



# La statistica nelle prove Invalsi

## LA REGRESSIONE

**Luca Scrucca**

**Dipartimento di Economia**

**Università degli Studi di Perugia**

**luca@stat.unipg.it**

**<http://www.stat.unipg.it/luca>**

Sala Sant'Anna, Scuola San Paolo

Viale Roma 15 – Perugia

11 Dicembre 2014

La regressione verso la mediocrità ...

Il modello di regressione lineare

oooooooooooooooo

Inferenza per il modello di regressione

ooooo

Shiny App interattiva

Estensioni

## Programma

- ✓ Dalla "regressione verso la mediocrità" alle Shiny App interattive: un breve sguardo storico al concetto di regressione
- ✓ Il modello di regressione lineare semplice
- ✓ Un approccio geometrico alla stima dei coefficienti di regressione
- ✓ Il punto di vista statistico al problema di stima dei parametri
- ✓ Interpretazione e utilizzo del modello di regressione
- ✓ L'utilizzo della grafica dinamica nella didattica
- ✓ Estensioni al modello di regressione lineare semplice

# La regressione verso la mediocrità ...

246

*Anthropological Miscellanea.*

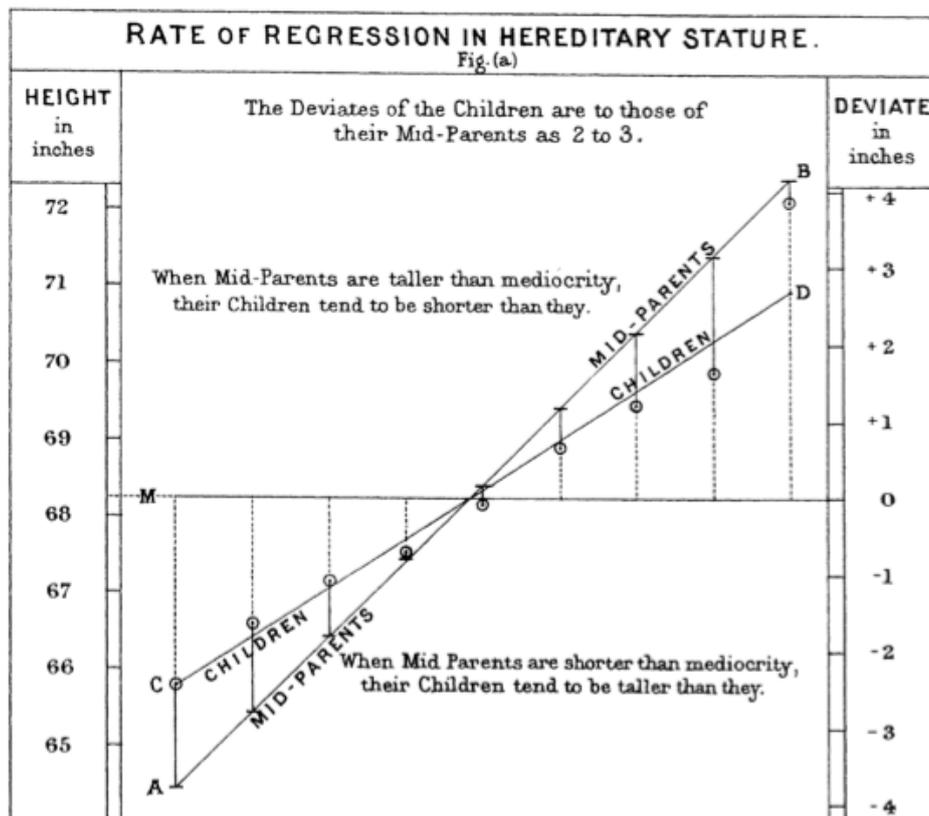
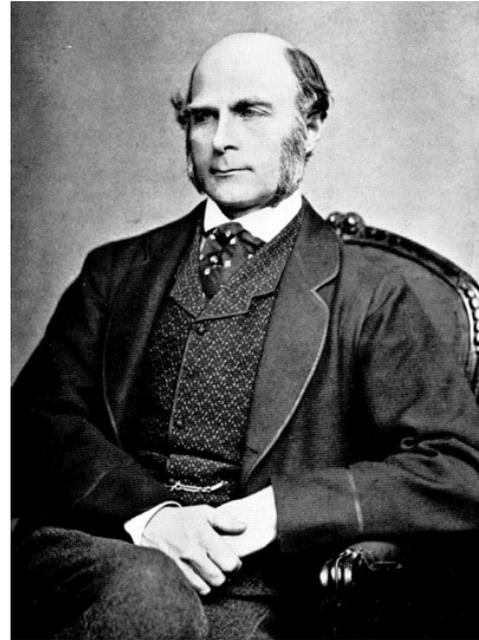
## ANTHROPOLOGICAL MISCELLANEA.

