

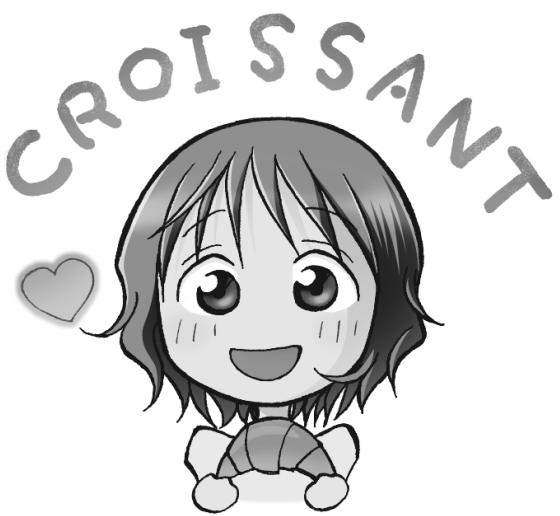
I MANGA DELLE SCIENZE

REGRESSIONE

ANALISI STATISTICA DEI DATI

SHIN TAKAHASHI
IROHA INOUE
TREND-PRO CO., LTD.





I MANGA DELLE SCIENZE

REGRESSIONE

ANALISI STATISTICA DEI DATI

SHIN TAKAHASHI
IROHA INOUE
TREND-PRO CO., LTD.



SOMMARIO

PREFAZIONE	IX
PROLOGO	1
1	
UN BEL BICCHIERE DI MATEMATICA	11
Gettiamo le fondamenta	12
Funzioni inverse	14
Esponenti e logaritmi	19
Regole per gli esponenti e i logaritmi	21
Calcolo differenziale	24
Matrici	37
Addizione di matrici	39
Moltiplicazione di matrici	40
Regole della moltiplicazione matriciale	43
Matrice identità e matrici inverse	44
Tipi di dati statistici	46
Verifica d'ipotesi	48
Misura della variabilità	49
Somma dei quadrati degli scarti	50
Varianza	50
Deviazione standard	51
Funzioni di densità di probabilità	52
Distribuzioni normali	53
Distribuzioni chi-quadro	54
Densità di probabilità: tavole di distribuzione	55
Distribuzioni F	57
2	
ANALISI DI REGRESSIONE SEMPLICE	61
Primi passi	62
Grafico dei dati	64
Equazione di regressione	66
Procedimento generale per l'analisi di regressione	68
Passo 1: tracciare il grafico di dispersione della variabile dipendente in funzione di quella indipendente. Se i punti si allineano, potrebbe esserci correlazione tra le variabili	69
Passo 2: calcolo dell'equazione di regressione	71

Passo 3: calcolo del coefficiente di correlazione (R) e valutazione della popolazione e delle ipotesi.....	78
Campioni e popolazioni	82
L'ipotesi di normalità.....	85
Passo 4: analisi della varianza	87
Passo 5: calcolo degli intervalli di confidenza.....	91
Passo 6: facciamo una previsione!.....	95
Quali sono i passi necessari?	100
Residui standardizzati.....	100
Interpolazione ed estrapolazione	102
Autocorrelazione	102
Regressione non lineare	103
Trasformazione delle equazioni non lineari in equazioni lineari.....	104
 3	
ANALISI DI REGRESSIONE MULTIPLA	107
Previsioni con più di una variabile	108
Equazione di una regressione multipla	112
Procedimento per l'analisi di regressione multipla.....	112
Passo 1: tracciare il grafico di dispersione di tutte le variabili predittive in funzione della variabile responso, per vedere se sembrano correlate	113
Passo 2: calcolo dell'equazione di regressione multipla.....	115
Passo 3: valutazione dell'accuratezza dell'equazione di regressione multipla	119
Il problema di R^2	122
R^2 corretto	124
Verifica d'ipotesi con la regressione multipla.....	127
Passo 4: test di analisi della varianza (ANOVA).....	128
Determinazione di S_{11} e S_{22}	132
Passo 5: calcolo degli intervalli di confidenza per la popolazione	133
Passo 6: facciamo una previsione!.....	136
Scelta della migliore combinazione di variabili predittive	138
Stima di popolazioni con l'analisi di regressione multipla	142
Residui standardizzati.....	143
Distanza di Mahalanobis	144
Passo 1	144
Passo 2	145
Passo 3	146
Uso dei dati categorici nell'analisi di regressione multipla	147
Multicollinearità	149
Determinazione dell'influenza relativa delle variabili predittive sulla variabile responso	149
 4	
ANALISI DI REGRESSIONE LOGISTICA.....	153
L'ultima lezione	154

Il metodo della massima verosimiglianza	160
Trovare la massima verosimiglianza con la funzione di verosimiglianza	163
Scegliere le variabili predittive	164
Analisi di regressione logistica in azione!	168
Procedura per l'analisi di regressione logistica.	168
Passo 1: tracciare un grafico di dispersione delle variabili predittive e della variabile risposta per vedere se sembrano correlate	169
Passo 2: calcolare l'equazione di regressione logistica	170
Passo 3: valutare l'accuratezza dell'equazione	173
Passo 4: svolgere il test d'ipotesi	178
Passo 5: prevediamo se verrà venduta la Norns Special	182
La regressione logistica nel mondo reale	190
Logit, odds ratio e rischio relativo	190
Logit	190
Odds ratio	191
Odds ratio aggiustato	192
Verifica d'ipotesi con gli odds	194
Intervallo di confidenza per l'odds ratio	194
Rischio relativo	195
APPENDICE	
CALCOLI DI REGRESSIONE CON UN FOGLIO ELETTRONICO	197
La costante di Eulero	198
Potenze	200
Logaritmi naturali	200
Prodotto di matrici	201
Matrici inverse	202
Calcolare una statistica chi-quadro da un <i>p</i> -value	204
Calcolare un <i>p</i> -value da una statistica chi-quadro	205
Calcolare una statistica F da un <i>p</i> -value	206
Calcolare un <i>p</i> -value da una statistica F	208
Coefficiente di regressione parziale di un'analisi di regressione multipla	209
Coefficiente di regressione di un'equazione di regressione logistica	210
INDICE	213

PREFAZIONE

Questo libro è un'introduzione all'analisi di regressione semplice, multipla e logistica.

L'analisi di regressione semplice e quella multipla sono metodi statistici per la previsione di valori; per esempio, con l'analisi di regressione semplice possiamo prevedere quante saranno le ordinazioni di tè freddo in base alla temperatura massima giornaliera e, con l'analisi di regressione multipla, le vendite mensili di un negozio in base alle sue dimensioni e alla distanza dalla stazione ferroviaria più vicina.

Con l'analisi di regressione logistica si prevedono probabilità, come quella di vendere un certo dolce in un dato giorno della settimana.

Questo libro è destinato agli studenti di matematica e statistica che hanno difficoltà a capire l'analisi di regressione, o a chiunque sia interessato a un'infarinatura di probabilità e previsioni statistiche.

Prima di iniziare occorrono alcune nozioni basilari, reperibili per esempio nel volume "Statistica" di questa stessa serie ("I Manga delle Scienze").

Il libro è così suddiviso:

- Capitolo 1: un bel bicchiere di matematica
- Capitolo 2: analisi di regressione semplice
- Capitolo 3: analisi di regressione multipla
- Capitolo 4: analisi di regressione logistica

Ogni capitolo si compone di alcune pagine di manga e una sezione di testo, un po' più tecnica. Il manga fornisce una panoramica dell'argomento, mentre nella sezione di testo si trovano ulteriori dettagli e definizioni utili.

Vorrei spendere qualche parola sul primo capitolo. Alcuni argomenti, come la derivazione e le operazioni con le matrici, saranno forse già noti a molti lettori, ma il capitolo li ripresenta nel contesto dell'analisi di regressione, per facilitare il seguito della lettura. Se per voi è solo un ripasso, benissimo. Se sono cose che avete studiato molto tempo fa, o forse mai, sarebbe consigliabile fare lo sforzo di capirle bene prima di andare avanti.

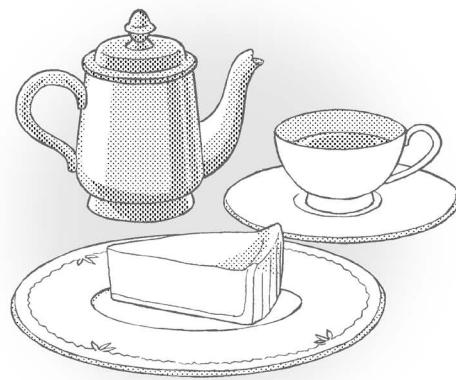
In questo libro i calcoli sono svolti in dettaglio. Chi è bravo in matematica dovrebbe riuscire a seguirli e capirli fino in fondo, gli altri avranno comunque una panoramica del metodo e potranno arrivare alla soluzione grazie alle istruzioni passo passo. Non c'è bisogno di intestardirsi a capire subito tutti i passaggi matematici. Leggete senza stressarvi, ma non saltate a piè pari lo svolgimento dei calcoli.

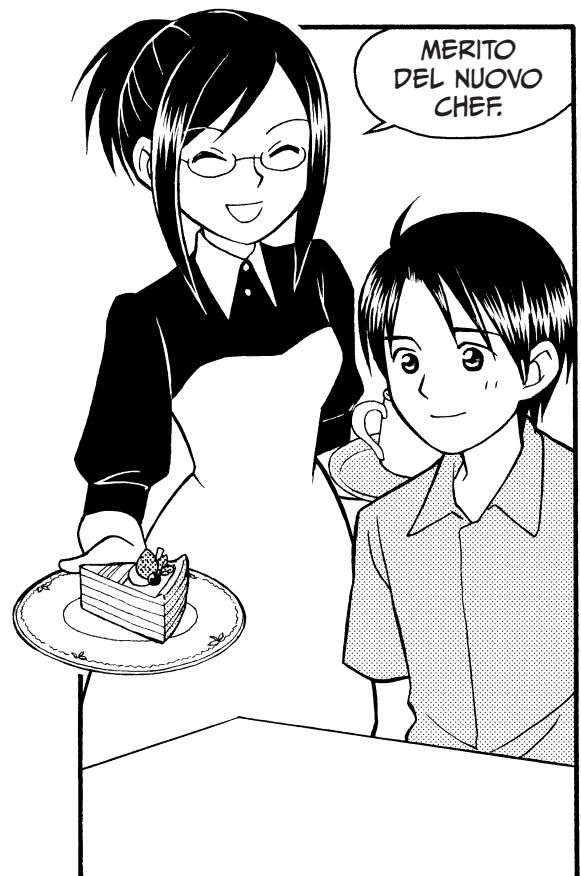
Abbiamo arrotondato alcune cifre per renderle più leggibili: rifacendo i calcoli autonomamente, quindi, potreste trovare valori diversi, anche se non di molto. Ci auguriamo non sia un problema.

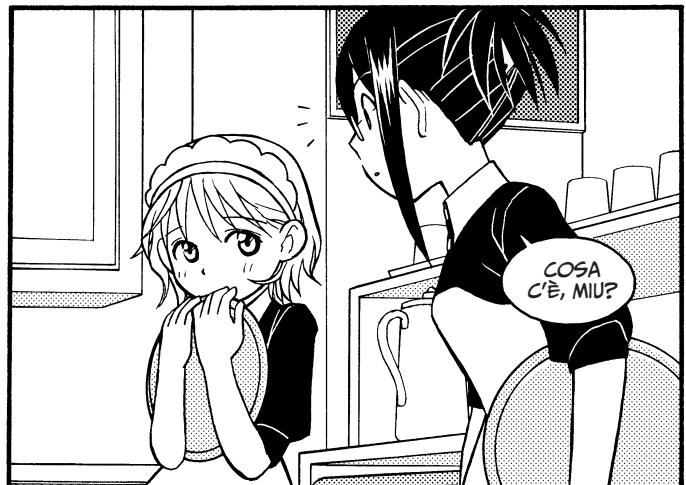
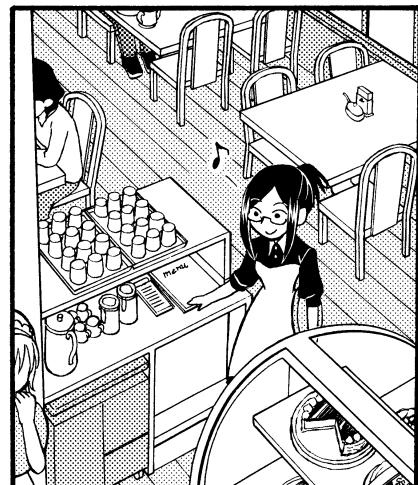
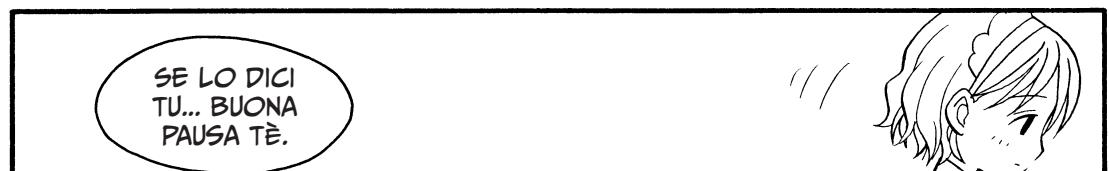
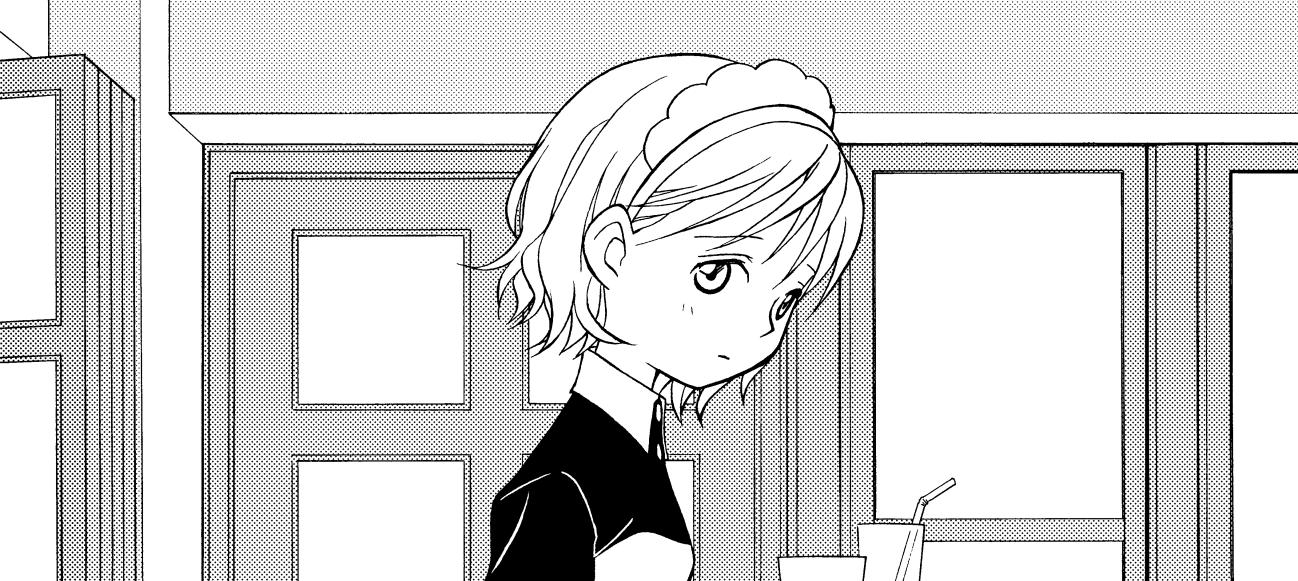
Vorrei ringraziare il mio editore, Ohmsha, per avermi dato l'opportunità di scrivere questo libro e la TREND-PRO per aver trasformato il mio manoscritto in un manga, grazie allo sceneggiatore re_akino e al disegnatore Iroha Inoue. Da ultimo, ma non meno importante, vorrei ringraziare il dottor Sakaori Fumitake del College of Social Relations, Rikkyo University, per i suoi consigli preziosissimi, ancor più numerosi che per il mio libro precedente. Vorrei esprimere anche a lui la mia profonda gratitudine.

