello di confidenza* α del test (NB confidence=fiducia) come una probabilità "sufficientemente piccola": molto spesso è $\alpha = 5\%$
- sulla base della distribuzione della statistica test τ sub H_0, si calcolano i confini tra:
  - *regione di accettazione*, dove sub H_0 τ cade con probabilità $1 - \alpha$, e
  - *regione di rifiuto*, dove τ ha una probabilità α ("molto piccola"!) di cadere se H_0 è vera

  si estrae il campione, si calcola il valore assunto da τ
  - se questo cade nella regione di accettazione, *non si rifiuta* l'ipotesi H_0
  - se cade nella regione di rifiuto, *si rifiuta* H_0

La verifica di ipotesi - esempio 1

Verifichiamo un'ipotesi sulla media di una popolazione (es. X=statura degli studenti). Assumiamo che nella popolazione X si distribuisca secondo una legge ignota la cui media sia il parametro μ, a sua volta incognito; e di essere in grado di estrarre dalla popolazione un campione casuale "abbastanza grande" (es. 100 unità).

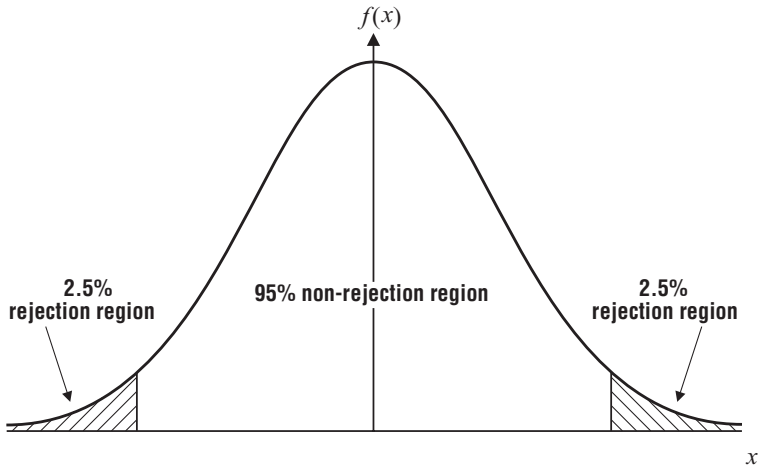
- vogliamo verificare $H_0 : \mu = 180$ al livello di confidenza del 5%

Scegliamo una statistica test di cui *sotto* H_0 conosciamo la distribuzione:

- per campioni "abbastanza grandi" la *media campionaria* $\bar{Y}$ si distribuisce come una Normale (th. Limite Centrale)
- essa è uno stimatore corretto, pertanto *sub* H_0 il suo valore atteso è 180
- disponiamo di uno stimatore per $ES_{\bar{Y}}$ sulla base del campione estratto, pertanto la distribuzione *sub* H_0 di τ è interamente descritta

La verifica di ipotesi - esempio 1 (cont.)

A questo punto i limiti della regione di accettazione coincidono con l'intervallo di confidenza al 5% per la media campionaria centrato su 180:



$$180 - z_{\frac{0.05}{2}} \cdot \hat{ES}_{\bar{Y}}; 180 + z_{\frac{0.05}{2}} \cdot \hat{ES}_{\bar{Y}}$$

Confrontiamo la media del campione effettivamente estratto con la distribuzione sub H_0: se cade nella regione di rifiuto, delle due l'una:

- H_0 è vera ma siamo stati molto sfortunati (errore di I specie)
- H_0 è falsa

Il test t

E' del tutto equivalente, ma più comodo, standardizzare la statistica test

- sottraendo il valore atteso sub H_0 in modo da centrare la distribuzione sullo zero
- dividendo per l'errore standard (stimato) in modo di scalare la varianza ad 1

Si ottiene così una statistica nota come $t - test$. Per una generica ipotesi $H_0 : \mu = m^*$

$$t = \frac{\hat{\mu} - m^*}{\hat{ES}(\hat{\mu})} \sim N(0, 1)$$

per campioni "abbastanza grandi". Altrimenti, per piccoli campioni, occorre affidarsi a una ulteriore ipotesi di normalità della popolazione di indagine. In questo caso,

$$t \sim t_{n-1}$$

Intervalli di confidenza e test di ipotesi

Usando un test t, e detti in generale t_{crit} i valori critici al livello α (p. es. $t_{crit} = z_{\frac{\alpha}{2}}$ in campioni "grandi"), H_0 non sarebbe rifiutata se la statistica test cade nella regione di "accettazione", ovvero se

$$-t_{crit} \leq \frac{\hat{\mu} - m^*}{ES(t)} \leq +t_{crit}$$

Equivalentemente,

$$-t_{crit} \times ES(\hat{\mu}) \leq \hat{\mu} - m^* \leq +t_{crit} \times ES(\hat{\mu})$$
$$\hat{\mu} - t_{crit} \times ES(\hat{\mu}) \leq m^* \leq \hat{\mu} + t_{crit} \times ES(\hat{\mu})$$

L'ipotesi nulla H_0 non sarà rifiutata al livello α se l'intervallo di confidenza stimato per il parametro incognito *contiene* il valore ipotizzato.

La verifica di ipotesi - esempio 2

Consideriamo la correlazione ρ tra X e Y (nella popolazione) e il suo stimatore campionario: il coefficiente di correlazione $r = \hat{\rho}$.
Abbiamo visto come, se X e Y sono normalmente distribuite e *(solo!) sotto l'ipotesi* $H_0 : \rho = 0$,

- r si distribuisca come una t di Student con $n - 2$ gradi di libertà (perché per calcolarlo stimo due parametri: $\bar{x}$ e $\bar{y}$)
- con un errore standard $ES(r) = \sqrt{\frac{1-r^2}{n-2}}$

Si può pertanto effettuare un test t dell'ipotesi di incorrelazione tra X e Y come segue:

$$\frac{r - 0}{ES(r)} = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$$

Ipotesi più generali del tipo $H_0 : r = r^*$ non sono testabili con questa statistica poiché la formula non vale in generale ma solo per $\rho = 0$.

Regressione

- La regressione è uno strumento fondamentale dell'analisi statistica.
- Consiste nel valutare la relazione tra una variabile *obiettivo* (solitamente chiamata *variabile dipendente*) e una o più *esplicative*.

Denotiamo la variabile dipendente con y e le k variabili esplicative con x_1, x_2, ..., x_k

- Nomi alternativi per le variabili y e x:

y	x
variable dipendente	regressori
variabile obiettivo	variabili esplicative

- Ci possono in generale essere numerose variabili x ma cominceremo col considerarne solo una.

Regressione e correlazione

Parlando di *correlazione* tra y e x, le trattiamo in maniera completamente simmetrica.

Nella regressione, invece, trattiamo la variabile dipendente (y) e le variabili esplicative (x) in modo molto differente.

La base filosofica del *modello di regressione* prevede un *processo generatore dei dati* (siamo realisti, non nominalisti)

Modello

L'idea di base è che le unità della popolazione (tutti i possibili campioni) siano generate da un *processo generatore dei dati* (DGP). Una descrizione formale del DGP prende il nome di *modello* e per noi avrà forma lineare del tipo:

$$Y = \beta X + u$$

Un modello è

- Una descrizione astratta e stilizzata della realtà...
- ...capace di riprodurne le caratteristiche cui siamo interessati.
- Un modo plausibile di generare i dati che stiamo osservando.

Operativamente, si cerca di costruire modelli che

- *spieghino* la maggior parte della variabilità nei dati osservati relativi al fenomeno di interesse,
- lasciando non spiegata solo una componente *non sistematica* detta *disturbo (o errore) casuale*.

Deams

A che serve un modello

Operativamente, se comprendiamo come "la nostra realtà è stata generata", saremo capaci di

- interpretarla
- riprodurla sotto condizioni diverse:
  - *previsione*
  - *what-if analysis*

Il modello sarà la formalizzazione della nostra teoria e la base per i tentativi di *falsificazione*, che prenderanno la forma di *test diagnostici* relativi ai vari aspetti del modello stesso (forma funzionale, proprietà degli errori, valori assunti dai parametri . . .)

Trovare l'interpolante ottimale

- Usiamo la generica equazione di una retta,

$$Y = a + bX$$

  per trovare la migliore interpolante dei nostri dati.

- Tuttavia, l'equazione $(Y = a + bX)$ è completamente deterministica.

- E' realistico? No. Pertanto aggiungiamo un *disturbo aleatorio*, u, all'equazione.

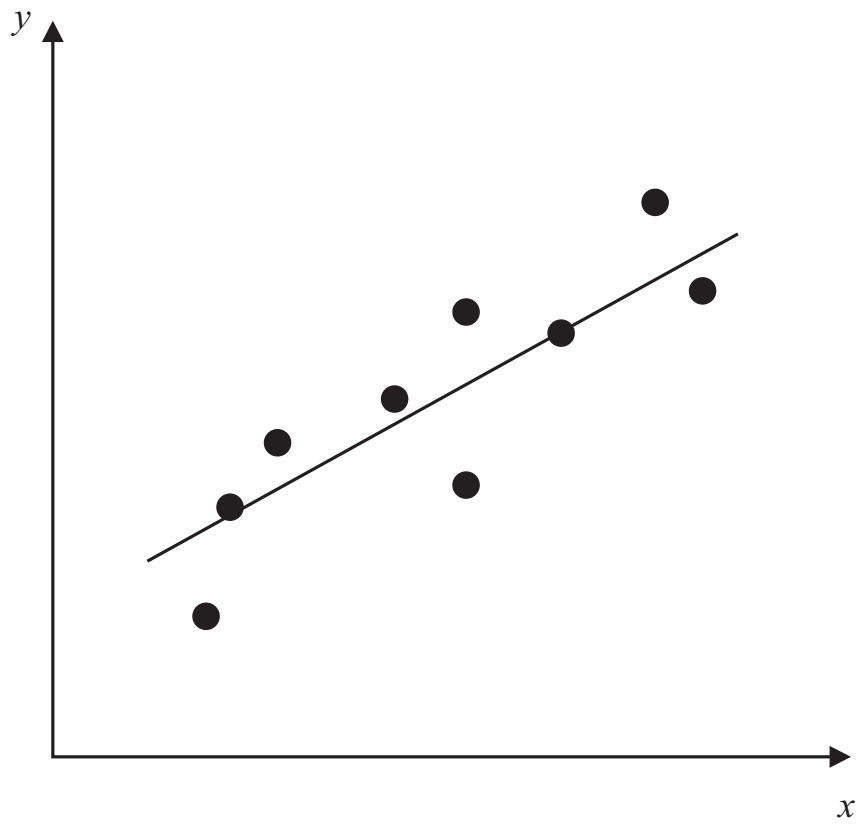
$$y_i = \alpha + \beta x_i + u_i$$

Perché includere un disturbo aleatorio?