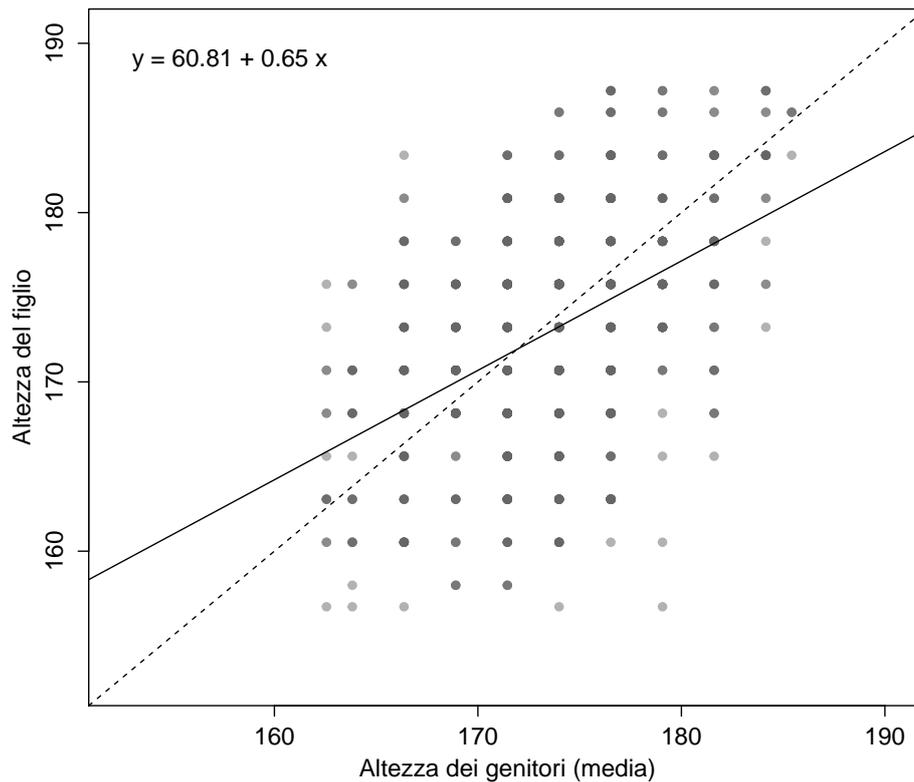
### REGRESSION *towards* MEDIOCRITY in HEREDITARY STATURE.

By FRANCIS GALTON, F.R.S., &c.

[WITH PLATES IX AND X.]

THIS memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.





## Il modello di regressione lineare

- ✓ Si considerino due caratteri quantitativi,  $X$  e  $Y$ , osservati su  $n$  unità statistiche di un collettivo.
- ✓ L'analisi di regressione suppone l'esistenza di una relazione funzionale tra la **variabile risposta** (o dipendente)  $Y$  e la **variabile esplicativa** (o indipendente)  $X$ , dove quest'ultima assume la veste di antecedente logico.

*Esempio:* Relazione tra fatturato e numero di addetti delle aziende.

*Esempio:* Relazione tra consumo pro-capite e reddito delle famiglie.

- ✓ L'obiettivo dell'analisi di regressione è capire se e come  $X$  influenza  $Y$  approssimando tale relazione tramite una funzione matematica  $y = f(x)$ .
- ✓ Nel caso del modello di regressione lineare la funzione  $f(x)$  è lineare.

## I dati

- ✓ I dati di partenza di un'analisi di regressione sono le osservazioni o i dati sperimentali su  $n$  unità statistiche.
- ✓ Una distribuzione doppia disaggregata può essere descritta dalla seguente tabella:

$X$	$Y$
$x_1$	$y_1$
$x_2$	$y_2$
$\vdots$	$\vdots$
$x_i$	$y_i$
$\vdots$	$\vdots$
$x_N$	$y_N$

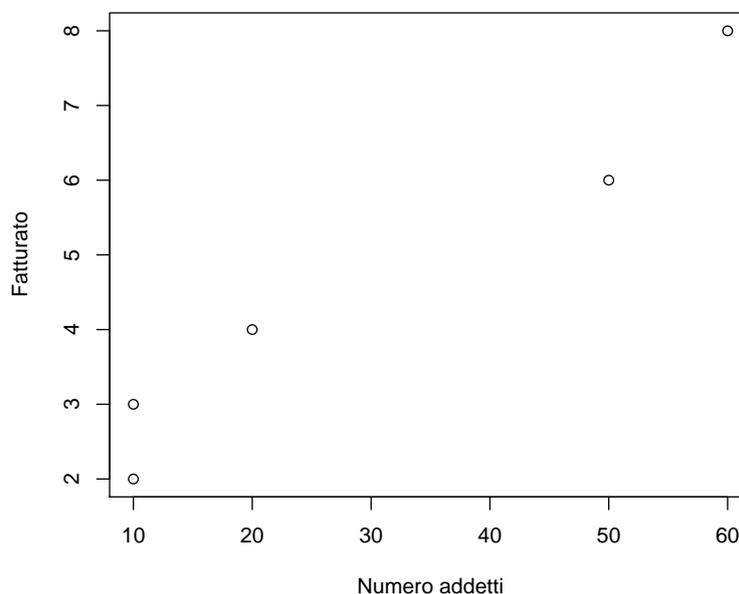
- ✓ Graficamente può essere rappresentata tramite un [grafico di dispersione](#), in cui si rappresentano in uno spazio cartesiano i punti di coordinate  $(x_i, y_i)$ ,  $i = 1, \dots, n$ .

