Università degli studi di Ferrara
Dipartimento di Matematica
A.A. 2019/2020 – I semestre

STATISTICA MULTIVARIATA

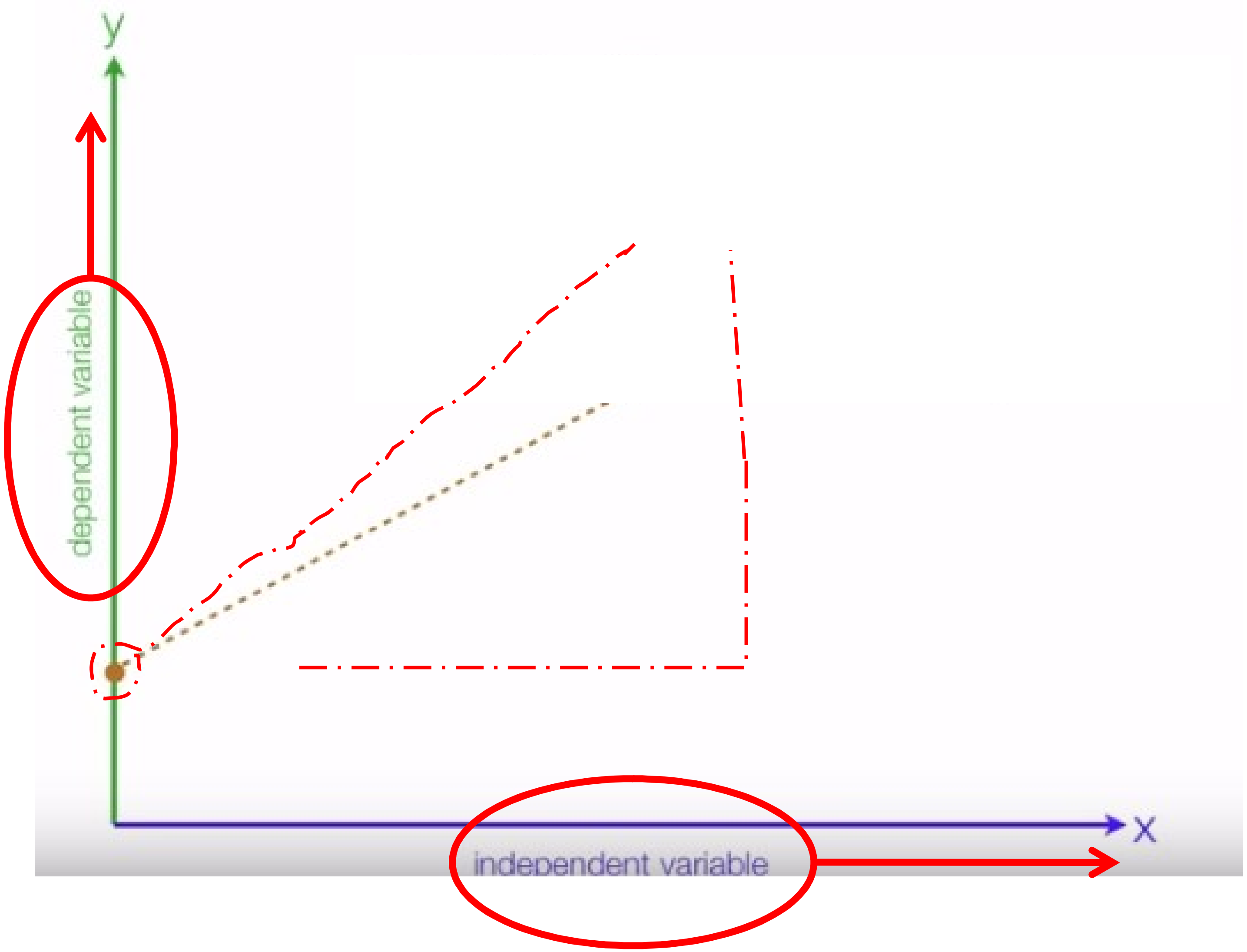
SSD MAT/06

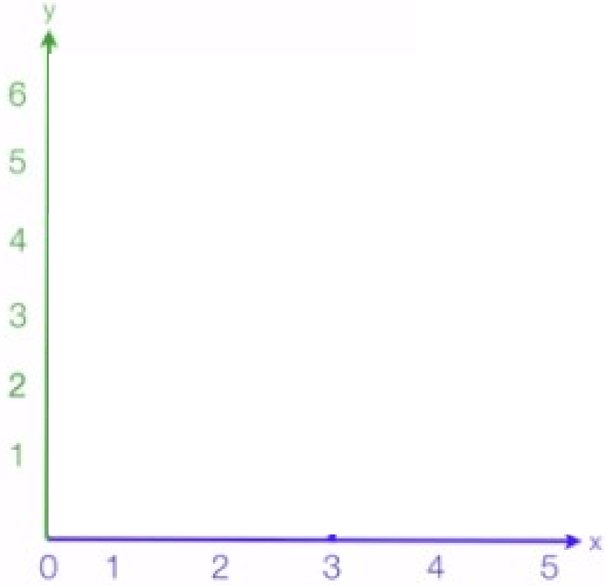
LEZION 3 - Regressione lineare: teoria, pratica e software (R)

Docente: Valentina MINI

valentina.mini@unife.it

RICEVIMENTO: lunedì 2-4 previa mail


y
dependent variable
x
independent variable



x	y
1	2
2	4
3	5
4	4
5	5

STRUTTURA DELLA LEZIONE

- RIPRENDIAMO I **CONCETTI DELLA REGRESSIONE LINEARE**
- **STIMA DEI COEFFICIENTI** DI REGRESSIONE b_0 E b_1 – APPROCCIO DEI MINIMI QUADRATI
- **INTERPRETAZIONE** DEI COEFFICIENTI DI REGRESSIONE
- LA **BONTA' DEL MODELLO**: COEFFICIENTE DI DETERMINAZIONE (Rq)
- **ERRORE** NELLA REGRESSIONE: L'ERRORE STANDARD DELLE STIME
- **4 ASSUNZIONI** ALLA BASE DEL MODELLO DI REGRESSIONE
- **SIGNIFICATIVITA'** NELLA POPOLAZIONE?
- INFERENZA SUL **COEFFICIENTE DI CORRELAZIONE**
- *ESERCITAZIONI*

Obiettivi

- Utilizzare regressione per stimare y

  (y = var dipendente, x = var indipendente)
- Stimare e sapere significato di coefficienti regressivi b_0 e b_1
- Valutare il rispetto delle assunzioni alla base della regressione
- Fare inferenza (su coefficienti e valori y).

modello base regressione lineare

Scopo: stimare una dipendente (y)

Variabile quantitativa a partire da dati osservati relativi ad un'altra

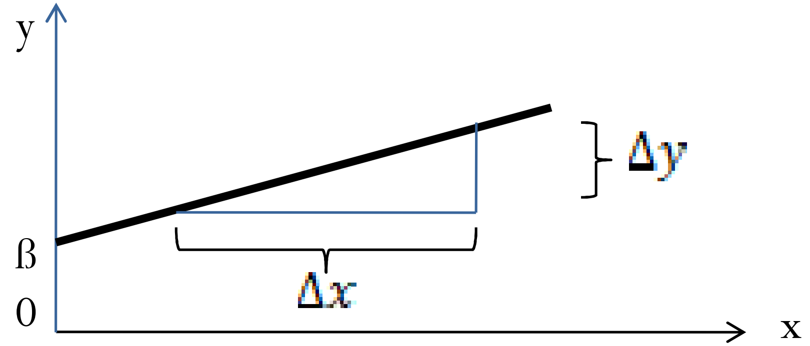
variabile detta indipendente o esplicativa (x)