- Il termine di errore (o disturbo aleatorio) u può dar conto di vari fenomeni:
  - Determinanti omessi di y_t
  - Errori di misura non modellizzabili di y_t
  - Influenze esogene su y_t che non possiamo includere nel modello

Determinare i coefficienti del modello

- Come determinare α e β?
- Cercansi α e β tali da rendere minime le distanze (verticali) tra i punti rappresentativi dei dati osservati e la retta stimata:



Ordinary Least Squares

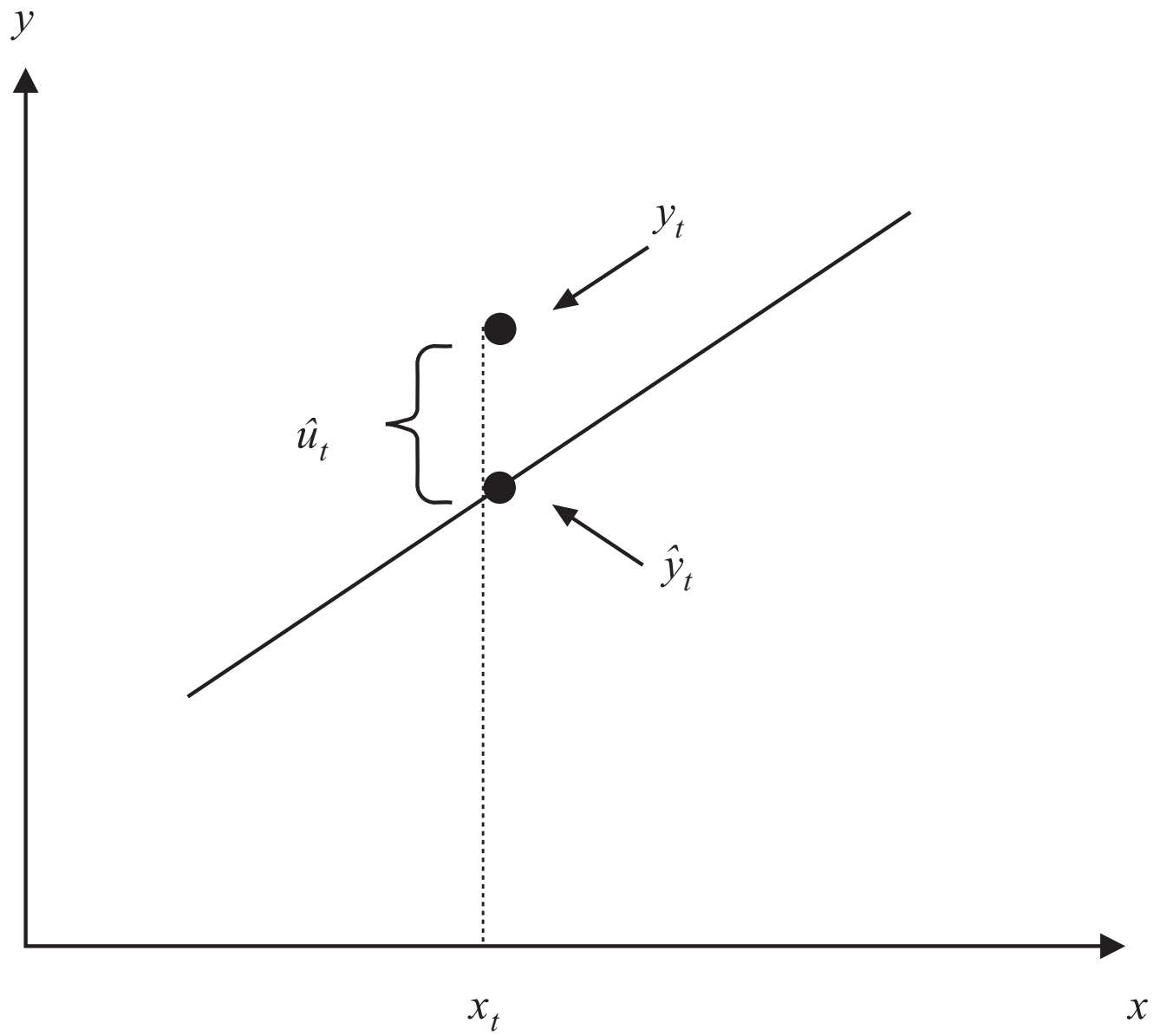
- Il metodo di stima più comune è noto come OLS (*ordinary least squares*, o minimi quadrati ordinari).
- Si minimizzano i quadrati delle distanze indicate in figura (da cui il nome).
- Più formalmente, siano

  y_t i valori osservati per ogni t

  $\hat{y}_t$ i valori corrispondenti (*stimati*) sulla retta di regressione

  $\hat{u}_t$ i residui, $\hat{u}_t = y_t - \hat{y}_t$

Valori osservati e stimati; residui



Minimi quadrati ordinari

- Cercansi i valori ottimi di $\hat{\alpha}$ e $\hat{\beta}$ tali da rendere minima la somma dei quadrati dei residui: $L = \sum_{t=1}^{5} \hat{u_t}^2$ che è la nostra *funzione di perdita*
- Ricordiamo che $\hat{u_t}$ è la differenza tra valori stimati e osservati, $y_t - \hat{y_t}$ ...
- ... ma $\hat{y_t} = \hat{\alpha} + \hat{\beta} x_t$ pertanto $L(\hat{\alpha}, \hat{\beta}) = \sum (y_t - \hat{y_t})^2$
- graficamente, minimizzare rispetto ai parametri la funzione di perdita L equivale a minimizzare i quadrati delle differenze tra valori osservati e retta stimata per ogni x_i

Derivazione dello stimatore OLS

E' $\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t$, pertanto sia

$$L = \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{T} (y_t - \hat{\alpha} - \hat{\beta} x_t)^2.$$

Minimizziamo L rispetto a $\hat{\alpha}$ e $\hat{\beta}$, perciò differenziamo L sub $\hat{\alpha}$ e $\hat{\beta}$

$$\frac{\partial L}{\partial \hat{\alpha}} = -2 \sum_t (y_t - \hat{\alpha} - \hat{\beta} x_t) = 0 \quad (1)$$

$$\frac{\partial L}{\partial \hat{\beta}} = -2 \sum_t x_t (y_t - \hat{\alpha} - \hat{\beta} x_t) = 0 \quad (2)$$

- *derivata della funzione composta*: $[g(f(z))]' = g'(f(z)) \cdot f'(z)$
- *derivata della somma*: $[\sum_t f_t(z)]' = \sum_t [f_t(z)]'$

Derivazione dello stimatore OLS (Cont'd)

Da (1),

$$\sum_t (y_t - \hat{\alpha} - \hat{\beta} x_t) = 0 \Leftrightarrow \sum y_t - T\hat{\alpha} - \hat{\beta} \sum x_t = 0$$

$\sum y_t = T\bar{y}$ e $\sum x_t = T\bar{x}$. Dunque

$$T\bar{y} - T\hat{\alpha} - T\hat{\beta}\bar{x} = 0 \text{ or } \bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0 \quad (3)$$

Da (2),

$$\sum_t x_t (y_t - \hat{\alpha} - \hat{\beta} x_t) = 0 \quad (4)$$

Da (3),

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad (5)$$

Derivazione dello stimatore OLS (Cont'd)

Sostituendo in (4) per $\hat{\alpha}$ da (5),

$$\sum_t x_t(y_t - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_t) = 0$$

$$\sum_t x_t y_t - \bar{y}\sum x_t + \hat{\beta}\bar{x}\sum x_t - \hat{\beta}\sum x_t^2 = 0$$

$$\sum_t x_t y_t - T\bar{x}\bar{y} + \hat{\beta}T\bar{x}^2 - \hat{\beta}\sum x_t^2 = 0$$

Mettendo in evidenza $\hat{\beta}$,

$$\hat{\beta}\left(T\bar{x}^2 - \sum x_t^2\right) = T\overline{xy} - \sum x_t y_t$$

$$\hat{\beta} = \frac{\sum x_t y_t - T\overline{xy}}{\sum x_t^2 - T\bar{x}^2} \quad e \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Lo stimatore OLS

Dunque in generale si ha

$$\hat{\beta} = \frac{\sum x_t y_t - T\overline{xy}}{\sum x_t^2 - T\bar{x}^2} \quad e \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

ma, si osservi, nel campione è
$\sum x_t y_t - T\overline{xy} = T(media(XY) - media(X) \cdot media(Y))$ e
$\sum x_t^2 - T\bar{x}^2 = T(media(X^2) - [media(X)]^2)$ pertanto

$$\hat{\beta} = \frac{Cov(XY)}{Var(X)}$$

Questo criterio di ottimalità, e gli stimatori che da esso prendono il nome, sono noti come OLS (da ordinary least squares).

Il modello di regressione come valore atteso condizionato

Dal punto di vista probabilistico, la parte sistematica del modello di regressione può essere vista come un modello per il *valore atteso condizionato* di Y *dato* $X = x$:

$$E(Y|x) = \alpha + \beta x$$

- In questo senso, ogni valore stimato (previsto) di y, $\hat{y}_i = E(Y|X = x_i)$ è visto come il valore atteso di Y se $X = x_i$
- La parte deterministica del modello è perciò detta anche *predittore lineare*

Stima e impiego del modello

- specificazione della componente deterministica (forma funzionale, variabili da includere) NB viene *imposta*
- specificazione delle caratteristiche dell'errore: NB vengono *ipotizzate, assunte*
- utilizzo dei dati campionari per stimare i parametri *subordinatamente alle ipotesi fatte*
- *critica* della validità del modello: i.e., della specificazione adottata e delle caratteristiche dello stimatore tramite *test diagnostici*
- stima della varianza dell'errore/disturbo aleatorio u e quindi della dispersione (varianza) dei parametri stimati $\hat{\alpha}, \hat{\beta}$
- interpretazione dei risultati
- impiego del modello

La bontà di adattamento del modello

Siamo interessati alla *bontà di adattamento* del nostro modello ai dati. Per valutarla usiamo un'altra *statistica*: il cosiddetto R^2.

