Statistica

Capitolo 12

Regressione Lineare Semplice

Obiettivi del Capitolo

Dopo aver completato il capitolo, sarete in grado di:

- Spiegare il significato del coefficiente di correlazione lineare e condurre una verifica di ipotesi su ρ
- Spiegare il modello di regressione lineare semplice
- Ottenere ed interpretare l'equazione della regressione lineare semplice per un insieme di dati
- Descrivere R^2 come una misura della capacità esplicativa del modello di regressione
- Comprendere le assunzioni su cui si basa l'analisi della regressione

Obiettivi del Capitolo

(continuazione)

Dopo aver completato il capitolo, sarete in grado di:

- Spiegare le misure di variazione e determinare se la variabile indipendente è significativa
- Determinare ed interpretare intervalli di confidenza per i coefficienti della regressione
- Usare l'equazione della regressione per fare previsioni
- Formare intervalli di previsione per i valori di Y in corrispondenza di un dato valore di X
- Usare l'analisi grafica per riconoscere potenziali problemi nell'analisi della regressione

Analisi della Correlazione

- L’analisi della Correlazione è usata per misurare la forza dell’associazione (relazione lineare) tra due variabili
  - La correlazione era già stata introdotta nel Capitolo 3
  - La correlazione riguarda solo la forza della relazione
  - La correlazione non implica un effetto causale

Analisi della Correlazione

- Il coefficiente di correlatione della popolazione è indicato con ρ
- Il coefficiente di correlazione campionario è

$$r = \frac{s_{xy}}{s_x s_y}$$

dove

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Verifica di Ipotesi sulla Correlazione

- Per verificare l'ipotesi nulla di assenza di correlazione,

$$H_0 : \rho = 0$$

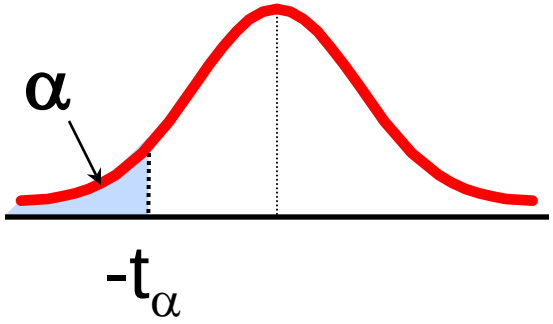
la statistica test ha una distribuzione t di Student con (n – 2) gradi di libertà:

$$T = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

Regole di Decisione

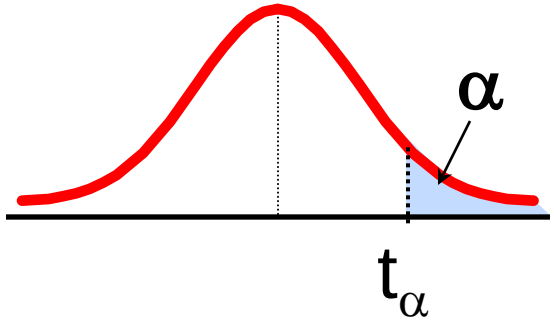
Verifica di Ipotesi sulla Correlazione

Test Unilaterale Sinistro:
H_0: $\rho = 0$
H_1: $\rho < 0$



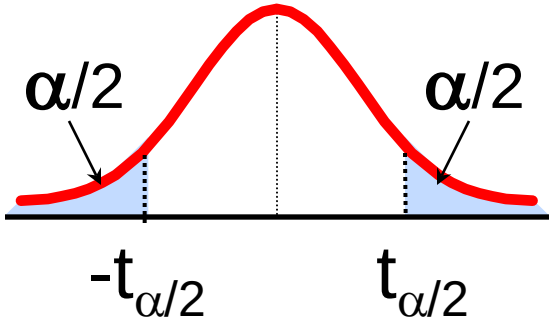
Rifiutare H_0 se $t < -t_{n-2,\ \alpha}$

Test Unilaterale Destro:
H_0: $\rho = 0$
H_1: $\rho > 0$



Rifiutare H_0 se $t > t_{n-2,\ \alpha}$

Test Bilaterale:
H_0: $\rho = 0$
H_1: $\rho \neq 0$



Rifiutare H_0 se $t < -t_{n-2,\ \alpha/2}$ o $t > t_{n-2,\ \alpha/2}$

Dove $t = \dfrac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$ ha n - 2 g.d.l.

Introduzione all'Analisi della Regressione

- L'analisi della Regressione viene usata per:
  - Prevedere il valore di una variabile dipendente sulla base del valore di almeno una variabile indipendente
  - Spiegare l'impatto di cambiamenti nella variabile indipendente sulla variabile dipendente

Variabile Dipendente: la variabile che desideriamo spiegare
(anche chiamata variabile endogena)

Variable Indipendente: la variabile usata per spiegare la variabile dipendente
(anche chiamata variabile esogena)

Modello di Regressione Lineare Semplice

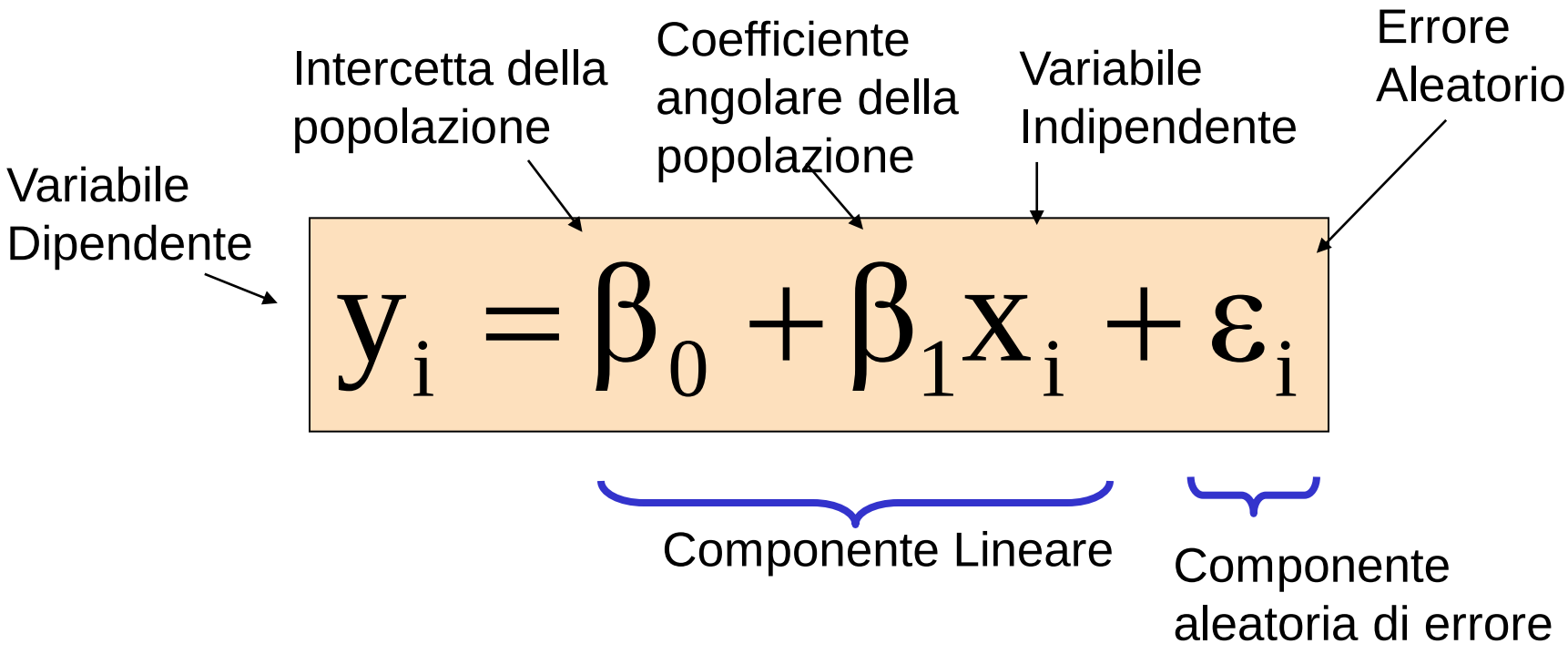
- La relazione tra X e Y è descritta da una funzione lineare
- Si assume che cambiamenti in Y siano causati da cambiamenti in X
- Equazione del modello di regressione lineare della popolazione

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

dove β_0 e β_1 sono i coefficienti del modello per la popolazione e ε_i è l'errore aleatorio.