IL MODELLO PIU' SEMPLICE: regressione lineare semplice

1x → per stimare e prevedere 1 y

Relazione lineare

GRAFICAMENTE: sul piano cartesiano è LINEA RETTA



modello base regressione lineare

FORMALMENTE:

$$Y_i = \text{ß}_0 + \text{ß}_1 x_i + e_i$$

Valore i-esimo della variabile dipendente

Intercetta sull’asse Y nella popolazione

Valore medio di y quando X=0

Coefficiente angolare nella popolazione

Variazione attesa nella y per variazioni unitarie in X

Valore osservato della variabile indipendente

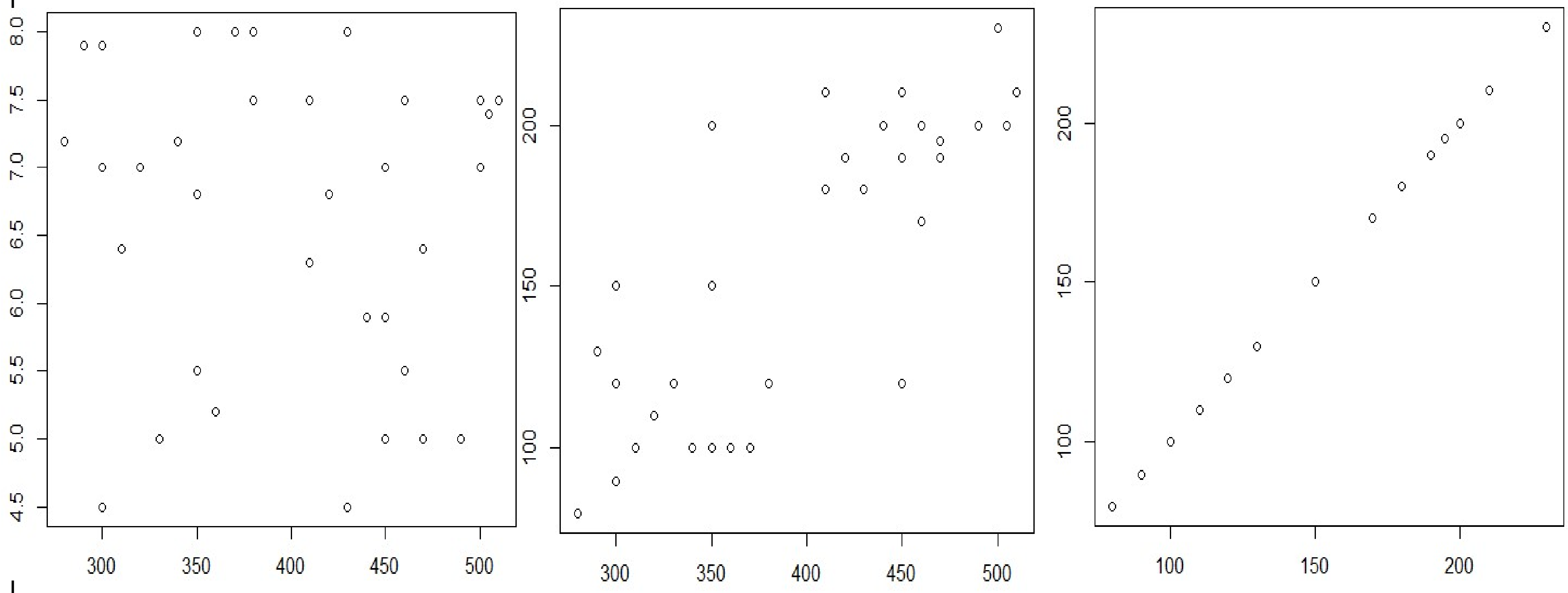
Errore nel valore y per l’i-esima osservazione

Scostamento casuale di Yi da quello atteso (sulla retta).

Residuo del modello

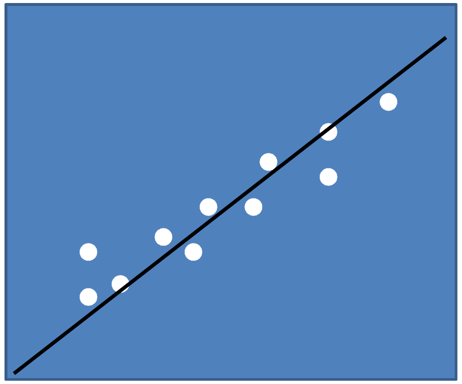
Simple Linear Regression Analysis: checking the relationship between two variables

- Examples of scatter plots

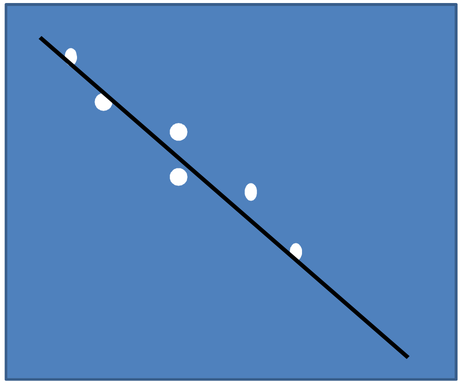


Simple Linear Regression Analysis: checking the relationship between two variables

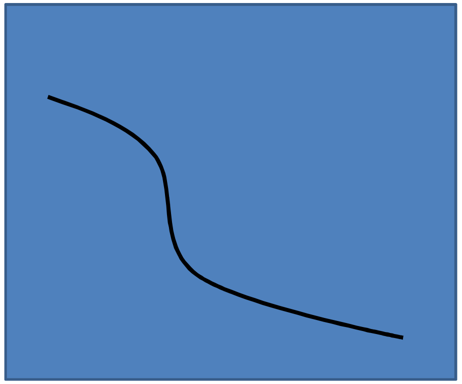
Using the scatter plot we do individuate the possible relationship between two observed variables



