Statistica per l'impresa

6.2 Correlazione e Regressione semplice

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
- ...

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).

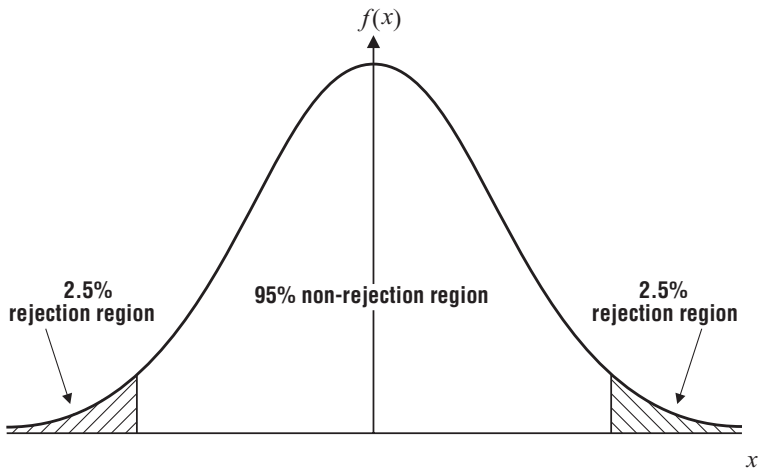
- vogliamo verificare $H_0 : \mu = 180$ al livello di confidenza del 5%

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 2

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.

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*.

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.