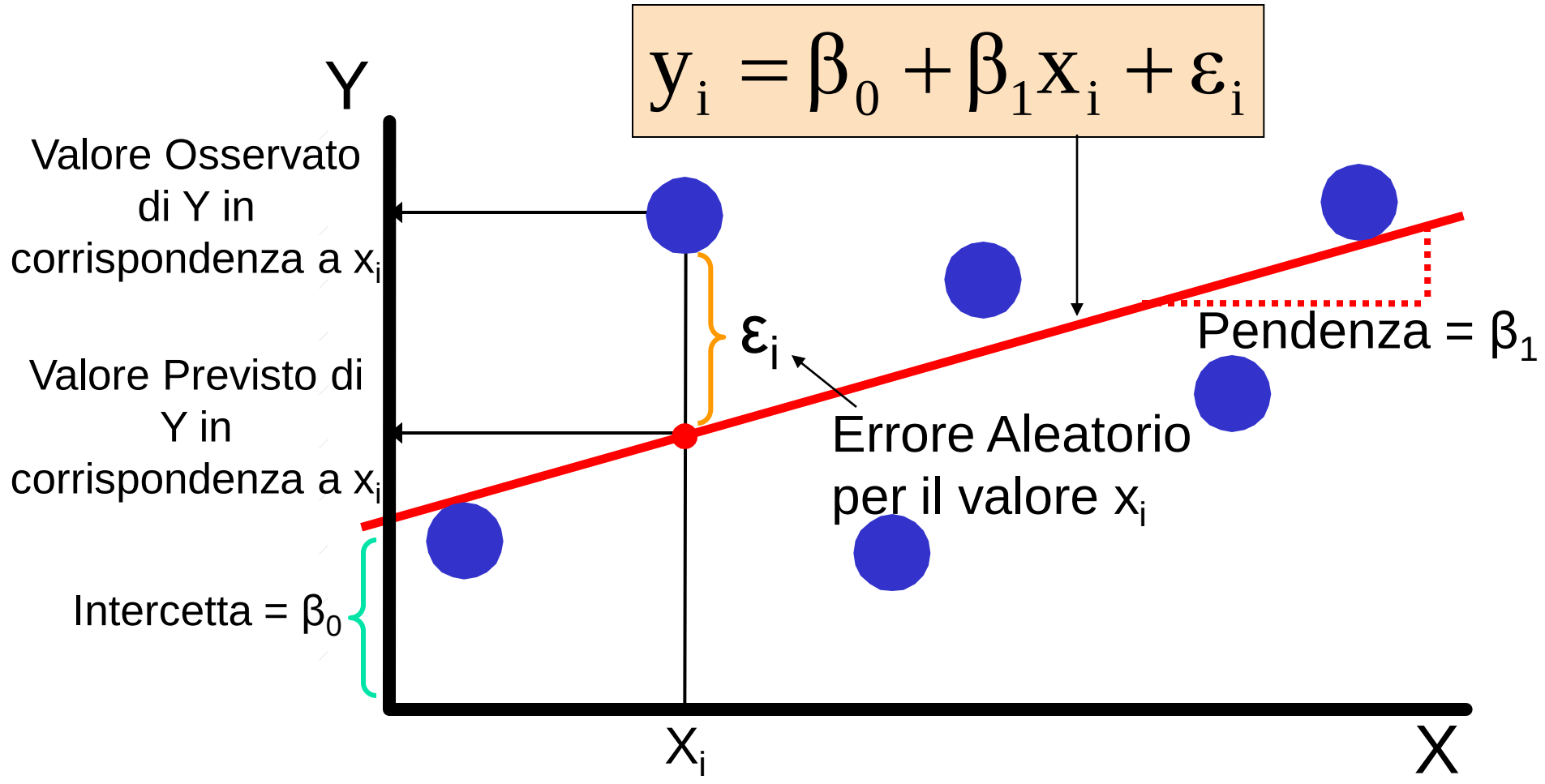
Modello di Regressione Lineare Semplice

Il modello di regressione per la popolazione:



Modello di Regressione Lineare Semplice

(continuazione)



Modello di Regressione Lineare Semplice

L'equazione della regressione lineare semplice fornisce una stima della retta di regressione per la popolazione

Valore stimato (o previsto) di y in corrispondenza della i-ma osservazione

Stima dell'intercetta

Stima del coefficiente angolare

Valore di x in corrispondenza della i-ma osservazione

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

Il residuo stimato e_i ha media zero

$$e_i = (y_i - \hat{y}_i) = y_i - (b_0 + b_1 x_i)$$

Stimatori dei Minimi Quadrati

- b_0 e b_1 sono ottenuti trovando i valori b_0 e b_1 che minimizzano la somma dei quadrati delle differenze tra y e $\hat{y}$:

$$\begin{aligned}\min \text{ SSE} &= \min \sum e_i^2 \\ &= \min \sum (y_i - \hat{y}_i)^2 \\ &= \min \sum [y_i - (b_0 + b_1 x_i)]^2\end{aligned}$$

Per ottenere gli stimatori dei coefficienti b_0 e b_1 che minimizzano SSE viene usato il calcolo differenziale

Stimatori dei Minimi Quadrati

(continuazione)

- Lo stimatore del coefficiente angolare è

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r_{xy}\frac{s_Y}{s_X}$$

- E la costante o intercetta è

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

- La retta di regressione passa sempre per il punto di coordinate $(\bar{x}, \bar{y})$

Determinazione dell'Equazione dei Minimi Quadrati

- I coefficienti b_0 e b_1 e gli altri risultati relativi alla regressione, forniti in questo capitolo, verranno ottenuti usando il computer
  - I calcoli a mano sono lunghi e ripetitivi
  - In Excel sono presenti molte funzioni statistiche
  - Si possono usare molti programmi applicativi di tipo statistico

Modello di Regressione Lineare: Assunzioni

- La forma della vera relazione è lineare (Y è una funzione lineare di X, più un errore aleatorio)
- I termini di errore, ε_i sono indipendenti dai valori di X
- I termini di errore sono variabili aleatorie con media 0 e varianza costante, σ^2

  (la proprietà di varianza costante è chiamata omoschedasticità)

$$E[\varepsilon_i] = 0 \quad \text{e} \quad E[\varepsilon_i^2] = \sigma^2 \qquad i = 1, \ldots, n$$

- I termini aleatori di errore, ε_i, non sono correlati fra loro, quindi

$$E[\varepsilon_i, \varepsilon_j] = 0 \qquad \text{per ogni } i \neq j$$

Interpretazione del Coefficiente Angolare e dell'Intercetta

- b_0 è il valore medio stimato di Y quando il valore di X è zero (se X = 0 appartiene all'intervallo di valori osservati per X)

- b_1 è la variazione stimata nel valore medio di Y relativa ad una variazione unitaria di X

Esempio: Regressione Lineare Semplice

- Un agente immobiliare vuole esaminare la relazione tra il prezzo di vendita di una casa e la sua superficie (misurata in piedi al quadrato)