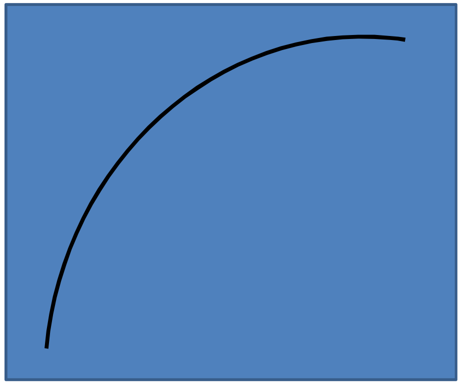
Linear positive relation



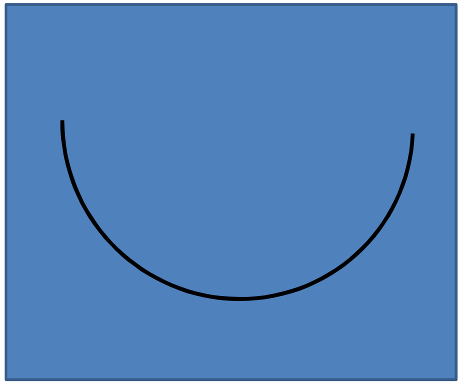
Linear negative relation



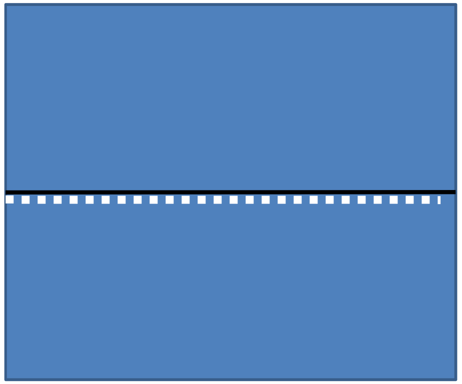
Non-linear relation



Exponential relation



U relation



Absence of relation

EX 1- ESEMPIO PRATICO

La Mini Market, una catena di negozi alimentari, per definire le nuove localizzazioni vuole procedere in modo sistematico, partendo dalla constatazione che l'ampiezza del punto vendita influenza il volume di vendite.

Il direttore commerciale chiede se esiste una procedura statistica per capire se effettivamente l'ampiezza dei punti vendita influenzi il volume di vendite.

DATI RACCOLTI:

14 negozi,

per ognuno si conosce estensione (mq) e

il volume di vendite annuali (milioni di Euro)

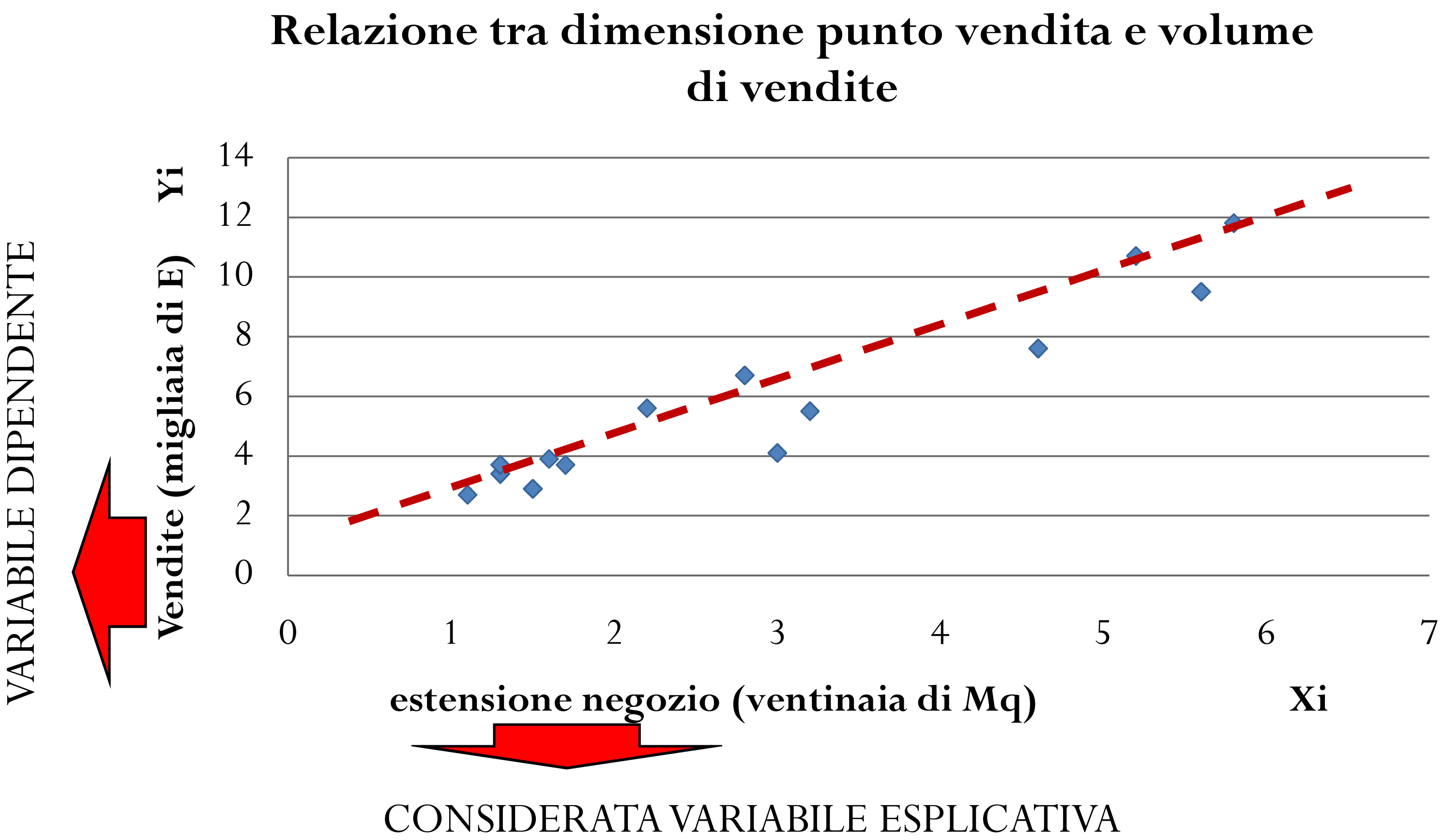
EX 1- esempio

ID negozio	Estensione (centinaia di Mq)	Vendite (migliaia di €)
1	1,7	3,7
2	1,6	3,9
3	2,8	6,7
4	5,6	9,5
5	1,3	3,4
6	2,2	5,6
7	1,3	3,7
8	1,1	2,7
9	3,2	5,5
10	1,5	2,9
11	5,2	10,7
12	4,6	7,6
13	5,8	11,8
14	3	4,1

Domanda: in fase esplorativa, cosa possiamo dire graficamente della relazione tra le due variabili in esame?

EX 1- esempio

1. PASSO: RAPPRESENTAZIONE GRAFICA DELLA RELAZIONE TRA LE VARIABILI



2. PASSO:
STIMARE I COEFFICIENTI DI REGRESSIONE

- Usando i dati campionari possiamo calcolare b_0 come stima di $ß_0$ e b_1 come stima di $ß_1$
- Con tali stime otteniamo il modello di regressione lineare semplice

$$\hat{y}_i = b_0 + b_1 x_1$$

- Approccio più comune: METODO DEI MINIMI QUADRATI
  - Minimizziamo la somma dei quadrati degli scarti tra i valori osservati e quelli stimati.