### Esempio: relazione tra numero di addetti e fatturato

È stato rilevato il fatturato (in milioni di €) ed il numero di addetti di 5 aziende:

Numero di addetti ( $X$ )	60	10	50	20	10
Fatturato ( $Y$ )	8	3	6	4	2

Il corrispondente grafico di dispersione è il seguente:

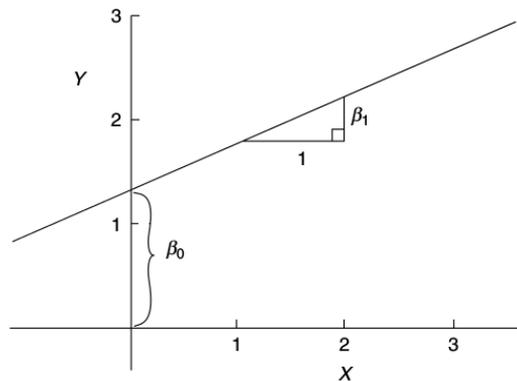


## Regressione lineare semplice

- ✓ Il modello di regressione lineare semplice consente di studiare la relazione tra  $Y$  e  $X$  tramite la funzione lineare

$$y = \beta_0 + \beta_1 x,$$

dove  $\beta_0$  (intercetta) e  $\beta_1$  (coefficiente angolare) sono chiamati **parametri** della retta interpolatrice:



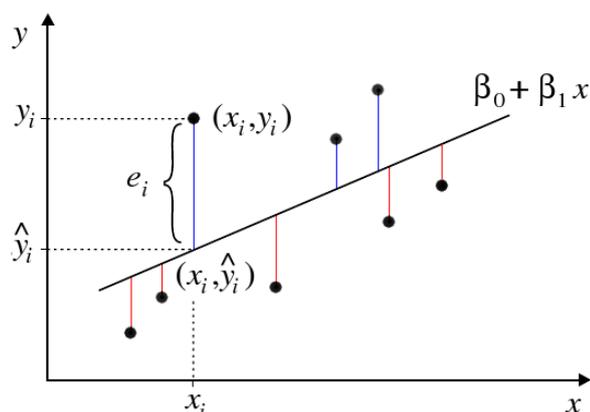
- ✓ I valori dei parametri incogniti  $(\beta_0, \beta_1)$  sono scelti in modo da ottimizzare l'adattamento della retta ai dati, ovvero minimizzando la distanza tra i valori osservati  $y_i$  e i valori  $\hat{y}_i$  che giacciono sulla retta.

## Il metodo dei minimi quadrati ordinari

- ✓ Metodo per la stima di parametri non noti introdotto da Legendre (1805) e Gauss (1809).
- ✓ Per un insieme di valori dei parametri  $(\hat{\beta}_0, \hat{\beta}_1)$ , si definisce valore teorico corrispondente alla  $i$ -esima osservazione la quantità

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- ✓ Il corrispondente errore o residuo è  $e_i = y_i - \hat{y}_i$



- ✓ La somma dei quadrati dei residui è una misura complessiva dell'errore che si commette interpolando i punti osservati con una retta:

$$S_q(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^n e_i^2$$

- ✓ Il **metodo dei minimi quadrati** consiste nel *minimizzare la somma dei quadrati dei residui* ( $S_q$ ) rispetto ai parametri  $\beta_0$  e  $\beta_1$ . Si può dimostrare che il minimo si raggiunge quando il coefficiente angolare e l'intercetta sono pari, rispettivamente, a

$$\hat{\beta}_1 = \frac{C_{XY}}{D_X}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

dove  $\bar{x}$  e  $\bar{y}$  sono le medie di  $X$  e di  $Y$ , mentre

$$C_{XY} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \text{codevarianza tra } X \text{ e } Y$$

$$D_X = \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{devianza di } X$$

- ✓ La retta di regressione di  $Y$  rispetto a  $X$  si ottiene sostituendo a  $\beta_0$  e  $\beta_1$  i corrispondenti valori trovati con il metodo dei minimi quadrati. Quindi l'equazione della **retta di regressione** è

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

*Esempio: relazione tra numero di addetti e fatturato (continua)*

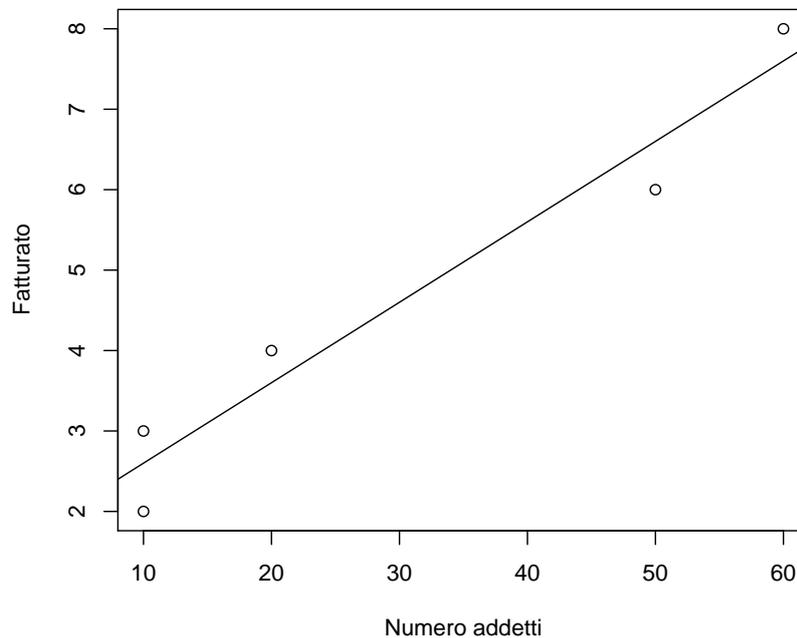
$i$	$x_i$	$y_i$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	60	8	900	102
2	10	3	400	32
3	50	6	400	28
4	20	4	100	6
5	10	2	400	52
Totale	150	23	2200	220