- Un modo di definire l' R^2 è dire che è il quadrato del coefficiente di correlazione tra y e $\hat{y}$ (si dimostra).
- Un modo ancora più utile di vedere le cose è il seguente:
  - vogliamo *spiegare* la variabilità di y attorno alla sua media $\bar{y}$, o *devianza*, che chiameremo *somma dei quadrati totali*, o *total sum of squares (TSS)*:

$$TSS = \sum_t (y_t - \bar{y})^2$$

  - la devianza *TSS* può essere divisa in due parti: la *devianza spiegata* o *ESS*, e la *devianza non spiegata* o *residua (RSS)*

Definizione di R^2

Misureremo la bontà di adattamento con $R^2 = \frac{ESS}{TSS}$

- Riesce

$$\begin{aligned} TSS &= ESS + RSS \\ \sum_t (y_t - \bar{y})^2 &= \sum_t (\hat{y}_t - \bar{y})^2 + \sum_t \hat{u}_t^2 \end{aligned}$$

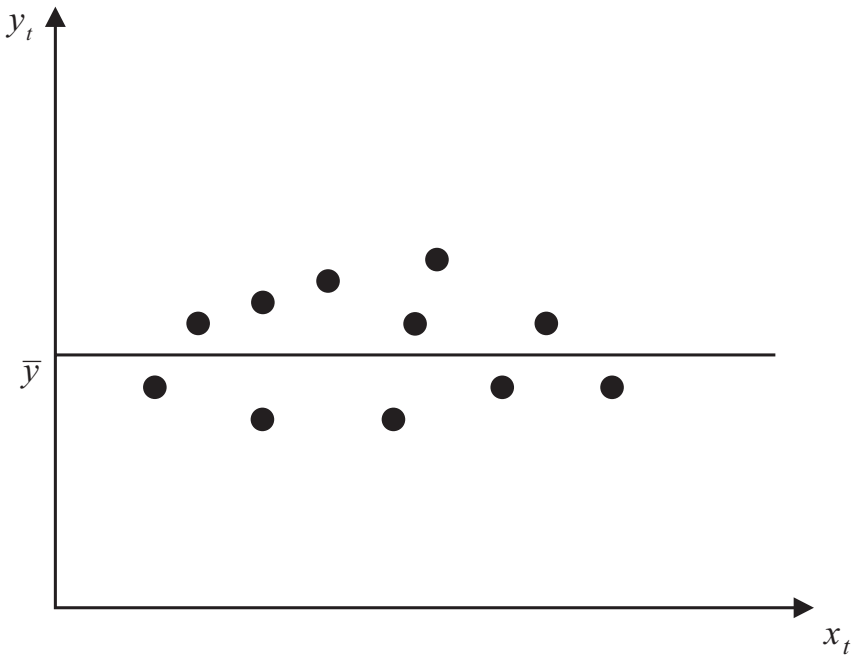
- pertanto

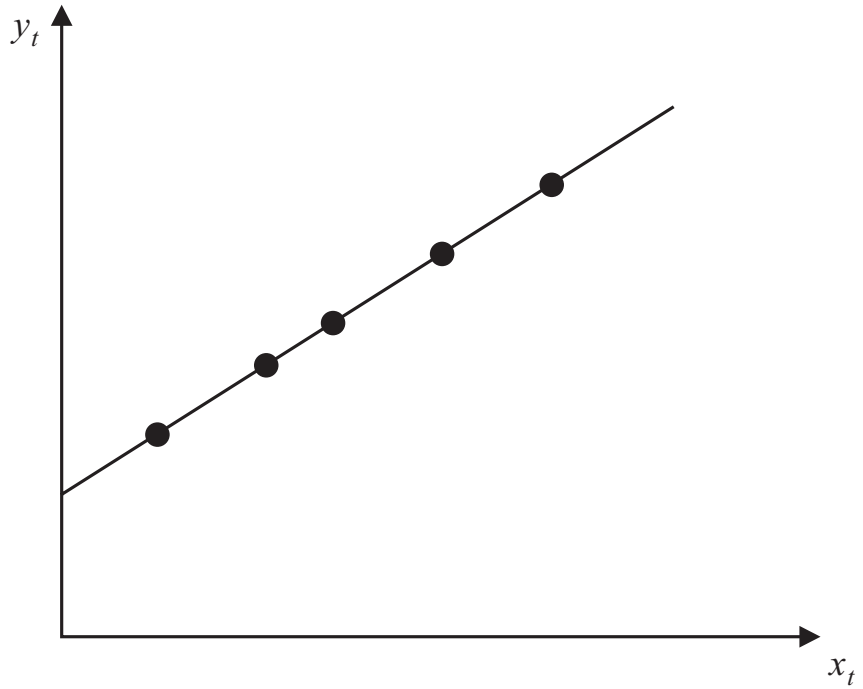
$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- Inoltre, R^2 è sempre compreso tra 0 e 1. Casi estremi:

/@ /@ /@ //@ /@ /@ /@ / $\begin{array}{lll} RSS = TSS & \text{i.e.} & ESS = 0 \quad \text{so} \\ ESS = TSS & \text{i.e.} & RSS = 0 \quad \text{so} \end{array}$

I casi limite: $R^2 = 0$ e $R^2 = 1$





Le ipotesi del modello lineare classico

- Il modello che abbiamo descritto è noto come *modello lineare classico*.

- Facciamo le seguenti ipotesi su u_t (gli errori non osservabili):

Notazione	Interpretazione
(1) $\mathrm{E}(u_t) = 0$	Gli errori hanno valore atteso nullo
(2) $\mathrm{var}(u_t) = \sigma^2$	La varianza degli errori è costante e finita
(3) $\mathrm{cov}(u_i, u_j) = 0$	Gli errori non sono correlati tra loro
(4) $\mathrm{cov}(u_t, x_t) = 0$	Non c'è correlazione tra l'errore in t e la corrispondente variabile x_t

- Un'assunzione alternativa alla (4) e più restrittiva è che le x_t siano non-stocastiche, o *fissate in campioni ripetuti*.

Le ipotesi del modello lineare classico (Cont'd)

- Una quinta ipotesi è richiesta per fare inferenza riguardo ai parametri della popolazione (i "veri" α e β) sulla base delle stime campionarie $\hat{\alpha}$ e $\hat{\beta}$
- Ipotesi ulteriore:

  (5) u_t è normalmente distribuito (per ogni t)

Proprietà dello stimatore OLS

Se le ipotesi da (1) a (4) sono soddisfatte, gli stimatori OLS sono detti *BLU* (Best Linear Unbiased). Cosa significa?

(in inglese si aggiunge "E" per 'Estimators' in quanto *stimatori* del "vero" valore di α e β)

- 'Linear' – $\hat{\alpha}$ e $\hat{\beta}$ sono stimatori *lineari* (funzioni lineari dei dati nel campione)
- 'Unbiased' – il valore atteso di $\hat{\alpha}$ and $\hat{\beta}$ è uguale ai "veri" valori di α e β
- 'Best' – significa che lo stimatore OLS ha varianza minima nella classe degli stimatori lineari corretti; questo risultato prende il nome di Teorema di Gauss–Markov.

Consistenza/Correttezza/Efficienza

- Consistenza

  Gli stimatori dei minimi quadrati $\hat{\alpha}$ e $\hat{\beta}$ sono consistenti, ovvero le stime convergeranno ai veri valori dei parametri al divergere a infinito della dimensione del campione. Servono le ipotesi $E(x_t u_t) = 0$ e $Var(u_t) = \sigma_t^2 < \infty$ per dimostrarlo. La consistenza implica che

$$\lim_{T \to \infty} \Pr\left[|\hat{\beta} - \beta| > \delta\right] = 0 \quad \forall\, \delta > 0$$

- Correttezza

  Gli stimatori dei minimi quadrati $\hat{\alpha}$ e $\hat{\beta}$ sono corretti. Ovvero, $E(\hat{\alpha}) = \alpha$ e $E(\hat{\beta}) = \beta$. Pertanto, in media le stime saranno uguali ai veri valori. La dimostrazione richiede a sua volta che $E(u_t) = 0$.

- Efficienza

Consistenza/Correttezza/Efficienza (Cont'd)

Uno stimatore $\hat{\beta}$ del parametro β è detto efficiente se è corretto ed ha la minima varianza tra gli stimatori corretti. Se uno stimatore è efficiente, stiamo minimizzando la probabilità che la stima si allontani dal vero valore di β.

Precisione e Standard Errors

- Le stime $\hat{\alpha}$ e $\hat{\beta}$ dipendono dal campione usato per la stima.
- Ricordate che gli stimatori di α e β dai dati del campione sono

$$\hat{\beta} = \frac{\sum x_t y_t - T\overline{xy}}{\sum x_t^2 - T\bar{x}^2} \quad \textit{and} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Precisione e Standard Errors (Cont'd)

- Ci serve una misura dell'affidabilità, o precisione, degli stimatori $\hat{\alpha}$ e $\hat{\beta}$. La precisione della stima è misurata dal suo errore standard (ES). Date le assunzioni (1)–(4), si può dimostrare che gli ES sono:

$$ES(\hat{\alpha}) = s\sqrt{\frac{\sum x_t^2}{T\sum (x_t - \bar{x})^2}} = s\sqrt{\frac{\sum x_t^2}{T\left(\left(\sum x_t^2\right) - T\bar{x}^2\right)}}$$

$$ES(\hat{\beta}) = s\sqrt{\frac{1}{\sum (x_t - \bar{x})^2}} = s\sqrt{\frac{1}{\sum x_t^2 - T\bar{x}^2}}$$

dove s è l'errore standard stimato *degli errori* (come si stima?)

Stimare la varianza del termine di errore

- La varianza della variabile aleatoria u_t è data da

$$\text{Var}(u_t) = \text{E}[(u_t)\text{-E}(u_t)]^2$$

che (è $E(u) = 0$ per ipotesi) si riduce a

$$\text{Var}(u_t) = \text{E}(u_t^2)$$

- Potremmo stimarla usando la media di u_t^2 nel campione:

$$\sigma_u^2 = \frac{1}{T}\sum u_t^2$$

- Purtroppo u_t non è osservabile. Possiamo usare la controparte campionaria di u_t, che è $\hat{u}_t$:

$$s^2 = \frac{1}{T}\sum \hat{u}_t^2$$

ma questa è uno stimatore distorto di σ_u^2.