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Trasformata in equazione a due incognite in modo da minimizzare lo scostamento tra le due variabili dipendenti (osservata e modellata)
- Determinazione di b_1 (SSXY/SSX)
- Determinazione di b_0 (Y-b_1X)

2. PASSO:
STIMARE I COEFFICIENTI DI REGRESSIONE

$$b_1 = ssxy/ssx$$

SSXY=

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

SSX=

$$\sum_{1=i}^{n}(x_i - \bar{x})^2 = \sum_{1=i}^{n} x_i^2 - \frac{(\sum_{1=i}^{n} x_i)^2}{n}$$

$$b_0 = \bar{Y} - b_1 \bar{x}$$

$$\bar{Y} = \frac{\sum_{1=i}^{n} y_i}{n}$$

$$\bar{x} = \frac{\sum_{1=i}^{n} x_i}{n}$$

2. PASSO:
STIMARE I COEFFICIENTI DI REGRESSIONE

5 "quantità" per calcolare b_1 e b_0

- n
- $\sum_{i=1}^{n} x$
- $\sum_{i=1}^{n} y$
- $\sum_{i=1}^{n} xx$
- $\sum_{i=1}^{n} xy$

Seguendo il nostro esempio del mini market calcoliamo i coefficienti di regressione per i nostri dati, utilizzando il metodo dei minimi quadrati

Interpretazione dei coefficienti applicati alla realtà osservata

EX 1- esempio

ID negozio	Mq (in centinaia) X	€ fatturato (in migliaia) y	X^2	X*Y
1	1,7	3,7	2,89	6,29
2	1,6	3,9	2,56	6,24
3	2,8	6,7	7,84	18,76
4	5,6	9,5	31,36	53,2
5	1,3	3,4	1,69	4,42
6	2,2	5,6	4,84	12,32
7	1,3	3,7	1,69	4,81
8	1,1	2,7	1,21	2,97
9	3,2	5,5	10,24	17,6
10	1,5	2,9	2,25	4,35
11	5,2	10,7	27,04	55,64
12	4,6	7,6	21,16	34,96
13	5,8	11,8	33,64	68,44
14	3	4,1	9	12,3
14	**40,9**	**81,8**	**157,41**	**302,3**

EX 1- esempio

$b_1 = SSXY/SSX =$

$SSXY = 302.3-(40.9*81.8)/14 = 63.3271$

$SSX = 157- (40.9*40.9)/14 = 37.92358$

$= 63.3271/37.9235 = 1.6699$

$b_0 = (81.8/14) – 1.6699*(40.9/14)$

$= 5.843-4.8785 =$

$= 0.9645$

MODELLO STIMATO

Y = 0.9645+1.6699 Xi

ID negozio	Mq (in centinaia) X	€ fatturato (in migliaia) y	MODELLO STIMATO
1	1,7	3,7	3.8
2	1,6	3,9	3.64
3	2,8	6,7	5.64
4	5,6	9,5	10.31
5	1,3	3,4	3.13
6	2,2	5,6	4.64
7	1,3	3,7	3.13
8	1,1	2,7	…
9	3,2	5,5	…
10	1,5	2,9	…
11	5,2	10,7	…
12	4,6	7,6	…
13	5,8	11,8	…
14	3	4,1	5.97
14	**40,9**	**81,8**	

EX 1- esempio

INTERPRETAZIONE DEI COEFFICIENTI DI REGRESSIONE

b_1 = per ogni incremento unitario della variabile X, si stima che il valore medio di Y cresca di 1.9966 unità. Ovvero:

aumentando l'ampiezza del negozio di 1 centinaio di mq, si stima che le vendite annuali aumentino di 1.966 milioni di Euro.

→*Il coefficiente angolare è quella parte di volume di vendite che si stima dipendente dalla grandezza del negozio*.

b_0= valore medio di Y quando x=0. Siccome un negozio non può di 0 mq → b_0 è nel nostro caso di scarsa interpretazione pratica; infatti il il valore di Xi è

FUORI DAL RANGE!
COSA SIGNIFICA?

EX 1- esempio

LA PREVISIONE

Quando utilizziamo un modello di regressione per scopi predittivi è necessario considerare **SOLO l'intervallo rilevante per i valori della var. dipendente**; tale intervallo corrisponde al **range di osservazioni di X (Xmin; Xmax)**

Non possiamo estrapolare info "sopra o sotto al range", in quanto la relazione al di fuori dei valori osservati potrebbe cambiare.

Nostro esempio:

Non possiamo estrapolare valori Y di vendite al di sotto di valori della x pari a 1.1 centinaia di Mq e al di sopra di 580 mq.

Esercizi

- Esercizi in lettura consigliata Levin et al. N.12.1 p 439
- Esercizi in lettura consigliata Levin et al. N.12.2 p 439

Es. È stato stimato un modello di regressione ottenendo:

$$Y_i = 4+8X_i$$

a) Interpretare il significato di intercetta b_0 (4)
b) Interpretare il significato di coefficiente angolare b_1(8)
c) Se i valori osservati di x appartengono all'intervallo(2;25), si chiede di usare il modello per stimare/predire valori Y quando X è uguale a
   a) 3
   b) -3
   c) 0
   d) 24

LA “BONTÀ” DEL MODELLO

Quanto il modello stimato si adatta ai dati?
Quanto “buono” è il modello stimato?

Per valutarlo ricorriamo al **COEFFICIENTE DI DETERMINAZIONE R^2**

R^2 = $\dfrac{\text{devianza di regressione}}{\text{devianza totale}}$

→ (variabilità spiegata dal modello)
→ (variabilità totale della Y)

= misura la porzione di variabilità della Y spiegata dalla relazione con la X all’interno del modello

= $\dfrac{SSR}{SST}$

→ somma degli scarti
→ somma degli scarti totali

Vi è una parte di devianza residua dovuta a fattori esterni alla relazione tra X e Y e non spiegata dal modello (SSE).

formalmente: la variazione totale è composta da due parti

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares

$$SST = \sum (Y_i - \bar{Y})^2 \qquad SSR = \sum (\hat{Y}_i - \bar{Y})^2 \qquad SSE = \sum (Y_i - \hat{Y}_i)^2$$

dove:

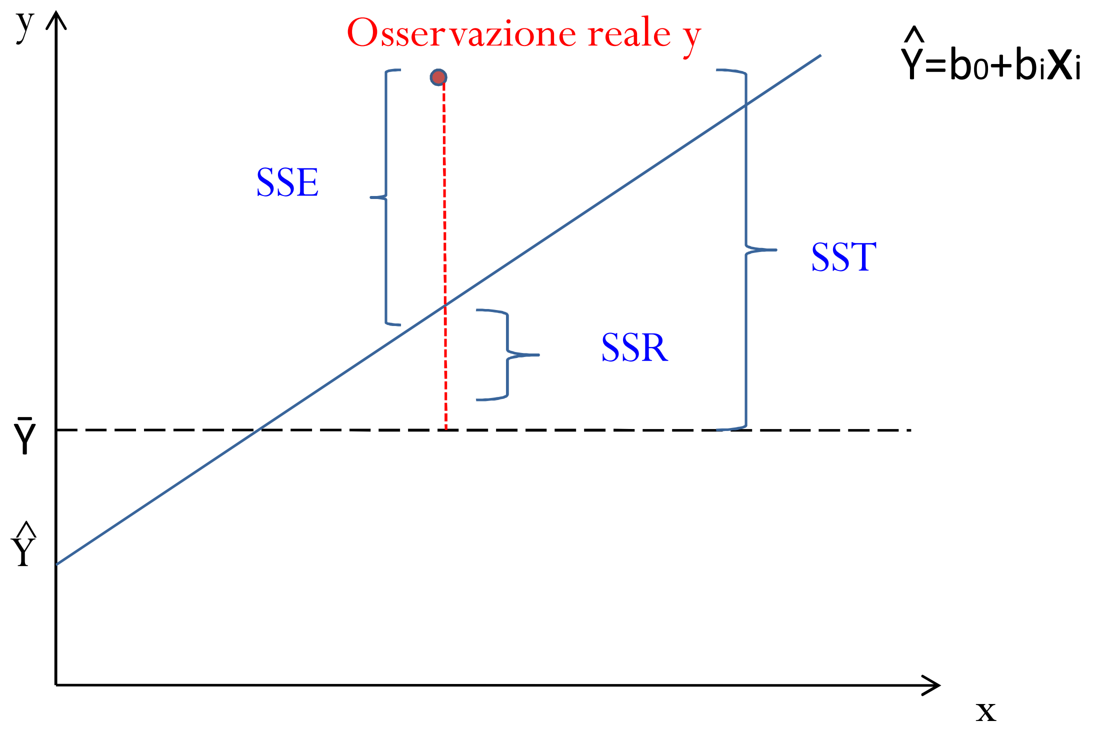
$\bar{Y}$ = Mean value of the dependent variable