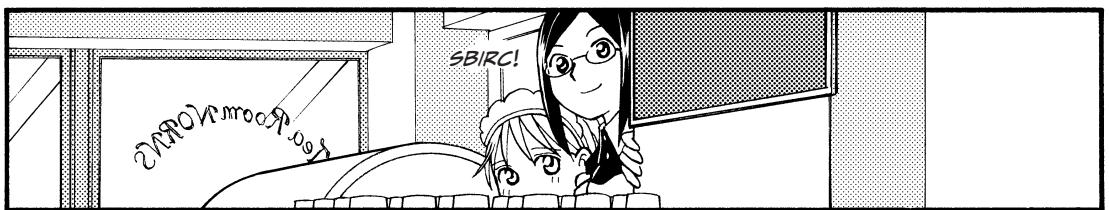
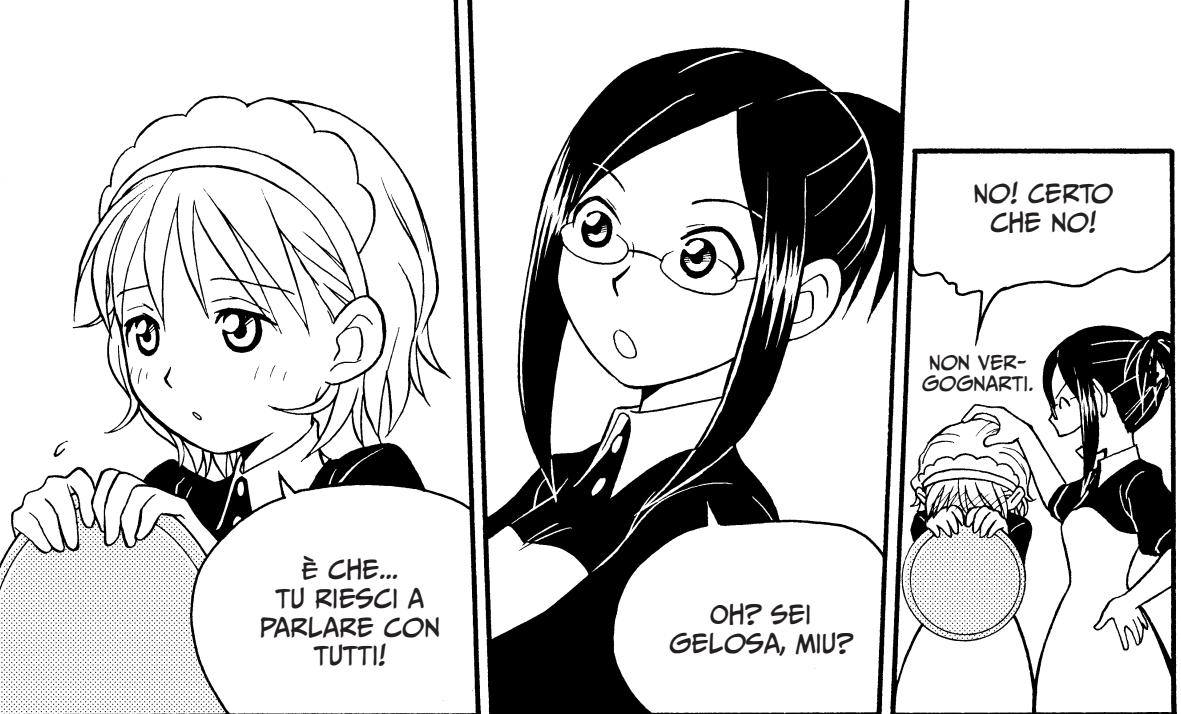
SHIN TAKAHASHI

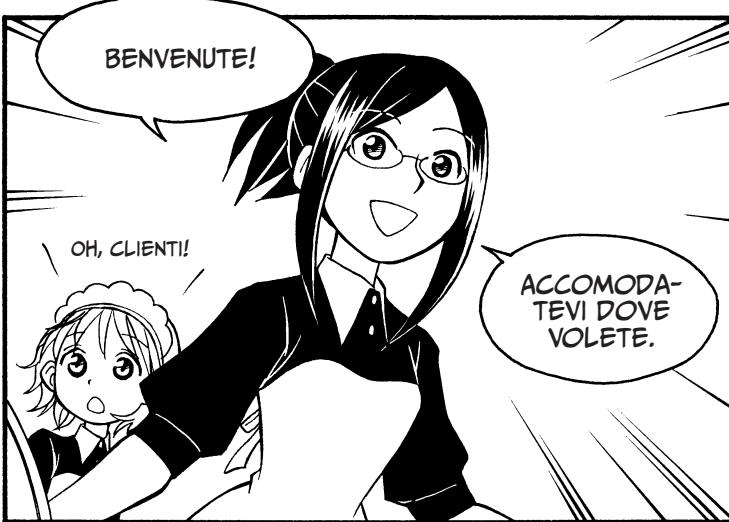
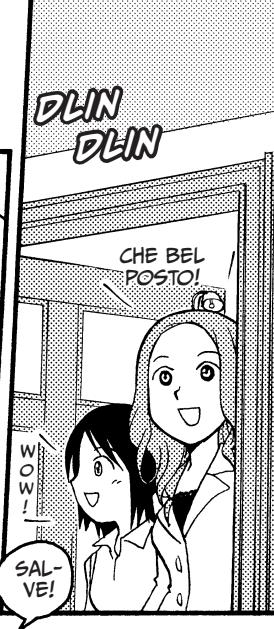
PROLOGO
ANCORA UN PO' DI TÈ?

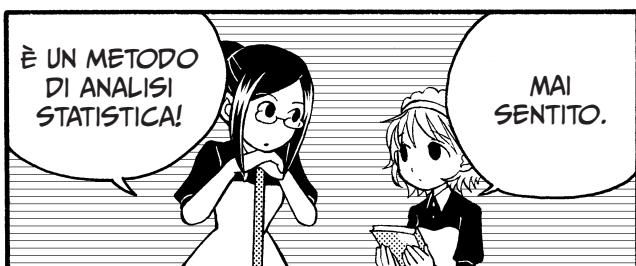
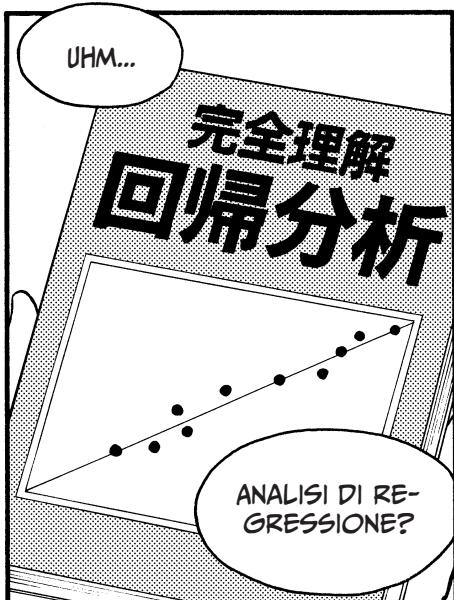


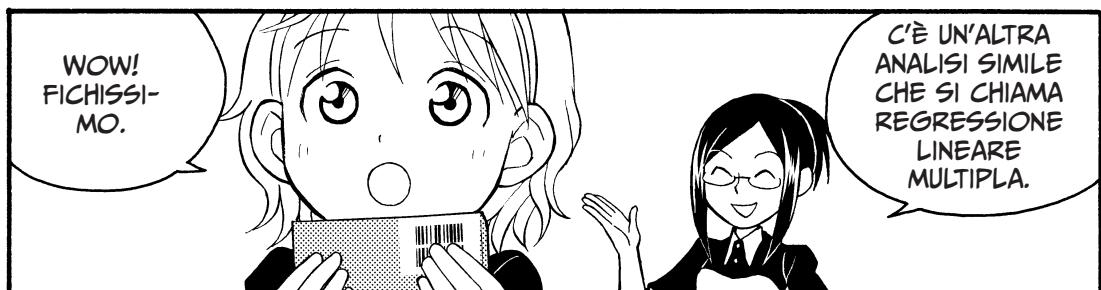
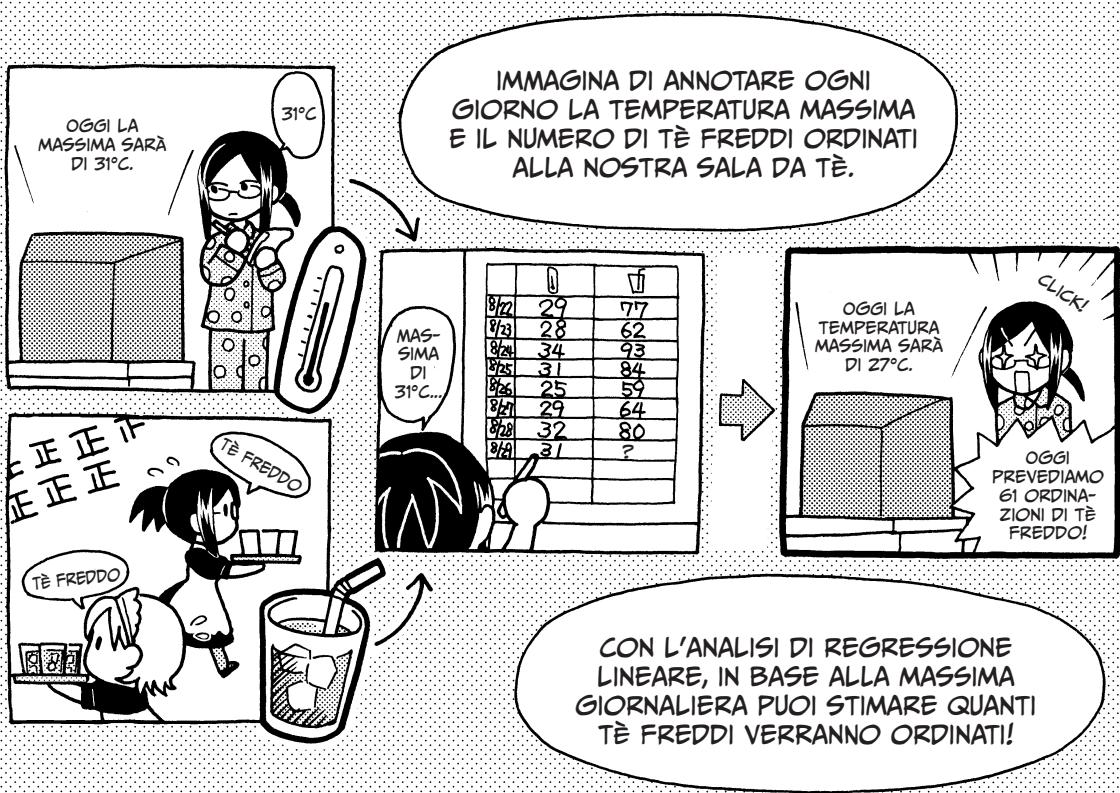












FACCIA
MOS UN
ESEMPIO DI ANALISI DI
REGRESSIONE LINEARE
MULTIPLA.

TIZIO È IL DIRETTORE DI UNA
CATENA DI NEGOZI. OLTRE
A MONITORARE LE VENDITE,
RACCOLGIE I SEGUENTI DATI
PER CIASCUN NEGOZIO:

- DISTANZA DAL NEGOZIO
CONCORRENTE PIÙ VICINO;
- NUMERO DI CASE NEL
RAGGIO DI UN CHILOMETRO;
- SPESSE DI PUBBLICITÀ.

Negozi	Distanza dal negozi concorrente più vicino (m)	Numero di case nel raggio di un chilometro	Spese di pubblicità (¥)	Vendite (¥)
A	○○○	○○○	○○○	○○○
B	△△△	△△△	△△△	△△△
C	□□□	□□□	□□□	□□□
	:	:	:	:



VALUTANDO
L'APERTURA DI UN
NUOVO NEGOZIO...



...PUÒ STIMARNE LE
VENDITE IN BASE
ALL'INFLUENZA DEI
VARI FATTORI SULLE
VENDITE DEI NEGOZI
ESISTENTI.



CI SONO ANCHE ALTRI
METODI, COME L'ANALISI
DI REGRESSIONE
LOGISTICA.

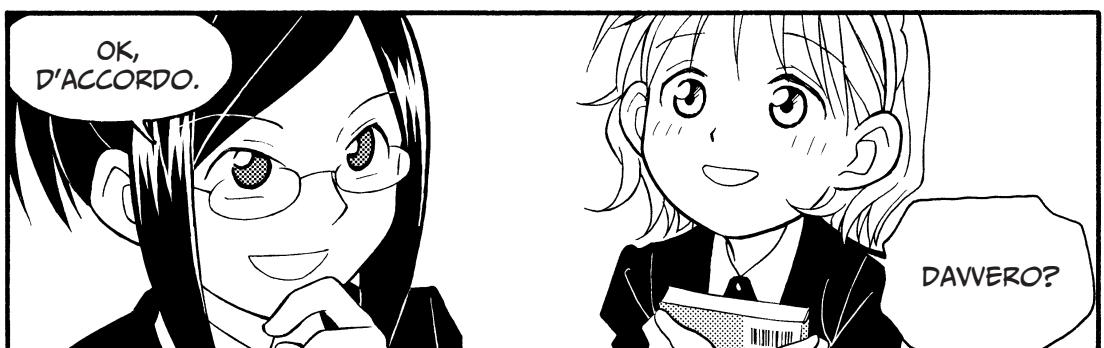
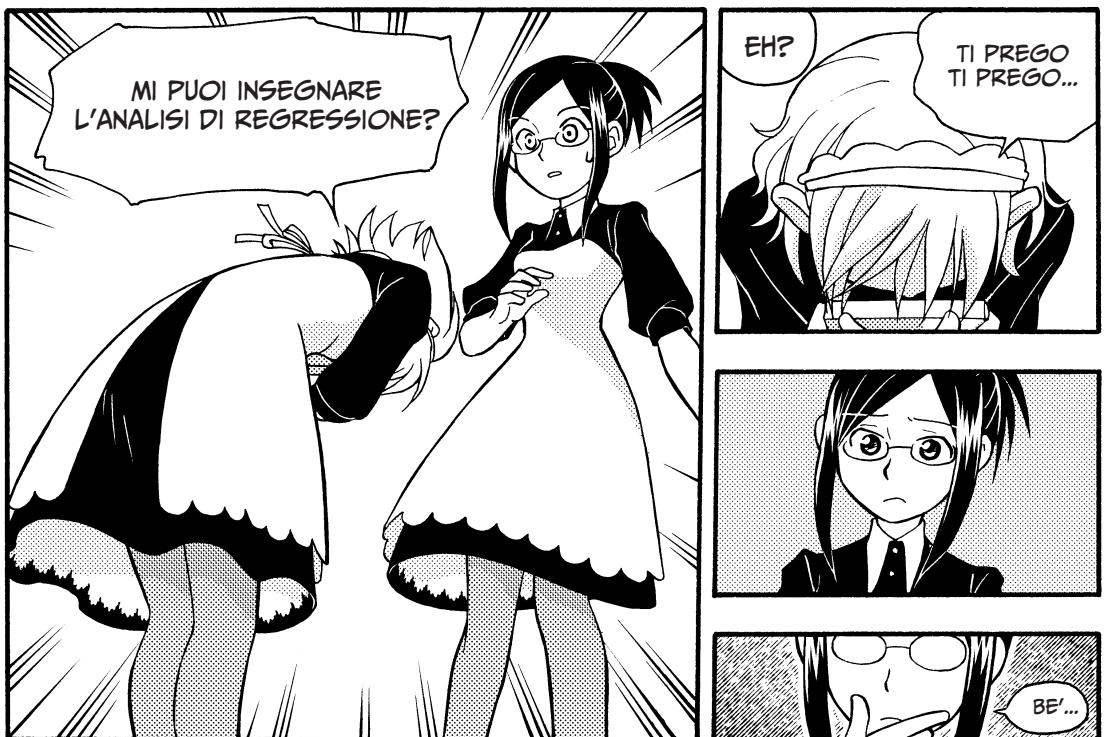


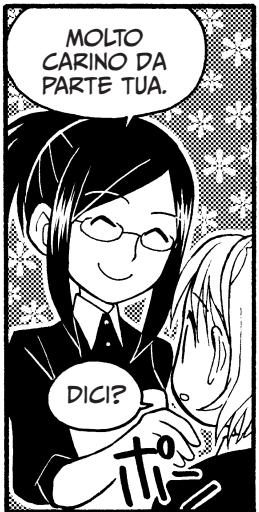
SE
STUDIO
QUESTO
LIBRO...



FORSE...







BE', EHM...
PENSO CHE LO CONSERVERÒ CON CURA FINCHÉ NON TORNA.

DICI?

SONO SICURA CHE GLI VERRÀ IN MENTE.

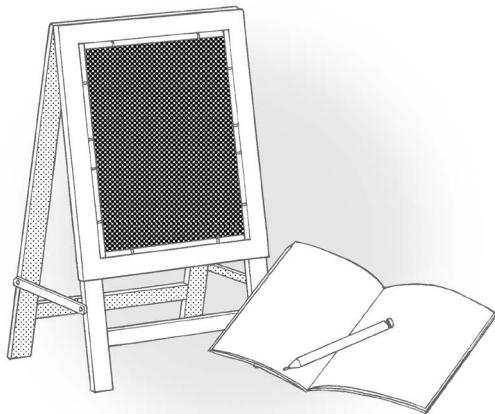
ALLORA A DOMANI!

OK... BUONANOTTE...



1

UN BEL BICCHIERE
DI MATEMATICA



GETTIAMO LE FONDAMENTA

FINALMENTE IL
CAPO È ANDATO VIA,
ANDIAMOCENE
ANCHE NOI!