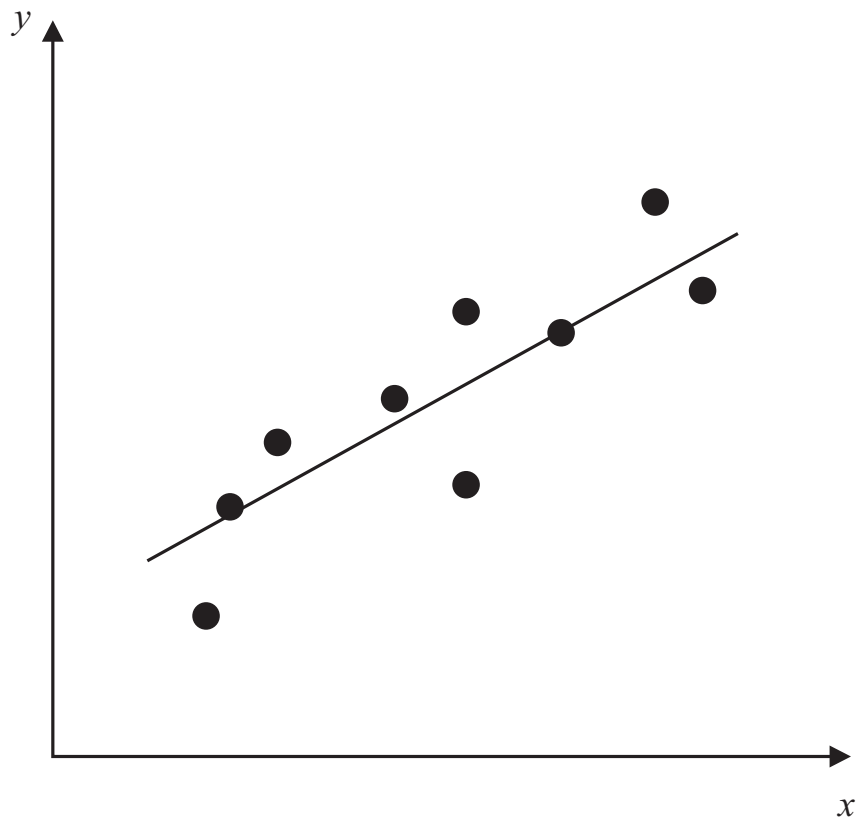
$$y_i = \alpha + \beta x_i + u_i$$

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

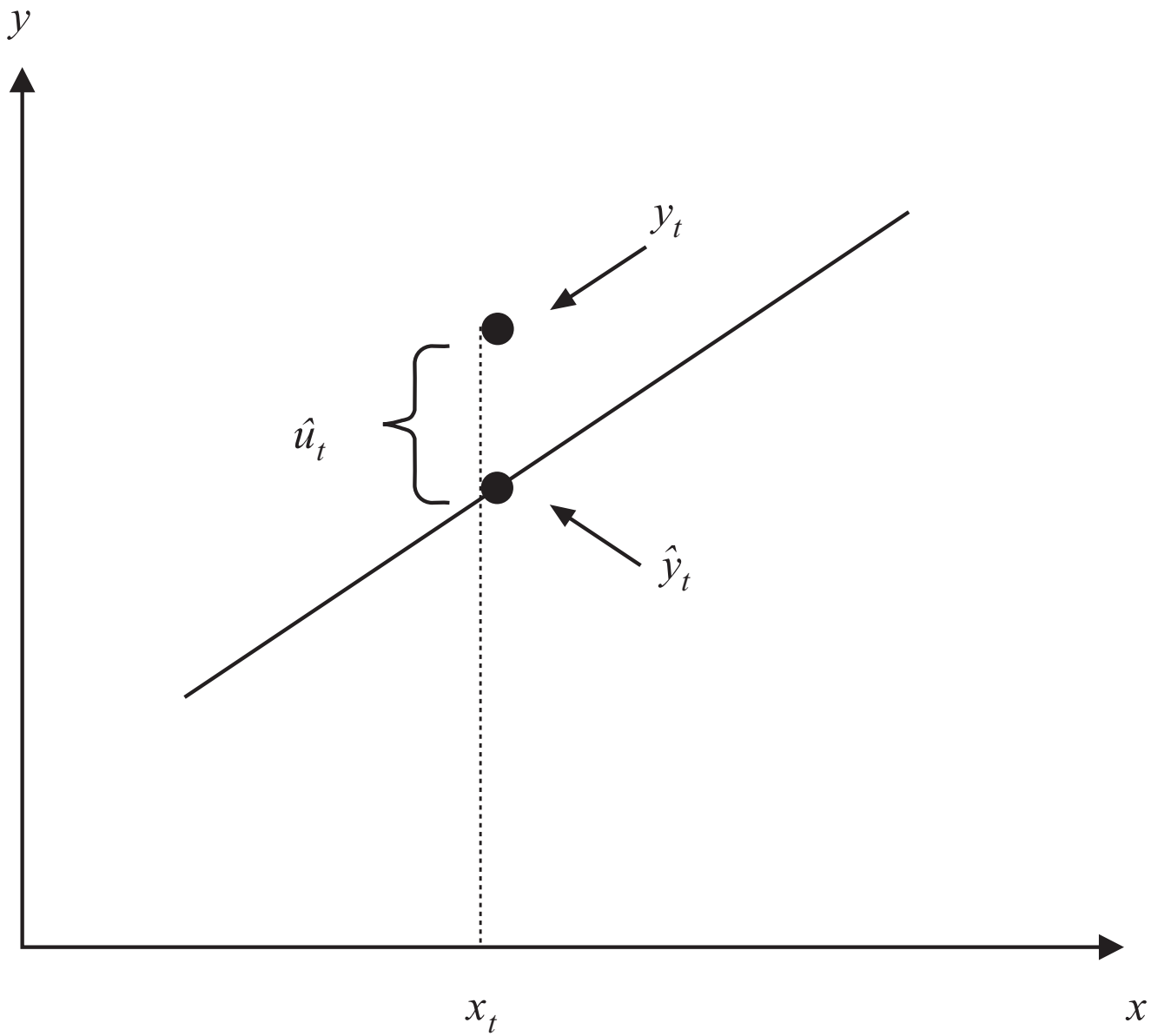
- Il metodo di stima più comune è noto come OLS (*ordinary least squares*, o minimi quadrati ordinari).
- Si minimizzano i quadrati delle distanze indicate in figura (da cui il nome).
- Più formalmente, siano

  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 $\hat{y}_t = \hat{\alpha} + \beta x_t$ 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 $\hat{\beta}$

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

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

usando

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

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

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

Da (2),

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

Da (3),

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{5}$$

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).

The Assumptions Underlying the CLRM

- The model which we have used is known as the classical linear regression model.
- We observe data for x_t, but since y_t also depends on u_t, we must be specific about how the u_t are generated.
- We usually make the following set of assumptions about the u_t's (the unobservable error terms):

Technical notation	Interpretation
(1) $\mathrm{E}(u_t) = 0$	The errors have zero mean
(2) $\mathrm{var}(u_t) = \sigma^2$	The variance of the errors is constant and finite over all values of x_t
(3) $\mathrm{cov}(u_i, u_j) = 0$	The errors are linearly independent of one another
(4) $\mathrm{cov}(u_t, x_t) = 0$	There is no relationship between the error and corresponding x variate

The Assumptions Underlying the CLRM (Cont'd)

- An alternative assumption to (4), which is slightly stronger, is that the x_t's are non-stochastic or fixed in repeated samples.
- A fifth assumption is required if we want to make inferences about the population parameters (the actual α and β) from the sample parameters ($\hat{\alpha}$ and $\hat{\beta}$)
- Additional assumption

  (5) u_t is normally distributed

Properties of the OLS Estimator

- If assumptions (1) through (4) hold, then the estimators and determined by OLS are known as Best Linear Unbiased Estimators (BLUE).

