LA REGRESSIONE LINEARE

ANALISI DELLA DIPENDENZA

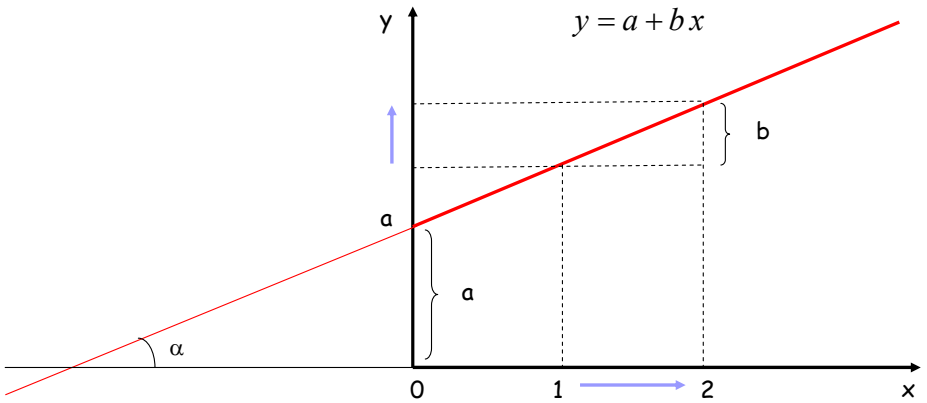
- **La Regressione Lineare**
- Quando tra due variabili c'è una relazione di dipendenza, si può cercare di prevedere il valore di una variabile in funzione del valore assunto dall'altra.
- Questo ha significato in senso stretto quando si ipotizza una relazione di causalità tra la variabile **indipendente**, su cui si agisce, e quella **dipendente**, su cui si vuole produrre un effetto.
- Volendo costruire un **modello statistico** per prevedere Y in funzione di X, si pone la questione di quale relazione funzionale ipotizzare tra la variabile indipendente X e la variabile dipendente Y.
- Il modello più semplice di relazione tra due variabili è quello lineare di primo grado, rappresentato da una retta, la cui equazione è :

$$y = a + b\,x$$

- Una volta determinata la retta, il modello permetterà di stimare il valore della variabile Y sulla base del valore assunto dalla X
- Per ottenere un buon modello, e quindi delle buone previsioni, occorre determinare la retta che meglio *descrive* i punti osservati: in pratica, si tratta di determinare i due coefficienti a e b che compaiono nell'equazione della retta: y = a + b x

ANALISI DELLA DIPENDENZA

- Ripassiamo un po' di geometria (e di trigonometria ...):
  - il parametro **a** è l'**intercetta** della retta con l'asse delle ordinate
  - il parametro **b** è il **coefficiente angolare**: misura l'inclinazione della retta (è la tangente dell'angolo α formato dalla retta con l'asse delle ascisse)
- In pratica:
  - a ci dice quanto vale Y quando X vale 0
  - b ci dice di quanto aumenta Y all'aumentare di una unità di X



ANALISI DELLA DIPENDENZA

- Se i punti fossero solo due, o fossero tutti allineati, determinare la retta interpolante sarebbe facile (per due punti passa una sola retta), ma sfortunatamente nella realtà i dati osservativi non sono mai esattamente allineati
- Per determinare la retta di regressione che meglio descrive i dati osservati è allora necessario stabilire un criterio *statistico* di valutazione della bontà del modello:

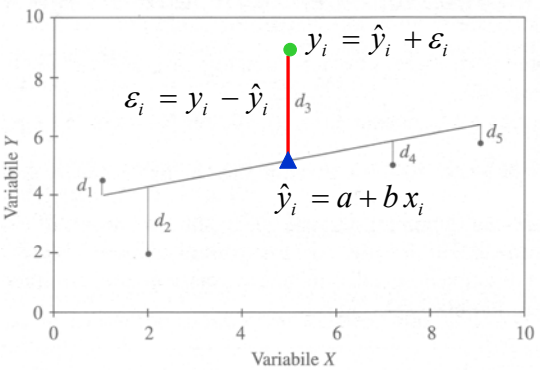
$$\hat{y}_i = a + b\,x_i \qquad \forall i$$

- L'errore che si commette prevedendo ciascun Y osservato con il modello, può essere misurato come differenza tra il dato reale e quello previsto:

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

cioè, per ciascuna osservazione i, si commette un errore

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

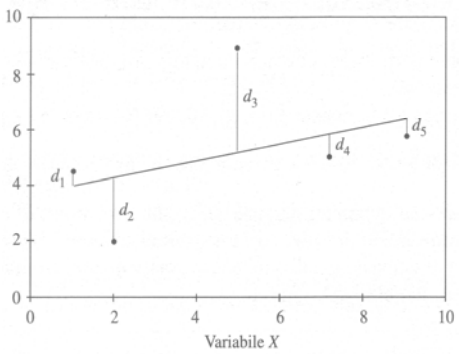


ANALISI DELLA DIPENDENZA

- **Il Metodo dei Minimi Quadrati**
- Il criterio detto dei **minimi quadrati** prevede di valutare la bontà del modello sulla base della somma dei quadrati di tutti errori di stima commessi:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - b\,x_i)^2 = \min$$

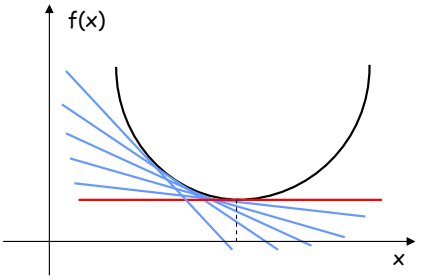
- La retta migliore, secondo questo criterio, è quella che minimizza la somma dei quadrati degli scarti dei valori stimati da quelli osservati, detti anche **residui** della regressione.
- Perché proprio il quadrato dei residui ?
  - per evitare che residui positivi e negativi si compensino
  - il valore assoluto è matematicamente più scomodo da gestire e non sempre porta ad una soluzione univoca
  - il quadrato dà peso maggiore agli scarti più grandi, che sono anche quelli che ci disturbano di più: è meglio fare tanti piccoli errori che non un errore molto grosso



ANALISI DELLA DIPENDENZA

- **Ricerca del minimo di una funzione**
- Il problema è determinare i coefficienti a e b del modello in modo da minimizzare la quantità :

$$\sum_{i=1}^{n} (y_i - a - b\,x_i)^2 = \min$$



- La soluzione di questo problema in matematica è semplice: si tratta di trovare il minimo di una funzione f(x)
- La soluzione ad un problema di ricerca del minimo di una funzione si determina trovando i punti "di svolta" della funzione, in cui la curva cambia andamento (concavità) : in un punto di svolta, la retta tangente alla curva risulta orizzontale
- L'inclinazione della retta tangente ad una curva in un punto è data dalla derivata prima della f(x) in quel punto, indicata come: f'(x)
- Quindi nel punto di svolta la derivata prima f'(x) della funzione f(x) si annulla
- Allora per determinare i punti di minimo di una funzione f(x) è necessario imporre che valga zero la sua derivata prima rispetto a x, in questo modo si determina il punto x in cui vale:

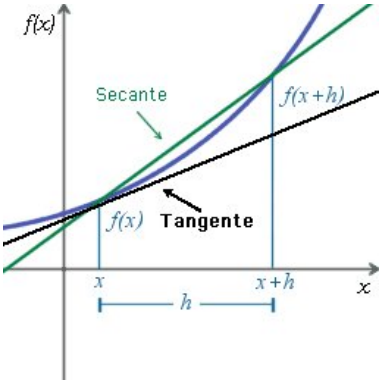
$$f'(x) = \frac{\partial f(x)}{\partial x} = 0$$