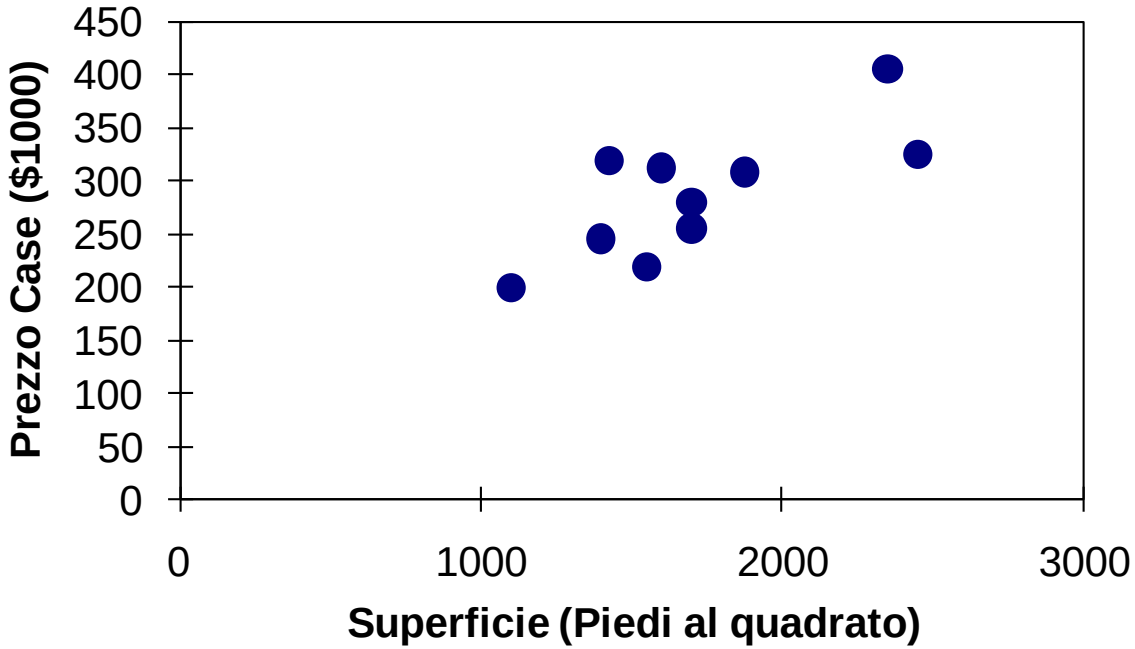
- Viene selezionato un campione casuale di 10 case
  - Variabile dipendente (Y) = prezzo case in $1000
  - Variabile indipendente (X) = superficie in piedi al quadrato

Dati Campionari per il Modello del Prezzo delle Case

Prezzo Case in $1000 (Y)	Superficie in piedi al quadrato (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

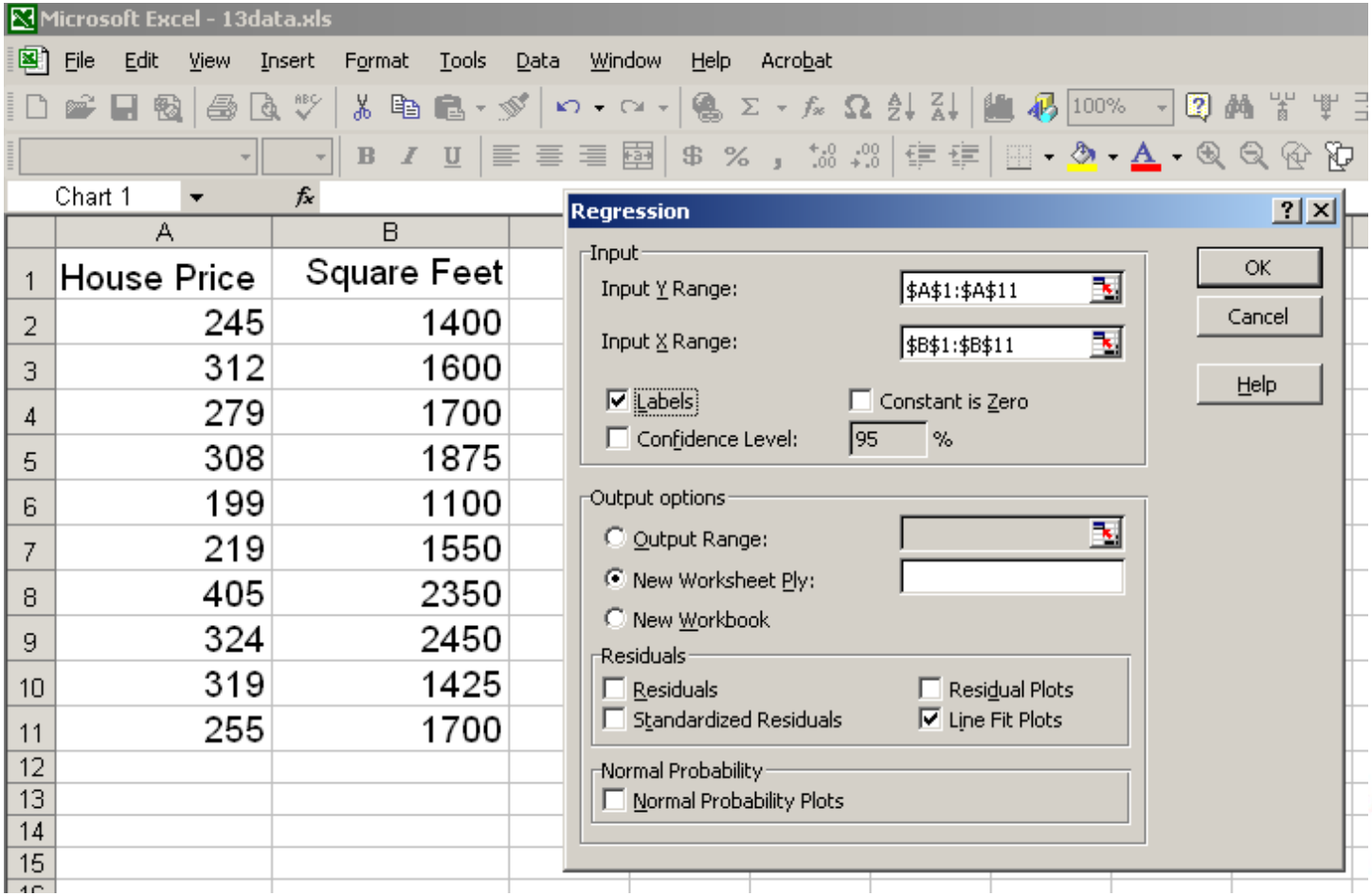
Rappresentazione Grafica

Modello del prezzo delle case: grafico di dispersione



Regressione con Excel

- Dati / Analisi Dati / Regressione



Output Excel

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

L'equazione della regressione è:

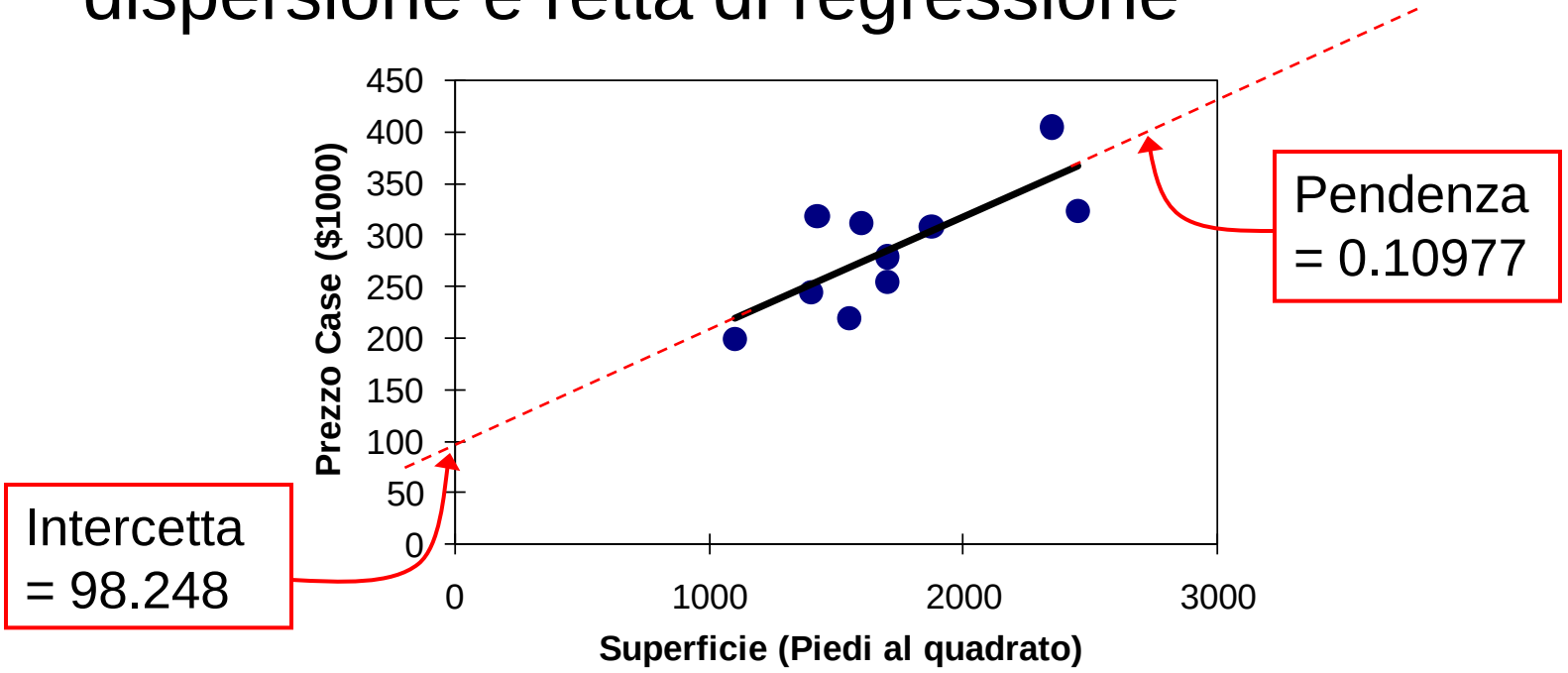
$$\widehat{\text{prezzo case}} = 98.24833 + 0.10977\,(\text{superficie})$$

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**

Rappresentazione Grafica

- Modello del prezzo delle case: grafico di dispersione e retta di regressione



$$\widehat{\text{prezzo case}} = 98.24833 + 0.10977\,(\text{superficie})$$

Interpretazione dell'Intercetta, b_0

$$\widehat{\text{prezzo case}} = 98.24833 + 0.10977\,(\text{superficie})$$

- b_0 è il valore medio stimato di Y quando il valore di X è zero (se X = 0 appartiene all'intervallo dei valori osservati per X)
  - Qui non ci sono case di 0 piedi al quadrato, quindi b_0 = 98.24833 indica solo che, per le case con una superficie compresa nell'intervallo dei valori osservati, $98248.33 è la porzione del prezzo che non è spiegata dalla superficie

Interpretazione del Coefficiente Angolare, b_1

$$\widehat{\text{prezzo case}} = 98.24833 + \boxed{0.10977}(\text{superficie})$$

- b_1 misura la variazione stimata nel valore medio di Y relativa ad una variazione unitaria di X
  - Qui $\boxed{b_1 = .10977}$ indica che il prezzo medio di una casa cresce , in media, di .10977($1000) = $109.77 per ogni piede al quadrato aggiuntivo nella superficie

Misure di Variabilità

- La variabilità complessiva di Y è composta da due parti:

$$SST = SSR + SSE$$

Somma dei Quadrati Totale	Somma dei Quadrati della Regressione	Somma dei Quadrati degli Errori
$SST = \sum (y_i - \bar{y})^2$	$SSR = \sum (\hat{y}_i - \bar{y})^2$	$SSE = \sum (y_i - \hat{y}_i)^2$

dove:

$\bar{y}$ = Valore medio della variabile dipendente

y_i = Valori osservati per la variabile dipendente

$\hat{y}_i$ = Valore previsto per Y in corrispondenza di un dato valore x_i

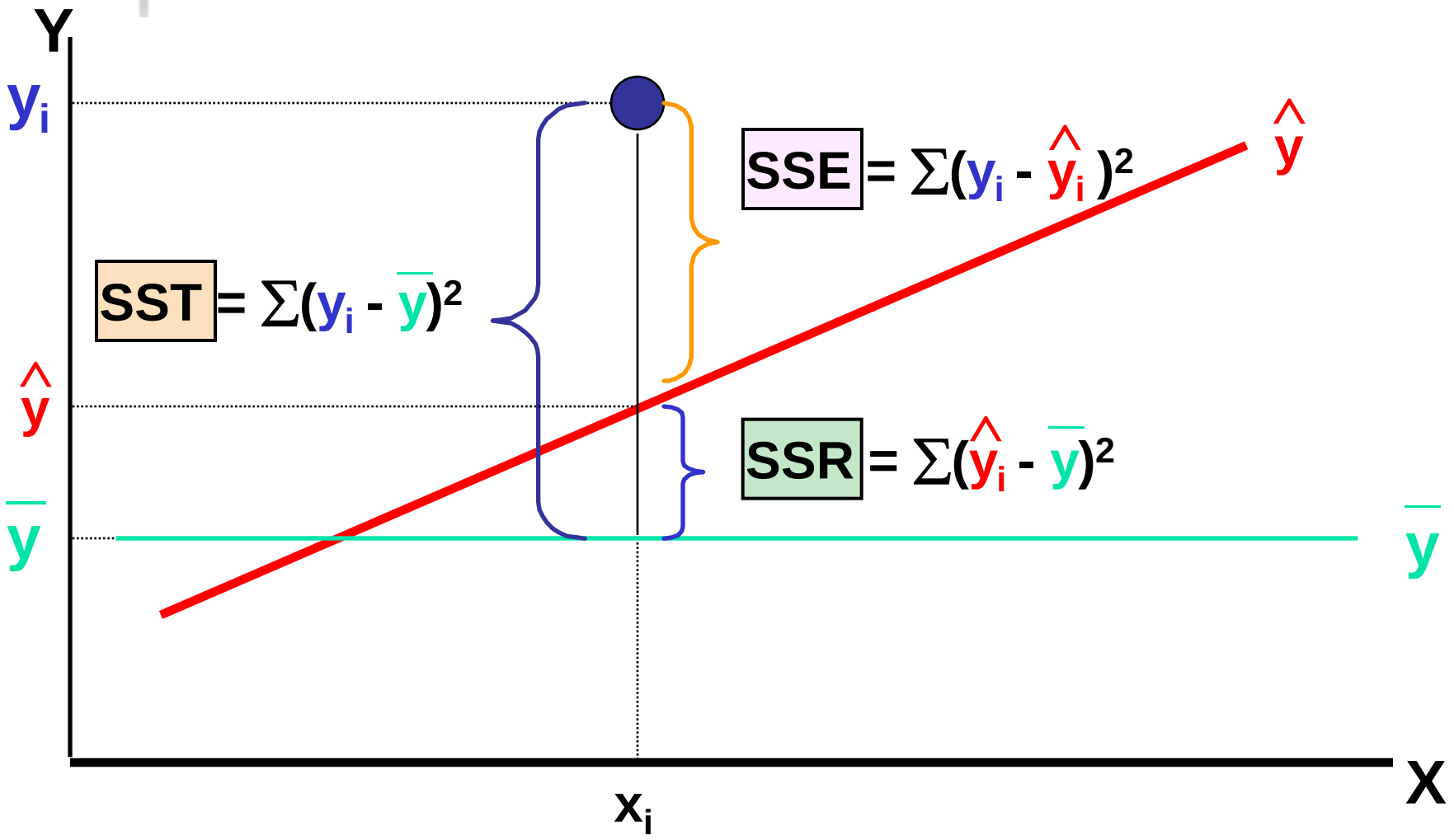
Misure di Variabilità

(continuazione)