OUF!

EHM, RISA,
POTREMMO
INIZIARE
LE LEZIONI
STASERA?

SUL
SERIO?
VUOI CO-
MINCIARE
ADESSO?!?

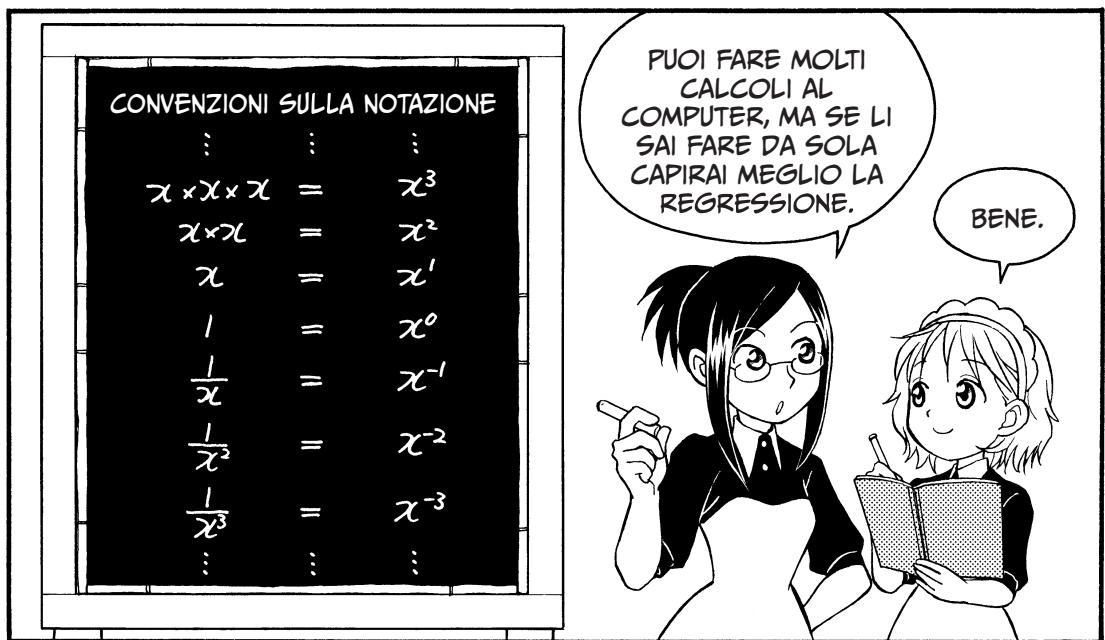
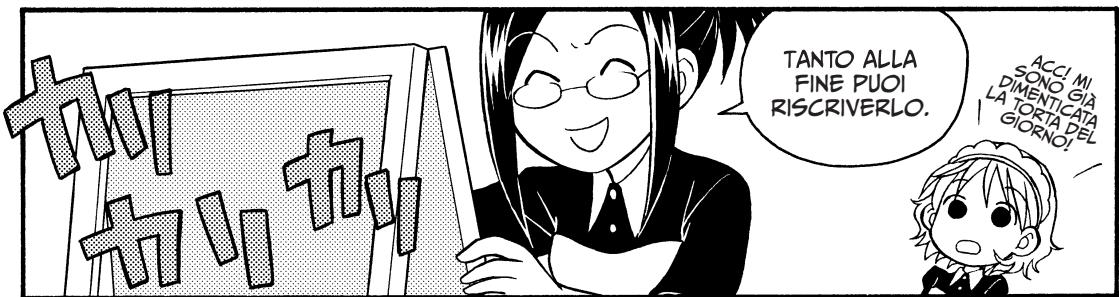
SISSI

NON TI
FACEVO COSÌ
ANSIOSA DI IMPA-
RARE! IN GENERE
A LEZIONE
DORMI.

LO SO,
È SOLO
CHE...

SCUSA...

NON
VOLEVO
CRITICARTI!



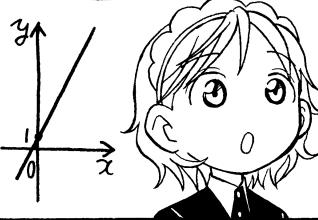
FUNZIONI INVERSE

$$y = 2x + 1$$

PRIMA DI TUTTO TI SPIEGHERÒ LE FUNZIONI INVERSE, FACENDO L'ESEMPIO DELLA FUNZIONE LINEARE $y = 2x + 1$.

SE x È ZERO,
QUANTO VALE
 y ?

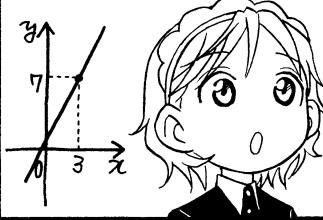
$$\begin{aligned} y &= 2x + 1 \\ &= 2 \times 0 + 1 \\ &= 0 + 1 \\ &= 1 \end{aligned}$$



VALE 1.

E SE x È 3?

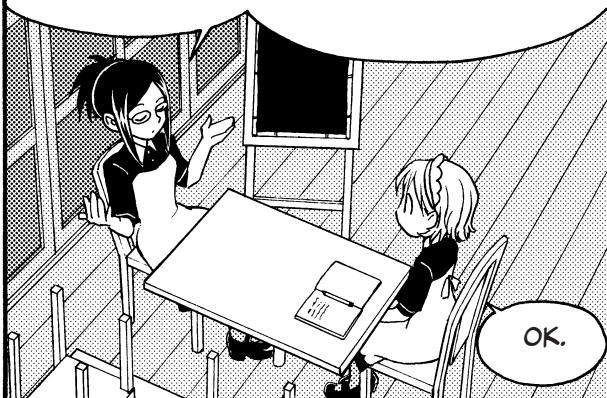
$$\begin{aligned} y &= 2x + 1 \\ &= 2 \times 3 + 1 \\ &= 6 + 1 \\ &= 7 \end{aligned}$$



y VALE 7.

IL VALORE DI
 y DIPENDE
DA QUELLO
DI x .

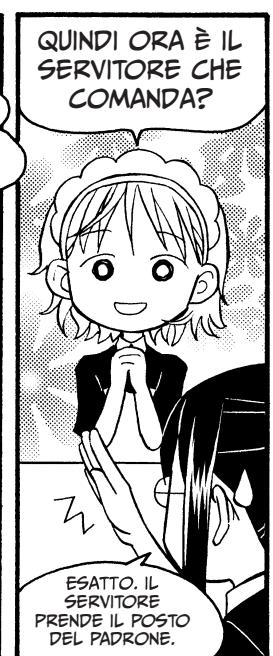
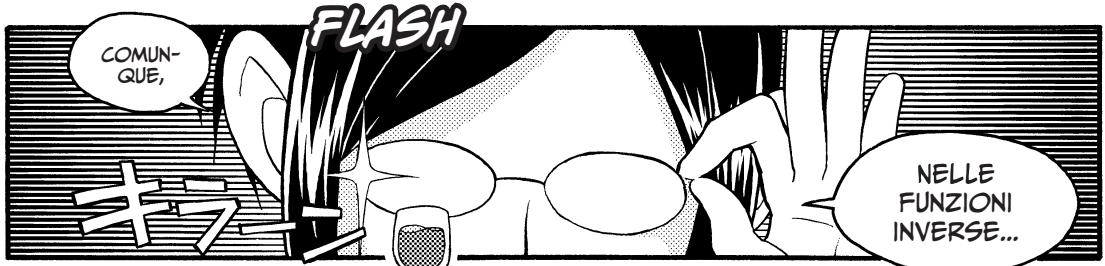
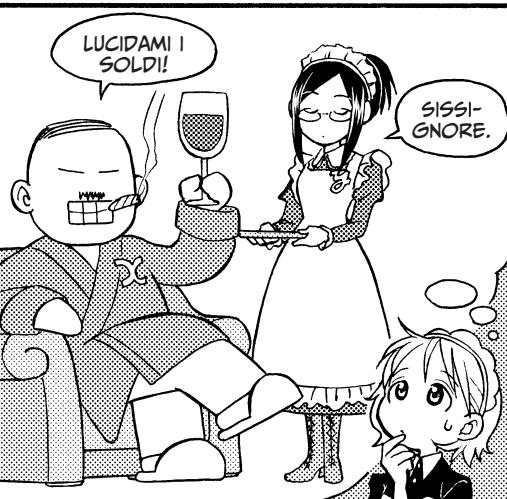
QUINDI y SI CHIAMA
VARIABILE RESPONSO O
VARIABILE DIPENDENTE,
E x È IL PREDITTORE O
VARIABILE INDIPENDENTE.



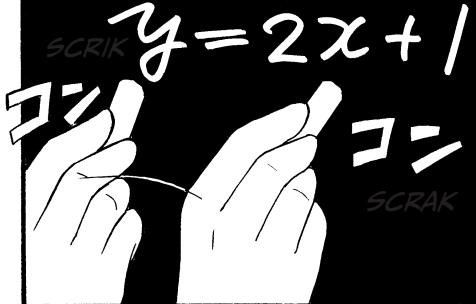
OK.

È COME SE x FOSSE IL
PADRONE DI y .





NELL'ESEMPIO DI
 $y = 2x + 1$, LA FUNZIONE
INVERSA È ALLORA...



$$\begin{array}{c} y = 2x + 1 \\ \downarrow \qquad \downarrow \\ x = 2y + 1 \end{array}$$

...QUELLA IN CUI y
E x SI SCAMBIANO
DI POSTO.

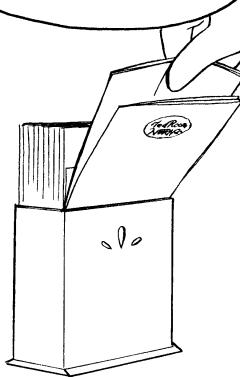


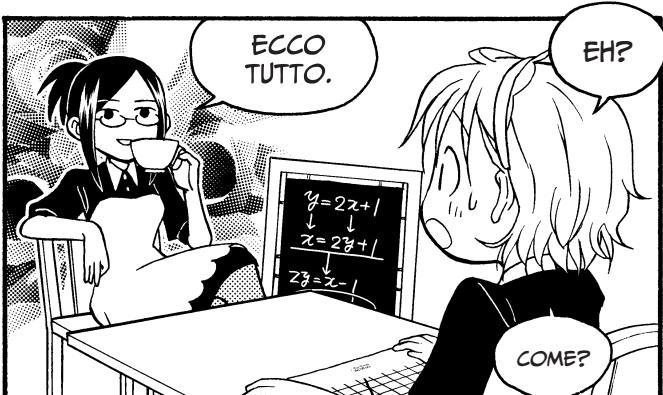
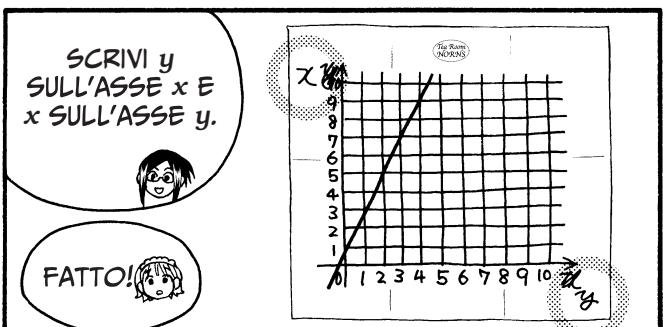
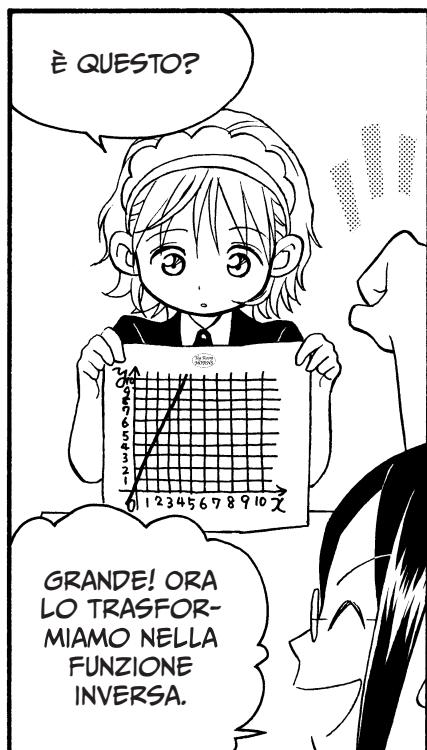
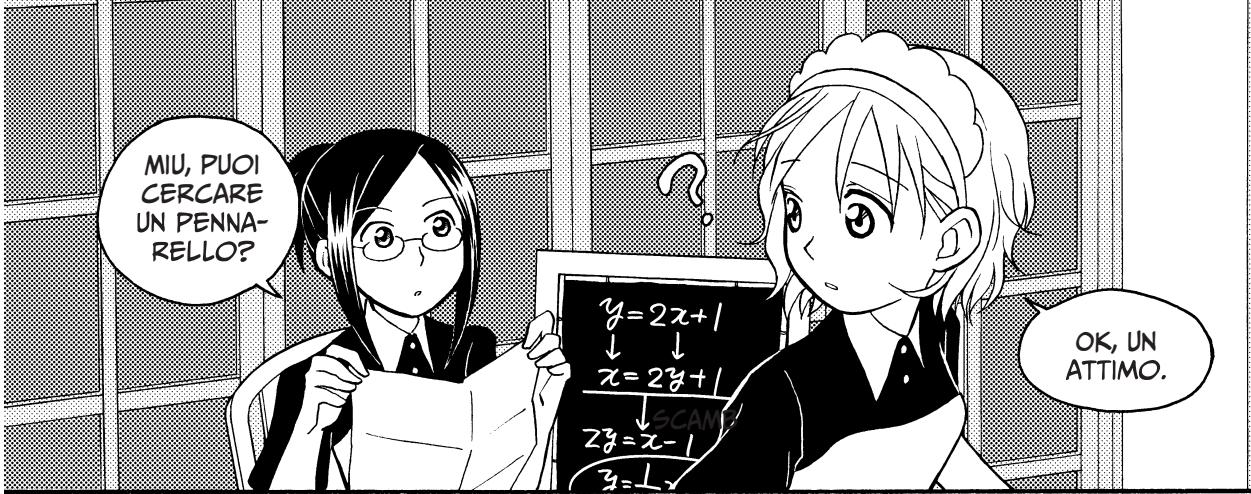
$$\begin{array}{c} \downarrow \qquad \downarrow \\ x = 2y + 1 \\ \hline \text{SCAMBIO} \\ \downarrow \\ 2y = x - 1 \\ \hline y = \frac{1}{2}x - \frac{1}{2} \end{array}$$

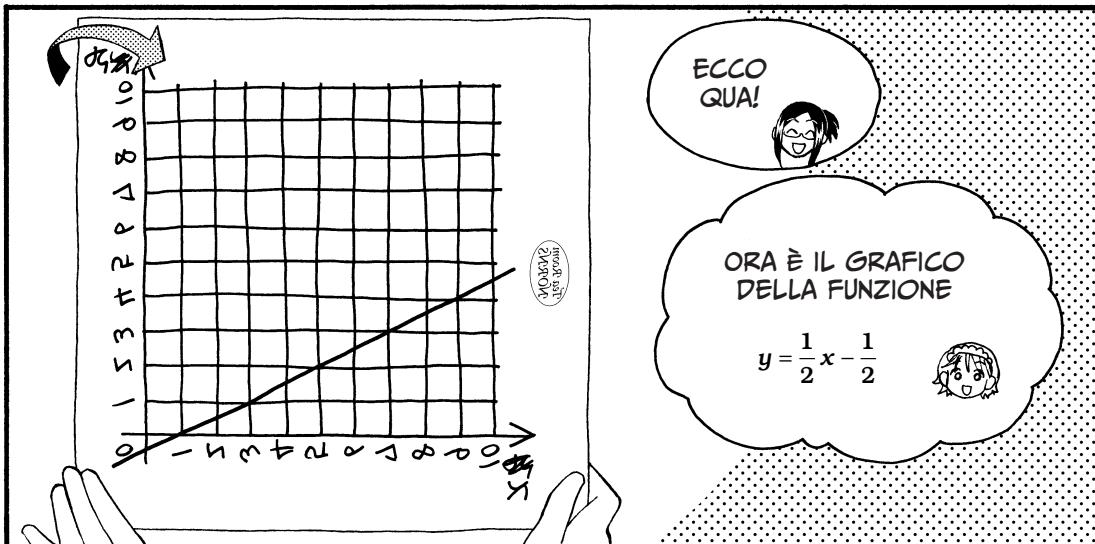
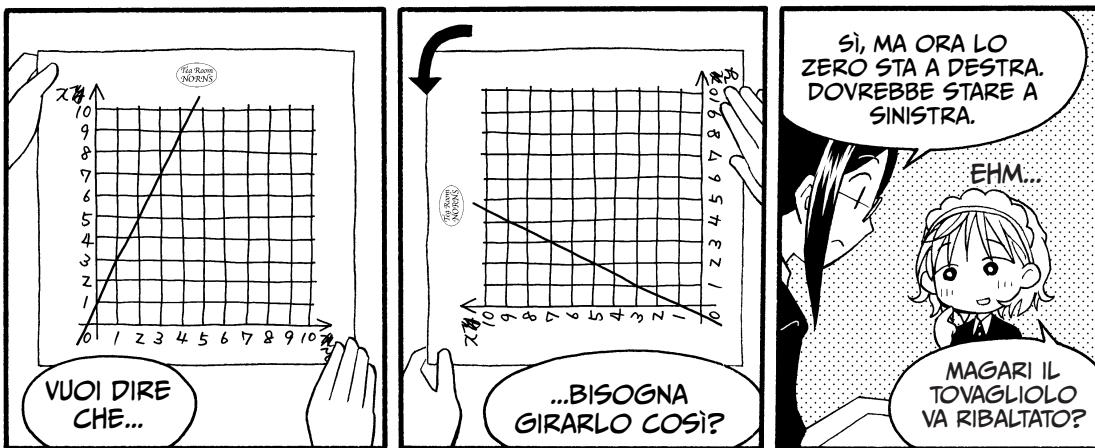
RISCRIVIAMO
LA FUNZIONE
IN QUESTO
MODO.