  What does the acronym stand for?

- 'Estimator' – $\hat{\alpha}$ and $\hat{\beta}$ are estimators of the true value of α and β
- 'Linear' – $\hat{\alpha}$ and $\hat{\beta}$ are linear estimators
- 'Unbiased' – on average, the actual values of $\hat{\alpha}$ and $\hat{\beta}$ will be equal to their true values
- 'Best' – means that the OLS estimator $\hat{\beta}$ has minimum variance among the class of linear unbiased estimators; the Gauss–Markov theorem proves that the OLS estimator is best.

Consistency/Unbiasedness/Efficiency

- Consistent

  The least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ are consistent. That is, the estimates will converge to their true values as the sample size increases to infinity. Need the assumptions $E(x_t u_t) = 0$ and $Var(u_t) = \sigma^2 < \infty$ to prove this. Consistency implies that

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

- Unbiased

  The least squares estimates of $\hat{\alpha}$ and $\hat{\beta}$ are unbiased. That is $E(\hat{\alpha}) = \alpha$ and $E(\hat{\beta}) = \beta$. Thus on average the estimated value will be equal to the true values. To prove this also requires the assumption that $E(u_t) = 0$. Unbiasedness is a stronger condition than consistency.

Consistency/Unbiasedness/Efficiency (Cont'd)

- Efficient

  An estimator $\hat{\beta}$ of parameter β is said to be efficient if it is unbiased and no other unbiased estimator has a smaller variance. If the estimator is efficient, we are minimising the probability that it is a long way off from the true value of β.

Precision and Standard Errors

- Any set of regression estimates of and are specific to the sample used in their estimation.
- Recall that the estimators of α and β from the sample parameters ($\hat{\alpha}$ and $\hat{\beta}$) are given by

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

Precision and Standard Errors (Cont'd)

- What we need is some measure of the reliability or precision of the estimators ($\hat{\alpha}$ and $\hat{\beta}$). The precision of the estimate is given by its standard error. Given assumptions (1)–(4) above, then the standard errors can be shown to be given by

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

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

where s is the estimated standard deviation of the residuals.

Estimating the Variance of the Disturbance Term

- The variance of the random variable $u\ t$ is given by

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

  which reduces to

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

- We could estimate this using the average of u_t^2:

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

- Unfortunately this is not workable since u_t is not observable. We can use the sample counterpart to u_t, which is $\hat{u}_t$:

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

  But this estimator is a biased estimator of σ^2.

Estimating the Variance of the Disturbance Term (cont'd)

- An unbiased estimator of σ is given by

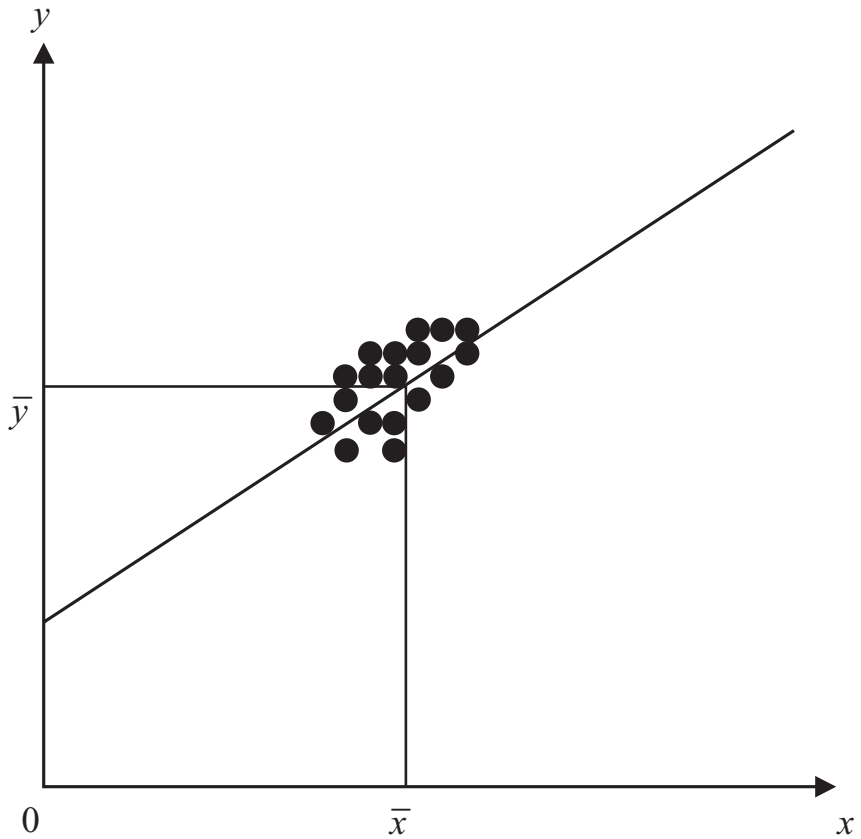
$$s = \sqrt{\frac{\sum \hat{u}_t^2}{T-2}}$$