Y_i = Observed value of the dependent variable

$\hat{Y}_i$ = Predicted value of Y for the given X_i value

LA "BONTÀ" DEL MODELLO

RAPPRESENTAZIONE GRAFICA



$$SST = SSE + SSR$$

$R^2 = SSR/SST$ → possiamo esprimerlo in %

EX 1- esempio

R^2= SSR/SST

SSR = SOMMATORI A $(\hat{Y}-\bar{Y})^2$

SST<<SSR+SSE = SOMMATORIA $(y_i-\bar{Y})^2$

$\hat{Y}$	$\hat{Y}-\bar{Y}$	$(\hat{Y}-\bar{Y})^2$ SSR	$y_i-\bar{Y}$	$(y_i-\bar{Y})^2$ SST
3,8	3,8-5,84= -2,04	4,16	-2,14	4,58
3,64	-2,2	4,84	-1,94	3,76
5,64	-0,2	0,04	0,86	0,79
10,31	4,47	19,98	3,66	13,39
3,13	-2,71	7,34	-2,44	5,95
4,64	-1,2	1,44	-0,24	0,06
3,13	-2,71	7,34	-2,14	4,58
2,8	-3,04	9,24	-3,14	9,86
6,3	0,46	0,21	-0,34	0,11
3,47	-2,37	5,62	-2,94	8,64
9,65	3,81	14,52	4,86	23,62
8,65	2,81	7,89	1,76	3,1
10,65	4,81	23,13	5,96	35,52
5,97	0,13	0,01	-1,74	3,03
		105,76		116,99

EX 1- esempio

R^2= 105.72/116.99=0.903669=0.904

INTERPRETAZIONE di R^2

→ Il 90.4% della variabilità delle vendite dell'azienda "mini Market" è spiegata dalle dimensioni del punto vendita

→ Vi è una significativa relazione lineare tra le due variabili osservate.

→ Solo il (1-0.904=)9.6% è variabilità dovuta ad altri fattori che non dipendono dalla relazione tra x (dimensione) e y (vendite).

Dove troviamo R^2 nell'output del software

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$R^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

L’ERRORE NEL MODELLO

Il metodo dei minimi quadrati identifichi la retta interpolante più vicina ai dati campionari osservati (minimo errore)

TUTTAVIA

La retta <u>non</u> rappresenta un modello perfetto

→ Ci saranno dati attorno alla retta di regressione (non tutti sulla retta)

→ ERRORE STANDARD DELLE STIME: variabilità dei valori osservati (Y_i) rispetto ai valori stimati ($\hat{Y}$)

Errore standard:

$$S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

NEL NOSTRO ESEMPIO:

SSE= somma$(Y_i-\hat{Y})^2$

n=14 → (n-2=12)

→ Syx= 0.966

INTERPRETAZIONE

Errore standard è 0.966, quindi pari a 966’000 Euro.

→ La deviazione media le vendite annuali osservate e quelle stimate dal modello è approssimativamente pari a 966’000 Euro.

LE 4 ASSUNZIONI BASE DEL MODELLO DI REGRESSIONE

!!! SE VIOLATE COMPROMETTONO LA VALIDITA' DEI RISULTATI !!!

1. Linearità = assumiamo che la relazione tra variabili studiate **sia lineare**
2. Indipendenza dei residui (e_i) = i residui devono **essere indipendenti** (nb. per rilevazioni periodiche) rispetto alla var. esplicativa
3. Normalità dei residui (ei) = assumiamo che i residui siano **distribuiti N** per ogni valore di x
4. Omoschedasticità (delle varianze dei e_i)= la varianza dei residui deve essere **costante** per tutti i valori di x.

ANALISI GRAFICA DEI RESIDUI PER VERIFICARE LA VALIDITA' DELLE ASSUNZIONI

EX 1- esempio

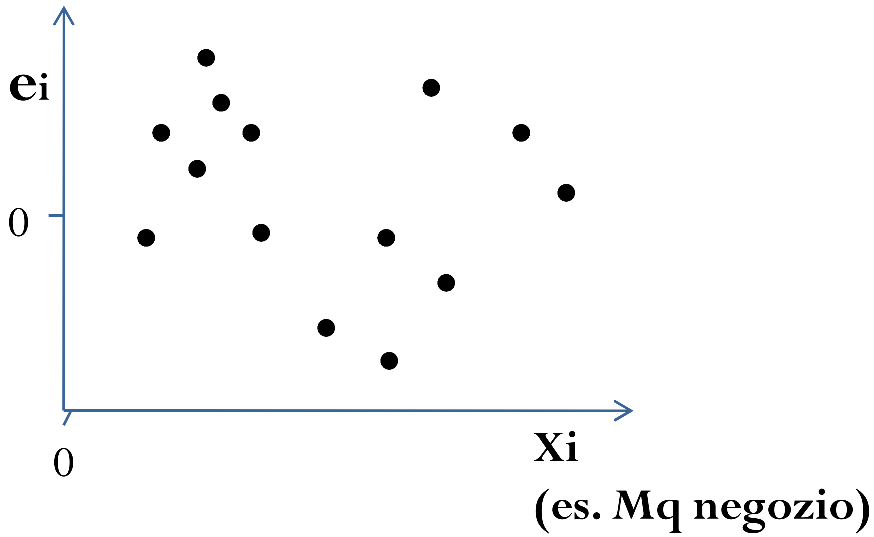
ANALISI GRAFICA DEI RESIDUI PER VERIFICARE LA VALIDITA' DELLE ASSUNZIONI

Nel modello i residui sono:

$$\varepsilon_i = y_i - \hat{y}_i$$

ovvero la differenza tra i valori osservati e quelli stimati

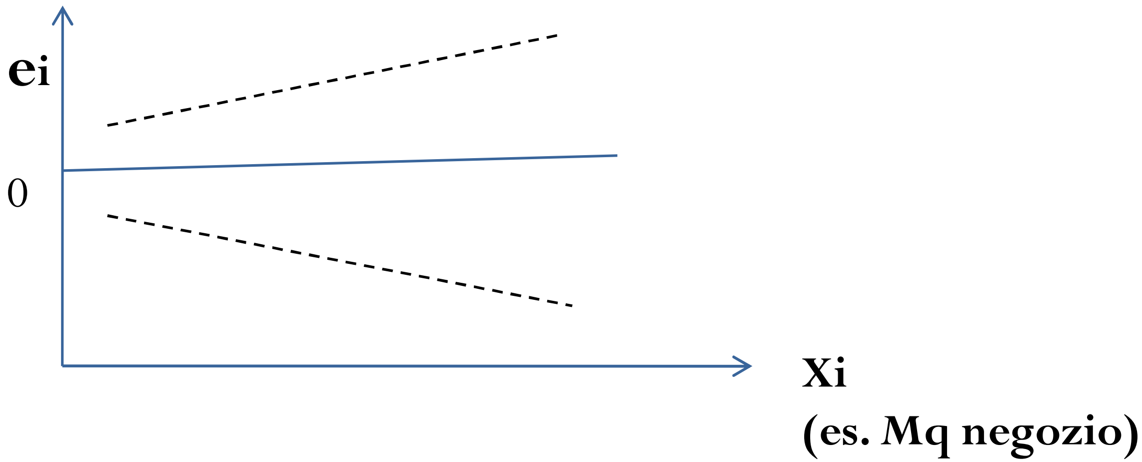
→ Analisi grafica tra e_i e X_i



1. Linearità = non devono esserci pattern tra **X** ed **e**
2. *Indipendenza = solo per raccolte consecutive*
3. Normalità = indagata attraverso l'istogramma dei residui
4. Omoschedasticità = al variare della X non vi è una differenza nella variazione dell'errore

ANALISI GRAFICA DEI RESIDUI PER VERIFICARE LA VALIDITA' DELLE ASSUNZIONI

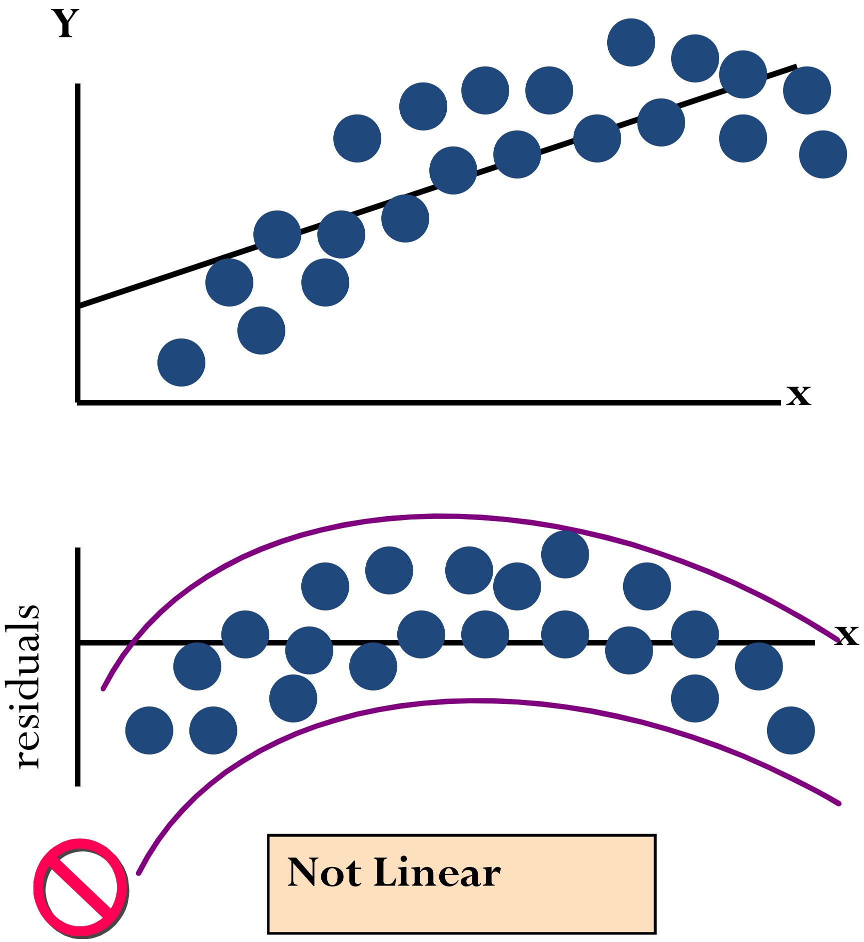
Esempio di violazione della OMOSCHEDASTICITA'

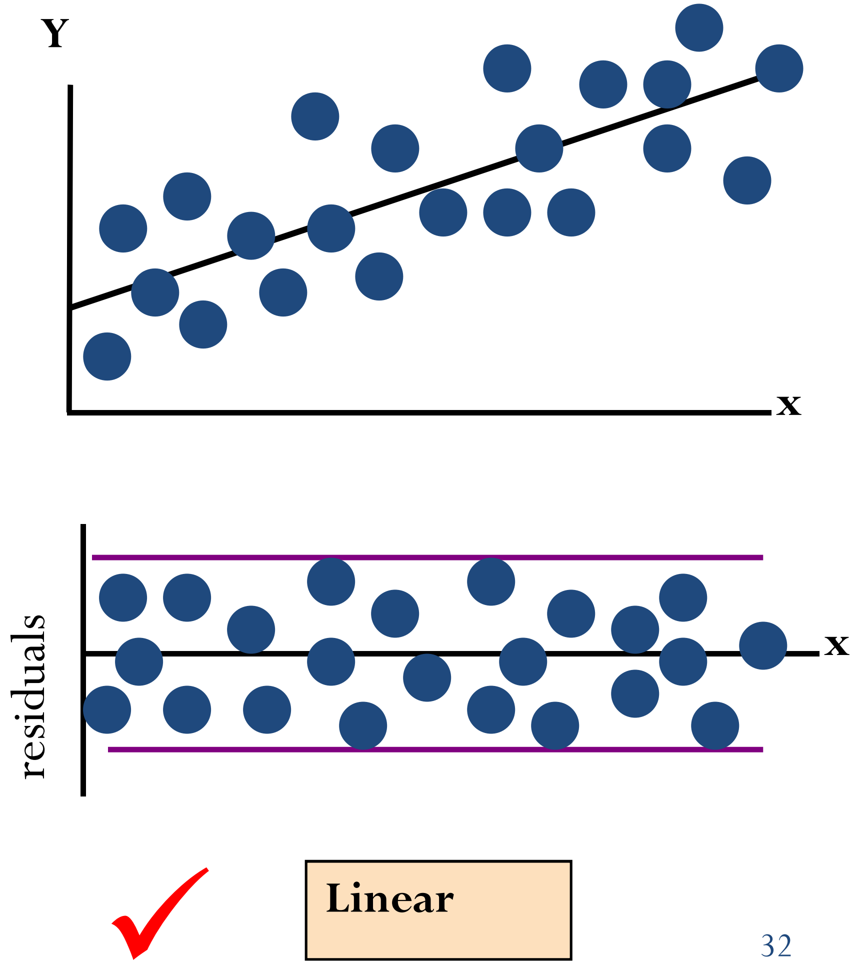


La variabilità dell'errore e_i aumenta in modo molto marcato all'aumentare di X_i → mancanza di omogeneità nella variabilità di e_i per ogni livello di X_i