MATEMATICA

- **La Derivata di una Funzione**
- La derivata di una funzione reale di variabile reale f(x) nel punto x è definita come il limite per h tendente a 0 del rapporto incrementale:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

cioè dell'incremento descritto dalla funzione quando la variabile X varia da x ad (x+h), diviso l'incremento h stesso



- Il rapporto incrementale rappresenta il coefficiente angolare della retta secante che interseca la curva della funzione f(x) nei punti x e x+h
- Quando l'incremento h tende a 0, la retta secante tende (cioè si avvicina fino) a coincidere con la tangente alla curva nel punto x : quindi il limite del rapporto incrementale ci fornisce il coefficiente angolare della retta tangente nel punto x
- La derivata, cioè proprio questo limite, rappresenta dunque l'inclinazione della retta tangente alla funzione f(x) nel punto x

DERIVATE ELEMENTARI

$$\frac{d}{dx} a = 0$$

$$\frac{d}{dx} x = 1$$

$$\frac{d}{dx} a\,x = a$$

$$\frac{d}{dx} x^a = a\,x^{a-1}$$

$$\frac{d}{dx} \frac{1}{x} = -\frac{1}{x^2}$$

$$\frac{d}{dx} \sqrt{x} = \frac{1}{2\sqrt{x}}$$

$$\frac{d}{dx} e^x = e^x$$

$$\frac{d}{dx} a^x = a^x \ln a \quad (a > 0)$$

$$\frac{d}{dx} \ln x = \frac{1}{x} \quad (x > 0)$$

$$\frac{d}{dx} \sin x = \cos x$$

$$\frac{d}{dx} \cos x = -\sin x$$

REGOLE DI DERIVAZIONE

$$\frac{d}{dx}\left(f(x)+g(x)\right)=f'(x)+g'(x).$$

$$\frac{d}{dx}\left(cf(x)\right)=cf'(x).$$

$$\frac{d}{dx}\left(f(x)g(x)\right)=f'(x)g(x)+f(x)g'(x).$$

$$\frac{d}{dx}\left(\frac{1}{f(x)}\right)=-\frac{f'(x)}{f(x)^2}.$$

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right)=\frac{g(x)f'(x)-f(x)g'(x)}{g(x)^2}.$$

$$\frac{d}{dx}\left(f(g(x))\right)=f'(g(x))g'(x).$$

ANALISI DELLA DIPENDENZA

- **Stima dei parametri ai minimi quadrati**
- Nel nostro caso vogliamo minimizzare una funzione f(a,b) rispetto ad a e b: dovremo imporre che siano nulle le *derivate prime parziali* della funzione rispetto ad a e b:

$$\sum_{i=1}^{n}(y_i - a - b x_i)^2 = \min$$

- Derivando la funzione prima rispetto ad a e poi rispetto a b, e ponendo a zero tali derivate, si ottiene un sistema di due equazioni di primo grado in due incognite (a e b):

$$\begin{cases} \dfrac{\partial\sum(y_i - a - b x_i)^2}{\partial a} = 0 \\ \dfrac{\partial\sum(y_i - a - b x_i)^2}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} -2\sum(y_i - a - b x_i) = 0 \\ -2\sum(y_i - a - b x_i)\,x_i = 0 \end{cases} \Rightarrow$$

$$\begin{cases} \sum(y_i - a - b x_i) = 0 \\ \sum(y_i - a - b x_i)\,x_i = 0 \end{cases}$$

Sistema di equazioni normali di regressione

- Risolvendo il sistema rispetto ad a e b, si determinano le stime ai minimi quadrati dei parametri della retta di regressione

ANALISI DELLA DIPENDENZA

- **Soluzione del sistema di equazioni normali**

$$\begin{cases} \sum (y_i - a - b x_i) = 0 \\ \sum (y_i - a - b x_i) x_i = 0 \end{cases} \Rightarrow \begin{cases} \sum y_i - na - b \sum x_i = 0 \\ \sum x_i y_i - a \sum x_i - b \sum x_i^2 = 0 \end{cases} \Rightarrow$$

$$\begin{cases} n\bar{y} - na - bn\bar{x} = 0 \\ \sum x_i y_i - an\bar{x} - b \sum x_i^2 = 0 \end{cases} \Rightarrow \begin{cases} a = \bar{y} - b\,\bar{x} \\ \sum x_i y_i - an\bar{x} - b \sum x_i^2 = 0 \end{cases} \Rightarrow$$

$$\sum x_i y_i - (\bar{y} - b\,\bar{x}) n\bar{x} - b \sum x_i^2 = 0$$

$$\sum x_i y_i - n\bar{x}\,\bar{y} + nb\,\bar{x}^2 - b \sum x_i^2 = 0$$

$$\frac{\sum x_i y_i}{n} - \frac{n\bar{x}\,\bar{y}}{n} + b\left(\frac{n\bar{x}^2}{n} - \frac{\sum x_i^2}{n}\right) = 0$$

$$Cov(x,y) = M(x,y) - \bar{x}\,\bar{y}$$
$$V(x) = M(x^2) - \bar{x}^2$$

$$\frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y} = -b\left(\bar{x}^2 - \frac{\sum x_i^2}{n}\right) \Rightarrow b = \frac{\frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y}}{\frac{\sum x_i^2}{n} - \bar{x}^2} = \frac{Cov(x,y)}{\sigma^2(x)}$$

ANALISI DELLA DIPENDENZA

- **I coefficienti di regressione**
- I parametri del modello vengono chiamati anche **coefficienti di regressione**:

$$\hat{b} = \frac{\sigma_{xy}}{\sigma_x^2}$$
$$\hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

Equazione della retta di regressione:

$$y = \hat{a} + \hat{b}\,x$$

- Bisogna calcolare prima il valore di b e poi quello di a. Il "cappello" sopra a e b sottolinea che si tratta delle *stime*, ai minimi quadrati, dei parametri del modello.
- Il metodo di calcolo più veloce per b è utilizzare l'espressione che compare nell'ultimo passaggio della soluzione del sistema di equazioni normali:

$$\hat{b} = \frac{Cov(x,y)}{\sigma^2(x)} = \frac{M(x,y) - \bar{x}\,\bar{y}}{M(x^2) - \bar{x}^2} = \frac{\frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y}}{\frac{\sum x_i^2}{n} - \bar{x}^2}$$

che ha il vantaggio di non richiedere il calcolo di scarti
(in pratica utilizza le formule indirette per la varianza e la covarianza)

ANALISI DELLA DIPENDENZA

- **Interpretazione dei coefficienti di regressione**
- Il segno del coefficiente di regressione dipende da quello della covarianza: $\hat{b} = \frac{\sigma_{xy}}{\sigma_x^2}$ e indica quindi se la relazione è diretta o inversa
- Ricordiamo che, nell'equazione della retta $y = \hat{a} + \hat{b}\,x$
  b rappresenta il coefficiente angolare, cioè l'inclinazione della retta

b > 0	relazione diretta, cioè y cresce al crescere di x	la retta è crescente, inclinazione positiva
b = 0	assenza di relazione (lineare), y non varia al variare di x	la retta è orizzontale, inclinazione nulla
b < 0	relazione inversa, cioè y diminuisce all'aumentare di x	la retta è decrescente, inclinazione negativa