$$\bar{x} = \frac{150}{5} = 30, \quad \bar{y} = \frac{23}{5} = 4.6, \quad D_X = 2200, \quad C_{XY} = 220$$

$$\hat{\beta}_1 = \frac{220}{2200} = 0.1 \quad \hat{\beta}_0 = 4.6 - 0.1 \times 30 = 1.6$$

*Esempio:* Quindi, l'equazione della retta di regressione del fatturato ( $Y$ ) sul numero di addetti ( $X$ ) è

$$\hat{y} = 1.6 + 0.1x$$



## Indice di determinazione

- ✓ Si può dimostrare che vale la seguente **decomposizione della devianza totale**:

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

devianza totale

devianza spiegata

devianza residua

- ✓ La devianza residua è sempre non negativa ( $D_R \geq 0$ ) e quanto più è vicina a 0, tanto migliore è la bontà di adattamento della retta ai punti.
- ✓ Anche la devianza spiegata dalla regressione è una quantità sempre non negativa ( $D_S \geq 0$ ) e quanto più è vicina a 0 e tanto peggiore è la bontà di adattamento della retta ai punti.

- ✓ Per misurare la bontà di adattamento della retta ai punti si può utilizzare l'indice di determinazione:

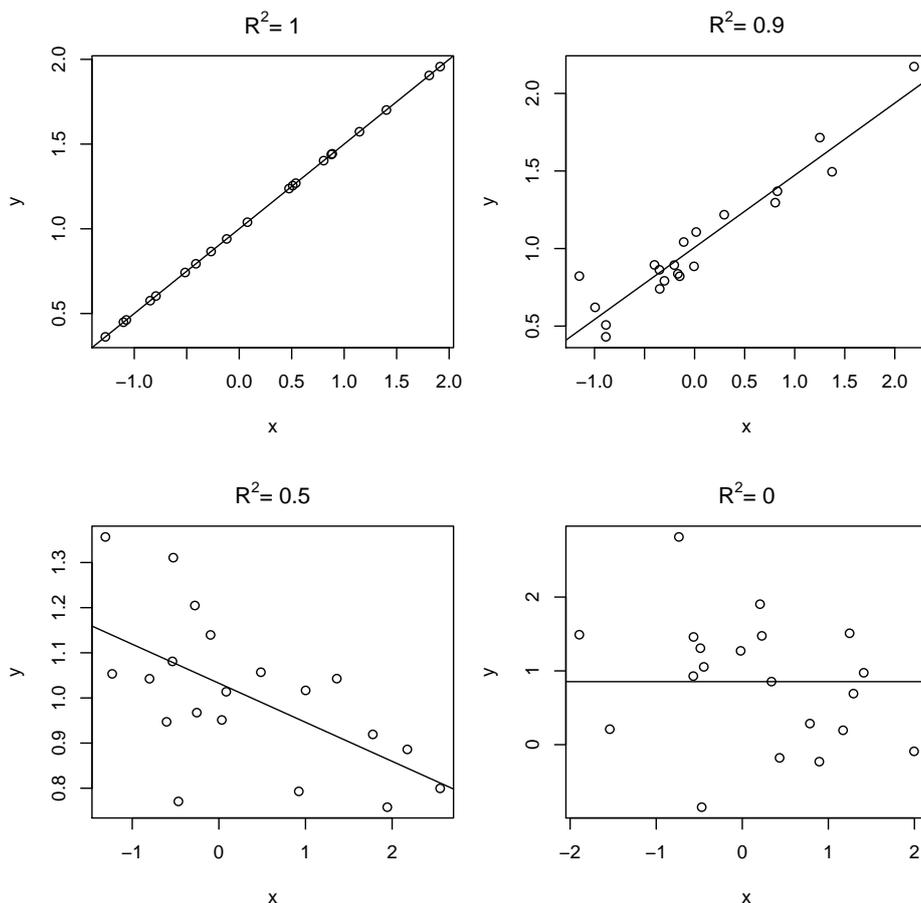
$$R^2 = \frac{D_S}{D_Y} = 1 - \frac{D_R}{D_Y}$$

- ✓ L'indice di determinazione assume valori nell'intervallo  $[0, 1]$ . Per la sua interpretazione si consideri che:

$R^2 = 0$  → assenza di relazione statistica di tipo lineare di  $Y$  da  $X$   
( $D_S = 0, D_R = D_Y$ )

$R^2 > 0$  → dipendenza lineare di  $Y$  da  $X$   
( $0 < D_S < D_Y$ )

$R^2 = 1$  → perfetta dipendenza lineare di  $Y$  da  $X$   
( $D_S = D_Y, D_R = 0$ )



Esempio: relazione tra numero di addetti e fatturato (continua) Per il modello di equazione

$$\hat{y} = 1.6 + 0.1x$$

si ottiene la seguente tabella:

$i$	$x_i$	$y_i$	$(y_i - \bar{y})^2$	$\hat{y}_i$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
1	60	8	11.56	7.6	9	0.16
2	10	3	2.56	2.6	4	0.16
3	50	6	1.96	6.6	4	0.36
4	20	4	0.36	3.6	1	0.16
5	10	2	6.76	2.6	4	0.36
Totale	150	23	23.20	23	22	1.20

$$\bar{y} = 4.6, \quad D_Y = 23.2, \quad D_S = 22, \quad D_R = 1.2$$

Quindi:

$$R^2 = \frac{22}{23.2} = 1 - \frac{1.2}{23.2} = 0.9483$$

L'indice di determinazione indica un ottimo adattamento della retta di regressione alla nuvola dei punti.