Stimare la varianza del termine di errore (cont'd)

- Uno stimatore corretto di σ è dato da

$$s = \sqrt{\frac{\sum \hat{u}_t^2}{T-2}}$$

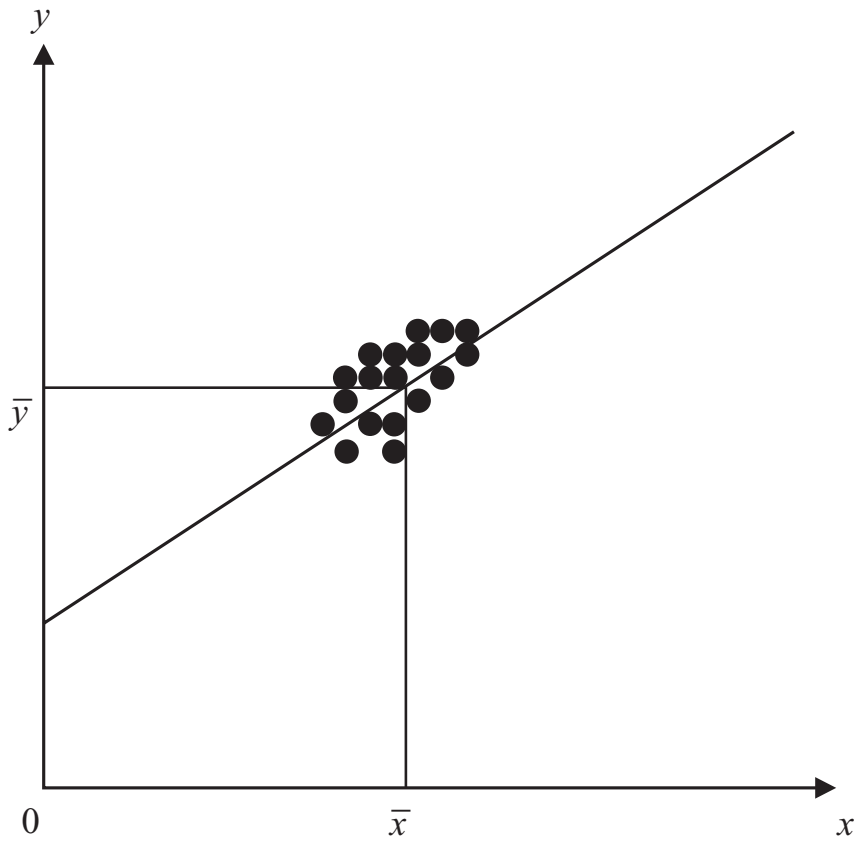
dove $\sum \hat{u}_t^2$ è la somma dei quadrati dei residui, T è la dimensione del campione e 2 è il numero dei regressori (parametri stimati).

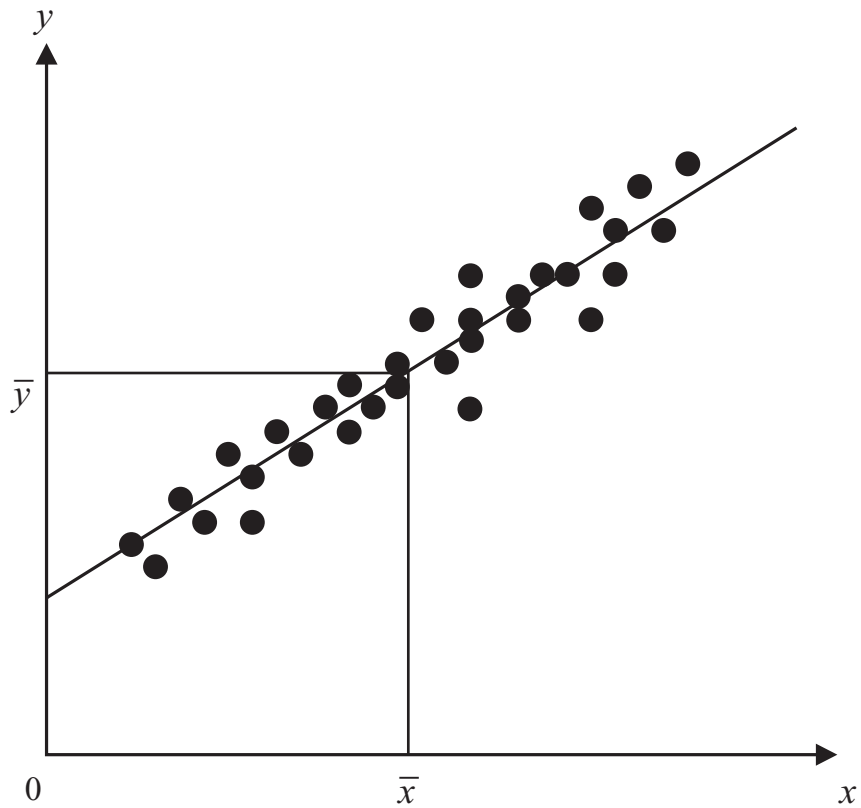
- Alcuni commenti sugli stimatori degli ES

1. Sia ES($\hat{\alpha}$) che ES($\hat{\beta}$) dipendono da s^2 (o s). Al crescere della varianza s^2, cresce la dispersione degli errori attorno alla loro media e pertanto cresce la dispersione di y attorno alla sua media.
2. La somma dei quadrati degli scarti di x attorno alla loro media appare in entrambe le formule. Maggiore questa somma, minori risultano le varianze dei coefficienti.

Alcuni commenti sugli stimatori degli ES

Considerate che succede se $\sum (x_t - \bar{x})^2$ è grande o piccolo:





Alcuni commenti sugli stimatori degli ES (Cont'd)

1. Maggiore la dimensione del campione, T, minori saranno le varianze dei coefficienti. T appare esplicitamente in ES($\hat{\alpha}$) e implicitamente in ES($\hat{\beta}$).

   T appare implicitamente perché la somma $\sum (x_t - \bar{x})^2$ va da $t = 1$ a T.

2. Il termine $\sum x_t^2$ appare in ES($\hat{\alpha}$).

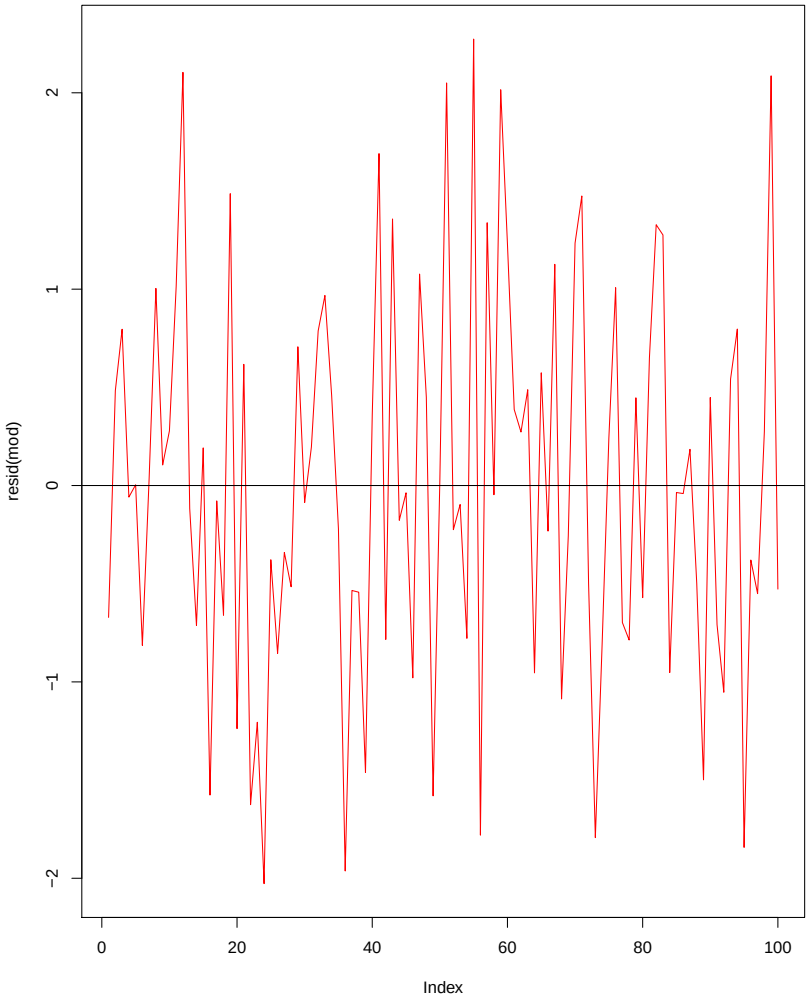
   La ragione è che $\sum x_t^2$ misura la distanza dei punti dall'asse y.

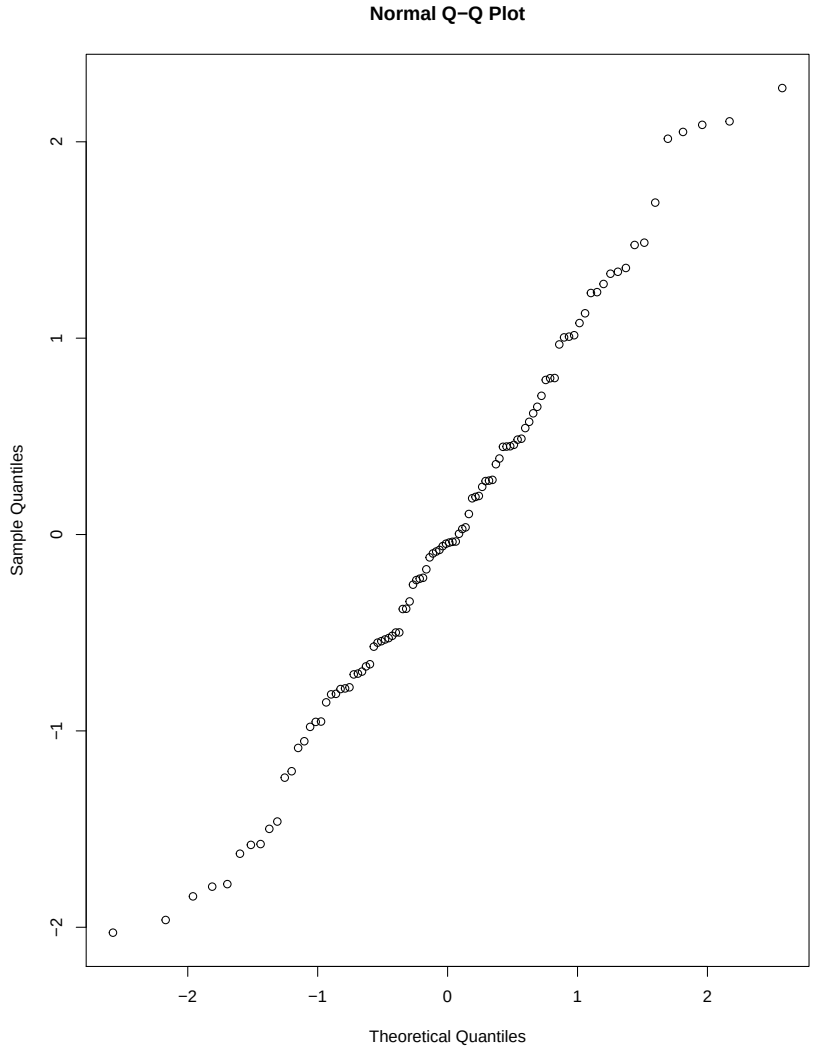
Analisi grafica dei residui

Vi sono numerosi test diagnostici sotto forma di *statistiche test* che è possibile impiegare per valutare le ipotesi fatte in partenza sugli errori. Noi ci limiteremo all'analisi *grafica* dei residui, che spesso è sufficiente per dare una prima indicazione di eventuali problemi *(Si veda l'es. Cap6.2)* .