HAI PORTATO LA x
DALL'ALTRA PARTE
E HAI DIVISO PER
2, COSÌ ORA y
STA DA SOLO.

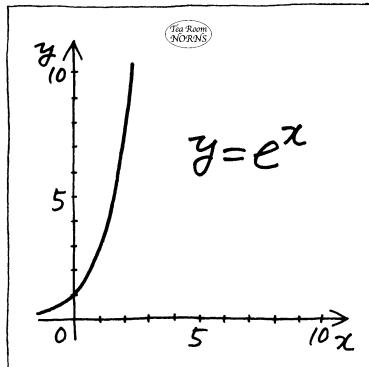
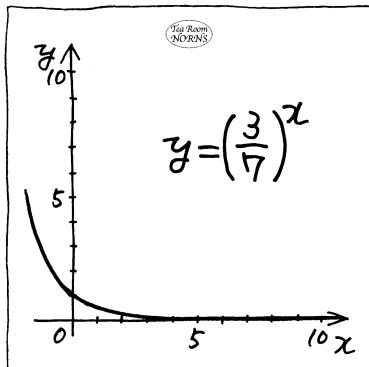
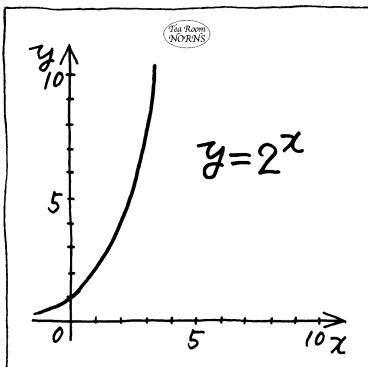
GIUSTO. PER CAPIRE
A CHE SERVE,
DISEGNIAMO IL
GRAFICO.







ESPOVENTI E LOGARITMI



OK...
PASSIAMO
AL PROSSIMO
ARGOMENTO. QUESTE
SI CHIAMANO FUNZIONI
ESPOVENTIALI.

PASSANO TUTTE
PER IL PUNTO (0, 1),
PERCHÉ QUAISIASI
NUMERO ELEVATO
ALLA POTENZA
ZERO FA 1.



GIUSTO!
ORA, HAI
MAI VISTO
QUESTA e ?

e È LA BASE DEI
LOGARITMI NATURALI
E VALE CIRCA 2,7182.
SI CHIAMA NUMERO
(O COSTANTE)
DI EULERO.

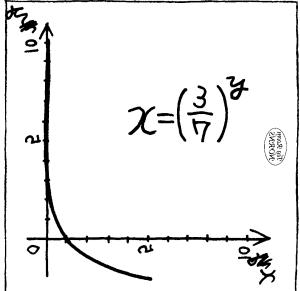
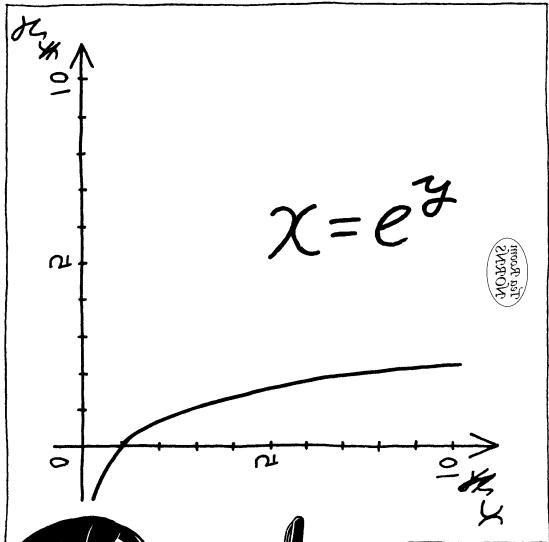
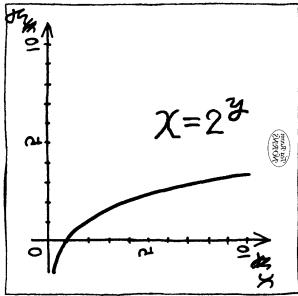
GIÀ SENTITO.

$y = e^x$

$\log_e y = x$
È L'INVERSA
DELL'EQUA-
ZIONE ESPO-
NENZIALE
 $y = e^x$.

AH! DI
NUOVO LE
FUNZIONI
INVERSE!

RIBALTI!



$y = e^x$ È LA FUNZIONE INVERSA DI $y = \log x$ CHE È DETTA FUNZIONE LOGARITMO NATURALE.

UN ALTRO RIBALTO-MENTO!

PER TROVARE L'INVERSA DI $y = e^x$, SCAMBIAVI LE VARIABILI x E y E NE CALCOLIAMI IL LOGARITMO PER ISOLARE y . SEMPLIFICANDO, $\log_e(e^y)$ È SEMPLICEMENTE y !

$$y = e^x$$

LE VARIABILI SI SCAMBIANO DI POSTO!

$$x = e^y$$

ABBIANO RIBALTATO L'EQUAZIONE PER RIPORTARE y A SINISTRA.

$$\Downarrow$$

$$y = \log_e x$$

ORA RIVEDIAMO LE REGOLE PER LE FUNZIONI ESPOENZIALI E LOGARITMICHE.

TIENILE A MENTE, SARANNO UTILI PIÙ AVANTI!



PREndo APPUNTI!

REGOLE PER GLI ESPONENTI E I LOGARITMI

1. REGOLA DEGLI ESPOVENTI

$(e^a)^b$ È UGUALE A $e^{a \times b}$.



Proviamo a fare i conti. Controlliamo che $(e^a)^b$ e $e^{a \times b}$ siano uguali nel caso $a = 2$ e $b = 3$.

$$(e^2)^3 = \underbrace{e^2 \times e^2 \times e^2}_3 = \underbrace{(e \times e) \times (e \times e) \times (e \times e)}_3 = \underbrace{e \times e \times e \times e \times e \times e}_6 = e^{2 \times 3}$$

Ciò significa inoltre che $(e^a)^b = e^{a \times b} = (e^b)^a$.

2. REGOLA DEL QUOZIENTE

$\frac{e^a}{e^b}$ È UGUALE A e^{a-b} .

Verifichiamo ancora i conti. Controlliamo che $\frac{e^a}{e^b}$ e e^{a-b} siano uguali nel caso $a = 3$ e $b = 5$.

$$\frac{e^3}{e^5} = \frac{e \times e \times e}{e \times e \times e \times e \times e} = \frac{e \times e \times e}{e \times e \times e \times e \times e} = \frac{1}{e^2} = e^{-2} = e^{3-5}$$

3. REGOLA DELLA CANCELLAZIONE DEGLI ESPONENZIALI

a È UGUALE A $\log_e(e^a)$.



Come accennato a pagina 20, $y = \log_e x$ e $x = e^y$ si equivalgono. Prima di tutto ricordiamo cos'è un logaritmo. Chiamiamo x il valore della funzione esponenziale di n in base b , cioè b elevato all'esponente n . Il logaritmo inverte questa operazione: in altri termini, il logaritmo in base b del valore x è uguale all'esponente n .

La base b è e , e il valore x corrisponde a e^a , quindi $e^n = e^a$ e $n = a$.

Perciò $b^n = x$ significa anche $\log_b x = n$.

↑ ↑ ↑
base valore esponente

4. REGOLA DELLA ESPONENZIAZIONE

$\log_e(a^b)$ È UGUALE A $b \times \log_e(a)$.



Verifichiamo che $\log_e(a^b)$ e $b \times \log_e(a)$ siano uguali. Iniziamo applicando la regola degli esponenti alla base e e all'esponente $b \times \log_e(a)$:

$$e^{b \times \log_e(a)} = (e^{\log_e(a)})^b$$

Poiché e è l'inversa di \log_e , al membro di destra si riduce $e^{b \times \log_e(a)}$ a:

$$e^{b \times \log_e(a)} = a^b$$

Ora sfruttiamo la regola secondo cui $b^n = x$ significa anche $\log_b x = n$, dove:

$$\begin{aligned} b &= e \\ x &= a^b \\ n &= b \times \log_e(a) \end{aligned}$$

Questo significa che possiamo riscrivere $e^{b \times \log_e(a)} = a^b$, e concludere che $\log_e(a^b) = b \times \log_e(a)$.



5. REGOLA DEL PRODOTTO

$\log_e(a) + \log_e(b) \in$
UGUALE A $\log_e(a \times b)$.

Verifichiamo che $\log_e(a) + \log_e(b)$ e $\log_e(a \times b)$ siano uguali. Di nuovo, useremo la regola secondo cui $b^n = x$ significa anche $\log_b x = n$.

Iniziamo definendo $e^m = a$ e $e^n = b$. Ora, grazie alla regola del prodotto di esponenziali, abbiamo $e^m e^n = e^{m+n} = a \times b$. A questo punto possiamo calcolare il logaritmo di entrambi i membri,

$$\log_e(e^{m+n}) = \log_e(a \times b),$$

che al primo membro si riduce a

$$m + n = \log_e(a \times b).$$

Sappiamo inoltre che $m + n = \log_e a + \log_e b$, quindi ovviamente

$$\log_e(a) + \log_e(b) = \log_e(a \times b).$$

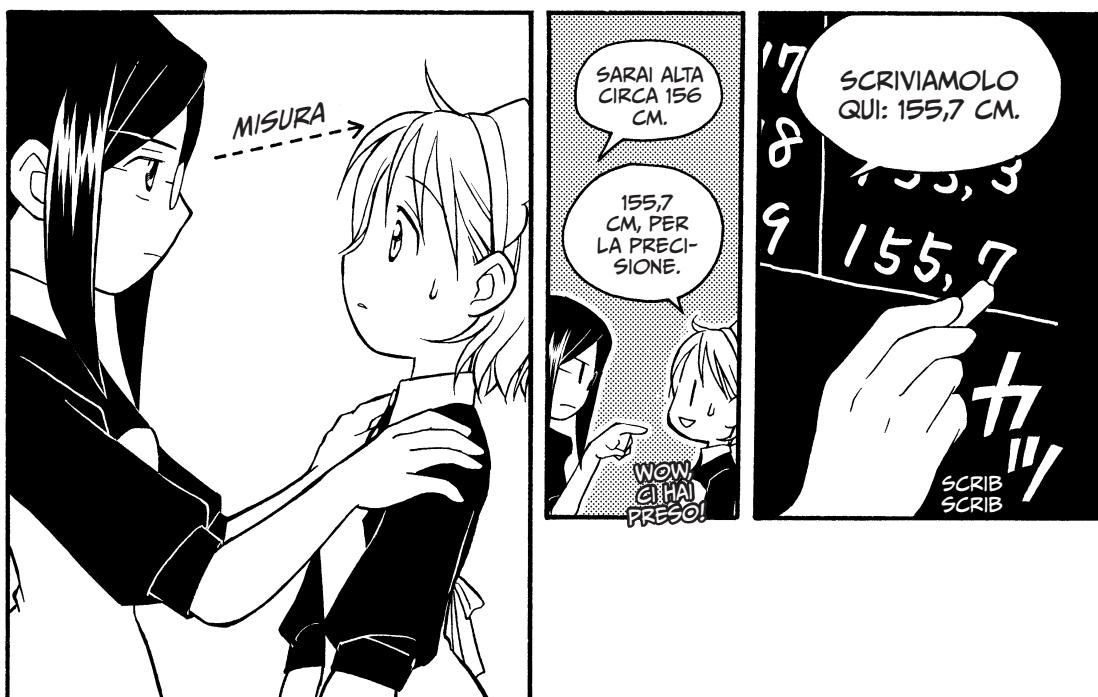
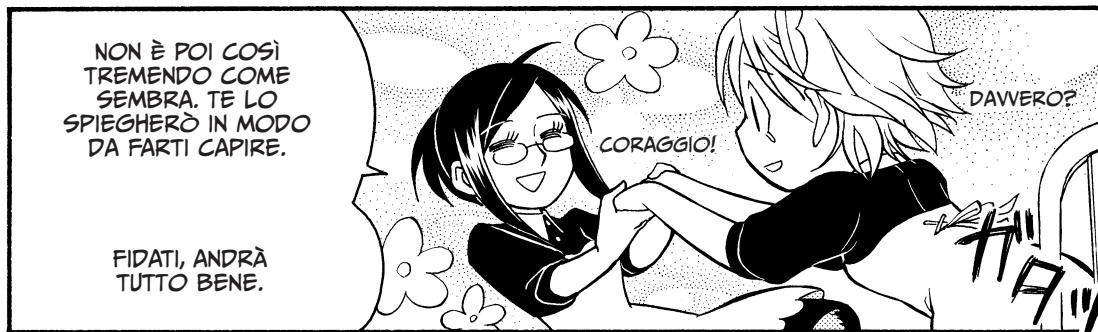
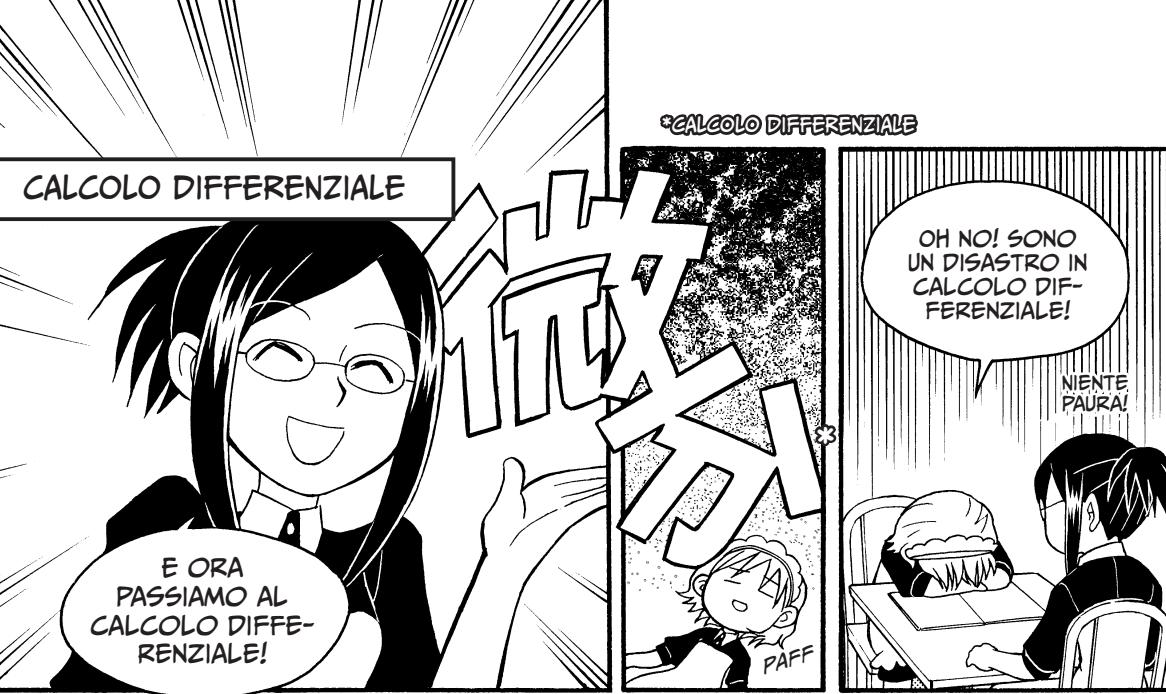
QUI HO RIASSUNTO LE REGOLE VISTE FINORA.

REGOLA 1	$(e^a)^b$ e e^{ab} sono uguali.
REGOLA 2	$\frac{e^a}{e^b}$ è uguale a e^{a-b} .
REGOLA 3	a è uguale a $\log_e(e^a)$.
REGOLA 4	$\log_e(a^b)$ è uguale a $b \times \log_e(a)$.
REGOLA 5	$\log_e(a) + \log_e(b)$ è uguale a $\log_e(a \times b)$.