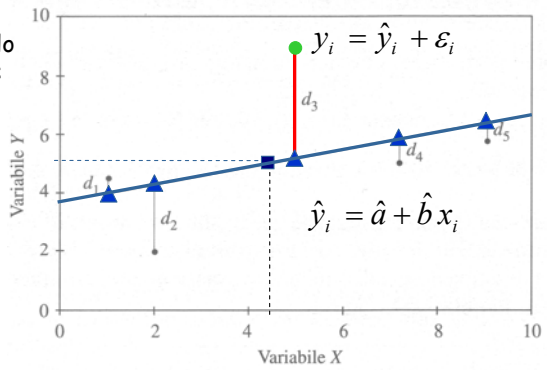
- Il valore assoluto di b indica di quanto varia la Y al variare di una unità della X
- Il coefficiente a rappresenta l'intercetta della retta con l'asse Y:
  indica quanto vale Y quando X vale 0; quando a = 0, la retta passa per l'origine degli assi cartesiani, cioè per il punto di coordinate (0, 0)

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **Proprietà della regressione ai minimi quadrati**
- Stimati i parametri, possiamo produrre i valori previsti dal modello in corrispondenza degli x osservati:

$$\hat{y}_i = \hat{a} + \hat{b}\,x_i \quad \forall i$$



- La retta di regressione passa per il baricentro del sistema, infatti quando x assume il valore x medio:

$$\hat{y}(\bar{x}) = \hat{a} + \hat{b}\,\bar{x} =$$
$$= (\bar{y} - \hat{b}\,\bar{x}) + \hat{b}\,\bar{x} = \bar{y}$$

- Per ciascun x(i) possiamo ora calcolare anche l'*errore* commesso prevedendo Y con il modello, come scarto tra il valore osservato e quello stimato: questi errori sono detti **residui** di regressione

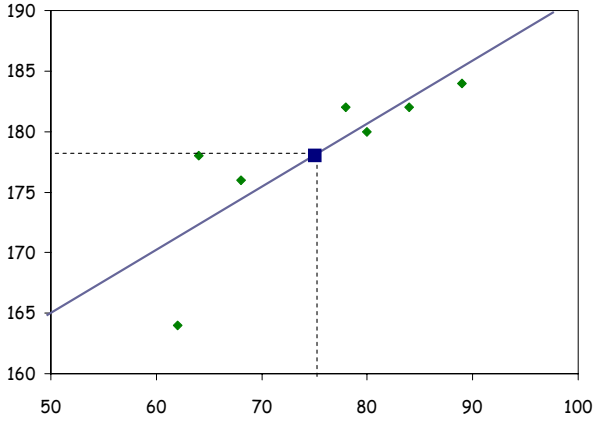
$$\varepsilon_i = y_i - \hat{y}_i \quad \forall i$$

- I residui sono importanti per la valutazione della bontà del modello, cioè la sua capacità di adattarsi ai dati osservati

ANALISI DELLA DIPENDENZA

- Esercizio. Determiniamo la retta di regressione ai minimi quadrati per la relazione tra le variabili Y=altezza e X=peso:

i	x(i)	y(i)
1	62	164
2	64	178
3	68	176
4	75	178
5	78	182
6	80	180
7	84	182
8	89	184



$$\hat{b} = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

- Osserviamo che tutto quello che ci serve per determinare i coefficienti di regressione della retta, lo abbiamo in realtà già calcolato per il coefficiente di correlazione

ANALISI DELLA DIPENDENZA

- Riprendiamo il prospetto di calcolo del coefficiente di regressione:

i	x(i)	y(i)	x(i)-Mx	y(i)-My	[x(i)-Mx]²	[y(i)-My]²	(x(i)-Mx)(y(i)-My)
1	62	164	-13,00	-14,00	169,00	196,00	182,00
2	64	178	-11,00	0,00	121,00	0,00	0,00
3	68	176	-7,00	-2,00	49,00	4,00	14,00
4	75	178	0,00	0,00	0,00	0,00	0,00
5	78	182	3,00	4,00	9,00	16,00	12,00
6	80	180	5,00	2,00	25,00	4,00	10,00
7	84	182	9,00	4,00	81,00	16,00	36,00
8	89	184	14,00	6,00	196,00	36,00	84,00
Totale	600	1424	0	0	650,00	272,00	338,00
Media	75,00	178,00			81,25	34,00	42,25

$$\sigma_x^2 = 81{,}25 \quad \sigma_y^2 = 34 \quad \sigma_{xy} = 42{,}25$$

$$b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{42{,}25}{81{,}25} = 0{,}52 \qquad a = \bar{y} - b\,\bar{x} = 178 - 0{,}52 \cdot 75 = 139$$

- Il valore di b ci dice che, al variare di una unità di x, il valore di y varia di 0,52
- Ora possiamo disegnare la retta di regressione, determinandone due punti a scelta

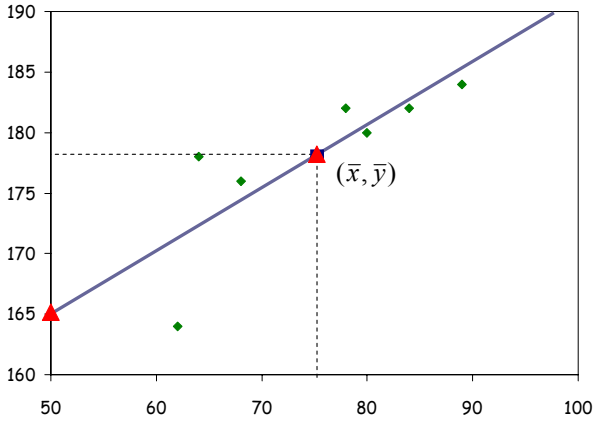
ANALISI DELLA DIPENDENZA

- Finalmente scriviamo l'equazione della retta di regressione:

$$\hat{b} = \frac{\sigma_{xy}}{\sigma_x^2} = 0{,}52$$

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x} = 139$$

$$\hat{y} = 139 + 0{,}52\,x$$

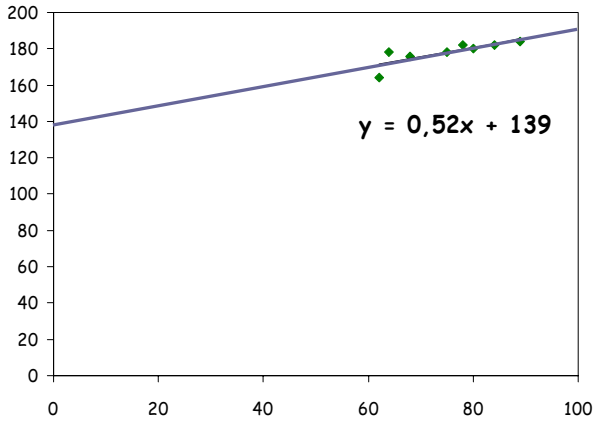


- Per disegnare la retta di regressione, è sufficiente determinarne due punti a scelta:
  - uno in realtà lo conosciamo già: è il punto medio (baricentro) del sistema
  - per l'altro scegliamo ad es. x = 50, da cui y = 139 + 0,52 * 50 = 165
- A vederla così non sembra che l'intercetta sia 139: questo dipende dal fatto che i punti sono molto lontani dall'origine e noi abbiamo "tagliato" la figura