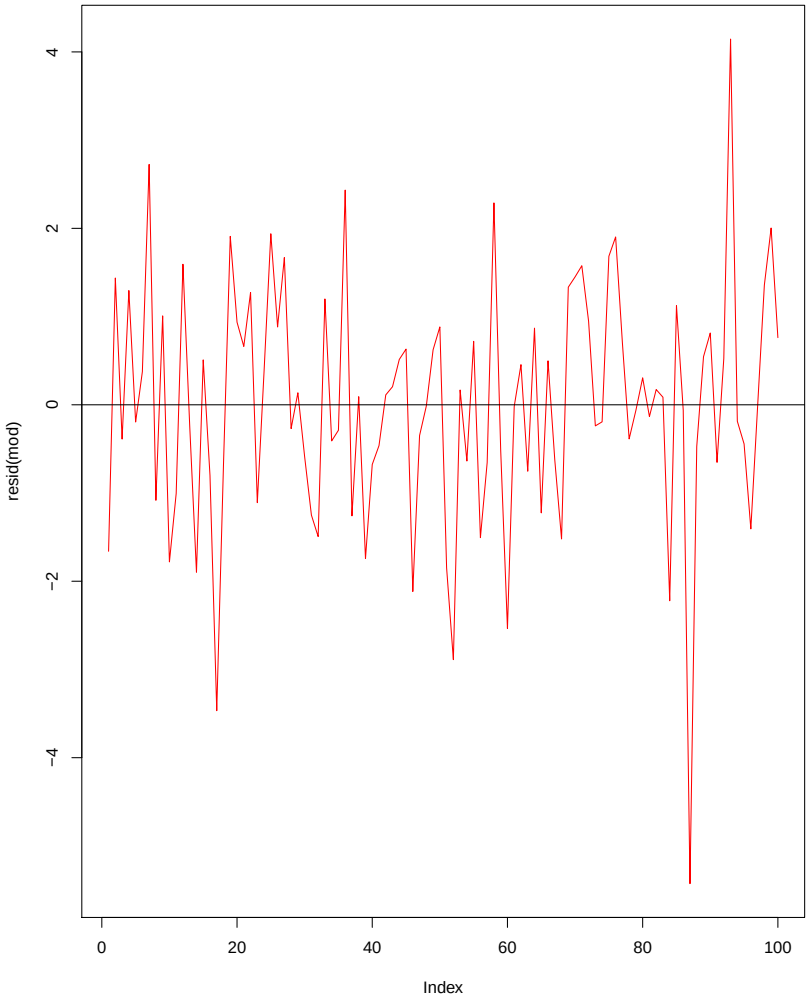
- Normalità dei residui: il c.d. *Q-Q (normal) plot* confronta i quantili empirici della distribuzione dei residui con quelli teorici di una normale. Idealmente i punti si dispongono su una retta.
- Omoschedasticità: il *residual plot* non dovrebbe mostrare variazioni di ampiezza (*sensibile a come si ordinano le osservazioni!*)
- Autocorrelazione: il *residual plot* non dovrebbe mostrare residui "simili" ai precedenti (es. sequenze di valori positivi, o negativi; sequenze di valori "grandi")
- Variabili omesse, cambiamenti strutturali: il *residual plot* non dovrebbe mostrare andamenti "sistematici"