- SST = somma dei quadrati totale
  - Misura la variazione dei valori y_i rispetto alla loro media, $\bar{y}$
- SSR = somma dei quadrati della regressione
  - Spiega la variabilità che può essere attribuita alla relazione lineare tra X e Y
- SSE = somma dei quadrati degli errori
  - Variabilità attribuibile a fattori diversi dalla relazione lineare tra X e Y

Misure di Variabilità

(continuazione)



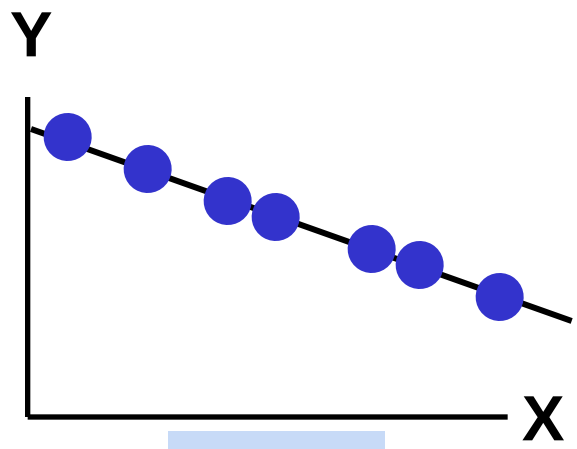
Coefficiente di Determinazione, R^2

- Il coefficiente di determinatione è la porzione di variabilità totale della variabile dipendente che è spiegata dalla variazione della variabile indipendente
- Il coefficiente di determinazione è anche chiamato R-quadrato ed è denotato con R^2

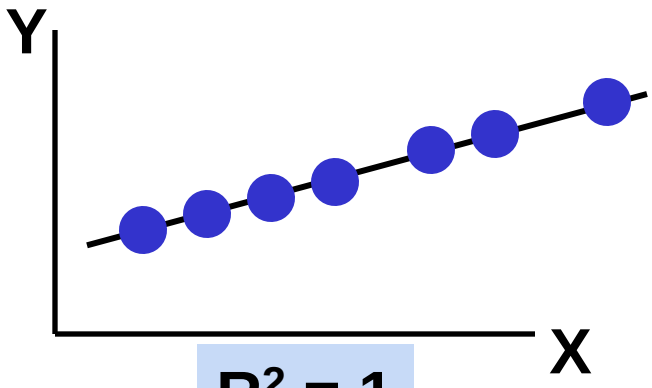
$$R^2 = \frac{SSR}{SST} = \frac{\text{somma dei quadrati della regressione}}{\text{somma dei quadrati totale}}$$

notare: $0 \le R^2 \le 1$

Esempi di Valori Approssimati di R^2



$R^2 = 1$



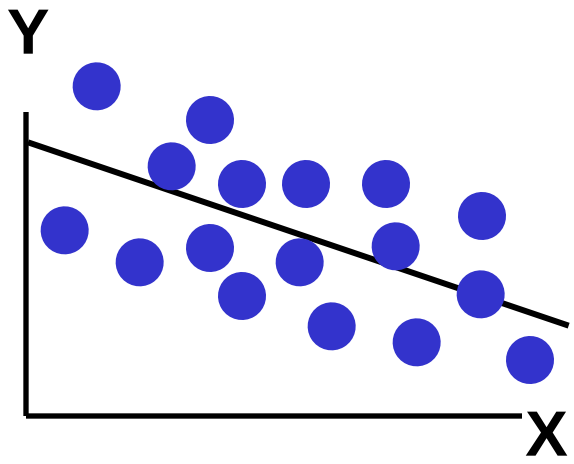
$R^2 = 1$

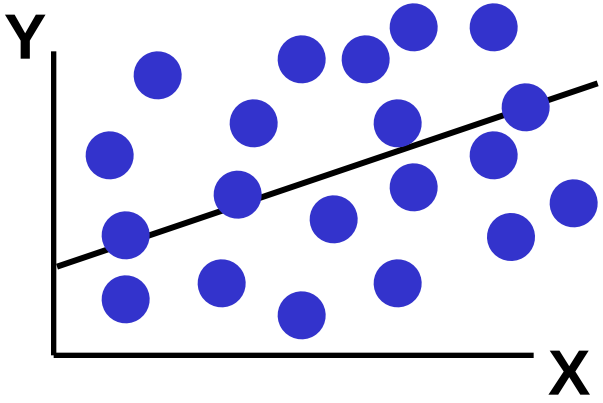
$R^2 = 1$

Relazione lineare perfetta tra X e Y:

100% della variabilità di Y è spiegata dalla variabilità di X

Esempi di Valori Approssimati di R^2



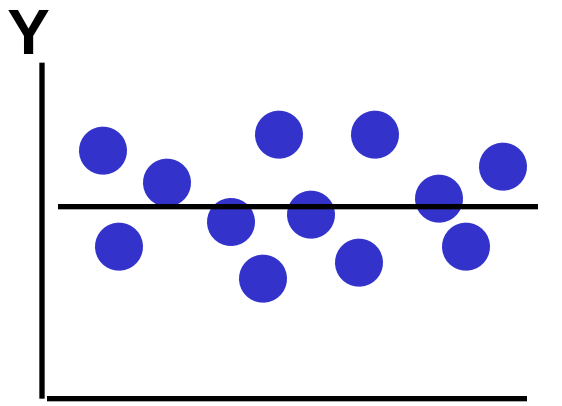


$0 < R^2 < 1$

Relazione lineare debole tra X e Y:

parte ma non tutta la variabilità di Y è spiegata dalla variabilità di X

Esempi di Valori Approssimati di R^2



$R^2 = 0$

$R^2 = 0$

Non esiste relazione lineare tra X e Y:

il valore di Y non dipende linearmente da X. (La variabilità di Y non è per nulla spiegata dalla variabilità di X)

Output Excel

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% della variazione del prezzo delle case è spiegato dalla variazione della superficie

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

Correlazione e R^2

- Il coefficiente di determinazione, R^2, per una regressione semplice, è uguale al quadrato del coefficiente di correlazione

$$R^2 = r_{xy}^2$$

Stima della Varianza del Modello

- Uno stimatore per la varianza del modello è

$$\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}$$

- La divisione per $n - 2$ invece di $n - 1$ deriva dal fatto che il modello di regressione lineare semplice usa due stime per i parametri, b_0 e b_1, invece di una

$$s_e = \sqrt{s_e^2}$$

è chiamato errore standard della stima (o del modello)