ANALISI DELLA DIPENDENZA

- Se disegniamo il grafico e la retta di regressione a partire dall'origine degli assi, vediamo che tutto torna:

$$\hat{y} = 139 + 0{,}52\,x$$



- La retta interseca l'asse delle ordinate proprio nel punto 139
- L'inclinazione apparente della retta nel disegno dipende dalla scala degli assi: abbiamo usato due scale diverse per l 'asse X e per l'asse Y
- Avevamo già calcolato il coefficiente di correlazione

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y} = 0{,}8039$$

- Un coefficiente di correlazione pari a 0,8 ci dice che i punti sono abbastanza ben allineati lungo la retta, ma ora vediamo meglio come valutare la capacita del modello di descrivere i dati osservati

ANALISI DELLA DIPENDENZA

- **Valutazione del modello**
- La bontà dell'interpolazione fornita dal modello si giudica sulla base della dispersione dei punti **osservati** intorno alla retta di regressione:
  - l'interpolazione è esatta quando la retta passa per tutti i punti osservati, che devono essere quindi perfettamente allineati
  - più i punti osservati si discostano dalla retta -> minore è la validità del modello, che diventa sempre meno efficace, fino a perdere praticamente di significato
- Il modello stimato può essere valutato, coerentemente con il criterio che è stato utilizzato per costruirlo, sulla base della somma dei quadrati degli scarti tra valori osservati e valori stimati, cioè la somma dei quadrati dei residui di regressione:

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

- I residui della regressione lineare ai minimi quadrati godono di 4 importanti proprietà: tutte queste proprietà derivano dalle equazioni normali, ovvero dalle condizioni che sono state imposte per determinare i parametri della retta di regressione, in modo da minimizzare la somma dei quadrati dei residui stessi

ANALISI DELLA DIPENDENZA

- **Le 4 Proprietà dei Residui di Regressione**

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)\, x_i = 0$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)\, \hat{y}_i = 0$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$$

$$\hat{y}_i = a + b\, x_i$$

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

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

- Le prime due proprietà derivano direttamente, anzi non sono altro che le condizioni imposte dalle equazioni normali di regressione:

$$\begin{cases} \sum (y_i - a - b\, x_i) = 0 \\ \sum (y_i - a - b\, x_i)\, x_i = 0 \end{cases} \quad \textit{ovvero} \quad \begin{cases} \sum (y_i - \hat{y}_i) = 0 \\ \sum (y_i - \hat{y}_i)\, x_i = 0 \end{cases}$$

ANALISI DELLA DIPENDENZA

- Dimostriamo le altre due:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)\,\hat{y}_i = 0$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(a + b\,x_i) = a\sum_{i=1}^{n}(y_i - \hat{y}_i) + b\sum_{i=1}^{n}(y_i - \hat{y}_i)x_i = 0$$

sono proprio le prime due proprietà = 0

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i - \bar{y}\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$$

questa è la terza proprietà = 0

- Teniamole presenti: fra un attimo ci torneranno utili

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- Esercizio. Verifica empirica delle 4 proprietà dei residui di regressione ai minimi quadrati per il modello appena stimato:

$$\hat{y} = 139 + 0{,}52\,x$$

i	x(i)	y(i)	y(i)^	y^-My	y-y^	(y-y^) x	(y-y^) y^	(y-y^)(y^-My)
1	62	164	171,24	-6,76	-7,24	-448,8800	-1239,7776	48,9424
2	64	178	172,28	-5,72	5,72	366,0800	985,4416	-32,7184
3	68	176	174,36	-3,64	1,64	111,5200	285,9504	-5,9696
4	75	178	178,00	0,00	0,00	0,0000	0,0000	0,0000
5	78	182	179,56	1,56	2,44	190,3200	438,1264	3,8064
6	80	180	180,60	2,60	-0,60	-48,0000	-108,3600	-1,5600
7	84	182	182,68	4,68	-0,68	-57,1200	-124,2224	-3,1824
8	89	184	185,28	7,28	-1,28	-113,9200	-237,1584	-9,3184
Totale	600	1424	1424	0,00	0,00	0,0000	0,0000	0,0000
Media	75,00	178,00	178,00					

ANALISI DELLA DIPENDENZA

- **Devianza (e Varianza) Residua**
- Il modello viene valutato sulla base della somma dei quadrati degli scarti tra valori osservati e valori stimati:

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

- Questa quantità è detta **devianza residua**, in quanto è la *devianza dei residui* di regressione (che ricordiamo hanno media nulla)
- Il termine **devianza** indica semplicemente la varianza *non divisa* per n, quindi tutto quello che diciamo per la devianza vale anche per la varianza
- La quantità:

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

$$M(\hat{y}) = M(\hat{a} + \hat{b}\,x) = M(\bar{y} - \hat{b}\,\bar{x} + \hat{b}\,x) =$$
$$= M(\bar{y}) + \hat{b}\,M(x - \bar{x}) = \bar{y}$$

- prende invece il nome di **devianza spiegata** dal modello (o devianza di regressione), in quanto *devianza dei valori stimati* dal modello: rappresenta la parte di variabilità della Y descritta dal modello di regressione
- Devianza spiegata e devianza residua sono legate tra loro e con la **devianza totale** della variabile dipendente

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **Scomposizione della Devianza (e della Varianza) di Y**
- La devianza della variabile Y può essere scomposta in due componenti:

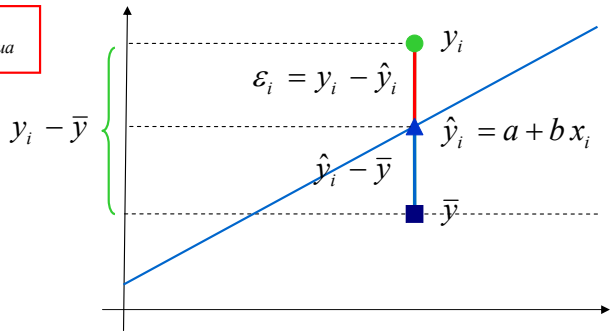
$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

cioè come somma della **devianza residua** e della **devianza spiegata** dal modello

- Devianza spiegata e residua sono quindi due quantità complementari e la loro somma è pari alla devianza totale della Y. Lo stesso vale anche per la varianza, basta dividere tutto per n:

$$\sigma_y^2 = \sigma_{modello}^2 + \sigma_{residua}^2$$

- Capiamo il significato della scomposizione:



ANALISI DELLA DIPENDENZA

$$(a+b)^2 = a^2 + b^2 + 2ab$$

- Dimostrazione:

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

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

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

= 0 per la quarta proprietà dei residui di regressione

devianza residua + devianza spiegata dal modello

- Lo stesso vale ovviamente anche per la varianza, basta dividere tutto per n, quindi possiamo scrivere anche la scomposizione della varianza totale di Y:

$$\sigma_y^2 = \sigma_{modello}^2 + \sigma_{residua}^2$$

- In virtù di questa scomposizione, si può costruire un indice standardizzato (compreso tra 0 e 1) per valutare la bontà del modello di regressione

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **Il Coefficiente di Determinazione R^2**
- Per valutare la bontà del modello si introduce il **coefficiente di determinazione R^2** rapporto tra la devianza del modello e la devianza totale della y:

$$R^2 = \frac{\sigma_{modello}^2}{\sigma_y^2} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