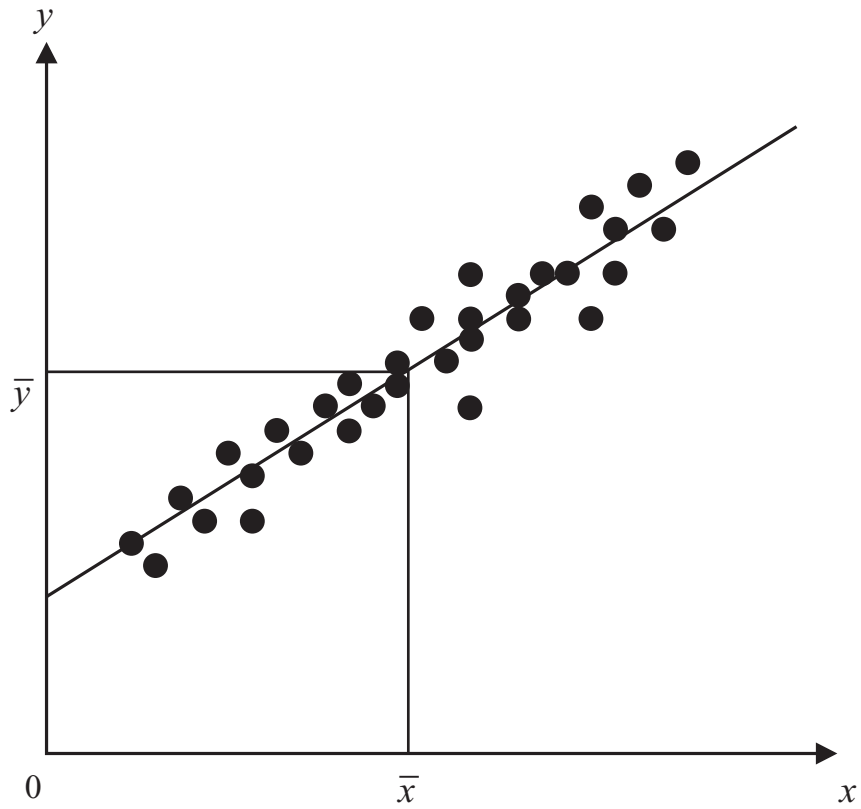
where $\sum \hat{u}_t^2$ is the residual sum of squares and T is the sample size.

- Some Comments on the Standard Error Estimators
  1. Both SE($\hat{\alpha}$) and SE($\hat{\beta}$) depend on s^2 (or s). The greater the variances^2, then the more dispersed the errors are about their mean value and therefore the more dispersed y will be about its mean value.
  2. The sum of the squares of x about their mean appears in both formulae. The larger the sum of squares, the smaller the coefficient variances.

Some Comments on the Standard Error Estimators

Consider what happens if $\sum (x_t - \bar{x})^2$ is small or large:





1. The larger the sample size, T, the smaller will be the coefficient variances. T appears explicitly in SE($\hat{\alpha}$) and implicitly in SE($\hat{\beta}$).

Some Comments on the Standard Error Estimators (Cont'd)

T appears implicitly since the sum $\sum (x_t - \bar{x})^2$ is from $t = 1$ to T.

2. The term $\sum x_t^2$ appears in the SE($\hat{\alpha}$).

   The reason is that $\sum x_t^2$ measures how far the points are away from the y-axis.