Output Excel

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

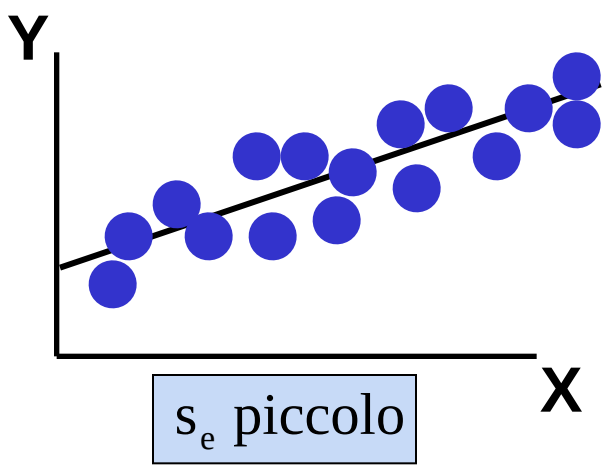
$$s_e = 41.33032$$

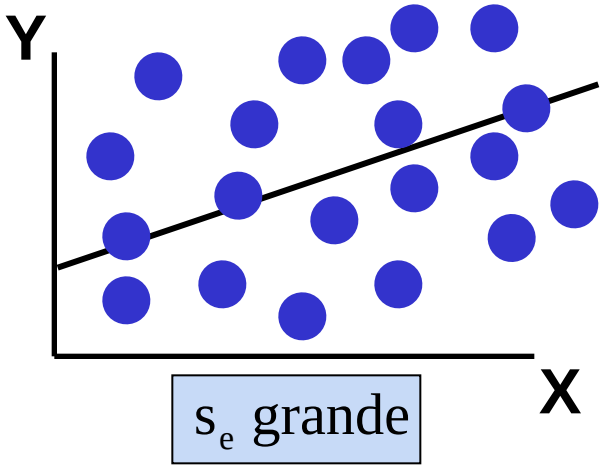
ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	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

Confronto fra Errori Standard

s_e è una misura della variazione dei valori osservati di Y rispetto alla retta di regressione





L'ordine di grandezza di s_e dovrebbe essere sempre giudicato in relazione all'ordine di grandezza dei valori campionari di Y

i.e., s_e = \$41.33 (in \$1000) è moderatamente piccolo relativamente ai prezzi delle case compresi nell'intervallo \$200 - \$300 (in \$1000)

Inferenza sul Modello di Regressione

- La varianza del coefficiente angolare della retta di regressione (b_1) è stimata da

$$s_{b_1}^2 = \frac{s_e^2}{\sum (x_i - \bar{x})^2} = \frac{s_e^2}{(n-1)s_x^2}$$

dove:

s_{b_1} = Errore standard del coefficiente angolare b_1

$s_e = \sqrt{\frac{SSE}{n-2}}$ = Errore Standard della stima

Output Excel

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

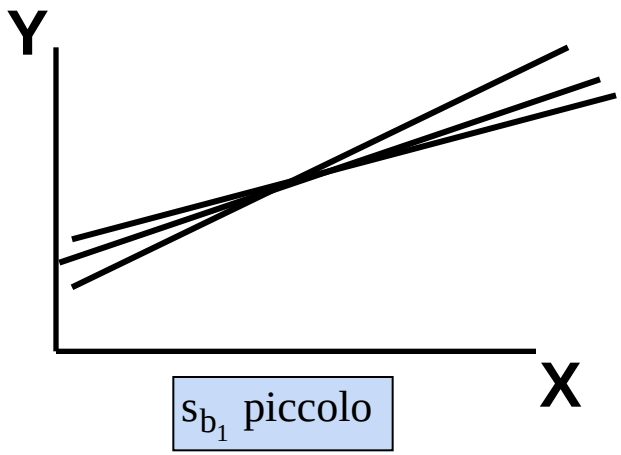
$$s_{b_1} = 0.03297$$

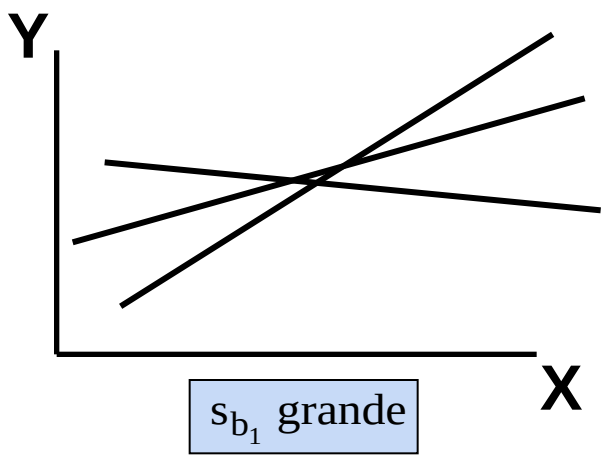
ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	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

Confronto fra Errori Standard del Coefficiente Angolare

s_{b_1} è una misura della variazione del coefficiente angolare della retta di regressione per diversi possibili campioni





Inferenza sul Coefficiente Angolare: Test T

- Test T sul coefficiente angolare della popolazione
  - C'è una relazione lineare tra X e Y?
- Ipotesi nulla e alternativa

H_0: $\beta_1 = 0$ (non esiste una relazione lineare)
H_1: $\beta_1 \neq 0$ (esiste una relazione lineare)

- Statistica test

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

$$\text{g.d.l.} = n - 2$$

dove:

b_1 = coefficiente angolare della regressione

β_1 = pendenza ipotizzata

s_{b1} = errore standard del coefficiente angolare

Esempio: Test T sul Coefficiente Angolare

(continuazione)

Prezzo Case in $1000 (Y)	Superficie in Piedi al quarato (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Stima dell'Equazione della Retta di Regressione:

$$\widehat{\text{prezzo case}} = 98.25 + 0.1098\,(\text{superficie})$$

Il coefficiente angolare del modello è 0.1098

La superficie di una casa ne influenza il prezzo di vendita?

Esempio: Test T sul Coefficiente Angolare

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Dall'output Excel:

b_1 S_{b_1}

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Esempio: Test T sul Coefficiente Angolare

(continuazione)

Statistica test: **t = 3.329**

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

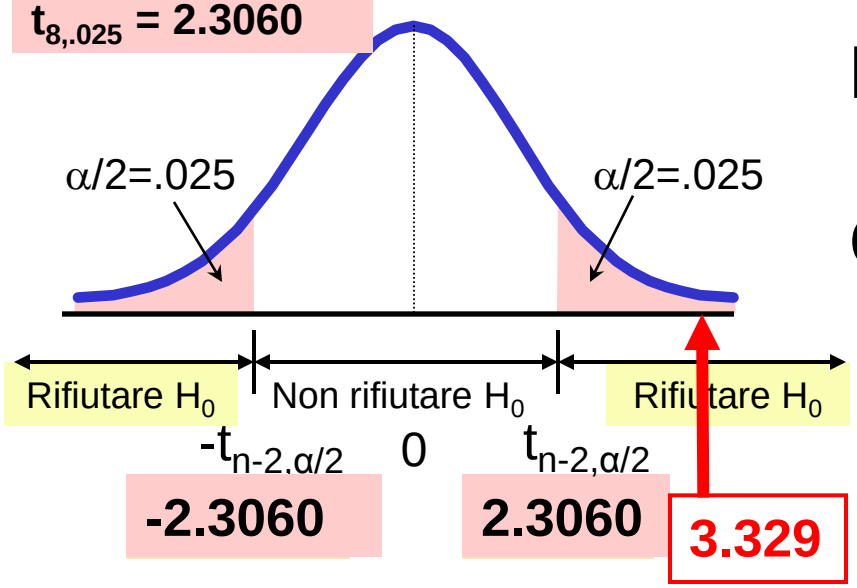
Dall'output Excel:

b_1 S_{b_1} t

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

g.d.l. = 10-2 = 8

$t_{8,.025}$ = 2.3060



Decisione:

Rifiutare H_0

Conclusione:

Ci sono sufficienti evidenze che la superficie influenzi il prezzo della casa

Esempio: Test T sul Coefficiente Angolare:

(continuazione)

P-value = **0.01039**

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Dall'output Excel:

p-value

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

Questo è un test bilaterale, quindi il p-value è

$P(t > 3.329)+P(t < -3.329) = 0.01039$

(per 8 g.d.l.)