Di fatto si può sostituire la base dei logaritmi naturali e con qualsiasi numero reale positivo d . Sapresti ripetere le dimostrazioni usando la base d ?

CALCOLO DIFFERENZIALE



ETÀ E ALTEZZA DI MIU

ETÀ	ALTEZZA
4	100,1
5	107,2
6	114,1
7	121,7
8	126,8
9	130,9
10	137,5
11	143,2
12	149,4
13	151,1
14	154,0
15	154,6
16	155,0
17	155,1
18	155,3
19	155,7

QUESTA TABELLA
RIPORTA LA TUA
ALTEZZA DAI 4 ANNI
FINO A OGGI.



COME HAI AVUTO
QUESTI DATI?!!?

TOP
SECRET.

MI SONO
INVENTATA
TUTTO! SHH.

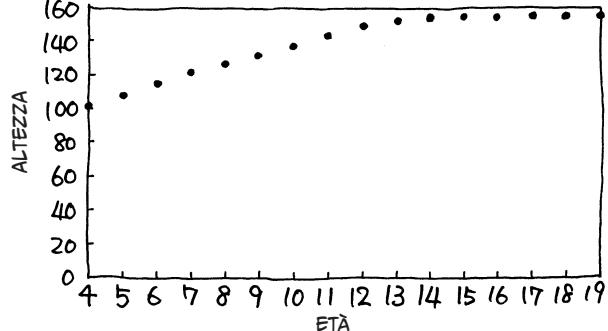
TRASFORMA
QUESTI DATI IN
UN GRAFICO DI
DISPERSIONE.

OK, UN
ATTIMO.

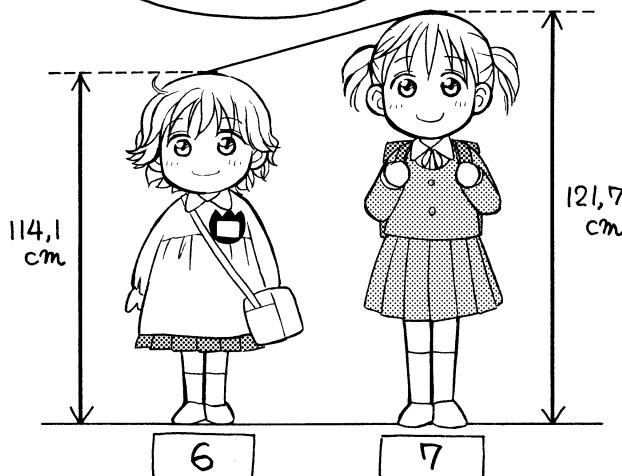
COSÌ?

BENIS-
SIMO.

GRAFICO DI DISPERSIONE
DELL'ETÀ E DELL'ALTEZZA DI MIU



CONFRONTIAMO
LA TUA ALTEZZA
A 6 E 7 ANNI.



SEI
CRESCIUTA,
MIU!

OH SÌ,
PAPA!

IN UN ANNO,
TRA I 6 E I 7
ANNI, SONO
CRESCIUTA
DI 7,6 CM
(121,7 – 114,1).

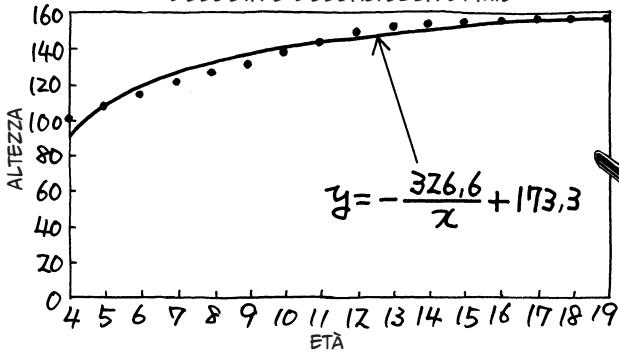
IL FATTO È
QUESTO.

GROSSO MODO, IL
LEGAME TRA LA TUA
ALTEZZA E L'ETÀ DA 4
A 19 ANNI...

$$y = -\frac{326.6}{x} + 173,3$$

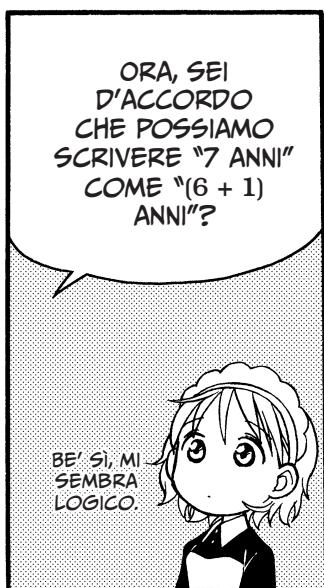
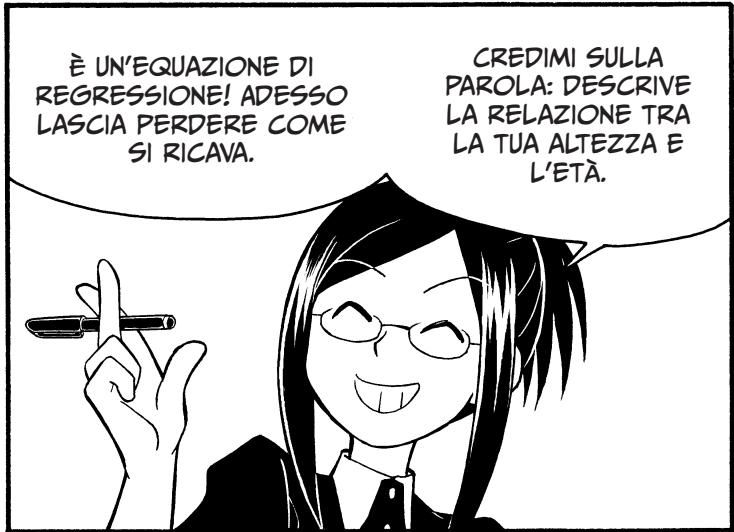
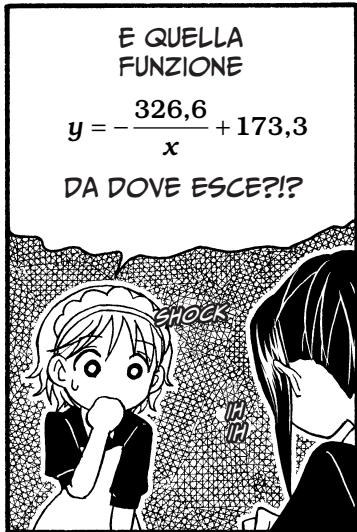
...SI PUÒ DESCRIVERE
CON QUESTA FUNZIONE.

GRAFICO DI DISPERSIONE
DELL'ETÀ E DELL'ALTEZZA DI MIU



LA CURVA
RAPPRESENTA
QUESTA
FUNZIONE.





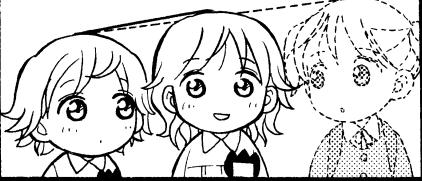
POSSIAMO MISURARE IL TASSO DI CRESCITA IN CENTIMETRI ALL'ANNO, VISTO CHE TRASCORRE UN ANNO TRA LE DUE ETÀ CONSIDERATE.

$$\frac{\left(-\frac{326,6}{(6+1)} + 173,3\right) - \left(-\frac{326,6}{6} + 173,3\right)}{1} \text{ CM/ANNO}$$

AH! HAI DIVISO LA FORMULA PRECEDENTE PER 1 PERCHÉ L'INTERVALLO DI TEMPO È UN ANNO.

ORA CONSIDERIAMO LA CRESCITA IN SEI MESI.

6 $6\frac{1}{2}$ 7



COME SI SCRIVE L'ETÀ DI SEI ANNI E MEZZO IN TERMINI DEL NUMERO 6?

VEDIAMO...
 $(6 + 0,5)$ ANNI?

GIUSTO!

LA CRESCITA IN 0,5 ANNI, TRA L'ETÀ DI 6 ANNI E QUELLA DI $(6 + 0,5)$ ANNI...

E QUESTA È LA CRESCITA ANNUALE TRA L'ETÀ DI 6 ANNI E QUELLA DI $(6 + 0,5)$ ANNI.

ALTEZZA A
 $(6 + 0,5)$

ALTEZZA A 6 ANNI

$$\left(-\frac{326,6}{(6+0,5)} + 173,3\right) - \left(-\frac{326,6}{6} + 173,3\right)$$

...SI PUÒ SCRIVERE COSÌ.

CHIARO.



$$\frac{\left(-\frac{326,6}{(6+0,5)} + 173,3\right) - \left(-\frac{326,6}{6} + 173,3\right)}{0,5} \text{ CM/ANNO}$$

STAVOLTA HAI DIVISO LA FORMULA PER 0,5 PERCHÉ L'INTERVALLO È DI SEI MESI. CI SONO!



INFINE...

CONSIDERIAMO
LA CRESCITA IN UN
PERIODO DI TEMPO
ESTREMAMENTE
BREVE.

MISURIAMO,
MISURIAMO, DI
CONTINUO!

OH, MIU, STAI
CRESCENDO IN
FRETTISSIMA!



P-PAPÀ?

IN MATE-
MATICA, LA
VARIAZIONE
SI RAPPRE-
SENTA CON
IL SIMBOLÒ
Δ (DELTA).

DESCRIVE IL BREVISSIMO
INTERVALLO DI TEMPO TRA
L'ISTANTE IN CUI COMPI 6 ANNI E
QUELLO SUBITO DOPO. CON LA
NOSTRA EQUAZIONE POSSIAMO
CALCOLARE LA CRESCITA IN
QUESTO INTERVALLO
DI TEMPO.

$$\left(-\frac{326,6}{(6+\Delta)} + 173,3 \right) - \left(-\frac{326,6}{6} + 173,3 \right)$$

COSÌ.

OH!

QUESTO SIGNIFICA CHE
POSSIAMO DESCRIVERE
COSÌ "LA CRESCITA ANNUALE
TRA L'ISTANTE IN CUI COMPI
6 ANNI E SUBITO DOPO":

D'ACCORDO.

$$\frac{\left(-\frac{326,6}{(6+\Delta)} + 173,3 \right) - \left(-\frac{326,6}{6} + 173,3 \right)}{\Delta} \text{ CM/ANNO}$$

ADESSO SEGUIMI:
QUEST'EQUAZIONE
SI SEMPLIFICA IN UN
BATTER D'OCCHIO!



$$\frac{\left(-\frac{326,6}{(6+\Delta)} + 173,3 \right) - \left(-\frac{326,6}{6} + 173,3 \right)}{\Delta}$$

$$= \frac{-\frac{326,6}{(6+\Delta)} + \frac{326,6}{6}}{\Delta}$$

$$= \frac{\frac{326,6}{6} - \frac{326,6}{(6+\Delta)}}{\Delta}$$

$$= \frac{326,6 \times \left(\frac{1}{6} - \frac{1}{(6+\Delta)} \right)}{\Delta}$$

$$= \frac{326,6 \times \frac{(6+\Delta)-6}{6(6+\Delta)}}{\Delta}$$

$$= \frac{326,6 \times \frac{\Delta}{6(6+\Delta)}}{\Delta}$$

$$= 326,6 \times \frac{\Delta}{6(6+\Delta)} \times \frac{1}{\Delta}$$

$$= 326,6 \times \frac{1}{6(6+\Delta)}$$

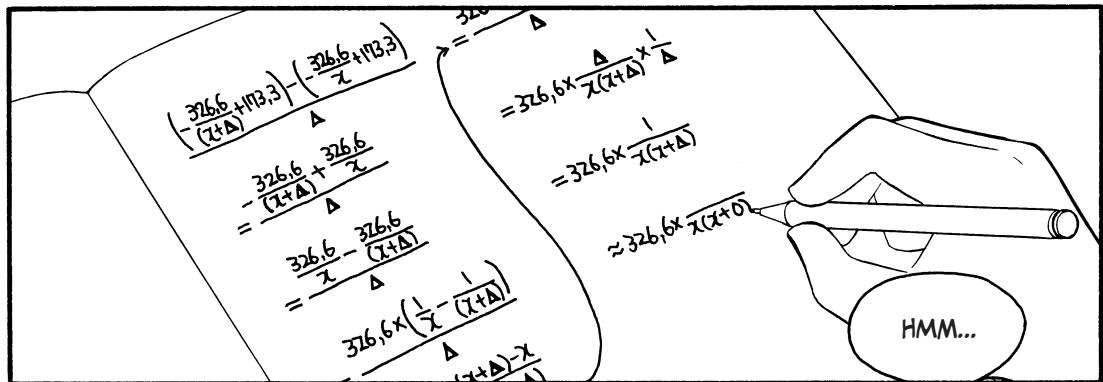
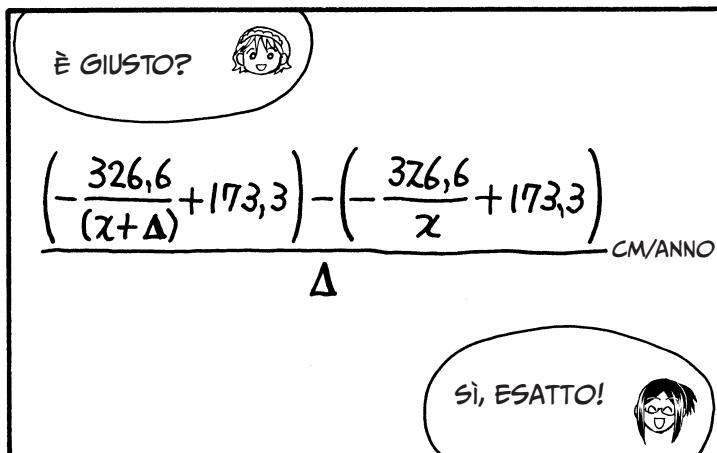
$$\approx 326,6 \times \frac{1}{6(6+0)} = 326,6 \times \frac{1}{6^2}$$

NELL'ULTIMO PASSAGGIO
HO MESSO $\Delta = 0$ PERCHÉ IL
TEMPO TRASCORSO IN PRATICA
È NULLO.



FINORA RIESCI A
SEGUIRE? QUESTO
CALCOLO È LUNGO
MA NON DIFFICILIS-
SIMO, NO?

INFATTI, PENSO
DI RIUSCIRCI.



LA SOLUZIONE È
 $326,6 \times \frac{1}{x^2}$.

QUELLO CHE STAI
FACENDO HA
UN NOME BEN
PRECISO.

BRAVISSIMA.

SI CHIAMA DERIVAZIONE,
QUELLA DEL CALCOLO
DIFFERENZIALE. ORA ABBIANO UNA
FUNZIONE CHE DESCRIVE IL TUO
TASSO DI CRESCITA!

HO STUDIATO IL
CALCOLO DIFFERENZIALE!

TRA L'ALTRO,
LE DERIVATE SI
INDICANO CON
L'APICE ('') O
COME

$$\frac{dy}{dx}.$$

$$\frac{dy}{dx} = 326,6 \times \frac{1}{x^2}$$

$$y' = 326,6 \times \frac{1}{x^2}$$

IL SIMBOLÒ
PER "Y PRIMO"
SEMBRA UN
APOSTROFO
LUNGO!

ORA TI SFIDO A DERIVARE
ALTRÉ FUNZIONI. CHE NE
DICI?

ACCETTO
LA SFIDA!

DERIVA $y = x$ RISPETTO A x .



$$\frac{(x + \Delta) - x}{\Delta} = \frac{\Delta}{\Delta} = 1 \quad \text{QUINDI} \quad \frac{dy}{dx} = 1$$

IL TASSO DI VARIAZIONE È COSTANTE!

DERIVA $y = x^2$ RISPETTO A x .



$$\frac{(x + \Delta)^2 - x^2}{\Delta} = \frac{x^2 + 2x\Delta + \Delta^2 - x^2}{\Delta} = \frac{(2x + \Delta)\Delta}{\Delta} = 2x + \Delta$$

$$\approx 2x + 0 = 2x \quad \text{QUINDI} \quad \frac{dy}{dx} = 2x$$

DERIVA $y = \frac{1}{x}$ RISPETTO A x .



$$\frac{\frac{1}{x + \Delta} - \frac{1}{x}}{\Delta} = \frac{\frac{x - (x + \Delta)}{(x + \Delta)x}}{\Delta} = \frac{-\Delta}{(x + \Delta)x} \times \frac{1}{\Delta} = \frac{-1}{(x + \Delta)x}$$

$$\approx \frac{-1}{(x + 0)x} = \frac{-1}{x^2} = -x^{-2} \quad \text{QUINDI} \quad \frac{dy}{dx} = -x^{-2}$$

DERIVA $y = \frac{1}{x^2}$ RISPETTO A x .