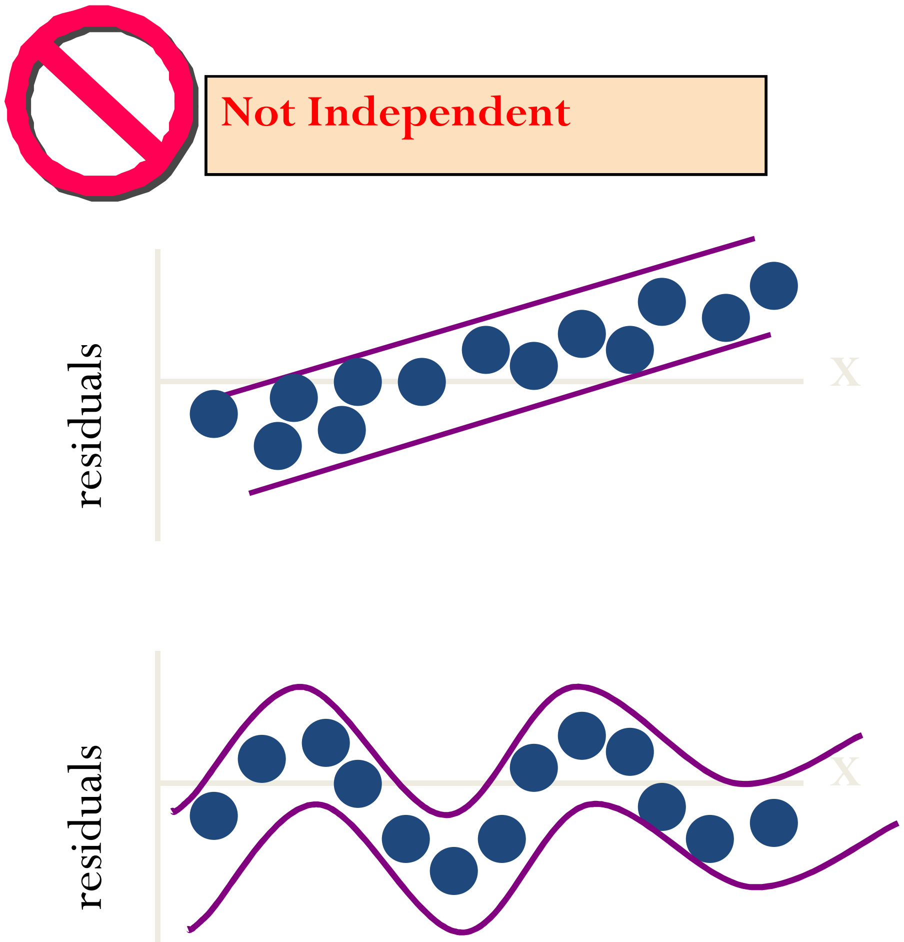
LINEARITA'