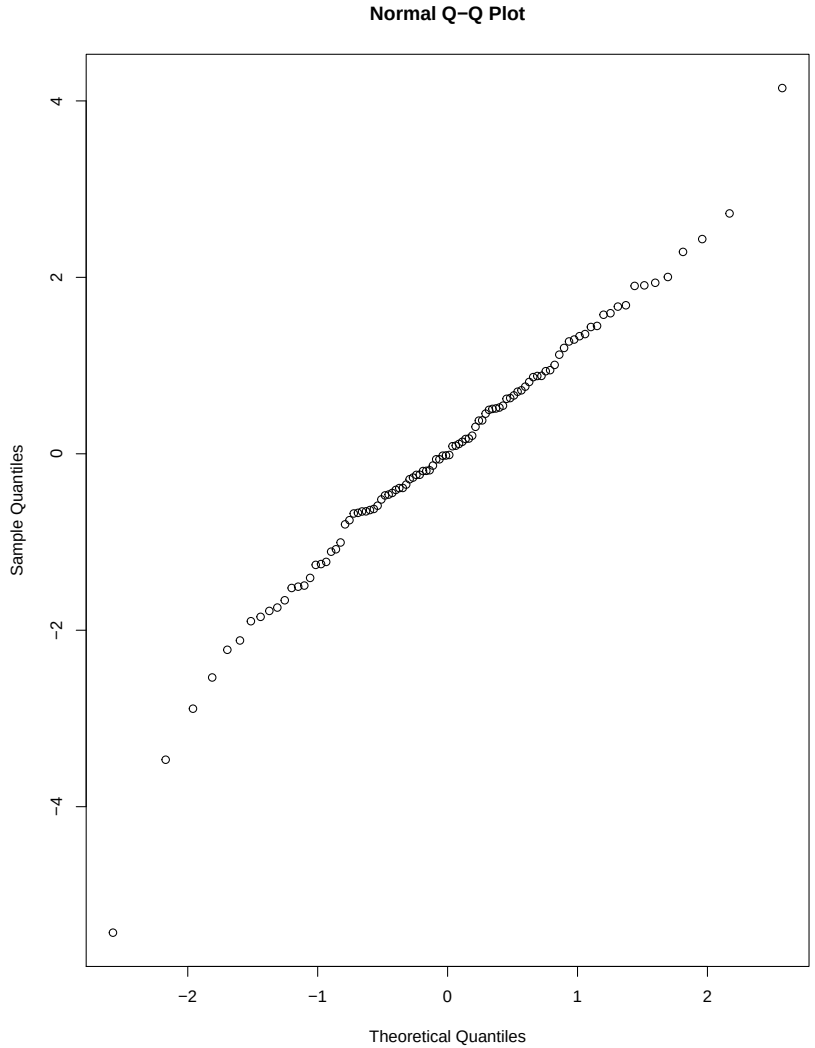
Residui normali ($u \sim N$) e Q-Q plot



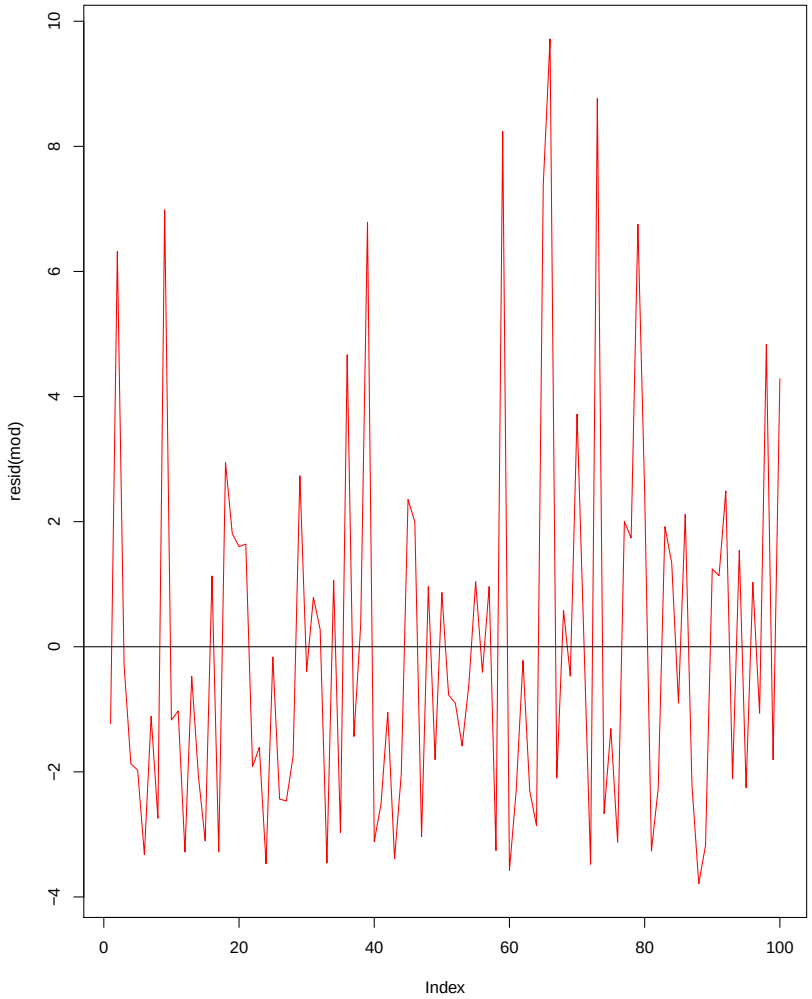


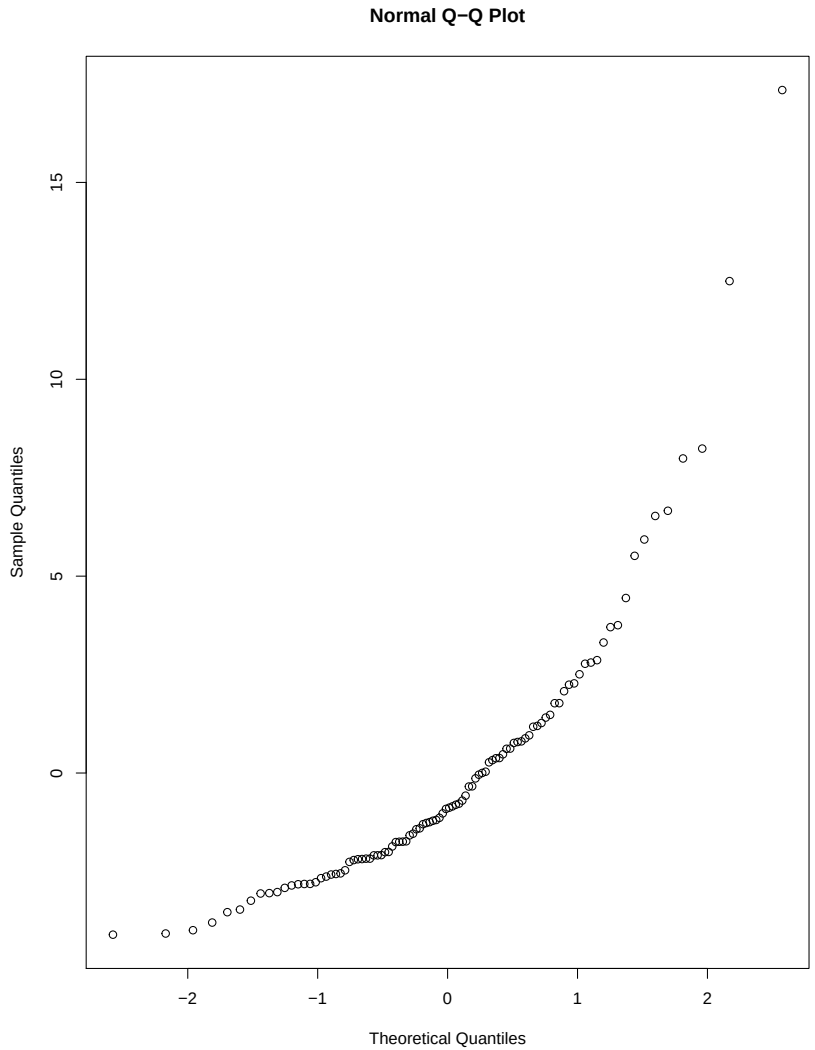
“Code grosse” ($u \sim$ t-Student) e Q-Q plot