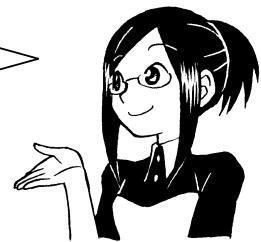
$$\begin{aligned}
 & \frac{\frac{1}{(x+\Delta)^2} - \frac{1}{x^2}}{\Delta} \\
 &= \frac{\left(\frac{1}{x+\Delta}\right)^2 - \left(\frac{1}{x}\right)^2}{\Delta} \\
 &= \frac{\left(\frac{1}{x+\Delta} + \frac{1}{x}\right)\left(\frac{1}{x+\Delta} - \frac{1}{x}\right)}{\Delta} \\
 &= \frac{\frac{x+(x+\Delta)}{(x+\Delta)x} \times \frac{x-(x+\Delta)}{(x+\Delta)x}}{\Delta} \\
 &= \frac{\frac{2x+\Delta}{(x+\Delta)x} \times \frac{-\Delta}{(x+\Delta)x}}{\Delta}
 \end{aligned}$$

$$\begin{aligned}
 & \frac{2x+\Delta}{(x+\Delta)x} \times \frac{-\Delta}{(x+\Delta)x} \times \frac{1}{\Delta} \\
 &= \frac{-(2x+\Delta)}{[(x+\Delta)x]^2} \\
 &\approx \frac{-(2x+0)}{[(x+0)x]^2} \\
 &= \frac{-2x}{x^4} \\
 &= \frac{-2}{x^3} \\
 &= -2x^{-3} \\
 \text{QUINDI } & \frac{dy}{dx} = -2x^{-3}
 \end{aligned}$$

DA QUESTI ESEMPI PUOI DEDURRE CHE DERIVANDO
 $y = x^n$ RISPETTO A x , IL RISULTATO È $\frac{dy}{dx} = nx^{n-1}$.



DERIVA $y = (5x - 7)^2$ RISPETTO A x .



$$\begin{aligned} & \frac{\{5(x + \Delta) - 7\}^2 - (5x - 7)^2}{\Delta} \\ &= \frac{[\{5(x + \Delta) - 7\} + (5x - 7)][\{5(x + \Delta) - 7\} - (5x - 7)]}{\Delta} \\ &= \frac{[2(5x - 7) + 5\Delta] \times 5\Delta}{\Delta} \\ &= [2(5x - 7) + 5\Delta] \times 5 \\ &\approx [2(5x - 7) + 5 \times 0] \times 5 \\ &= 2(5x - 7) \times 5 \end{aligned}$$

$$\text{QUINDI } \frac{dy}{dx} = 2(5x - 7) \times 5$$

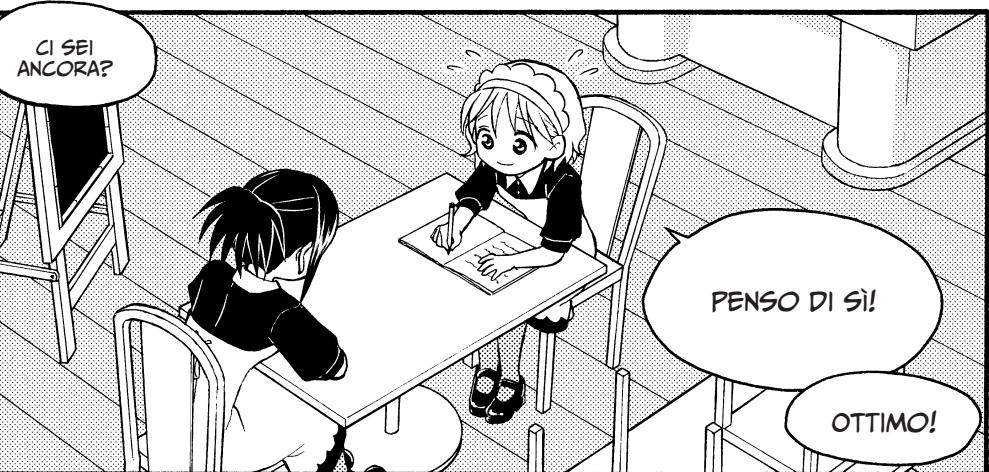
SE DERIVI $y = (ax + b)^n$ RISPETTO A x ,

$$\text{IL RISULTATO È } \frac{dy}{dx} = n(ax + b)^{n-1} \times a.$$



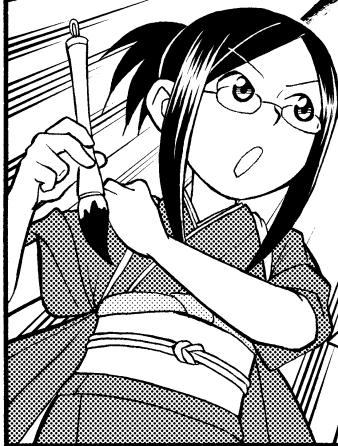
ECCO QUALCHE ALTRA DERIVATA
CHE SI USA SPESO:

- LA DERIVATA DI $y = e^x$, $\frac{dy}{dx} = e^x$.
- LA DERIVATA DI $y = \log x$, $\frac{dy}{dx} = \frac{1}{x}$.
- LA DERIVATA DI $y = \log(ax + b)$, $\frac{dy}{dx} = \frac{a}{ax + b}$.
- LA DERIVATA DI $y = \log(1 + e^{ax+b})$,
$$\frac{dy}{dx} = a - \frac{a}{1 + e^{ax+b}}.$$



MATRICI

行
列*



L'ULTIMA COSA
CHE VEDREMO
STASERA SONO
LE MATRICI.

* MATRICI

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

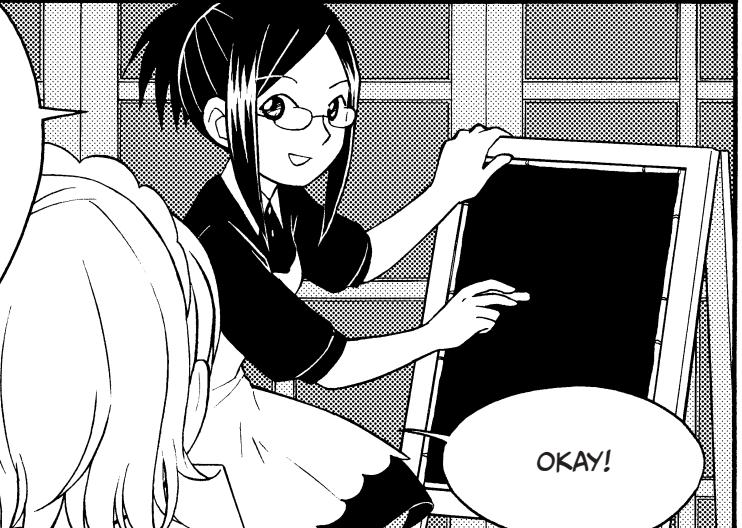
...

LE MATRICI
SEMBRANO
PALAZZINE FATTE DI
NUMERI.



SEI TUTTA
AGITATA.
RILASSATI!

IN MATEMATICA,
UNA MATRICE È
UNA DISPOSIZIONE
RETTOANGOLARE DI UNA
SERIE DI NUMERI. ORA
RIVEDREMO LE REGOLE
PER L'ADDIZIONE, LA
MOLTIPLICAZIONE E
L'INVERSIONE DI MATRICI.
PRENDI BENE GLI
APPUNTI, OK?



OKAY!

LE MATRICI AIUTANO A SCRIVERE LE EQUAZIONI RAPIDAMENTE. COME PER GLI ESPONENTI, I MATEMATICI USANO UNA NOTAZIONE BEN PRECISA.

$$\begin{cases} x_1 + 2x_2 = -1 \\ 3x_1 + 4x_2 = 5 \end{cases} \text{ SI PUÒ SCRIVERE COME } \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 5 \end{pmatrix}$$



$$\begin{matrix} x_1 + 2x_2 \\ 3x_1 + 4x_2 \end{matrix} \text{ SI PUÒ SCRIVERE COME } \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

ESEMPIO

$$\begin{cases} k_1 + 2k_2 + 3k_3 = -3 \\ 4k_1 + 5k_2 + 6k_3 = 8 \\ 10k_1 + 11k_2 + 12k_3 = 2 \\ 13k_1 + 14k_2 + 15k_3 = 7 \end{cases} \quad \begin{array}{l} \text{si può} \\ \text{scrivere} \\ \text{come} \end{array} \quad \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \\ k_3 \end{pmatrix} = \begin{pmatrix} -3 \\ 8 \\ 2 \\ 7 \end{pmatrix}$$

Se i valori delle espressioni sono ignoti, è possibile riscriverle in forma matriciale:

$$\begin{cases} k_1 + 2k_2 + 3k_3 \\ 4k_1 + 5k_2 + 6k_3 \\ 7k_1 + 8k_2 + 9k_3 \\ 10k_1 + 11k_2 + 12k_3 \\ 13k_1 + 14k_2 + 15k_3 \end{cases} \quad \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \\ k_3 \end{pmatrix}$$

Diciamo che le matrici hanno *righe* e *colonne*, proprio come fossero tabelle. Ogni numero della matrice è detto *elemento*.

RIASSUMENDO

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1q}x_q = b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2q}x_q = b_2 \\ \dots \\ a_{p1}x_1 + a_{p2}x_2 + \dots + a_{pq}x_q = b_p \end{cases}$$

si può
scrivere
come

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1q} \\ a_{21} & a_{22} & \dots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pq} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$$

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1q}x_q \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2q}x_q \\ \dots \\ a_{p1}x_1 + a_{p2}x_2 + \dots + a_{pq}x_q \end{cases}$$

si può
scrivere
come

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1q} \\ a_{21} & a_{22} & \dots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pq} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{pmatrix}$$

ADDITIONE DI MATRICI

ORA TI SPIEGO COME SOMMARE LE MATRICI.

CONSIDERIAMO QUESTO CASO: $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} + \begin{pmatrix} 4 & 5 \\ -2 & 4 \end{pmatrix}$

ORA BASTA SOMMARE I NUMERI CHE SI TROVANO NELLE POSIZIONI CORRISPONDENTI: QUELLO IN ALTO A SINISTRA CON QUELLO IN ALTO A SINISTRA, E COSÌ VIA.

$$\begin{pmatrix} 1+4 & 2+5 \\ 3+(-2) & 4+4 \end{pmatrix} = \begin{pmatrix} 5 & 7 \\ 1 & 8 \end{pmatrix}$$

SI POSSONO SOMMARE SOLTANNO LE MATRICI CHE HANNO LE STESSE DIMENSIONI, CIOÈ LO STESSO NUMERO DI RIGHE E COLONNE.



ESERCIZIO RISOLTO N. 1

Quanto vale $\begin{pmatrix} 5 & 1 \\ 6 & -9 \end{pmatrix} + \begin{pmatrix} -1 & 3 \\ -3 & 10 \end{pmatrix}$?

SOLUZIONE

$$\begin{pmatrix} 5 & 1 \\ 6 & -9 \end{pmatrix} + \begin{pmatrix} -1 & 3 \\ -3 & 10 \end{pmatrix} = \begin{pmatrix} 5+(-1) & 1+3 \\ 6+(-3) & (-9)+10 \end{pmatrix} = \begin{pmatrix} 4 & 4 \\ 3 & 1 \end{pmatrix}$$

ESERCIZIO RISOLTO N. 2

Quanto vale $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{pmatrix} + \begin{pmatrix} 7 & 2 & 3 \\ -1 & 7 & -4 \\ -7 & -3 & 10 \\ 8 & 2 & -1 \\ 7 & 1 & -9 \end{pmatrix}$?

SOLUZIONE

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{pmatrix} + \begin{pmatrix} 7 & 2 & 3 \\ -1 & 7 & -4 \\ -7 & -3 & 10 \\ 8 & 2 & -1 \\ 7 & 1 & -9 \end{pmatrix} = \begin{pmatrix} 1+7 & 2+2 & 3+3 \\ 4+(-1) & 5+7 & 6+(-4) \\ 7+(-7) & 8+(-3) & 9+10 \\ 10+8 & 11+2 & 12+(-1) \\ 13+7 & 14+1 & 15+(-9) \end{pmatrix} = \begin{pmatrix} 8 & 4 & 6 \\ 3 & 12 & 2 \\ 0 & 5 & 19 \\ 18 & 13 & 11 \\ 20 & 15 & 6 \end{pmatrix}$$

RIASSUMENDO

Ecco due matrici generiche.

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1q} \\ b_{21} & b_{22} & \cdots & b_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ b_{p1} & b_{p2} & \cdots & b_{pq} \end{pmatrix}$$

Possiamo sommarle,

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1q} \\ b_{21} & b_{22} & \cdots & b_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ b_{p1} & b_{p2} & \cdots & b_{pq} \end{pmatrix}$$

così:

$$\begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1q} + b_{1q} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2q} + b_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} + b_{p1} & a_{p2} + b_{p2} & \cdots & a_{pq} + b_{pq} \end{pmatrix}$$

E naturalmente, la sottrazione funziona allo stesso modo. Basta sottrarre i numeri nelle posizioni corrispondenti!

MOLTIPLICAZIONE DI MATRICI

PASSIAMO ALLA MOLTIPLICAZIONE DI MATRICI! FUNZIONA IN MANIERA DIVERSA DALLA SOMMA E SOTTRAZIONE. È PIÙ FACILE DA SPIEGARE CON UN ESEMPIO, QUINDI PROVIAMO A MOLTIPLICARE QUESTE DUE MATRICI:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \end{pmatrix}$$



INIZIAMO RICAVANDO LA PRIMA COLONNA DELLA MATRICE FINALE. MOLTIPLICHIAMO GLI ELEMENTI CORRISPONDENTI DELLA PRIMA RIGA DELLA MATRICE DI SINISTRA E DELLA PRIMA COLONNA DELLA MATRICE DI DESTRA, E SOMMIAMO I PRODOTTI, POI FACCIAO ALTRETTANTO CON GLI ELEMENTI DELLA SECONDA RIGA DELLA MATRICE DI SINISTRA E DELLA PRIMA COLONNA DELLA MATRICE DI DESTRA:

$$1x_1 + 2x_2$$

$$3x_1 + 4x_2$$

PER RICAVARE LA SECONDA COLONNA DELLA MATRICE FINALE, RIPETIAMO IL PROCEDIMENTO CON LA SECONDA COLONNA DELLA MATRICE DI DESTRA, OTTENENDO:

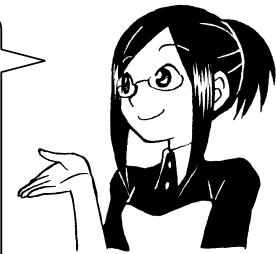
$$1y_1 + 2y_2$$

$$3y_1 + 4y_2$$

QUINDI IL RISULTATO FINALE È:

$$\begin{pmatrix} 1x_1 + 2x_2 & 1y_1 + 2y_2 \\ 3x_1 + 4x_2 & 3y_1 + 4y_2 \end{pmatrix}$$

NELLA MOLTIPLICAZIONE MATRICIALE, PRIMA SI MOLTIPLICA E POI SI SOMMA PER ARRIVARE AL RISULTATO. FACCIAO QUALCHE ESERCIZIO.



ESERCIZIO RISOLTO N. 1

Quanto vale $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 4 & 5 \\ -2 & 4 \end{pmatrix}$?

Sappiamo come moltiplicare gli elementi e poi sommarli per ottenere il risultato. Moltiplicando prendiamo la matrice di sinistra, riga per riga, e la moltiplichiamo per quella di destra.*

SOLUZIONE

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 4 \\ -2 \end{pmatrix} = \begin{pmatrix} 1 \times 4 + 2 \times (-2) \\ 3 \times 4 + 4 \times (-2) \end{pmatrix} = \begin{pmatrix} 0 \\ 4 \end{pmatrix} \quad \text{Prima colonna}$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 \\ 4 \end{pmatrix} = \begin{pmatrix} 1 \times 5 + 2 \times 4 \\ 3 \times 5 + 4 \times 4 \end{pmatrix} = \begin{pmatrix} 13 \\ 31 \end{pmatrix} \quad \text{Seconda colonna}$$

$$\text{Quindi la soluzione è } \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 4 & 5 \\ -2 & 4 \end{pmatrix} = \begin{pmatrix} 0 & 13 \\ 4 & 31 \end{pmatrix}.$$

*OSSERVATE CHE LA MATRICE RISULTANTE AVRÀ LO STESSO NUMERO DI RIGHE DELLA PRIMA MATRICE E LO STESSO NUMERO DI COLONNE DELLA SECONDA.

ESERCIZIO RISOLTO N. 2

Quanto vale $\begin{pmatrix} 1 & 2 \\ 4 & 5 \\ 7 & 8 \\ 10 & 11 \end{pmatrix} \begin{pmatrix} k_1 & l_1 & m_1 \\ k_2 & l_2 & m_2 \end{pmatrix}$?