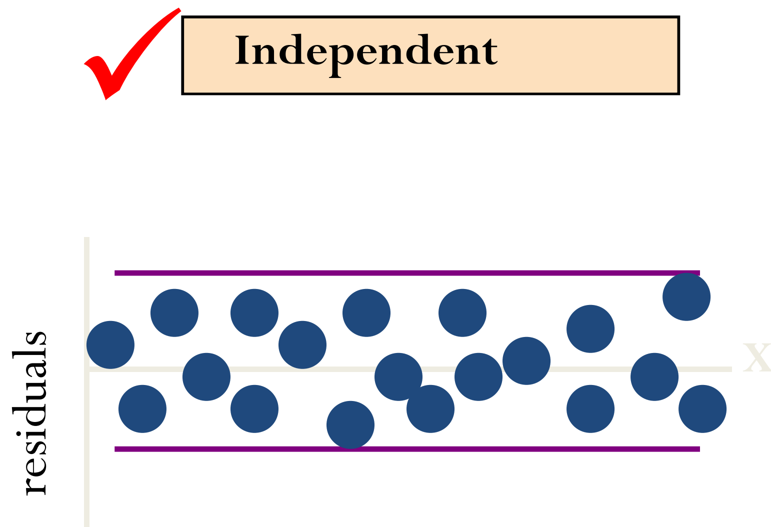
Analysis of residuals



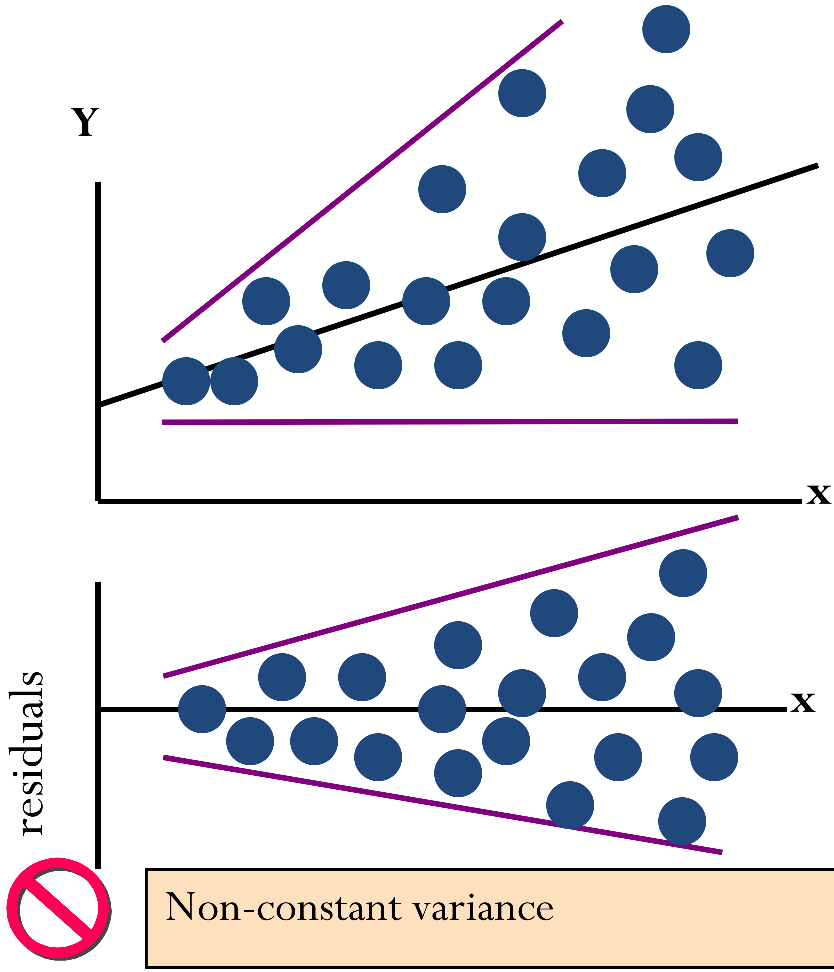


INDIPENDENZA

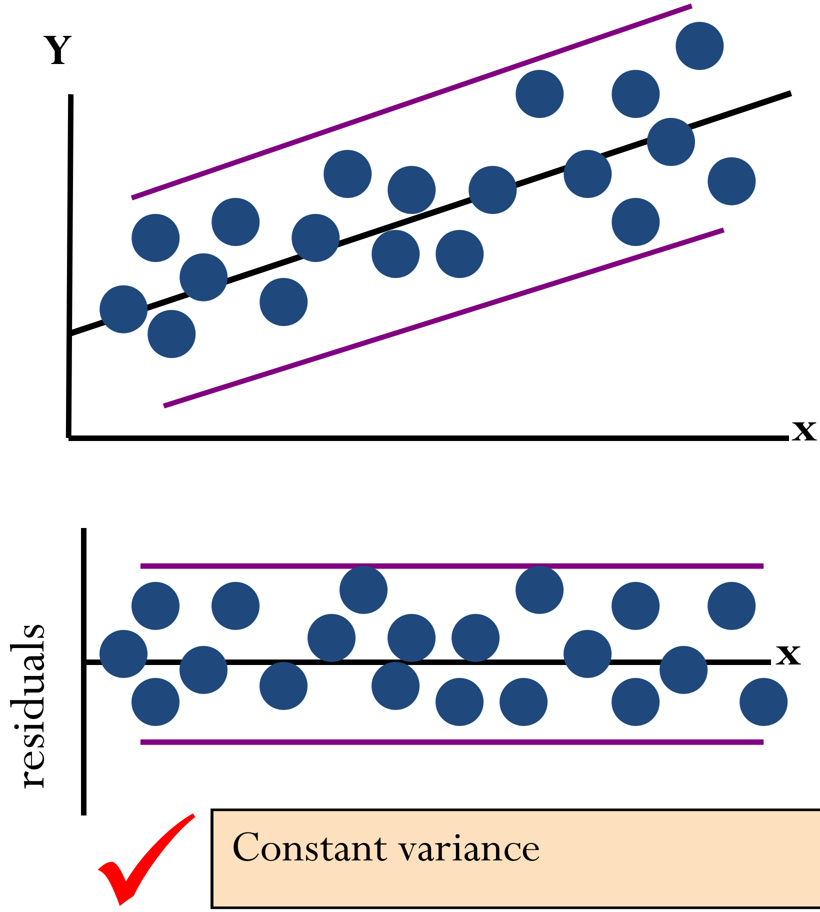




OMOSCHEDASTICITA'



Non-constant variance



Constant variance

NORMALITA'

- ISTOGRAMMA DEI RESIDUI
- COSTRUIRE UN Normal Probability Plot dei residui

…punto della situazione…

Regressione

→ Metodo dei minimi quadrati per determinare b_0 e b_1

→ Calcolo coefficiente di determinazione (R^2)

→ Calcolo errore standard (S_{yx})

→ Analisi dei residui per verificare le assunzioni base (relazione tra e_i - x_i)

→ **SE** le assunzioni **sono rispettate** → ci chiediamo

LA RELAZIONE TRA X E Y
NELLA POPOLAZIONE
E' SIGNIFICATIVA?

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE

Partendo da un campione per fare inferenza abbiamo calcolato le stime.

Per verificare che

tali stime stiano “significativamente rispecchiando la realtà”,

dobbiamo esaminare **SE** nella realtà (nella popolazione)

la relazione lineare **è significativa**

IMPOSTIAMO LE IPOTESI:

- H_0 = ipotesi nulla
- H_1 = ipotesi alternativa

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE

Per verificare che la relazione tra X e Y nella popolazione sia significativa, dobbiamo verificare l'ipotesi che **ß1=0**

Ovvero

A) Se il coefficiente angolare ß1=0 →

al variare di x non vi è alcuna variazione di Y

$Y_i = ß_0 + 0$ → retta costante parallela a x

B) Se ß1≠ 0 → vi è una variazione in Y al variare di X

PROCEDIAMO CON IL TEST DI IPOTESI

Se si rifiuta H_0 vi è evidenza empirica dell'esistenza della relazione lineare tra X e Y nella popolazione

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE

Elementi da utilizzare:

Valore Critico (V.C.) e t-stat

SE t-stat **>** valore critico → SI RIFIUTA H_0 (esiste una relazione lineare tra X e Y)

T-stat= (b1-ß1)/Sb1

Dove :
b1= coefficiente angolare campionario
ß1=valore ipotizzato nella popolazione
(=0 per H0)
Sb1= errore standard del coefficiente angolare campionario

VALORE CRITICO $t_{a/2}$

-Con a fissato a priori
-Con gradi di libertà n-2

→ Tavola E3 di t-student

EX 1- esempio

Verifichiamo che vi sia evidenza empirica per la relazione lineare tra *Mq di dimensione* e *volume delle vendite* dei punti vendita.

A priori fissiamo un livello di significatività a=0.05 (95%)

A) Valore critico:
   A) n-2= 12 gradi di libertà
   B) a=0.05 → a/2 = 0.025
   C) TAVOLA E3 → vc = 2.1788
B) t-stat:
   A) B1= 1.6699 ß1=0 (imposto per H0)
   B) Syx= 0.9664
   C) Calcolo di SSX (sotto radice)
   D) Otteniamo Sb1= 0.9664/RQ SSX = 0.1569

   → t-stat = (1.6699-0)/0.1569 = 10.64

EX 1- esempio

t-stat (10.64)> vc (2.1788)

→RIFIUTO H0

→Esiste evidenza empirica che

a livello di significatività 0.05 (confidenza al 95%)

esiste una relazione lineare tra X e Y nella popolazione

INFERENZA 2:
IL COEFFICIENTE DI CORRELAZIONE

La forza relativa di un legame tra due variabili quantitative è data dal

$$S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

$$\varepsilon_i = y_i - \hat{y}_i$$

$$r = cov\,\frac{xy}{SxSy}$$

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

$$-1 \leq r \geq +1$$

Quando dati relativi alla popolazione il coefficiente di correlazione è indicato con ρ.

INFERENZA 2:
IL COEFFICIENTE DI CORRELAZIONE

Graficamente

$\rho = -1$ $\rho = 0$ $\rho = +1$

INFERENZA 2:
IL COEFFICIENTE DI CORRELAZIONE

Per verificare che ci sia una relazione lineare statisticamente significativa tra X e Y nella popolazione dobbiamo testare le ipotesi seguenti:

$$H0: \ \rho = 0$$

$$H1: \ \rho \neq 0$$

Si procede con il calcolo di v.c e t-stat → confronto

Se t-stat> v.c → rifiuto H0

→ Se rifiuto H0 esiste una significativa correlazione tra x e y nella popolazione

INFERENZA 2:
IL COEFFICIENTE DI CORRELAZIONE

$$Vc\colon \frac{a}{2} \quad (n-2)$$

$$t - stat = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}}$$

$$dove \;\; r = +\sqrt{R^2} \; se \; b1 > 0; \quad -\sqrt{R^2} \; se \; b1 < 0$$

APPLICAZIONE AL NOSTRO ESEMPIO “MINI MARKET”

Letture consigliate

Levin, Krenbiel, Berenson (2010) Statistica. *Quinta edizione*. *Cap 12*

ESERCITAZIONI
(tipologia esame)

Q. La differenza tra analisi di regressione lineare multipla e regressione lineare semplice dipende da…:

a. Il numero di variabili dipendenti
b. Il numero di variabili esplicative
c. Il numero di equazioni di regressione

Q. Quale delle seguenti non è una assunzione tipica sugli errori del modello di regressione lineare classico:

a. Indipendenza
b. Normalità
c. Eteroschedasticità

Q. I residui del modello di regressione sono:

a. Le differenze tra valori osservati e stimati della variabile dipendente
b. Le differenze tra valori osservati e stimati delle variabili esplicative
c. Le differenze tra valori osservati e media campionaria della variabile dipendente

Laboratorio in R

Si apra l'ambiente R e

si seguano le indicazioni date dalla docente