## Previsione

- ✓ Una volta calcolati i parametri della retta di regressione  $(\hat{\beta}_0, \hat{\beta}_1)$  è possibile "prevedere" <sup>1</sup> il valore di  $Y$  in corrispondenza di un nuovo valore di  $X$ , indicato con  $x_*$ . Il valore previsto si calcola come:

$$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$$

- ✓ In generale, si parla di **interpolazione** se il valore  $x_*$  per il quale si calcola il valore previsto è compreso nel range dei valori osservati di  $X$ , viceversa si parla di **estrapolazione**. In quest'ultimo caso il valore calcolato presuppone che la relazione tra  $X$  e  $Y$  rimanga costante anche al di fuori dei valori osservati.
- ✓ La previsione è particolarmente utilizzata nell'analisi di serie storiche quando, ad esempio, si vuole prevedere un aggregato economico in base ai valori del passato osservati per un certo periodo di tempo.

<sup>1</sup>Il termine inglese sarebbe *predict*, diverso dal concetto di previsione espresso dal termine *forecast*. Purtroppo, in lingua italiana entrambi i termini sono tradotti come "previsione".

*Esempio: serie storica esportazioni delle esportazioni di merci (dati in milioni di euro) per l'Italia dal 2001 al 2005:*

Anno	2001	2002	2003	2004	2005
Esportazioni	266434	266561	261898	281348	292011

Al fine di interpolare la serie delle esportazioni rispetto all'anno di riferimento si definisce (per semplicità di calcoli) la variabile indipendente  $X = \text{Anno} - 2000$ , mentre  $Y$  è la serie delle esportazioni.

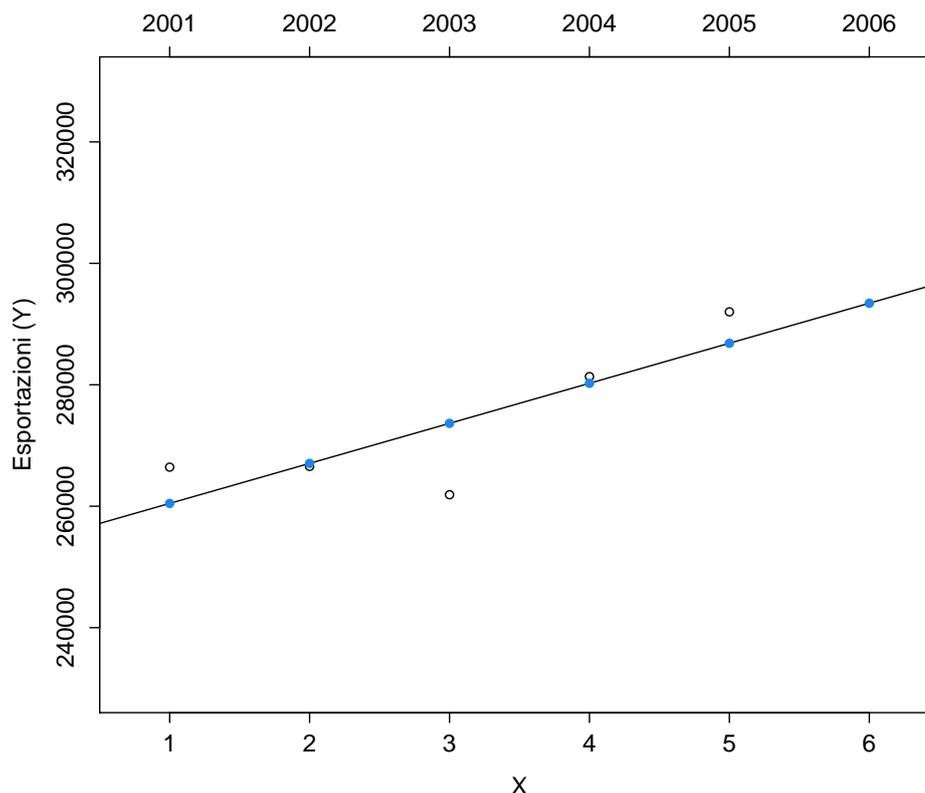
Dai calcoli si ottiene:  $\bar{x} = 3$ ,  $\bar{y} = 273650$ ,  $D_X = 10$ ,  $C_{XY} = 65941$ . Quindi:  $\hat{\beta}_1 = 65941/10 = 6594.1$  e  $\hat{\beta}_0 = 273650 - 6594.1 \times 3 = 253867.7$ , da cui l'equazione della retta

$$y = 253868.7 + 6594.1x$$

Si supponga di voler prevedere il valore per l'anno 2006 sulla base della retta calcolata. Essendo l'anno 2006 al di fuori del range di valori osservati (è un valore futuro rispetto alla serie storica osservata) si tratta di estrapolazione. Tale valore si può calcolare dall'equazione della retta per  $x_* = 2006 - 2000 = 6$ , cioè

$$y_* = 253868.7 + 6594.1 \times 6 = 293433.3$$

La serie storica delle esportazioni e la corrispondente retta interpolatrice, nonché il valore previsto per l'anno 2006 sono riportati nel grafico seguente.