SOLUZIONE

$$\begin{pmatrix} 1 & 2 \\ 4 & 5 \\ 7 & 8 \\ 10 & 11 \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \begin{pmatrix} k_1 + 2k_2 \\ 4k_1 + 5k_2 \\ 7k_1 + 8k_2 \\ 10k_1 + 11k_2 \end{pmatrix}$$

Moltiplichiamo le righe della prima matrice per la prima colonna della seconda.

$$\begin{pmatrix} 1 & 2 \\ 4 & 5 \\ 7 & 8 \\ 10 & 11 \end{pmatrix} \begin{pmatrix} l_1 \\ l_2 \end{pmatrix} = \begin{pmatrix} l_1 + 2l_2 \\ 4l_1 + 5l_2 \\ 7l_1 + 8l_2 \\ 10l_1 + 11l_2 \end{pmatrix}$$

Facciamo altrettanto con la seconda colonna.

$$\begin{pmatrix} 1 & 2 \\ 4 & 5 \\ 7 & 8 \\ 10 & 11 \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} m_1 + 2m_2 \\ 4m_1 + 5m_2 \\ 7m_1 + 8m_2 \\ 10m_1 + 11m_2 \end{pmatrix}$$

E con la terza colonna.

Per arrivare alla soluzione basta combinare i tre risultati qui sopra.

$$\begin{pmatrix} k_1 + 2k_2 & l_1 + 2l_2 & m_1 + 2m_2 \\ 4k_1 + 5k_2 & 4l_1 + 5l_2 & 4m_1 + 5m_2 \\ 7k_1 + 8k_2 & 7l_1 + 8l_2 & 7m_1 + 8m_2 \\ 10k_1 + 11k_2 & 10l_1 + 11l_2 & 10m_1 + 11m_2 \end{pmatrix}$$

REGOLE DELLA MOLTIPLICAZIONE MATRICIALE

MOLTIPLICANDO LE MATRICI BISOGNA RICORDARE TRE COSE.

- IL NUMERO DI COLONNE DELLA PRIMA MATRICE DEVE ESSERE UGUALE AL NUMERO DI RIGHE DELLA SECONDA.
- LA MATRICE RISULTANTE AVRÀ LO STESSO NUMERO DI RIGHE DELLA PRIMA MATRICE.
- LA MATRICE RISULTANTE AVRÀ LO STESSO NUMERO DI COLONNE DELLA SECONDA MATRICE.



È possibile moltiplicare le seguenti coppie di matrici? In caso positivo, quante righe e colonne avranno le matrici risultanti?

ESERCIZIO RISOLTO N. 1

$$\begin{pmatrix} 2 & 3 & 4 \\ -5 & 3 & 6 \end{pmatrix} \begin{pmatrix} 2 \\ -7 \\ 0 \end{pmatrix}$$

SOLUZIONE

Sì! La matrice risultante avrà 2 righe e 1 colonna:

$$\begin{pmatrix} 2 & 3 & 4 \\ -5 & 3 & 6 \end{pmatrix} \begin{pmatrix} 2 \\ -7 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \times 2 + 3 \times (-7) + 4 \times 0 \\ (-5) \times 2 + 3 \times (-7) + 6 \times 0 \end{pmatrix} = \begin{pmatrix} -17 \\ -31 \end{pmatrix}$$

ESERCIZIO RISOLTO N. 2

$$\begin{pmatrix} 9 & 4 & -1 \\ 7 & -6 & 0 \\ -5 & 3 & 8 \end{pmatrix} \begin{pmatrix} 2 & -2 & 1 \\ 4 & 9 & -7 \end{pmatrix}$$

SOLUZIONE

No. La prima matrice ha 3 colonne, ma la seconda matrice ha 2 righe. Queste matrici non si possono moltiplicare.

MATRICE IDENTITÀ E MATRICI INVERSE

L'ULTIMA COSA PER STASERA SONO LA
MATRICE IDENTITÀ E LE MATRICI INVERSE.

LA MATRICE IDENTITÀ È UNA MATRICE QUADRATA CHE HA SOLO NUMERI 1 LUNGO LA DIAGONALE, DALL'ANGOLO SUPERIORE SINISTRO A QUELLO INFERIORE DESTRO, E 0 IN OGNI ALTRA POSIZIONE.

ECCO UNA MATRICE IDENTITÀ 2×2 : $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

ED ECCO UNA MATRICE IDENTITÀ 3×3 : $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$



Alcune matrici quadrate (cioè, matrici che hanno lo stesso numero di righe e colonne) sono *invertibili*. Una matrice quadrata moltiplicata per la sua inversa dà una matrice identità della stessa forma e dimensione, quindi è facile dimostrare che una matrice è inversa di un'altra data.

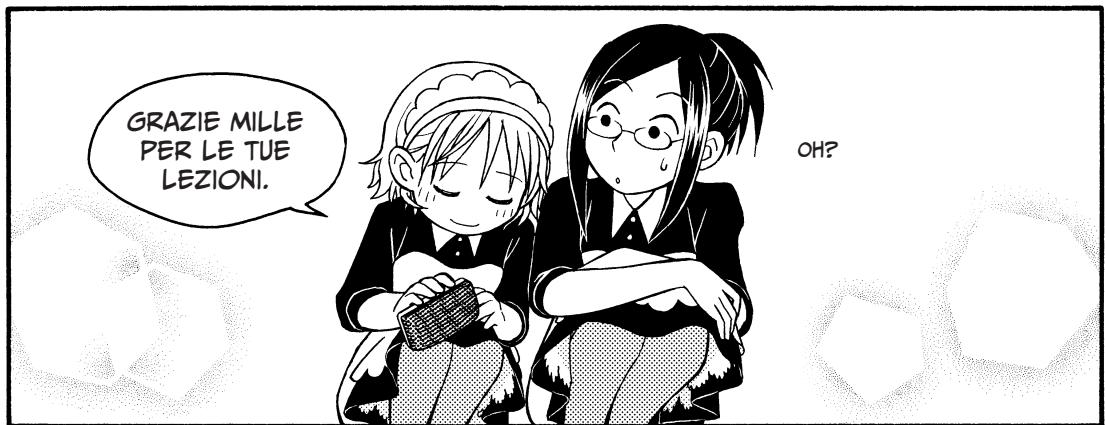
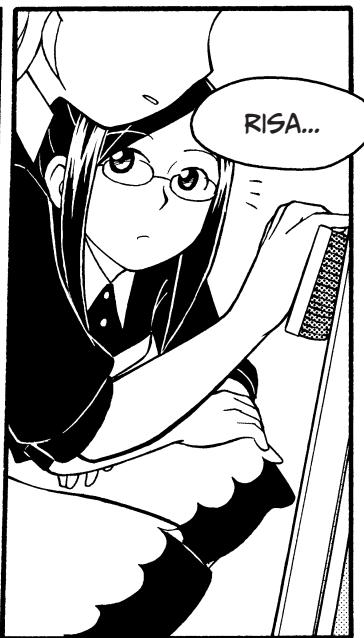
Per esempio:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} -2 & 1 \\ 1,5 & -0,5 \end{pmatrix} = \begin{pmatrix} 1 \times (-2) + 2 \times 1,5 & 1 \times 1 + 2 \times -0,5 \\ 3 \times (-2) + 4 \times 1,5 & 3 \times 1 + 4 \times -0,5 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Quindi $\begin{pmatrix} -2 & 1 \\ 1,5 & -0,5 \end{pmatrix}$ è l'inverso di $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$.

PSS! EHI, MIU,
SVEGLIATI.

PER OGGI
ABBIAMO
FINITO.



TIPI DI DATI STATISTICI

Dopo questa sorsata di matematica generale, è l'ora dell'ammazzacaffè: un bel bicchierino di *statistica*, una branca della matematica che si occupa di interpretare e analizzare i dati. Bando alle ciance e iniziamo.

I dati si classificano in due categorie: *dati numerici* o *cardinali* se è possibile misurarli numericamente, *dati categorici* in caso contrario. I dati numerici si chiamano anche *dati quantitativi*, e gli altri, *dati qualitativi*. Questa nomenclatura è soggettiva e varia a seconda degli specialisti e dei campi di studio. La Tabella 1-1 riporta alcuni esempi di dati numerici e categorici.

TABELLA 1-1: DATI NUMERICI E CATEGORICI

	Numero di libri letti al mese	Età (anni)	Luogo preferito di lettura	Sesso
Persona A	4	20	Treno	Femminile
Persona B	2	19	Casa	Maschile
Persona C	10	18	Caffè	Maschile
Persona D	14	22	Biblioteca	Femminile

Il *Numero di libri letti al mese* e l'*Età* sono esempi di dati numerici, mentre il *Luogo preferito di lettura* e il *Sesso* non sono in genere indicati da numeri. È tuttavia possibile convertire i dati categorici in numerici e viceversa. La Tabella 1-2 dà un esempio di conversione di dati numerici in categorici.

TABELLA 1-2: CONVERSIONE DI DATI NUMERICI IN CATEGORICI

	Numero di libri letti al mese	
		Numero di libri letti al mese
Persona A	4	Pochi
Persona B	2	Pochi
Persona C	10	Molti
Persona D	14	Molti

In questa tabella i valori da 1 a 5 sono stati inglobati nella categoria *Pochi*, quelli da 6 a 9 nella categoria *Medi* e quelli uguali o superiori a 10 nella categoria *Molti*. Possiamo fissare a nostra discrezione i limiti degli intervalli. Osservate che le tre categorie (*Pochi*, *Medi*, *Molti*) sono *ordinali*, cioè possiamo metterle in ordine: *Molti* libri sono più di *Medi*, che sono più di *Pochi*. Alcune categorie sono difficili da ordinare: per esempio, che fareste se fossero *Marrone*, *Viola*, *Verde*?

La Tabella 1-3 riporta un esempio di conversione di dati categorici in numerici.

TABELLA 1-3: CONVERSIONE DI DATI CATEGORICI IN NUMERICI

	Stagione preferita	Primavera	Estate	Autunno	Inverno
Persona A	Primavera	1	0	0	0
Persona B	Estate	0	1	0	0
Persona C	Autunno	0	0	1	0
Persona D	Inverno	0	0	0	1

In questo caso, abbiamo convertito il dato categorico *Stagione preferita*, con quattro categorie possibili (*Primavera*, *Estate*, *Autunno*, *Inverno*), in dati binari riportati su quattro colonne. Si tratta di dati binari perché possono assumere soltanto due valori: *Preferita* corrisponde a 1 e *Non preferita* corrisponde a 0.

È anche possibile rappresentare questi dati con tre sole colonne. Come mai possiamo lasciarne fuori una? Perché la stagione preferita di ogni intervistato si deduce comunque. Per esempio, se le prime tre colonne (*Primavera*, *Estate*, *Autunno*) sono 0, sappiamo che *Inverno* deve valere 1, benché sia omesso.

Nell'analisi di regressione multipla bisogna assicurarsi che i dati siano *linearmente indipendenti*; in altri termini, che in un insieme di colonne nessun sottoinsieme permetta di calcolarne esattamente un'altra. Spesso si ottiene l'indipendenza lineare eliminando l'ultima colonna di dati. Poiché l'affermazione seguente è vera, possiamo eliminare la colonna *Inverno* dalla Tabella 1-3:

$$(\text{Inverno}) = 1 - (\text{Primavera}) - (\text{Estate}) - (\text{Autunno})$$

È importante distinguere le variabili numeriche, ordinali e categoriche per svolgere correttamente l'analisi di regressione.

VERIFICA DI IPOTESI

Si usano spesso metodi statistici per verificare ipotesi scientifiche. Si chiama *ipotesi* un'affermazione che intende stabilire un legame fra alcune variabili o fra diverse proprietà di una singola variabile, descrivendo un concetto o un fenomeno dato. Si usa la verifica di ipotesi per stabilire se questa ipotesi è in accordo con i dati precedentemente raccolti.

Per svolgere la verifica di ipotesi non si formula un'ipotesi sola, ne occorrono due: l'*ipotesi nulla* (H_0) e l'*ipotesi alternativa* (H_a). L'ipotesi nulla è l'ipotesi iniziale che vogliamo confutare, in genere affermando che tra le variabili o le proprietà di una singola variabile sussiste una certa relazione (oppure nessuna). L'ipotesi alternativa è quella che stiamo cercando di dimostrare. Se i dati sono abbastanza diversi da quelli attesi nel caso in cui l'ipotesi nulla fosse vera, possiamo scartarla e accettare l'ipotesi alternativa. Consideriamo un esempio semplicissimo con le ipotesi seguenti:

H_0 : i bambini bevono in media 10 tazze di cioccolata calda al mese.

H_a : i bambini non bevono in media 10 tazze di cioccolata calda al mese.

In questo caso formuliamo ipotesi su una singola variabile – il numero di cioccolate calde al mese – per verificare che abbia una certa proprietà: una media pari a 10. Immaginate di osservare cinque bambini per un mese, concludendo che bevono rispettivamente 7, 9, 10, 11 e 13 tazze di cioccolata calda. Supponiamo che questi cinque bambini siano un *campione* rappresentativo della *popolazione* totale dei bambini che bevono cioccolata calda. La media per questi cinque bambini è di 10 tazze. In questo caso non possiamo dimostrare che l'ipotesi nulla sia falsa: essa propone un valore (10) pari alla media di questo campione.

Supponiamo però di osservare per un mese un campione di altri cinque bambini, che bevono rispettivamente 29, 30, 31, 32 e 35 tazze di cioccolata calda. La media per questi cinque bambini è di 31,4 tazze; di fatto, tutti i bambini sono lontanissimi dalla cifra di 10 sole tazze di cioccolata. Sulla base di questi dati concluderemo che l'ipotesi nulla va rifiutata.

In questo esempio abbiamo formulato ipotesi su una singola variabile: il numero di cioccolate calde bevute al mese da ogni bambino. Ma considerando due o più variabili, come nell'analisi di regressione, in genere l'ipotesi nulla prevede l'assenza di qualsiasi legame fra le variabili studiate, e l'ipotesi alternativa la presenza di un qualche legame.

MISURA DELLA VARIABILITÀ

Immaginate che Miu e Risa organizzino una gara di karaoke con alcune compagne di scuola, formando due squadre di cinque componenti. La Tabella 1-4 riporta i punteggi.

TABELLA 1-4: PUNTEGGI AL KARAOKE PER LE SQUADRE DI MIU E RISA

Componente	Punti	Componente	Punti
Miu	48	Risa	67
Yuko	32	Asuka	55
Aiko	88	Nana	61
Maya	61	Yuki	63
Marie	71	Rika	54
Media	60	Media	60

Esistono varie statistiche per descrivere il “centro” di un insieme di dati. La Tabella 1-4 riporta la *media* di entrambe le squadre, calcolata sommando i punteggi di ogni componente e dividendo per il numero di partecipanti. Entrambe le squadre hanno un punteggio medio di 60.

In alternativa, si può identificare il centro dell’insieme di dati con il dato che si trova in posizione centrale in un elenco ordinato. Questa è la *mediana* dei dati. Per determinarla, scriviamo i punteggi in ordine crescente (per la squadra di Miu abbiamo 32, 48, 61, 71 e 88); la mediana è il numero al centro della lista, cioè il punteggio di Maya, pari a 61. Si dà il caso che anche per la squadra di Risa la mediana sia 61, corrispondente al punteggio di Nana. Se le squadre fossero composte da un numero pari di partecipanti, la mediana sarebbe la media dei due punteggi centrali.

Le statistiche calcolate finora sembrano indicare che i due insiemi di punteggi siano equivalenti. Ma proviamo a disporli su una retta numerica (osservate la Figura 1-1); notate qualcosa?

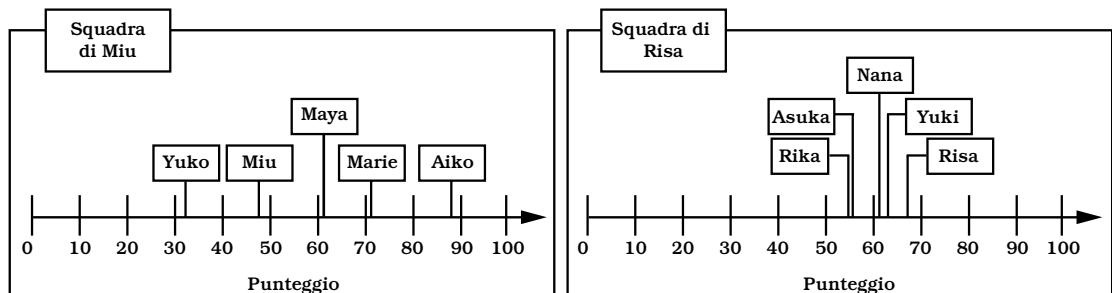


FIGURA 1-1: PUNTEGGI AL KARAOKE PER LE SQUADRE DI MIU E RISA SU UNA RETTA NUMERICA

I punteggi della squadra di Miu sono molto più “sparpagliati” di quelli della squadra di Risa. Si dice allora che i due insiemi di dati non hanno la stessa *variabilità*.