- R^2 misura la frazione di varianza (o di devianza) della Y spiegata dal modello
- R^2 si può scrivere anche come:

$$R^2 = \frac{\sigma_{modello}^2}{\sigma_y^2} = \frac{\sigma_y^2 - \sigma_{residua}^2}{\sigma_y^2} = 1 - \frac{\sigma_{residua}^2}{\sigma_y^2}$$

- Assume valori compresi tra 0 e 1:
  - vale 1 quando il modello spiega completamente la varianza della Y: i residui sono tutti nulli, cioè i punti sono perfettamente allineati
  - vale 0 quando la varianza descritta dal modello è nulla: questo accade quando la retta di regressione risulta parallela all'asse X, cioè

$$\hat{y}_i = \bar{y} \Rightarrow \sum(\hat{y}_i - \bar{y})^2 = 0$$

ANALISI DELLA DIPENDENZA

- Osserviamo che per calcolare R^2 è necessario calcolare tutti gli y(i) stimati per ogni x(i) osservato, e tutti gli scarti tra i valori stimati e M(y)

$$R^2 = \frac{\sigma^2_{modello}}{\sigma^2_y} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sigma^2_y} = 1 - \frac{\sigma^2_{residua}}{\sigma^2_y} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sigma^2_y}$$

- Oppure, con la seconda formulazione, dobbiamo calcolare tutti gli scarti tra gli y stimati e quelli osservati (cioè i residui)
- Possiamo però sviluppare R^2 anche in un altro modo:

$$R^2 = \frac{\sigma^2_{modello}}{\sigma^2_y} = \frac{V(\hat{y})}{V(y)} = \frac{V(\hat{a} + \hat{b}x)}{V(y)} = \frac{\hat{b}^2 V(x)}{V(y)} = \frac{\hat{b}^2 \sigma^2_x}{\sigma^2_y} =$$

$$= \left(\frac{\sigma_{xy}}{\sigma^2_x}\right)^2 \frac{\sigma^2_x}{\sigma^2_y} = \frac{\sigma^2_{xy}}{\sigma^4_x} \frac{\sigma^2_x}{\sigma^2_y} = \frac{\sigma^2_{xy}}{\sigma^2_x \sigma^2_y} = \left(\frac{\sigma_{xy}}{\sigma_x \sigma_y}\right)^2 = \boxed{\rho^2}$$

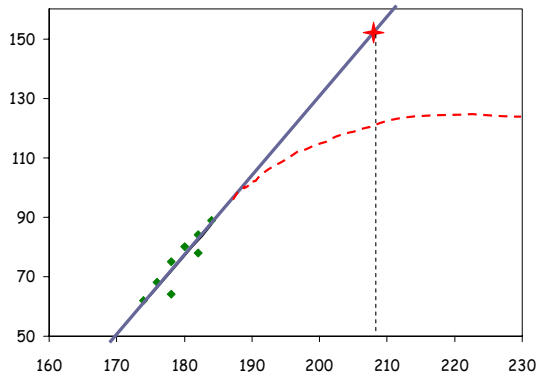
- Il coefficiente di determinazione R^2 (solo per il modello lineare di primo grado cioè per la retta di regressione) è uguale al quadrato del coefficiente di correlazione ρ^2
- Quindi il quadrato del coefficiente di correlazione ci fornisce la frazione di varianza spiegata dal modello, senza calcolare i residui e nemmeno gli y stimati

ANALISI DELLA DIPENDENZA

- **Considerazioni sull'applicabilità di un modello (lineare)**
- Quando R^2 è basso significa che il modello lineare non riesce a descrivere adeguatamente il fenomeno, non si adatta bene ai dati, e quindi anche le nostre *previsioni* saranno poco affidabili.
- Quando R^2 è elevato, diciamo prossimo a 1, il modello descrive bene i dati empirici, spiegando gran parte della variabilità osservata della variabile dipendente.
- La capacità *descrittiva* del modello è dunque buona, almeno relativamente ai dati osservati. Si può allora ritenere utilizzabile il modello a scopi *previsivi*?
- Il passaggio dalla descrizione dei dati osservati alla previsione di nuovi dati e comportamenti è un problema di generalizzabilità dei risultati di un modello (validità esterna)
- La possibilità di operare generalizzazioni si basa su assunzioni "ragionevoli" come:
  - "ordine" della natura: se le cose sono andate in un certo modo finora non c'è ragione che non funzionino così anche in futuro (e in passato: attualismo)
  - determinismo degli eventi: nessun evento avviene per caso, ma è determinato da precedenti eventi; ovvero, se le cose funzionano in un certo modo, la ragione c'è sempre, e quindi continueranno …
- Questi principi sono plausibili, ma non assicurati: nessuno può garantire che anche le prossime osservazioni si comporteranno allo stesso modo.

ANALISI DELLA DIPENDENZA

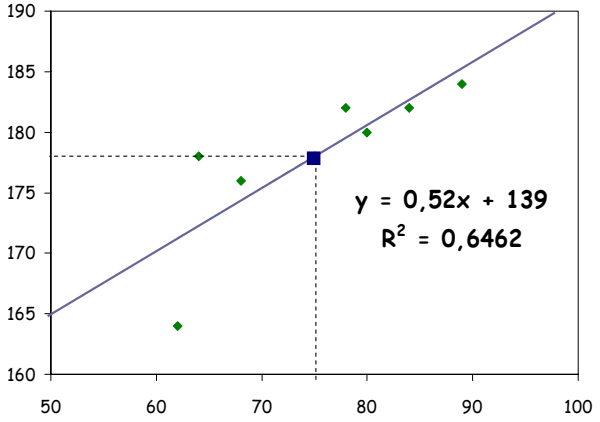
- **Considerazioni sull'applicabilità di un modello (lineare)**
- Assai più azzardato è utilizzare un modello per fare previsioni per valori x esterni al campo di variazione dei dati osservati, sui quali il modello è stato stimato: questo tipo di previsioni, dette *estrapolazioni*, portano spesso a conclusioni errate (es. sopravvalutare un trend di crescita positivo)
- Raramente i fenomeni sono lineari, oppure lo sono solo su una parte limitata del loro campo di esistenza: utilizzare un modello lineare è quasi sempre una *semplificazione* del fenomeno reale, che può risultare accettabile per determinate applicazioni
- I risultati del modello non devono essere accettati ciecamente: occorre sempre considerarne criticamente le possibili limitazioni



ANALISI DELLA DIPENDENZA

- Esercizio. Torniamo al nostro esempio, per valutare la bontà del modello stimato: determiniamo i punti stimati e calcoliamo la varianza spiegata e la varianza residua

i	x(i)	y(i)
1	62	164
2	64	178
3	68	176
4	75	178
5	78	182
6	80	180
7	84	182
8	89	184
Totale	600	1424
Media	75,00	178,00



- Il modello spiega circa il 65% della varianza complessiva della Y: possiamo ritenerci soddisfatti ?
- Nelle scienze sociali e comportamentali difficilmente si trovano modelli che spiegano molta varianza, perché i fenomeni sono complessi e dipendono da molti fattori diversi, inoltre le relazioni possono essere non lineari