Decisione: P-value < α quindi Rifiutare H_0

Conclusione:

Ci sono sufficienti evidenze che la superficie influenzi il prezzo della casa

Stima per Intervallo del Coefficiente Angolare

Stima per intervallo del coefficiente angolare:

$$b_1 - t_{n-2,\alpha/2} s_{b_1} < \beta_1 < b_1 + t_{n-2,\alpha/2} s_{b_1}$$

g.d.l. = *n* - 2

Output Excel per i prezzi delle case:

	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**

A livello di confidenza del 95%, l'intervallo di confidenza per il coefficiente angolare è (0.0337, 0.1858)

Stima per Intervallo del Coefficiente Angolare

(continuazione)

	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**

Siccome l'unità di misura per la variabile Prezzo delle Case è $1000, siamo confidenti al 95% che l'impatto medio sul prezzo delle case è fra $33.70 e $185.80 per ogni piede al quadrato

Questo intervallo di confidenza al 95% non include lo 0.

Conclusione: C'è una relazione significativa fra il prezzo delle case e la loro superficie, a livello di significatività .05

Test F su β_1

- Statistica Test:

$$F = \frac{MSR}{MSE}$$

con

$$MSR = \frac{SSR}{k}$$

$$MSE = \frac{SSE}{n-k-1}$$

La statistica test F ha una distribuzione F con *k* gradi di libertà del numeratore e (*n* – *k* - 1) gradi di libertà del denominatore

(*k* = numero di variabili indipendenti nel modello di regressione)

Output Excel

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

$$F = \frac{MSR}{MSE} = \frac{18934.9348}{1708.1957} = 11.0848$$

Con 1 e 8 gradi di libertà

p-value per il test F

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

Test F su β_1

(continuazione)

$H_0: \beta_1 = 0$

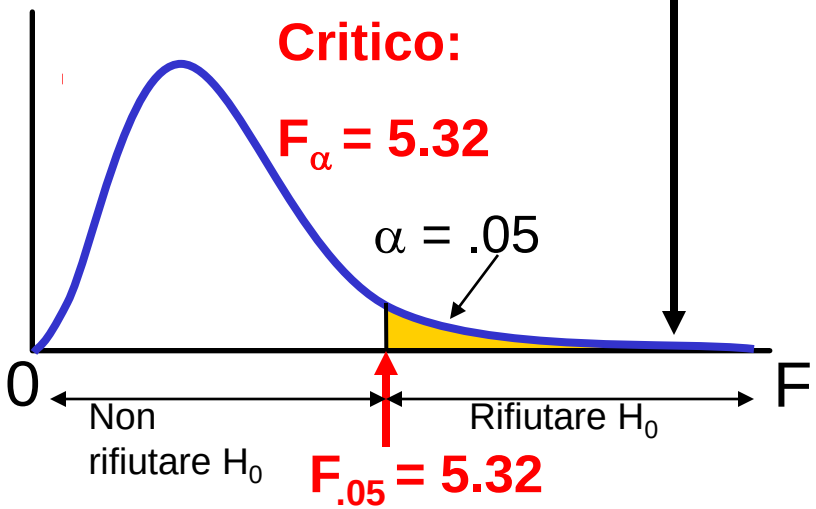
$H_1: \beta_1 \neq 0$

$\alpha = .05$

$gdl_1 = 1, gdl_2 = 8$

Valore Critico:

$F_\alpha = 5.32$



Statistica test:

$$F = \frac{MSR}{MSE} = 11.08$$

Decisione:

Rifiutare H_0 per $\alpha = 0.05$

Conclusione:

Ci sono sufficienti evidenze che la superficie delle case influenzi il loro prezzo di vendita

Previsione

- L'equazione della regressione può essere usata per prevedere un valore di Y in corrispondenza di un particolare valore di X

- Per uno specifico valore, x_{n+1} il valore previsto è

$$\hat{y}_{n+1} = b_0 + b_1 x_{n+1}$$

Previsione con l'Analisi della Regressione

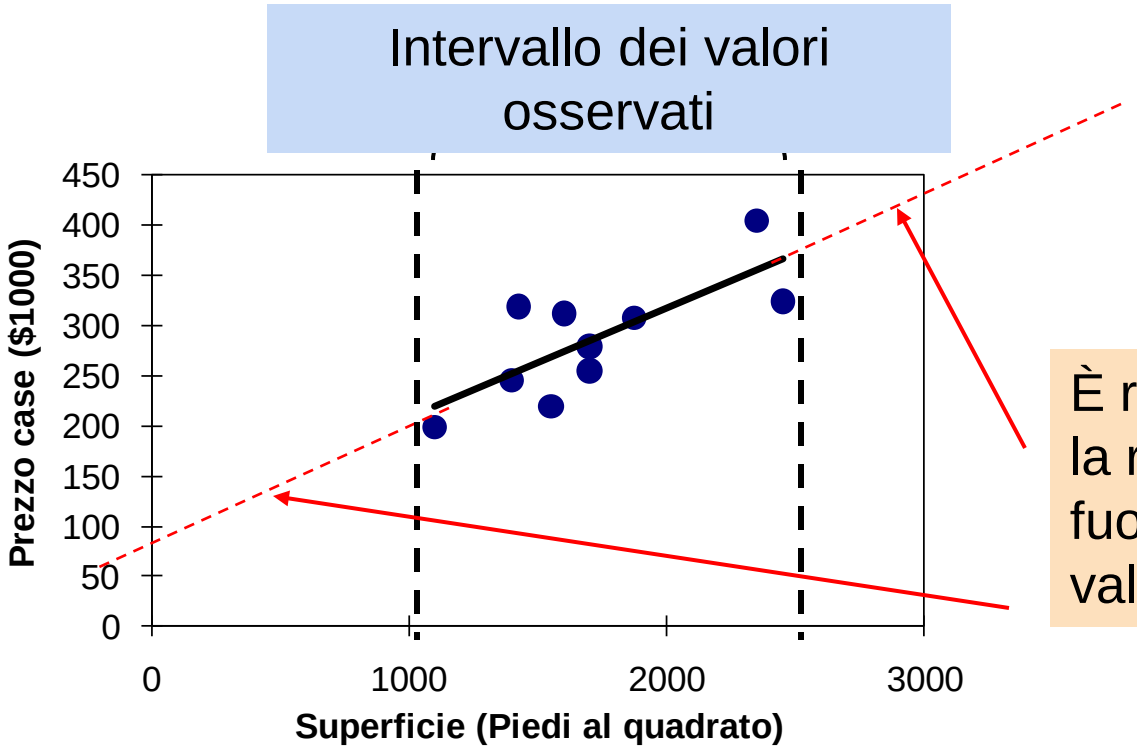
Prevedere il prezzo di una casa di 2000 piedi al quadrato:

$$\widehat{\text{prezzo case}} = 98.25 + 0.1098\,(\text{superficie})$$
$$= 98.25 + 0.1098(2000)$$
$$= 317.85$$

Il valore previsto per una casa di 2000 piedi al quadrato è 317.85($1000) = $317850

Intervallo dei Valori Osservati

- Quando si usa un modello di regressione per fare previsioni, la previsione deve essere fatta entro l'intervallo dei valori osservati



È rischioso estrapolare la regressione al di fuori dell'intervallo dei valori osservati per X