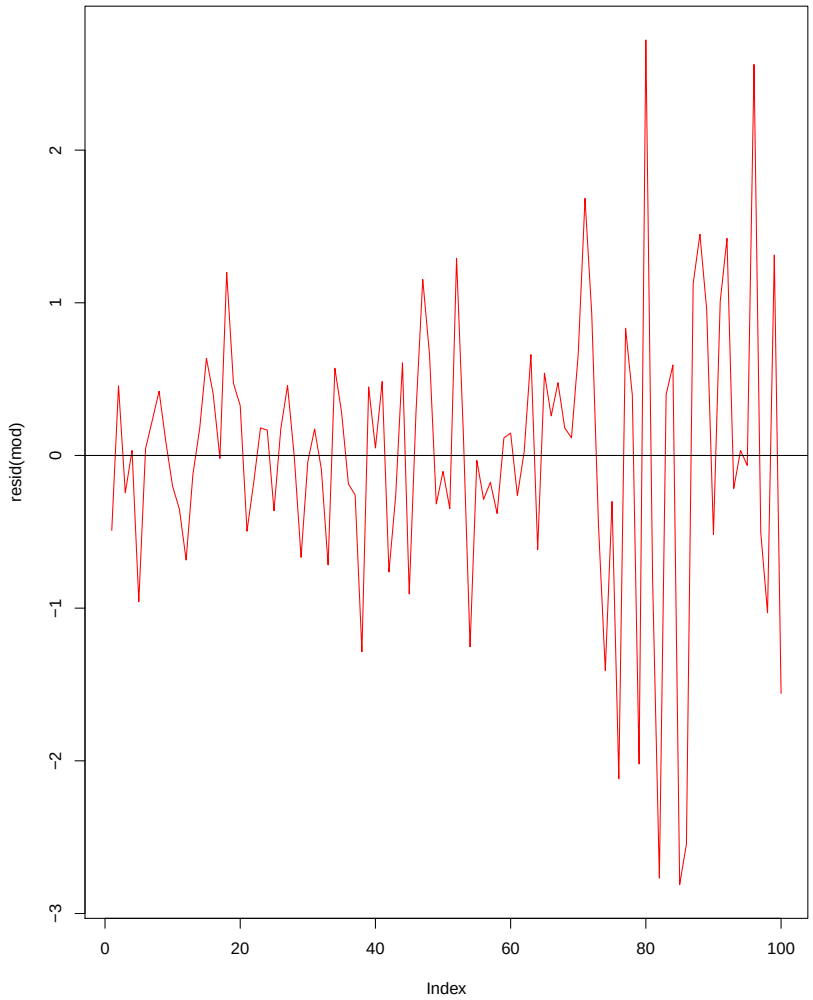


Asimmetria ($u \sim \chi^2$) e Q-Q plot

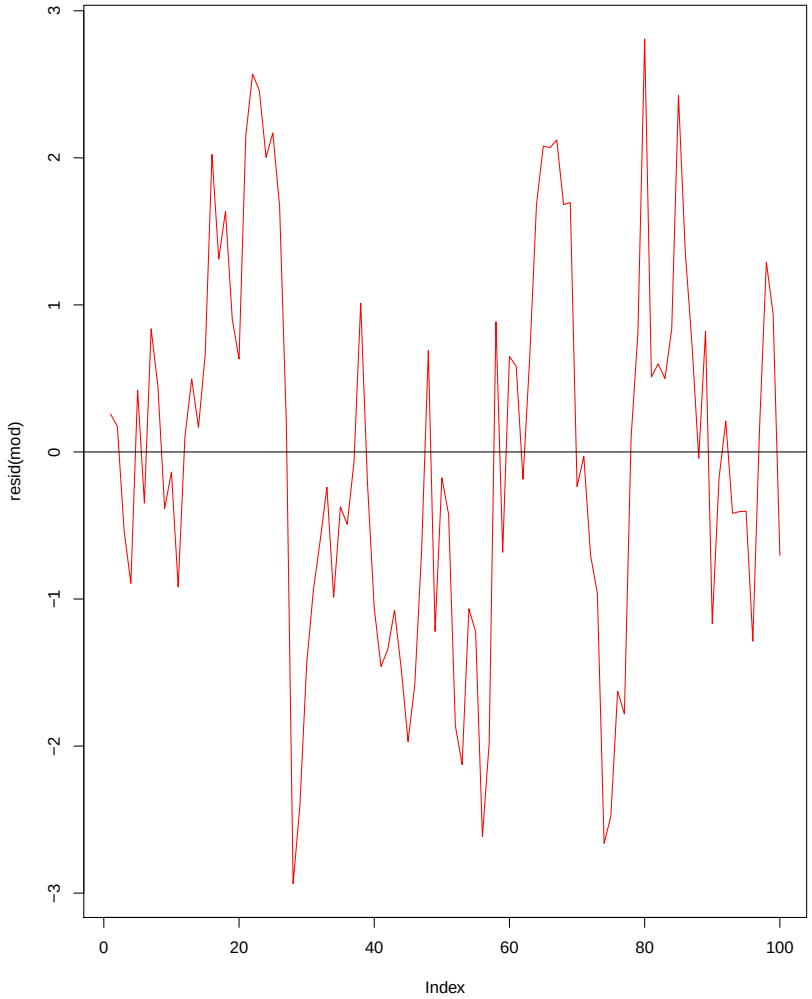


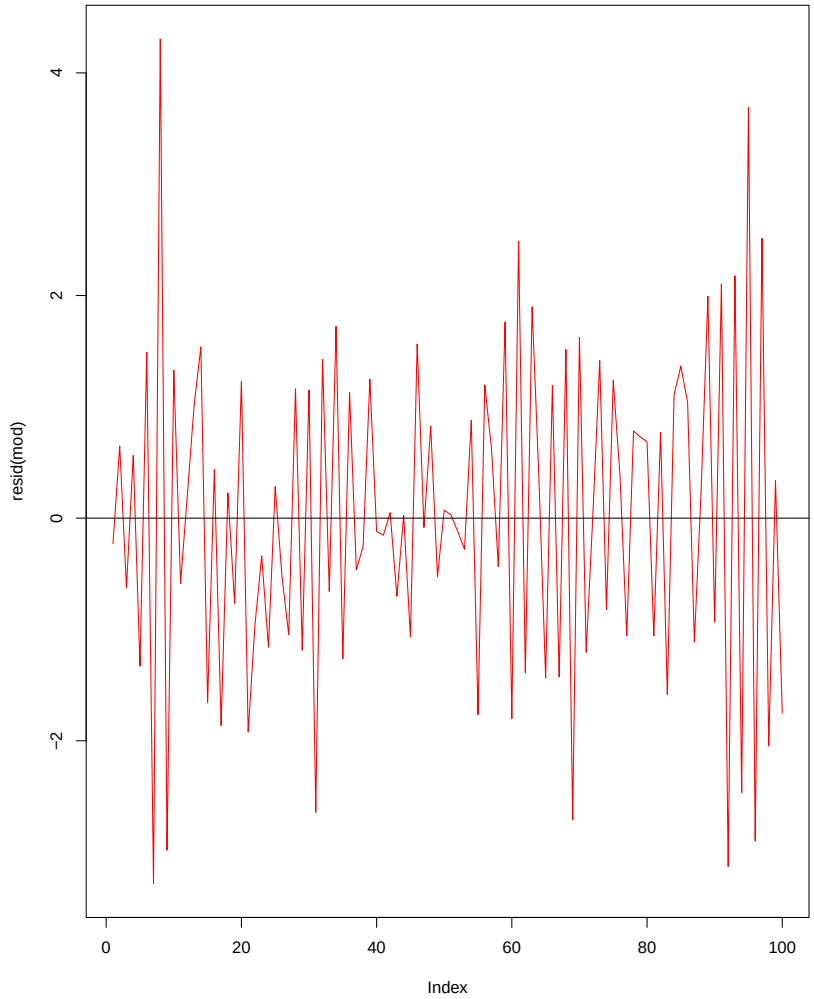


Eteroschedasticità

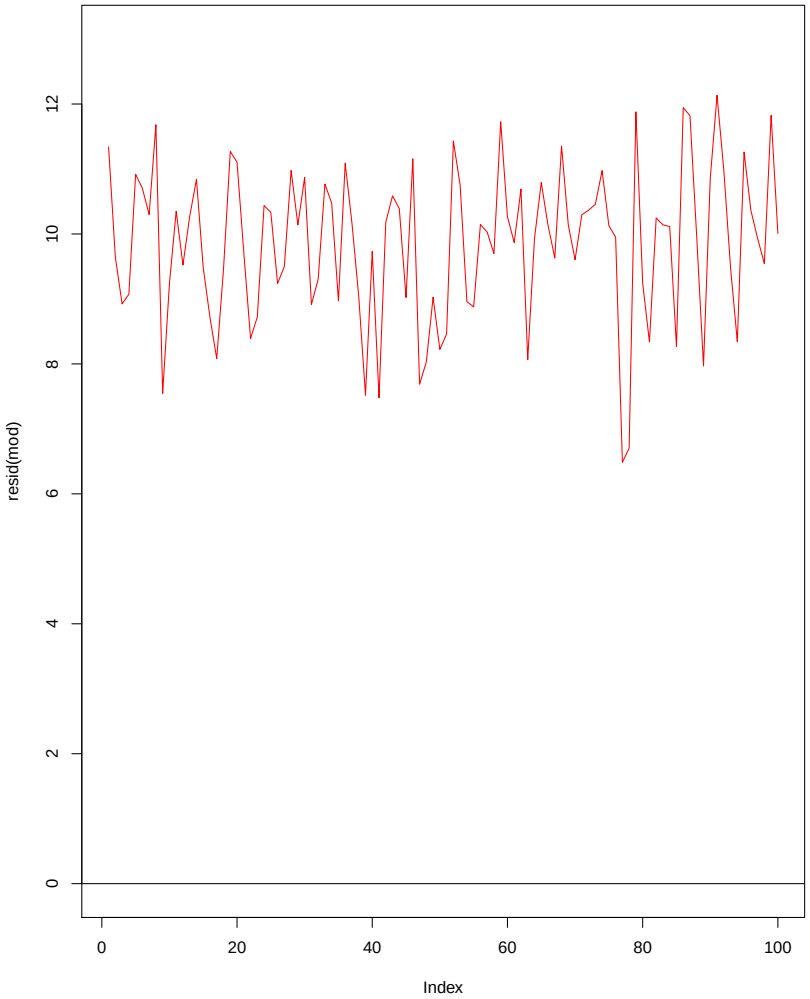


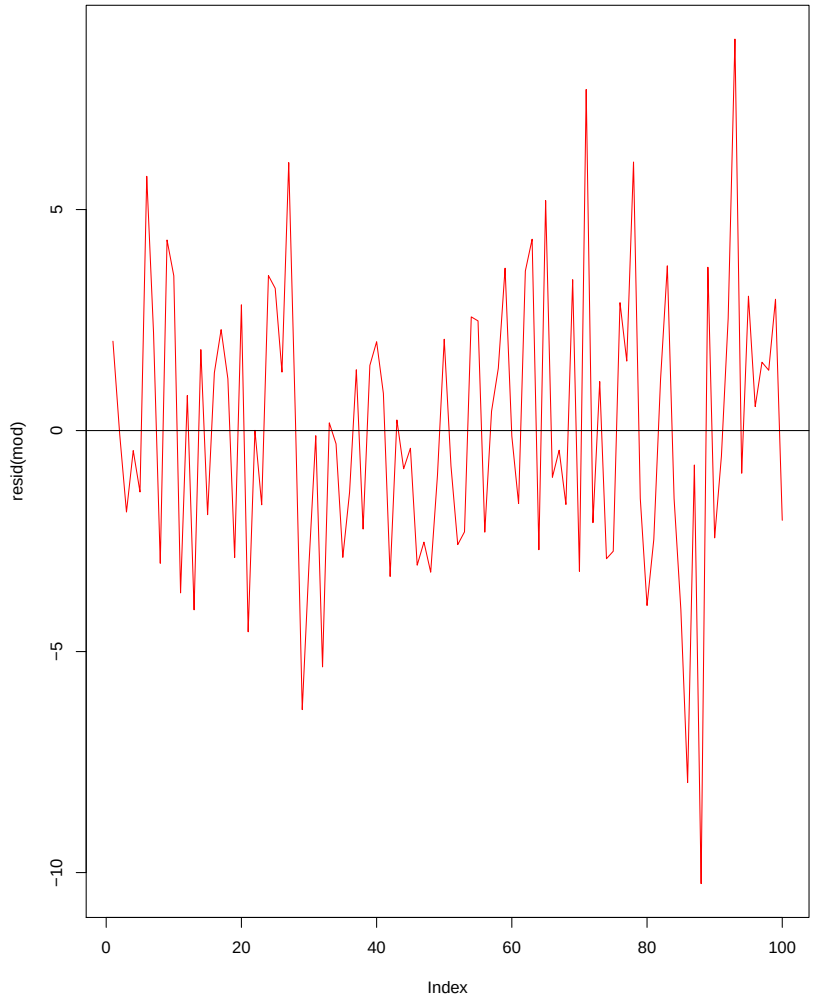
Autocorrelazione positiva vs. negativa



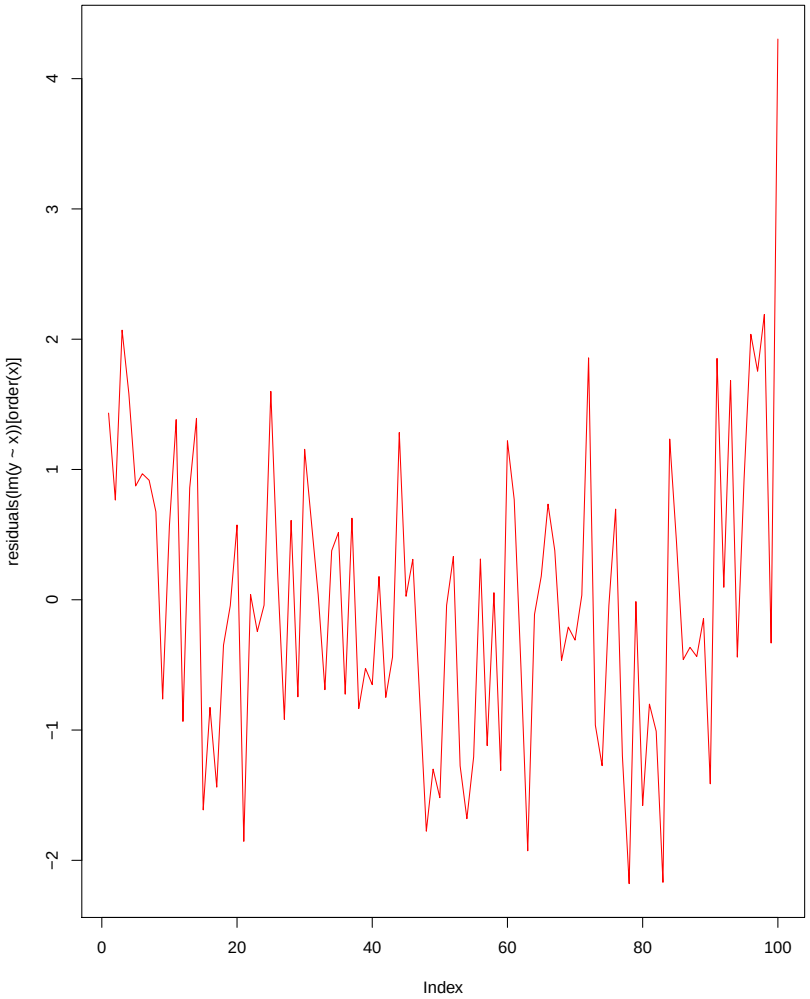


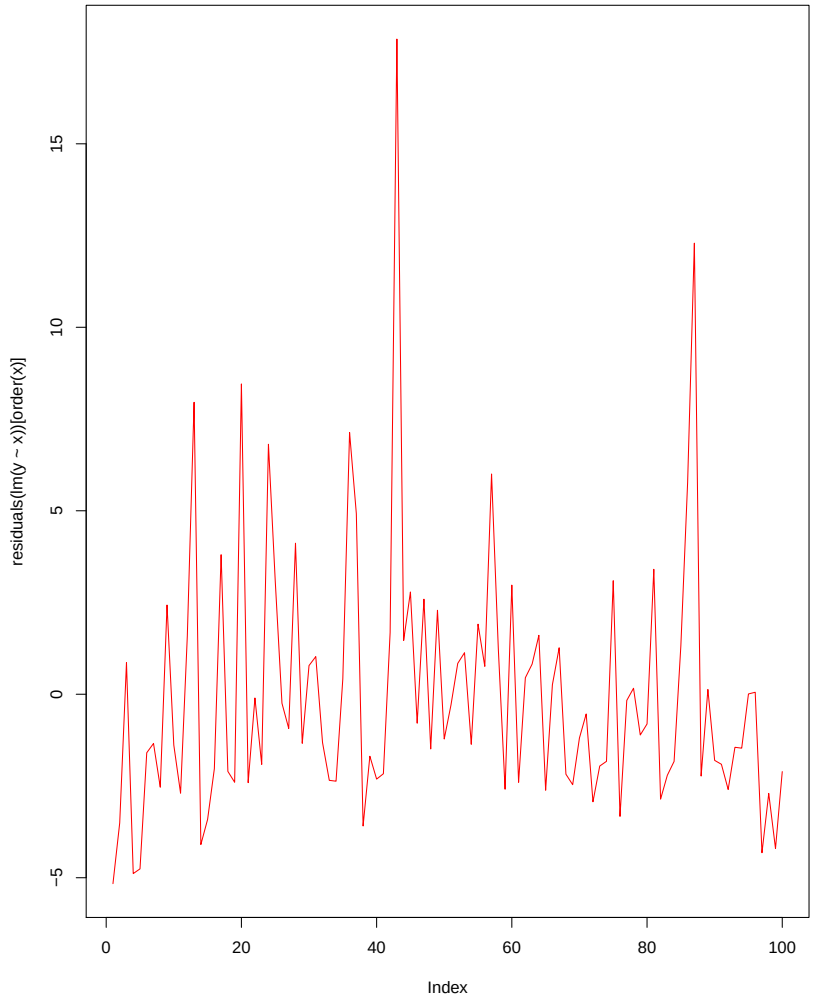
Variabili omesse: intercetta vs. variabile x_2





“Vero” modello non lineare: espon. vs. quadratico





La distribuzione di probabilità degli stimatori OLS

Assumiamo (5) che $u_t \sim N(0, \sigma^2)$

- Siccome gli stimatori OLS sono combinazioni lineari di variabili aleatorie:
  ovvero $\hat{\beta} = \sum w_t y_t$
- La somma di variabili Normali è a sua volta distribuita come una Normale, pertanto

$$\hat{\alpha} \sim N(\alpha, Var(\hat{\alpha}))$$
$$\hat{\beta} \sim N(\beta, Var(\hat{\beta}))$$

- E se non valesse (5), ovvero gli errori non fossero Normali? Sarà ancora Normale la distribuzione dei parametri?
- Sì, se valgono le altre ipotesi (1-4) e la dimensione del campione è *sufficientemente grande*.