ANALISI DELLA DIPENDENZA

- Prospetto di calcolo del coefficiente di determinazione R^2:

i	x(i)	y(i)	y^(i)	y^(i)-My	y(i)-y^(i)	y(i)-My	[y^(i)-My]²	[y(i)-y^(i)]²	[y(i)-My]²
1	62	164	171,24	-6,76	-7,24	-14,00	45,70	52,42	196,00
2	64	178	172,28	-5,72	5,72	0,00	32,72	32,72	0,00
3	68	176	174,36	-3,64	1,64	-2,00	13,25	2,69	4,00
4	75	178	178,00	0,00	0,00	0,00	0,00	0,00	0,00
5	78	182	179,56	1,56	2,44	4,00	2,43	5,95	16,00
6	80	180	180,60	2,60	-0,60	2,00	6,76	0,36	4,00
7	84	182	182,68	4,68	-0,68	4,00	21,90	0,46	16,00
8	89	184	185,28	7,28	-1,28	6,00	53,00	1,64	36,00
Totale	600	1424	1424	0,00	0,00	0,00	175,76	96,24	272,00
Media	75,00	178,00	178,00				21,97	12,03	34,00

$$\sigma^2_{modello} = 21{,}97 \qquad \sigma^2_{residua} = 12{,}03 \qquad \sigma^2_y = 34 = \sigma^2_{modello} + \sigma^2_{residua}$$

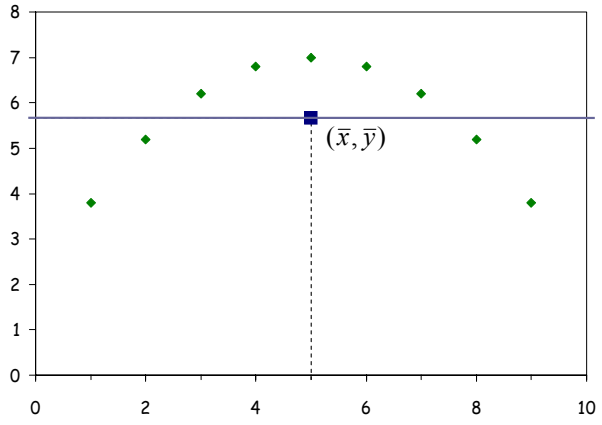
$$R^2 = \frac{\sigma^2_{modello}}{\sigma^2_y} = \frac{21{,}97}{34} = 0{,}6462 \qquad \textit{oppure} \quad R^2 = 1 - \frac{\sigma^2_{residua}}{\sigma^2_y} = 1 - \frac{12{,}03}{34} = 0{,}6462$$

- Conoscendo già il valore del coefficiente di correlazione, potevamo risparmiarci tutti questi calcoli: $R^2 = \rho^2_{xy} = (0{,}8039)^2 = 0{,}6462$

ANALISI DELLA DIPENDENZA

- Esercizio.
  Ritorniamo all'esempio di due variabili legate da una relazione quadratica esatta:
  $Y = 2 + 2\,x - 0{,}2\,x^2$

i	x(i)	y(i)
1	1	3,8
2	2	5,2
3	3	6,2
4	4	6,8
5	5	7
6	6	6,8
7	7	6,2
8	8	5,2
9	9	3,8
Totale	45	51
Media	5,00	5,67



- Dal grafico si evidenzia chiaramente la relazione quadratica tra le due variabili, che disegna una parabola con la concavità rivolta verso il basso
- Come risulta la retta di regressione ?

ANALISI DELLA DIPENDENZA

- Esempio. Un caso particolare di retta di regressione si verifica quando la stima di b risulta uguale a 0:

$$\begin{cases} \hat{b} = 0 \\ \hat{a} = \bar{y} - b\,\bar{x} = \bar{y} \end{cases} \Rightarrow \quad \hat{y}_i = \bar{y}$$

- La retta interseca l'asse delle ordinate nel punto M(y)
- La stima è sempre pari a M(y), per qualunque valore di x
- La varianza spiegata dal modello è nulla

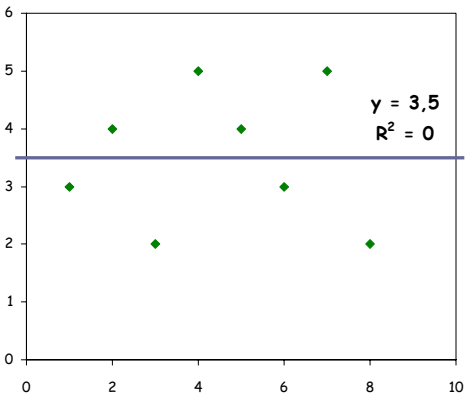
$$\sum (\hat{y}_i - \bar{y})^2 = 0$$

- La varianza residua è massima, cioè uguale alla varianza totale di Y

$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - \bar{y})^2$$

- Il coefficiente di determinazione R^2:

$$R^2 = \frac{\sigma^2_{regress}}{\sigma^2_y} = 1 - \frac{\sigma^2_{resid}}{\sigma^2_y} = \rho^2_{xy} = 0$$

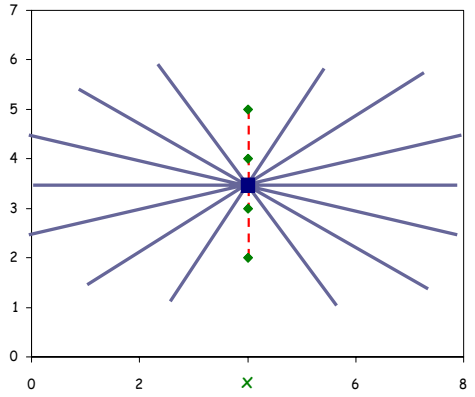


ANALISI DELLA DIPENDENZA

- **La Collinearità**
- Un problema, detto **collinearità**, si verifica quando i valori di X sono molto vicini tra loro: in questo caso la determinazione della retta di regressione risulta incerta
- Vediamo cosa succede nella situazione limite in cui tutti gli x(i) vengano a coincidere esattamente:

$$\hat{b} = \frac{Cov(x,y)}{\sigma^2(x)} = \frac{Cov(k,y)}{\sigma^2(k)} = \frac{0}{0}$$

- La stima di b assume una forma *indeterminata*
- Ciò equivale a dire che *qualunque* retta passante per il punto medio del sistema è la retta interpolante ai minimi quadrati
- Non dimentichiamo infatti che i residui vengono calcolati parallelamente all'asse Y



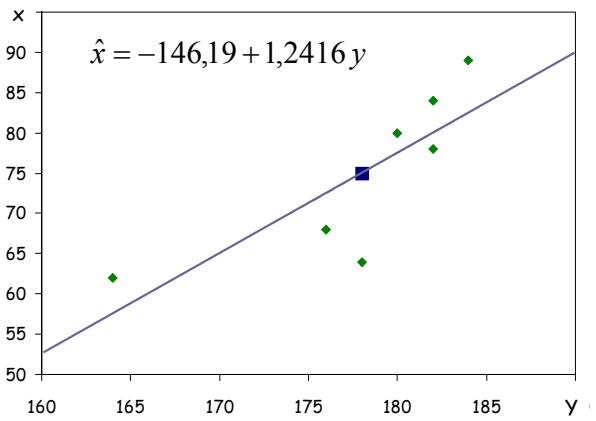
- In pratica non si verificherà mai il caso limite della coincidenza di tutti i valori di X, ma basta che siano molto vicini perché il calcolo di b risulti molto instabile, e quindi il modello sia privo di significato
- Significato: se la X osservata non varia, non può neanche spiegare la variabilità di Y

ANALISI DELLA DIPENDENZA

- **L'*altra* Retta di regressione**
- Cambiamo le carte in tavola, e questa volta stimiamo, sempre sugli stessi dati usati finora, il modello per prevedere X in funzione di Y:

$$\hat{x} = c + d\,y$$

i	x(i)	y(i)
1	62	164
2	64	178
3	68	176
4	75	178
5	78	182
6	80	180
7	84	182
8	89	184
Totale	600	1424
Media	75,00	178,00