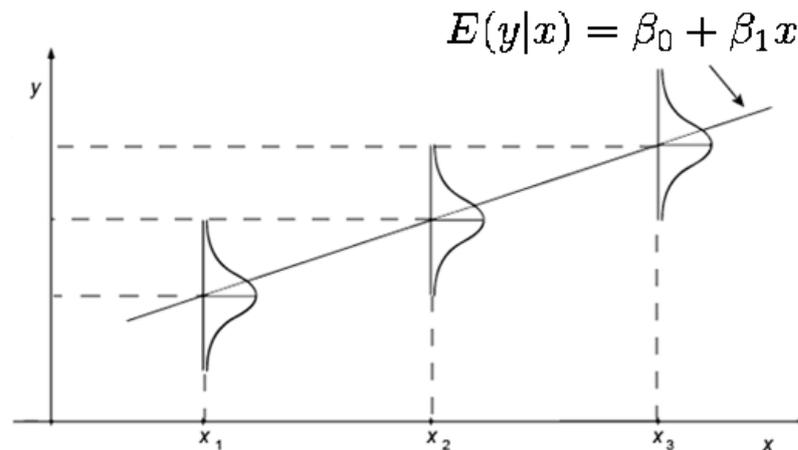
## Inferenza per il modello di regressione

- ✓ Nell'approccio inferenziale al modello di regressione si ipotizza l'esistenza di una **componente stocastica di errore** indipendente e Gaussiana:

$$y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

ovvero

$$y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$$



- ✓ Il metodo statistico più utilizzato per la stima di parametri incogniti va sotto il nome di **metodo della massima verosimiglianza**.
- ✓ Le stime di massima verosimiglianza (MV) che si ottengono sono equivalenti a quelle ottenute con il metodo dei minimi quadrati (che non assume  $\epsilon \sim N$ ).

- ✓ Inoltre,  $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}}$ .

- ✓ Lo stimatore di MV è il migliore tra gli stimatori lineari non distorti (BLUE = *best linear unbiased estimator*).
- ✓ Lo stimatore di MV è non distorto, asintoticamente con la minore varianza e con distribuzione Gaussiana.
- ✓ Quest'ultima proprietà consente di ottenere facilmente le procedure inferenziali classiche (intervalli di confidenza, verifica di ipotesi sui coefficienti, intervalli di previsione).

## Intervalli di confidenza

- ✓ Intervallo di confidenza al livello 95% per  $\beta_1$  ha limiti pari a

$$\hat{\beta}_1 \pm t_{0.05/2;n-2} \text{se}(\hat{\beta}_1)$$

dove  $\text{se}(\hat{\beta}_1) = \hat{\sigma} / \sqrt{D_X}$  e  $t_{0.05/2;n-2}$  è il quantile  $(1 - 0.05/2)$  della  $t$  di Student con  $n - 2$  gradi di libertà.

- ✓ Se  $n$  è sufficientemente grande,  $t_{0.05/2;n-2} = 1.96$ .
- ✓ Interpretazione: ipotizzando un campionamento ripetuto sotto le medesime condizioni, con probabilità del 95% il valore (fisso) del parametro incognito  $\beta_1$  sarà contenuto all'interno dei limiti (variabili) di confidenza calcolati.

## Verifica d'ipotesi

- ✓ Si consideri il sistema d'ipotesi  $H_0 : \beta_1 = 0$  vs  $H_1 : \beta_1 \neq 0$ .
- ✓ La statistica

$$t = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$$

deve essere confrontata con la distribuzione  $t$  di Student con  $n - 2$  gradi di libertà,  $t_{\alpha/2;n-2}$  per un prefissato  $\alpha$ .

- ✓ Per un dato livello di significatività  $\alpha$  (ad es. 0.05, 0.01, ecc.)
  - se  $|t| > t_{\alpha/2;n-2}$  si rifiuta l'ipotesi nulla  $H_0$ ,
  - altrimenti si accetta l'ipotesi nulla  $H_0$  e quindi il coefficiente angolare non è statisticamente differente da zero.

Analogamente,  $p$ -value maggiore del livello  $\alpha$  stabilito indica la non significatività del parametro  $\beta_1$  (*metodo del livello di significatività osservato*).