Testare ipotesi sui parametri: il t-test

Con riferimento al modello di regressione $y_t = \alpha + \beta x_t + u_t$ si immagini di voler testare l'ipotesi statistica $H_0 : \beta = \beta^*$ con β^* una costante (che in genere corrisponde a un'ipotesi economica di interesse).

Al solito, si cerca di esprimere l'ipotesi di interesse sotto forma di una statistica calcolabile, la cui distribuzione *sub* H_0 sia nota.
In generale,

- variabili Normali standard possono essere costruite a partire da $\hat{\alpha}$ e $\hat{\beta}$:

$$\frac{\hat{\alpha}-\alpha}{\sqrt{Var(\hat{\alpha})}} \sim N(0,1) \text{ e } \frac{\hat{\beta}-\beta}{\sqrt{Var(\hat{\beta})}} \sim N(0,1)$$

- ma Var(α) e Var(β) sono ignote, perciò si usano le stime $ES(\hat{\alpha})$, $ES(\hat{\beta})$

$$\frac{\hat{\alpha}-\alpha}{ES(\hat{\alpha})} \sim t_{T-2} \text{ e } \frac{\hat{\beta}-\beta}{ES(\hat{\beta})} \sim t_{T-2}$$

Testare ipotesi sui parametri: il t-test - 2

Per una specifica ipotesi $H_0 : \beta = \beta^*$

1. Si stimano $\hat{\alpha}$, $\hat{\beta}$ e $SE(\hat{\alpha})$, $SE(\hat{\beta})$
2. Si calcola la statistica test: $\frac{\hat{\beta}-\beta^*}{ES(\hat{\beta})}$
3. Si considera la distribuzione con cui confrontare il valore assunto dalla statistica test. come osservato, in questo caso sotto H_0 essa si distribuisce come una *t di Student* con *T-2* gradi di libertà.
4. Si sceglie un livello di significatività: di solito il 5%.
5. Dalle tavole della *t* (o da R: `> qt(1-0.025, df=T-2)`) si ottengono i valori critici che delimitano le regioni di rifiuto.
6. Se la statistica test finisce in una regione di rifiuto allora si rifiuta H_0, altrimenti non la si rifiuta.

t di Student vs. Normale

Il test t "esatto" in campioni finiti dipende dall'ip. (5): $u_t \sim N(0, \sigma_u^2)$.
Per campioni "grandi:

- la distribuzione degli stimatori tende alla Normale a prescindere da (5)
- la t tende alla Normale e le distinzioni svaniscono. Ma cosa vuol dire "grande"?

Esempi di valori critici:

Livello di significatività	$N(0,1)$	$t(40)$	$t(4)$
50%	0.00	0.00	0.00
5%	1.64	1.68	2.13
2.5%	1.96	2.02	2.78
0.5%	2.57	2.70	4.60

Naturalmente, $T - 2 = 4$ non è realistico. Per la maggior parte dei campioni, la differenza pratica è "piccola".

Intervalli di confidenza e test di ipotesi - 1

Per un parametro β e un dato livello di significatività (a cui corrisponde un valore critico t_{crit}), l'intervallo di confidenza è

$$(\hat{\beta} - t_{crit} \times ES(\hat{\beta}), \hat{\beta} + t_{crit} \times ES(\hat{\beta}))$$

- Se stimiamo che, per es., $\hat{\beta} = 0.93$, e che il suo "intervallo di confidenza" è (0.77, 1.09), questo significa che per il "vero" (ma ignoto) parametro è

$$Pr[\beta \in (0.77, 1.09)] = 0.95$$

- Con riferimento a un'ipotesi su β, p. es. $H_o : \beta = \beta^*$, se β^* cade fuori dall'intervallo di confidenza, si rifiuta H_0.

Intervalli di confidenza e test di ipotesi - 2

I due approcci (intervallo di confidenza e test di ipotesi) sono equivalenti:

- Test di ipotesi: non si rifiuta $H_0 : \beta = \beta^*$ se

$$-t_{crit} \leq \frac{\hat{\beta}-\beta^*}{ES(\hat{\beta})} \leq +t_{crit}$$

- Riordinando, se

$$-t_{crit} \times ES(\hat{\beta}) \leq \hat{\beta} - \beta^* \leq +t_{crit} \times ES(\hat{\beta}))$$
$$\hat{\beta} - t_{crit} \times ES(\hat{\beta}) \leq \beta^* \leq \hat{\beta} + t_{crit} \times ES(\hat{\beta}))$$

che è la regola nell'approccio dell'intervallo di confidenza.

Correlazione e regressione con più di due variabili

- La correlazione multipla non è che l'insieme di tutte le correlazioni tra le variabili, prese due a due
- La regressione multipla, invece, ha caratteristiche molto più interessanti:
  - l'effetto di tutte le variabili esplicative incluse viene valutato congiuntamente
  - i coefficienti pertanto rappresentano *effetti parziali* nel senso che essi rappresentano l'effetto di quella variabile *tenendo costanti tutte le altre*
  - si dice pertanto che nella regressione multipla l'effetto di ogni variabile viene valutato *controllando* per quello di ogni altra

Nel modello lineare

$$y_t = \alpha + \beta x_t + \gamma z_t + u_t$$

i coefficienti $\beta = \frac{dy}{dx}$, $\gamma = \frac{dy}{dz}$ rappresentano gli *effetti parziali* di ogni variabile esplicativa

Regressione semplice e regressione multipla

Da una a k variabili esplicative: il modello generale è

$$y_t = \beta_0 + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + u_t$$

- La specificazione della regressione multipla è una estensione di quella della regressione lineare semplice; le proprietà ipotizzate per il termine di errore sono, *mutatis mutandis*, le stesse. Poco da dire.
- La bontà di adattamento si valuta sempre con $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$
- la significatività di ogni variabile X_j può essere valutata *singolarmente* con il solito $t - test$

Stima di parametri e ES nella regressione multipla

Il calcolo di $\hat{\beta}_j$ e $ES(\hat{\beta}_j)$ per $j = 1, \ldots, k$ richiede l'algebra delle matrici. Posto y il vettore $T \times 1$ delle osservazioni campionarie riguardanti la variabile dipendente e X la matrice $T \times (k+1)$ delle osservazioni delle variabili esplicative (*inclusa l'intercetta*), è

$$\hat{\beta} = (X'X)^{-1}X'y$$

il vettore $(k+1) \times 1$ dei parametri stimati;

$$Var(\hat{\beta}) = s^2(X'X)^{-1}$$

(dove, al solito, $s^2 = \hat{\sigma}_u^2$) è la matrice $(k+1) \times (k+1)$ di covarianza di $\hat{\beta}$, dalla cui diagonale si estraggono gli ES:

$$ES(\hat{\beta}_j) = \sqrt{Var(\hat{\beta})_{jj}}$$

Variabili omesse nella regressione multipla

Se il "vero" DGP contiene una certa variabile, e.g.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

e il modello stimato no, e.g.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3$$

quest'ultimo risulta *incompleto*. Conseguenze:

- Gli stimatori dei parametri relativi ai regressori inclusi, $\hat{\beta}_1$ e $\hat{\beta}_3$, sono *distorti* e *inconsistenti* – a meno che la variabile omessa x_2 sia incorrelata con entrambe x_1 e x_3 (raramente succede)
- I relativi ES sono a loro volta distorti e inconsistenti
- Tutte le statistiche diagnostiche sono inconsistenti
- ... insomma il modello è da buttare *tout court*

Variabili ridondanti nella regressione multipla - 1

Se il "vero" DGP *non* contiene una certa variabile, e.g.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

e il modello stimato la include, e.g.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

quest'ultimo risulta *sovraparametrizzato*. Conseguenze:

- Sub 1) - 4), gli stimatori dei parametri $\hat{\beta}_j$ sono BLUE
- Il coefficiente della variabile ridondante verrà quindi stimato per quello che è: zero!
- Tutte le statistiche diagnostiche saranno valide, incluso il $t - test$ di significatività della variabile ridondante
- ... sulla base di tali statistiche potremo semplificare il modello.

Variabili ridondanti nella regressione multipla - 2

Insomma meglio abbondare. La sovraparametrizzazione ha, peraltro, alcune conseguenze *minori*:

- Gli $ES(\hat{\beta}_j)$ possono risultare sovrastimati (*inflazionati*)
- L'R^2 cresce *comunque* sia che si aggiunga al modello una variabile rilevante che una ridondante; pertanto esso non è un valido strumento di decisione. Si definisce a questo scopo il c.d. R^2 *corretto*:

$$\bar{R}^2 = \frac{ESS}{TSS}\frac{(T-1)}{(T-k-1)} = R^2\frac{T-1}{T-k-1}$$

  che aggiusta *ESS* e *TSS* per i rispettivi gradi di libertà, *penalizzando* le specificazioni più ricche (NB: $\bar{R}^2$ non è più compreso tra 0 e 1)

Le osservazioni qui sopra valgono anche nel caso non vi siano variabili ridondanti, ma semplicemente una certa correlazione tra le variabili esplicative (si parla di *multicollinearità*).

Significatività congiunta delle variabili

In un contesto di regressione multipla ha senso testare l'ipotesi che h variabili $X_j, \ldots X_{j+h-1}$ siano *congiuntamente* significative: ovvero confrontare la validità statistica del modello completo (U per *unrestricted*) con quello che le omette (R per *restricted*).

- La corrispondente ipotesi nulla $H_o : \beta_j = \ldots = \beta_{j+h-1} = 0$ può essere testata con la statistica

$$F = \frac{(RSS_R - RSS_U)}{RSS_U} \frac{(T-k-1)}{h} \sim F_{h,T-k-1}$$

- Nel *caso particolare* in cui si testi l'omissione di *tutti* i regressori tranne l'intercetta (modello completo vs. modello *vuoto*), si ottiene una statistica F tale che (è $ESS = RSS_0 - RSS$):

$$F = \frac{ESS}{RSS} \frac{(T-k-1)}{k} = \frac{R^2/k}{(1-R^2)/(T-k-1)} \sim F_{k,T-k-1}$$