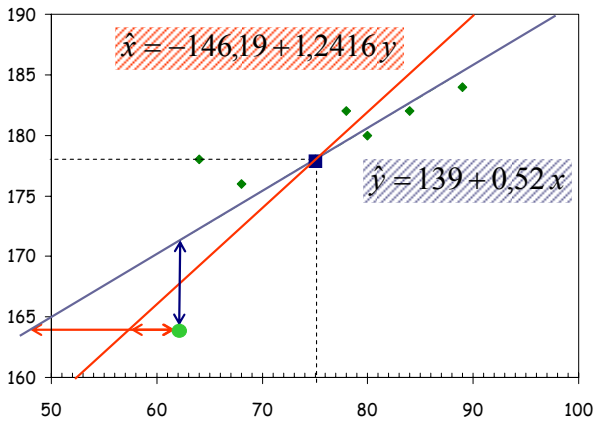


- Il grafico risulta rovesciato rispetto a prima, infatti questa volta troviamo la variabile Y sull'asse delle ascisse e la X sull'asse delle ordinate

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- Riportando i valori stimati dai due modelli su uno stesso grafico, ci accorgiamo che la nuova retta di regressione non coincide con quella stimata in precedenza
- Come si spiega ?
- Il metodo dei minimi quadrati valuta i residui lungo l'asse della variabile indipendente
- I residui calcolati lungo l'asse Y risultano in generale diversi da quelli calcolati lungo l'asse X
- Quindi si hanno sempre due rette di regressione, che in generale non coincidono



- La retta di regressione per stimare Y in funzione di X è stata ottenuta minimizzando la somma dei quadrati dei residui di Y, calcolati parallelamente all'asse Y
- La retta di regressione per stimare X in funzione di Y si ottiene minimizzando la somma dei quadrati dei residui di X, calcolati parallelamente all'asse X

ANALISI DELLA DIPENDENZA

- Consideriamo le due rette di regressione e vediamo cosa hanno in comune:

$$\left.\begin{array}{l} \hat{y} = a + b\,x \quad \Rightarrow \quad \hat{b} = \dfrac{\sigma_{x\,y}}{\sigma_x^2} \\ \\ \hat{x} = c + d\,y \quad \Rightarrow \quad \hat{d} = \dfrac{\sigma_{x\,y}}{\sigma_y^2} \end{array}\right\} \; R^2 = \rho_{x\,y}^2 = \left(\frac{\sigma_{x\,y}}{\sigma_x\,\sigma_y}\right)^2 = \rho_{y\,x}^2$$

- Le rette regressione ai minimi quadrati risultano diverse, ma il coefficiente di determinazione R^2 non cambia, essendo il quadrato del coefficiente di correlazione, che è una quantità simmetrica
- La frazione di variabilità di Y spiegata dalla regressione di Y in funzione di X è pari alla frazione di variabilità di X spiegata dalla regressione di X in funzione di Y
- C'è una relazione tra i due coefficienti di regressione b e d:

$$b \cdot d = \frac{\sigma_{x\,y}}{\sigma_x^2}\frac{\sigma_{x\,y}}{\sigma_y^2} = \frac{\sigma_{x\,y}^2}{\sigma_x^2\,\sigma_y^2} = \rho^2 \qquad quindi \;\; \rho = \sqrt{b\,d}$$

- Le due rette di regressione coincidono quando i punti osservati sono perfettamente allineati: in questa situazione r=1 e quindi:

$$\rho = 1 \;\Rightarrow\; b\,d = 1 \;\Rightarrow\; b = 1/d$$

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **Il Modello senza intercetta**
- Torniamo alla stima di Y in funzione di X, e consideriamo il modello ridotto eliminando il parametro a, ovvero una retta senza intercetta:

$$\hat{y} = b\,x$$

i	x(i)	y(i)
1	62	164
2	64	178
3	68	176
4	75	178
5	78	182
6	80	180
7	84	182
8	89	184
Totale	600	1424
Media	75,00	178,00

- La stima di questo modello è molto semplice: dobbiamo minimizzare la somma dei quadrati dei residui, che in questo caso sono una funzione di una sola variabile (b)

$$\sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b\,x_i)^2 = \min$$

- Per determinare il punto di minimo, deriviamo la funzione rispetto a b, ottenendo una sola equazione:

$$-2\sum_{i=1}^{n}(y_i - b\,x_i)\,x_i = 0$$

da cui:

$$\sum(x_i\,y_i - b\,x_i^2) = 0 \;\Rightarrow\; \sum x_i\,y_i - b\sum x_i^2 = 0$$

- La stima di b risulta diversa da quella del modello completo :

$$\hat{b} = \frac{\sum x_i\,y_i}{\sum x_i^2}$$

ANALISI DELLA DIPENDENZA

- Prospetto di calcolo del coefficiente di regressione:

i	x(i)	y(i)	x(i)²	x(i) y(i)
1	62	164	3844	10168
2	64	178	4096	11392
3	68	176	4624	11968
4	75	178	5625	13350
5	78	182	6084	14196
6	80	180	6400	14400
7	84	182	7056	15288
8	89	184	7921	16376
Totale	600	1424	45650	107138
Media	75,00	178,00		

$$\hat{b} = \frac{M(x \cdot y)}{M(x^2)} = \frac{107138}{45650} = 2{,}3469$$

$$\hat{y}_i = 2{,}3469\, x_i$$

- E' facile verificare che questa retta di regressione NON passa per il punto medio del sistema :

$$\hat{y}(\bar{x}) = 2{,}3469 \cdot 75 = 176{,}02 \neq 178$$

- Adesso vogliamo di valutare la bontà del modello: calcoliamo dunque la varianza residua e quella di regressione spiegata dal modello, e vediamo cosa succede ...

ANALISI DELLA DIPENDENZA

- Prospetto di calcolo del coefficiente di determinazione R²:

i	x(i)	y(i)	y^(i)	y^(i)-My	y(i)-y^(i)	y(i)-My	[y^(i)-My]²	[y^(i)-y(i)]²	[y(i)-My]²
1	62	164	145,51	-32,49	18,49	-14,00	1055,57	341,86	196,00
2	64	178	150,20	-27,80	27,80	0,00	772,59	772,59	0,00
3	68	176	159,59	-18,41	16,41	-2,00	338,85	269,22	4,00
4	75	178	176,02	-1,98	1,98	0,00	3,92	3,92	0,00
5	78	182	183,06	5,06	-1,06	4,00	25,62	1,13	16,00
6	80	180	187,76	9,76	-7,76	2,00	95,17	60,15	4,00
7	84	182	197,14	19,14	-15,14	4,00	366,47	229,32	16,00
8	89	184	208,88	30,88	-24,88	6,00	953,45	618,92	36,00
Totale	600	1424	1408,17	-15,83	15,83	0,00	3611,63	2297,10	272,00
Media	75,00	178,00	176,02				451,45	287,14	34,00

$$\sigma^2_{modello} = 451{,}45 \qquad \sigma^2_{residua} = 287{,}14 \qquad \sigma^2_y = 34 \neq \sigma^2_{modello} + \sigma^2_{residua}\ !!$$

$$R^2 = \frac{\sigma^2_{modello}}{\sigma^2_y} \neq 1 - \frac{\sigma^2_{residua}}{\sigma^2_y} \neq \rho^2$$