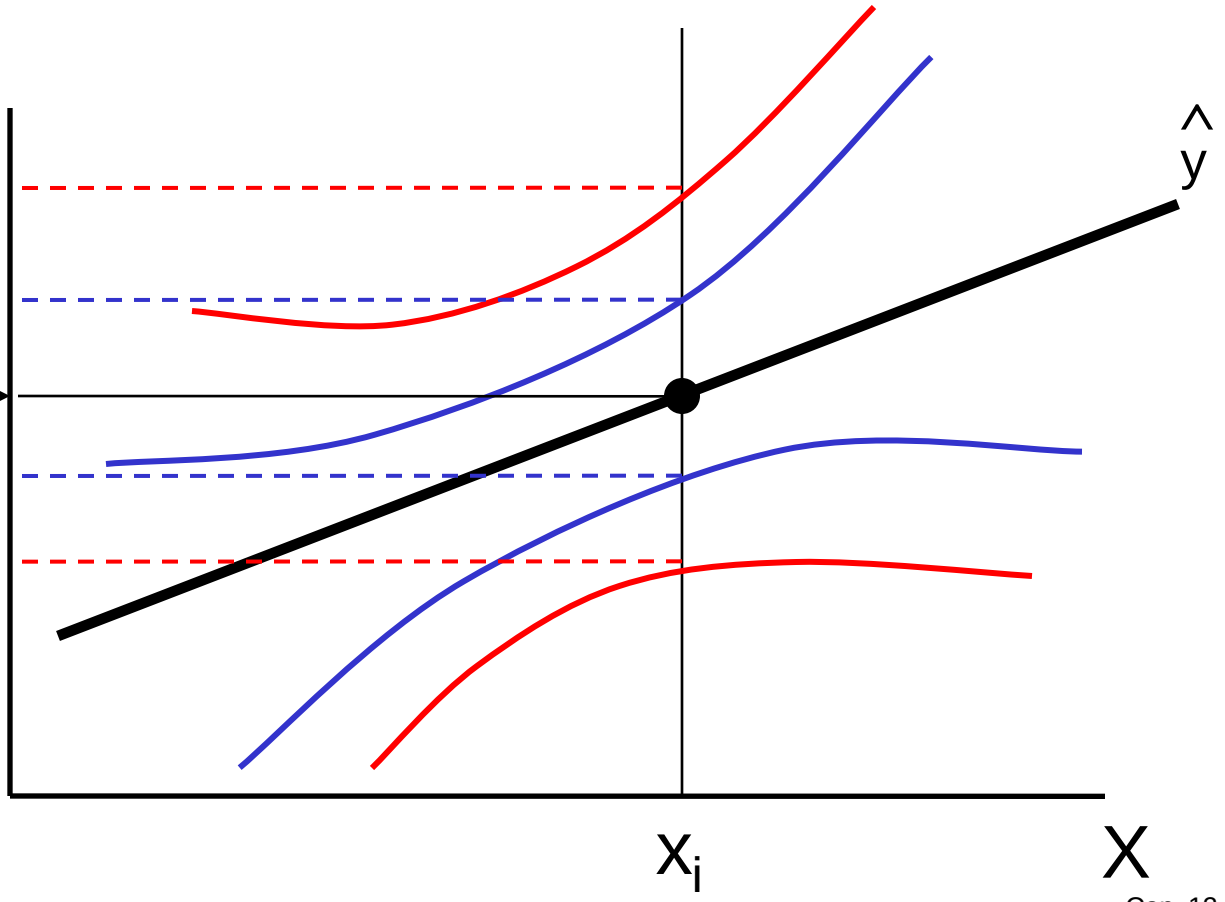
Previsione del Valore Medio e Previsione del Singolo Valore

Obiettivo: determinare degli intervalli sull'asse Y, per esprimere l'incertezza del valore di Y in corrispondenza ad un dato x_i

Intervallo di confidenza per la previsione del valore atteso di Y, dato x_i

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

Intervallo di confidenza per la previsione del singolo valore di Y, dato x_i



Intervallo di Confidenza per la Previsione del Valore Atteso di Y, dato X

Intervallo di confidenza per la previsione del **valore atteso di Y** in corrispondenza a x_i

Intervallo di Confidenza per $E(Y_{n+1} \mid x_{n+1})$:

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} s_e \sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$$

Notare che la formula include il termine $(x_{n+1} - \bar{x})^2$

quindi l'ampiezza dell'intervallo varia in funzione della distanza fra x_{n+1} e la media, $\bar{x}$

Intervallo di Confidenza per la Previsione di un Singolo Y, dato X

Intervallo di confidenza per la previsione del **valore osservato di Y** in corrispondenza a x_i

Intervallo di Confidenza per $\hat{y}_{n+1}$:

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} s_e \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$$

All'ampiezza dell'intervallo si somma questo termine, per includere l'incertezza aggiuntiva relativa alla previsione di un singolo valore

Esempio: Intervallo di Confidenza per la Previsione del Valore Atteso

Intervallo di Confidenza per $E(Y_{n+1}|x_{n+1})$

Trovare l'intervallo di confidenza al 95% per il prezzo medio delle case di 2000 piedi al quadrato

Prezzo Previsto $\hat{y}_i = 317.85$ ($1000)

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = 317.85 \pm 37.12$$

I limiti dell'intervallo di confidenza sono 280.66 e 354.90, cioè da $280660 a $354900

Esempio: Intervallo di Confidenza per la Previsione del Singolo Valore

Intervallo di Confidenza per $\hat{y}_{n+1}$

Trovare l'intervallo di confidenza al 95% per il prezzo di una singola casa di 2000 piedi al quadrato

Prezzo Previsto $\hat{y}_i = 317.85$ ($1000)

$$\hat{y}_{n+1} \pm t_{n-1,\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = 317.85 \pm 102.28$$

I limiti dell'intervallo di confidenza sono 215.50 e 420.07, cioè da $215500 a $420070

Intervalli di Confidenza per la Previsione con PHStat

- Utilizzare

  PHStat | Regression | Simple Linear Regression …

  - Selezionare l'opzione

    "Confidence and Prediction Interval for X ="

    e inserire i valori di X e il livello di confidenza desiderato

Intervalli di Confidenza per la Previsione con PHStat

(continuazione)

	A	B
1	**Confidence Interval Estimate**	
2		
3	**Data**	
4	**X Value**	**2000**
5	**Confidence Level**	**95%**
6		
7	Intermediate Calculations	
8	Sample Size	10
9	Degrees of Freedom	8
10	t Value	2.306006
11	Sample Mean	1715
12	Sum of Squared Difference	1571500
13	Standard Error of the Estimate	41.33032
14	h Statistic	0.151686
15	Average Predicted Y (YHat)	317.7838
16		
17	**For Average Predicted Y (YHat)**	
18	Interval Half Width	37.11952
19	**Confidence Interval Lower Limit**	**280.6643**
20	**Confidence Interval Upper Limit**	**354.9033**
21		
22	**For Individual Response Y**	
23	Interval Half Width	102.2813
24	**Prediction Interval Lower Limit**	**215.5025**
25	**Prediction Interval Upper Limit**	**420.0651**

Valori in Input

$\hat{y}$

Intervallo di confidenza per $E(Y_{n+1}|X_{n+1})$

Intervallo di confidenza per il singolo $\hat{y}_{n+1}$

Analisi Grafica

- Il modello di regressione lineare è basato sulla minimizzazione della somma dei quadrati degli errori
- Se ci sono outlier, il loro errore al quadrato, potenzialmente grande, può avere una forte influenza sulla retta di regressione ottenuta
- Esaminare sempre i dati graficamente, per individuare outlier e punti estremi
- Decidere, sulla base del vostro modello e della logica, se è opportuno tenere o rimuovere i punti estremi

Riepilogo del Capitolo

- Introdotto il modello di regressione lineare semplice
- Discusse la correlazione e le assunzioni su cui si basa la regressione lineare
- Discussa la stima dei coefficienti di regressione lineare
- Descritte le misure di variazione
- Descritta l'inferenza sul coefficiente angolare
- Discussi gli intervalli di confidenza per la previsione dei valori medi e per la previsione di singoli valori