Call:

```
lm(formula = Y ~ X, data = data)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-19.8247	-3.4700	0.1237	4.1500	15.0531

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	60.81149	7.13963	8.517	<2e-16 ***
X	0.64629	0.04114	15.711	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.686 on 926 degrees of freedom

Multiple R-squared: 0.2105, Adjusted R-squared: 0.2096

F-statistic: 246.8 on 1 and 926 DF, p-value: < 2.2e-16

Confidence intervals 95%:

	2.5 %	97.5 %
(Intercept)	46.7997529	74.8232204
X	0.5655602	0.7270209

## Intervallo di previsione

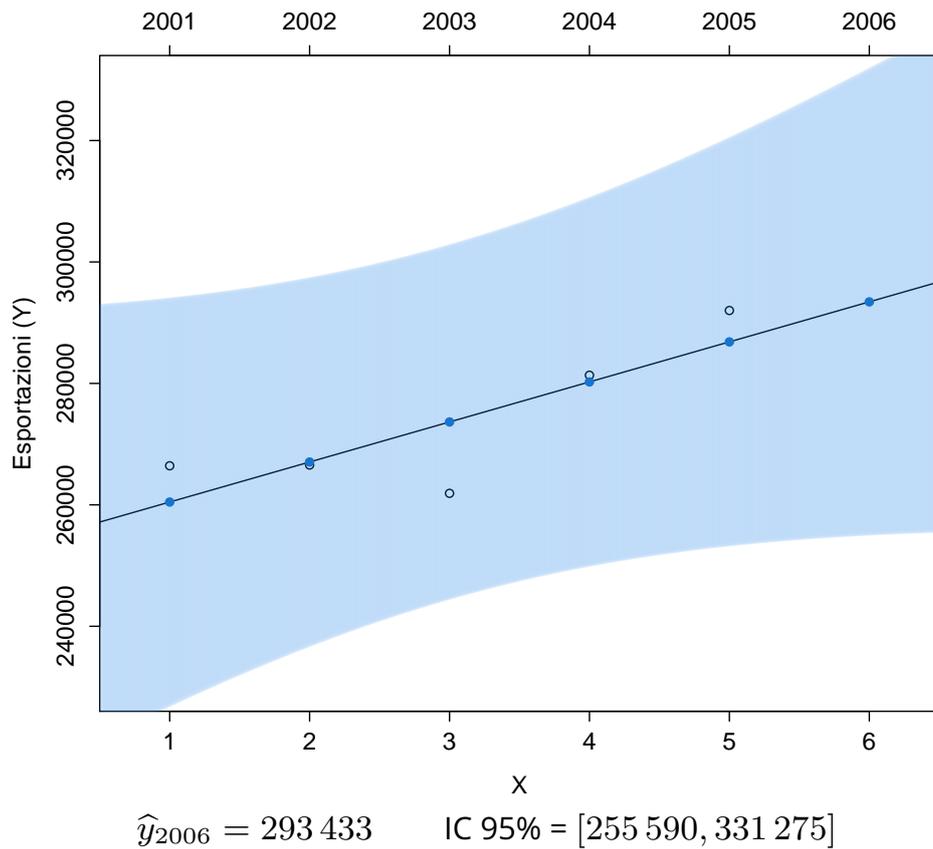
- ✓ Il valore previsto della variabile risposta per un dato valore  $x$  è dato dall'equazione

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- ✓ Assumendo che gli errori siano distribuiti normalmente, un intervallo di previsione approssimato al 95% è dato dalla formula

$$\hat{y} \pm t_{0.05/2; n-2} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{D_X}}$$

- ✓ Intervalli per diversi valori del livello di confidenza si ottengono sostituendo il valore  $t_{0.05/2; n-2}$  con l'opportuno quantile della v.c.  $t$  di Student con  $n - 2$  gradi di libertà (es., per un livello del 90% e  $n$  sufficientemente grande si utilizza il valore 1.645).
- ✓ Dalla formula per l'intervallo di previsione si può vedere che questi risultano più ampi per valori di  $x$  distanti dalla media.



## Shiny App interattiva



About R  
[What is R?](#)  
[Contributors](#)  
[Screenshots](#)  
[What's new?](#)

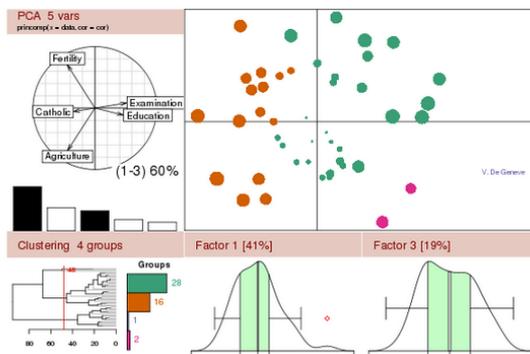
Download, Packages  
[CRAN](#)

R Project  
[Foundation](#)  
[Members & Donors](#)  
[Mailing Lists](#)  
[Bug Tracking](#)  
[Developer Page](#)  
[Conferences](#)  
[Search](#)

Documentation  
[Manuals](#)  
[FAQs](#)  
[The R Journal](#)  
[Wiki](#)  
[Books](#)  
[Certification](#)  
[Other](#)

Misc  
[Bioconductor](#)  
[Related Projects](#)  
[User Groups](#)  
[Links](#)

### The R Project for Statistical Computing



#### Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).
- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

#### News:

- **R version 3.1.2** (Pumpkin Helmet) has been released on 2014-10-31.
- **The R Journal Volume 6/1** is available.
- **useR! 2014**, took place at the University of California, Los Angeles, USA June 30 - July 3, 2014.
- **R version 3.0.3** (Warm Puppy) has been released on 2014-03-06.