- La varianza residua è maggiore della varianza totale di Y ...
- R² perde ogni significato, perché non vale più la scomposizione della varianza

ANALISI DELLA DIPENDENZA

- Rispetto al sistema di equazioni normali derivate per la stima dei parametri del modello completo, in pratica non è stata imposta la prima condizione :

$$\begin{cases} \sum (y_i - a - b x_i) = 0 \\ \sum (y_i - a - b x_i)\, x_i = 0 \end{cases} \Rightarrow \begin{cases} \sum y_i - na - b \sum x_i = 0 \\ \sum x_i y_i - a \sum x_i - b \sum x_i^2 = 0 \end{cases} \Rightarrow$$

$$\begin{cases} n\bar{y} - na - b\, n\bar{x} = 0 \\ \sum x_i y_i - a\, n\bar{x} - b \sum x_i^2 = 0 \end{cases} \Rightarrow \begin{cases} a = \bar{y} - b\,\bar{x} \\ \sum x_i y_i - a\, n\bar{x} - b \sum x_i^2 = 0 \end{cases}$$

- Come conseguenza immediata, non vale più la prima proprietà dei residui:
  - la somma dei residui di regressione del modello ridotto non risulta uguale a 0
  - non è nulla nemmeno la somma degli scarti dei valori stimati dalla M(y)
  - la media dei valori stimati non coincide con la media di Y
  - la retta non passa per il baricentro del sistema
- La seconda proprietà continua a valere perché è la condizione imposta dall'unica equazione normale, derivata per la stima ai minimi quadrati del modello ridotto

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **Proprietà dei residui di regressione nel modello senza intercetta**

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)\, x_i = 0$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)\, \hat{y}_i = 0$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$$

$$\hat{y}_i = b\, x_i$$
$$y_i = \hat{y}_i + \varepsilon_i$$
$$\varepsilon_i = y_i - \hat{y}_i$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)\, b\, x_i = b \sum_{i=1}^{n} (y_i - \hat{y}_i)\, x_i = 0$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)\hat{y}_i - \bar{y} \sum_{i=1}^{n} (y_i - \hat{y}_i) \neq 0$$

- Non valendo la quarta proprietà, non vale più nemmeno la scomposizione della devianza totale, perché il doppio prodotto non si annulla:

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

$$-2\bar{y} \sum (y_i - \hat{y}_i)$$

ANALISI DELLA DIPENDENZA

- Esempio.
  Verifica empirica delle 4 proprietà dei residui di regressione ai minimi quadrati:

i	x(i)	y(i)	y(i)^	y^-My	y-y^	(y-y^) x	(y-y^) y^	(y-y^)(y^-My)
1	62	164	145,51	-32,49	18,49	1146,3467	2690,4117	-600,7127
2	64	178	150,20	-27,80	27,80	1778,9168	4175,0184	-772,5940
3	68	176	159,59	-18,41	16,41	1115,7303	2618,5567	-302,0314
4	75	178	176,02	-1,98	1,98	148,4392	348,3785	-3,9172
5	78	182	183,06	5,06	-1,06	-82,8081	-194,3461	-5,3737
6	80	180	187,76	9,76	-7,76	-620,4425	-1456,1439	-75,6593
7	84	182	197,14	19,14	-15,14	-1272,0379	-2985,4018	-289,8930
8	89	184	208,88	30,88	-24,88	-2214,1445	-5196,4735	-768,1845
Totale	600	1424	1408,17	-15,83	15,83	0,0000	0,0000	-2818,3658
Media	75,00	178,00	176,02					-352,2957

- Verifichiamo che la varianza totale di Y è ottenibile come:

$$\sigma_y^2 = \sigma_{modello}^2 + \sigma_{residua}^2 + 2\frac{-\bar{y}\sum(y_i - \hat{y}_i)}{8} =$$
$$= 451{,}45 + 287{,}14 + 2\cdot(-178\cdot 15{,}83/8) =$$
$$= 451{,}45 + 287{,}14 + 2\cdot(-352{,}2957) = 34$$

ANALISI DELLA DIPENDENZA

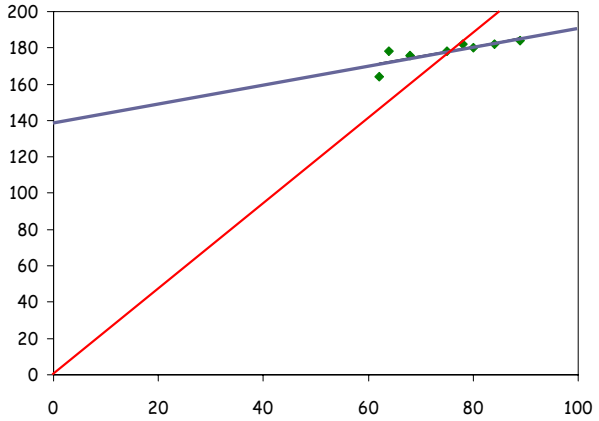
- Il modello ridotto, per la mancanza dell'intercetta, obbliga la retta a passare per l'origine degli assi:

$$\hat{y} = 2{,}3469\,x$$

- L'introduzione di questo vincolo peggiora la capacità del modello di adattarsi ai dati osservati, rispetto al modello completo

$$\hat{y} = 139 + 0{,}52\,x$$

- In generale, l'introduzione di un vincolo in un modello ne peggiora sempre la capacità di adattamento



- Possiamo dire che, a parità di predittori, l'aumento di un parametro migliora sempre (o almeno non peggiora) la capacità descrittiva di un modello, mentre l'eliminazione di un parametro la riduce
- Per migliorare l'adattamento e descrivere anche relazioni non lineari, potremmo ad esempio passare ad un modello di secondo grado, con tre parametri:

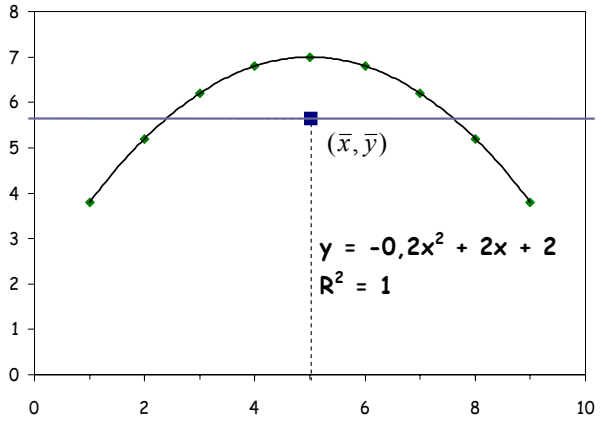
$$\hat{y} = a + b\,x + c\,x^2$$

ANALISI DELLA DIPENDENZA

- Stimiamo ai minimi quadrati un modello di regressione quadratico sui dati già analizzati in precedenza con il modello di primo grado (retta)

$$\hat{y} = a + b\,x + c\,x^2$$

i	x(i)	y(i)
1	1	3,8
2	2	5,2
3	3	6,2
4	4	6,8
5	5	7
6	6	6,8
7	7	6,2
8	8	5,2
9	9	3,8
Totale	45	51
Media	5,00	5,67



- In questo caso dovremo minimizzare la somma dei quadrati dei residui del modello rispetto ai tre parametri (a, b, c) ottenendo un sistema di 3 equazioni normali
- Risolvendo il sistema, troviamo un modello che si adatta perfettamente ai dati osservati: $R^2=1$