- **useR! 2015**, will take place at the University of Aalborg, Denmark, June 30 - July 3, 2015.



Welcome to RStudio - Open source and enterprise-ready professional software for R

[Download RStudio](#) [Discover Shiny](#)



**Powerful IDE for R**

RStudio IDE is a powerful and productive user interface for R. It's free and open source, and works great on Windows, Mac, and Linux.

[Learn More >](#)

**R Packages**

Our developers and expert trainers are the authors of several popular R packages, including ggplot2, plyr, lubridate, and others.

[Learn More >](#)

**Bring R to the web**

Shiny is an elegant and powerful web framework for building interactive reports and visualizations using R — with or without web development skills.

[Learn More >](#)

# Shiny

by RStudio

**A web application framework for R**

Turn your analyses into interactive web applications  
No HTML, CSS, or JavaScript knowledge required

[TUTORIAL](#) [ARTICLES](#) [GALLERY](#) [REFERENCE](#) [DEPLOY](#) [HELP](#)



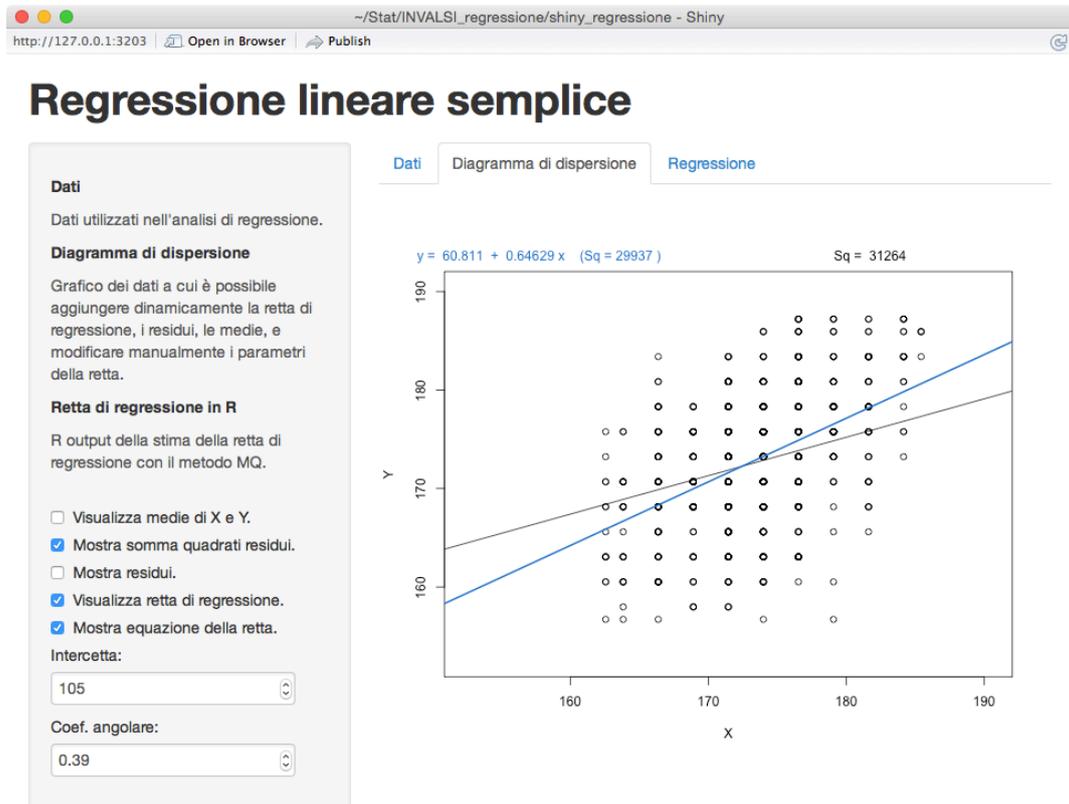

**Get inspired**  
(gallery)



**Get started**  
(tutorial)



**Go deeper**  
(articles)



## Estensioni al modello di regressione lineare

- ✓ Modello di regressione lineare multiplo

$$E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

- regressione polinomiale (quadratica, cubica, ecc.)
- variabili qualitative (fattori)
- modello di analisi della varianza
- modello di analisi della covarianza

- ✓ Modelli semi-parametrici

$$E(y|x) = m(x)$$

dove  $m(\cdot)$  è una funzione non parametrica stimata a partire dai dati (es. splines, lowess, kernel regression, ecc.)

✓ Modelli additivi

$$E(y|\mathbf{x}) = \beta_0 + m_1(x_1) + m_2(x_2) + \dots + m_p(x_p)$$

dove  $m_j(x_j)$  sono funzioni non parametriche stimate per ciascuna variabile  $x_j$  ( $j = 1, \dots, p$ ), oppure possono essere fissate opportunamente (es. funzione identità, polinomiale, ecc.).

✓ Modelli non lineari (nei parametri)

*Esempio: Modello logistico per la crescita della popolazione:*

$$E(y|x) = \frac{\beta_1}{1 + \exp\{\beta_2 + \beta_3 x\}}$$

dove  $x$  rappresenta il tempo,  $\beta_1$  è il limite superiore per la crescita della popolazione,  $\beta_2$  è la dimensione della popolazione al tempo 0, e  $\beta_3 (< 0)$  è il tasso di crescita della popolazione.

✓ Modelli lineari generalizzati (modello logistico per dati binari, modello di Poisson per dati conteggio, modello Gamma per dati di durata, ecc.)