Per misurare la variabilità esistono vari indici statistici, come la somma dei quadrati degli scarti, la varianza e la deviazione standard. Tutti questi indici hanno le seguenti caratteristiche:

- Misurano lo sparpagliamento dei dati rispetto alla media.
- Aumentano al crescere della variabilità dei dati.
- Sono sempre positivi; si annullano solo se i dati non variano affatto!

SOMMA DEI QUADRATI DEGLI SCARTI

La *somma dei quadrati degli scarti* è un indice che ricorre spesso nell’analisi di regressione. Si calcola in questo modo:

somma di $(\text{punteggio individuale} - \text{punteggio medio})^2$,

che matematicamente si esprime come

$$\sum (x - \bar{x})^2.$$

Da sola, la somma dei quadrati degli scarti si usa di rado, perché ha un serio difetto: il suo valore cresce all’aumentare del numero di dati. Avendo sempre più dati, infatti, la somma delle loro differenze dalla media non fa che aumentare.

VARIANZA

Per porre rimedio a questo problema si calcola la *varianza*:

$$\frac{\sum (x - \bar{x})^2}{n - 1}, \text{ dove } n = \text{numero di dati.}$$

Questa grandezza è anche detta *varianza campionaria corretta*, perché il denominatore è il numero di punti diminuito di 1.

Nell’analisi statistica di campioni di una popolazione, in genere si diminuisce di 1 il numero di punti per tener conto del fatto che non si sta considerando l’intera popolazione. In questo modo la varianza aumenta.

Il denominatore corretto è chiamato *numero di gradi di libertà*, perché rappresenta il numero di quantità libere di variare. In pratica corrisponde al numero di punti sperimentali (per esempio osservazioni o gruppi) diminuito di 1.

Se prendiamo la squadra di Miu o quella di Risa come campioni dell'intera popolazione di cantanti di karaoke, calcolandone le statistiche diremmo quindi che ci sono 4 gradi di libertà, perché ciascuna squadra conta cinque componenti. Diminuiamo di 1 il numero di componenti perché sono un piccolo sottoinsieme di tutti i cantanti di karaoke al mondo, e preferiamo sovrastimare la varianza in questo sottoinsieme.

La varianza non si misura nelle stesse unità dei dati osservati, ma in quelle unità elevate al quadrato; nel nostro caso, "punti al quadrato".

DEVIAZIONE STANDARD

Come la varianza, la *deviazione standard* misura quanto sono dispersi i dati sperimentali. La deviazione standard è semplicemente la radice quadrata della varianza:

$$\sqrt{\text{varianza}}$$

L'indice di variabilità usato più spesso è la deviazione standard, perché si misura nelle stesse unità dei dati originari. Per le nostre cantanti di karaoke, la deviazione standard si misura in "punti".

Calcoliamo la somma dei quadrati degli scarti, la varianza e la deviazione standard per la squadra di Miu (Tabella 1-5).

TABELLA 1-5: MISURA DELLA VARIABILITÀ DEI PUNTEGGI PER LA SQUADRA DI MIU

Indice di variabilità	Calcolo
Somma dei quadrati degli scarti	$(48 - 60)^2 + (32 - 60)^2 + (88 - 60)^2 + (61 - 60)^2 + (71 - 60)^2 \\ = (-12)^2 + (-28)^2 + 28^2 + 1^2 + 11^2 \\ = 1834$
Varianza	$\frac{1834}{5 - 1} = 458,8$
Deviazione standard	$\sqrt{458,8} = 21,4$

Adesso facciamo altrettanto per la squadra di Risa (Tabella 1-6).

TABELLA 1-6: MISURA DELLA VARIABILITÀ DEI PUNTEGGI PER LA SQUADRA DI RISA

Indice di variabilità	Calcolo
Somma dei quadrati degli scarti	$(67 - 60)^2 + (55 - 60)^2 + (61 - 60)^2 + (63 - 60)^2 + (54 - 60)^2 \\ = 7^2 + (-5)^2 + 1^2 + 3^2 + (-6)^2 \\ = 120$
Varianza	$\frac{120}{5 - 1} = 30$
Deviazione standard	$\sqrt{30} = 5,5$

Possiamo vedere che la deviazione standard per la squadra di Risa è di 5,5 punti, mentre quella della squadra di Miu è di 21,4 punti. I punteggi della squadra di Risa variano di meno: le sue componenti ottengono punteggi più omogenei.

FUNZIONI DI DENSITÀ DI PROBABILITÀ

La probabilità serve a creare modelli per descrivere eventi che non ammettono previsioni infallibili. Possiamo prevedere con precisione molti eventi futuri – per esempio, il fatto che finendo la benzina la macchina si fermerà, o quanto carburante occorre per portare un razzo su Marte – ma vari problemi di natura fisica, chimica, biologica, sociale e strategica sono talmente complicati che non c'è speranza di conoscere tutte le forze e le variabili in gioco.

Un esempio semplice è il lancio di una monetina. Non conosciamo i valori di tutte le variabili fisiche coinvolte in un singolo lancio: la temperatura, il momento angolare, la rotazione, le irregolarità della superficie su cui cade e via dicendo. Ci aspettiamo però che nel corso di molti lanci la varianza di tutti questi fattori si compensi, e che osserveremo uno stesso numero di teste o croci. La Tabella 1-7 mostra i risultati di un miliardo di lanci, in valore assoluto e in percentuale.

TABELLA 1-7: RISULTATI DI UN MILIARDI DI LANCI DI UNA MONETINA

	Numero di lanci	Percentuale di lanci
Teste	499.993.945	49,99939%
Croci	500.006.054	50,00061%
La monetina rimane in bilico	1	0,0000001%

Come prevedibile, le percentuali di teste e croci sono entrambe vicinissime al 50%. Possiamo riassumere quanto sappiamo su questi lanci di monetine con la funzione di densità di probabilità, $P(x)$:

$$P(\text{Testa}) = 0,5, P(\text{Croce}) = 0,5 \quad P(\text{Monetina in bilico}) < 1 \times 10^{-9}$$

Potremo in seguito applicare questa P a qualsiasi lancio di monetine. Ma che succede se giochiamo con un baro? Magari ha truccato la moneta in modo che ora $P(x)$ sia:

$$P(\text{Testa}) = 0,3, P(\text{Croce}) = .7, P(\text{Monetina in bilico}) = 0$$

Cosa prevediamo per un singolo lancio: sarà sempre croce? E quanto varrà la media dopo un miliardo di lanci?

Rispetto all'esempio della monetina, vari eventi hanno un numero assai maggiore di risultati possibili. Spesso i dati che vogliamo descrivere ammettono misurazioni continue. Un esempio di dato continuo è l'altezza di una persona: possiamo arrotondarla al metro, al centimetro, al millimetro... o al nanometro. Se i valori possibili dei dati descrivono uno spazio continuo, bisogna usare funzioni continue per rappresentare la probabilità degli eventi.

La funzione di densità di probabilità permette di calcolare la probabilità che i dati rientrino in un certo intervallo di valori. Se tracciamo il grafico di questa funzione, l'asse x rappresenta lo spazio degli eventi, cioè i possibili valori della variabile, e l'asse y è $f(x)$, cioè il valore della funzione densità di probabilità in x . L'area sottesa dalla curva tra due valori possibili equivale alla probabilità di ottenere un valore compreso nell'intervallo corrispondente.

DISTRIBUZIONI NORMALI

Un'importante funzione di densità di probabilità è la *distribuzione normale* (Figura 1-2), detta anche *curva a campana* per via della sua forma simmetrica, e usata nei modelli di molti eventi.

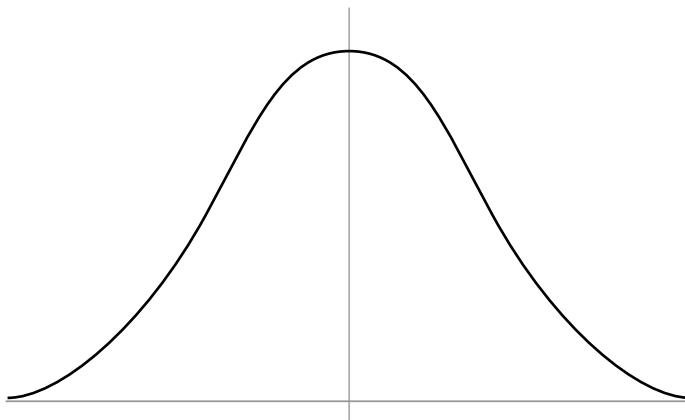


FIGURA 1-2: DISTRIBUZIONE NORMALE

La funzione di densità di probabilità per la distribuzione normale standard si può esprimere così:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

La media della funzione di distribuzione normale standard è zero. Tracciando il grafico notiamo che il *massimo* si trova in corrispondenza della media, cioè a $x = 0$. Le code della distribuzione seguono la sagoma di una campana, scendendo simmetricamente da entrambi i lati della media, e si estendono all'infinito, avvicinandosi all'asse x senza raggiungerlo mai. La distribuzione normale standard ha deviazione standard pari a 1. Poiché la media è zero e la deviazione standard è 1, questa distribuzione si indica anche con $N(0, 1)$.

L'area sottesa dall'intera curva è uguale a 1 (cioè 100%), perché la variabile assumerà certamente un qualche valore lungo la curva. I valori più lontani dalla media sono meno probabili, come segnala il fatto che la curva si abbassa. Avrete forse visto una curva del genere nel grafico che descrive la distribuzione dei voti agli esami. La maggior parte degli esaminandi ha un voto vicino alla media; in rari casi il voto è molto più alto, o molto più basso.

DISTRIBUZIONI CHI-QUADRO

Non sempre la distribuzione normale rappresenta un modello ottimale per i dati in questione. La *distribuzione chi-quadro* (χ^2) è una funzione di densità di probabilità che rappresenta la distribuzione della somma di grandezze al quadrato; torna quindi utile per stimare la variabilità. Ecco la funzione di densità chi-quadro:

$$f(x) = \begin{cases} \frac{1}{2^{\frac{k}{2}} \int_0^{\infty} x^{\frac{k}{2}-1} e^{-x} dx} \times x^{\frac{k}{2}-1} \times e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

La somma di quadrati non è mai negativa, e vediamo che $f(x)$ vale esattamente zero per gli x negativi. Quando la funzione densità di probabilità per una variabile x è quella mostrata qui sopra, si dice che “ x segue una distribuzione chi-quadro con k gradi di libertà”.

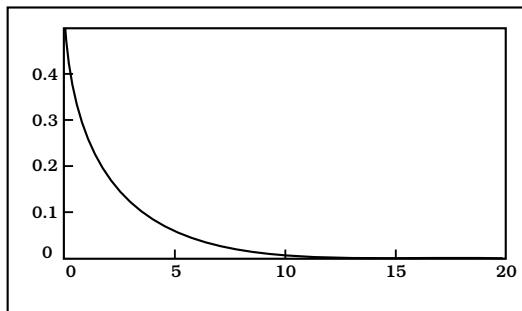
La distribuzione chi-quadro è legata alla distribuzione normale standard. Prendendo infatti un insieme di variabili casuali (o “aleatorie”) indipendenti Z_1, Z_2, \dots, Z_k che seguono tutte la stessa distribuzione normale standard e poi calcolando la somma dei quadrati in questo modo:

$$X = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

risulta che la variabile casuale X segue la distribuzione chi-quadro con k gradi di libertà. Useremo quindi la distribuzione chi-quadro per rappresentare la somma dei quadrati di un insieme di k variabili casuali normali.

Nella Figura 1-3 abbiamo tracciato i grafici di due funzioni di densità chi-quadro, con gradi di libertà pari rispettivamente a $k = 2$ e $k = 10$.

Se $k = 2$



Se $k = 10$

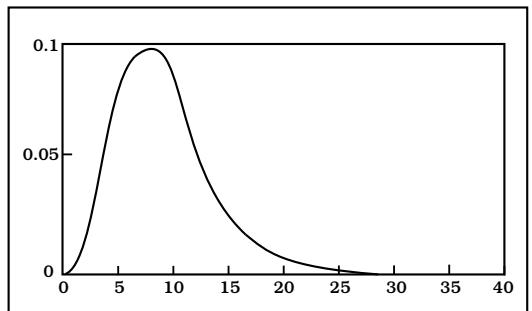


FIGURA 1-3: GRAFICI DELLA DENSITÀ DI PROBABILITÀ CHI-QUADRO PER 2 GRADI DI LIBERTÀ (A SINISTRA) E 10 GRADI DI LIBERTÀ (A DESTRA)

Osservate le differenze tra le due funzioni. Qual è il loro limite quando x va a infinito? Dov'è il loro massimo?

DENSITÀ DI PROBABILITÀ: TAVOLE DI DISTRIBUZIONE

Consideriamo un insieme di dati di una variabile X che segue una distribuzione chi-quadro, con 5 gradi di libertà. Dato un punto x , detto anche *valore critico* della distribuzione, per stabilire se la probabilità P che $X > x$ sia minore di una probabilità data dobbiamo integrare la funzione di densità. *Integrare* significa calcolare l'area sotto la parte corrispondente della curva, come mostrato dalla Figura 1-4.

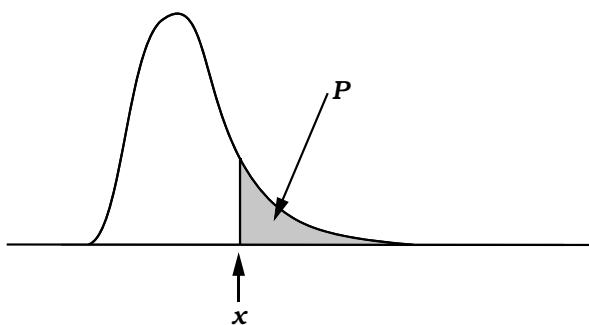


FIGURA 1-4: PROBABILITÀ P CHE UN VALORE X SUPERI IL VALORE CRITICO x DELLA DISTRIBUZIONE CHI-QUADRO

Poiché sono calcoli macchinosi da fare a mano, si ricorre al computer o alle tavole delle distribuzioni riportate nei libri, che riassumono molte caratteristiche delle funzioni di densità. Nel caso della distribuzione chi-quadro, la tavola di distribuzione fornisce il punto x tale che la probabilità che $X > x$ sia uguale a un valore P . Spesso gli statistici fissano il valore $P = 0,05$, che equivale a dire che un valore casuale di X sarà maggiore di x con una probabilità del 5%. Il valore di P è detto *p-value*.

Cerchiamo nella tavola di distribuzione della densità chi-quadro (Tabella 1-8) la casella corrispondente ai gradi di libertà e al *p-value* che ci interessano; otteniamo così il valore di χ^2 (la statistica test nel nostro caso). La probabilità di un chi-quadro di questa grandezza è uguale o minore della p in cima alla colonna.

TABELLA 1-8: TAVOLA DI DISTRIBUZIONE DELLA DENSITÀ CHI-QUADRO

gradi di libertà \ <i>p</i>	0,995	0,99	0,975	0,95	0,05	0,025	0,01	0,005
1	0,000039	0,0002	0,0010	0,0039	3,8415	5,0239	6,6349	7,8794
2	0,0100	0,0201	0,0506	0,1026	5,9915	7,3778	9,2104	10,5965
3	0,0717	0,1148	0,2158	0,3518	7,8147	9,3484	11,3449	12,8381
4	0,2070	0,2971	0,4844	0,7107	9,4877	11,1433	13,2767	14,8602
5	0,4118	0,5543	0,8312	1,1455	11,0705	12,8325	15,0863	16,7496
6	0,6757	0,8721	1,2373	1,6354	12,5916	14,4494	16,8119	18,5475
7	0,9893	1,2390	1,6899	2,1673	14,0671	16,0128	18,4753	20,2777
8	1,3444	1,6465	2,1797	2,7326	15,5073	17,5345	20,0902	21,9549
9	1,7349	2,0879	2,7004	3,3251	16,9190	19,0228	21,6660	23,5893
10	2,1558	2,5582	3,2470	3,9403	18,3070	20,4832	23,2093	25,1881

Per leggere la tabella, troviamo la riga che ci interessa cercando i k gradi di libertà nella prima colonna. Poi scegliamo un valore di p . Per esempio, avendo $k = 5$ gradi di libertà e scegliendo $p = 0,05$, ci serve l'intersezione della quinta colonna e della quinta riga (evidenziata nella Tabella 1-6): troviamo che $x = 11,0705$. Ciò significa che per una variabile casuale chi-quadro e 5 gradi di libertà, la probabilità di estrarre a caso un valore di X uguale o maggiore di 11,0705 è 0,05. In altri termini, l'area sottesa dalla curva e corrispondente a valori chi-quadro maggiori o uguali a 11,0705 è pari al 5% dell'area totale.

Osservando una variabile casuale chi-quadro con 5 gradi di libertà, la probabilità che assuma un valore maggiore di 6,1 è maggiore o minore di 0,05?

DISTRIBUZIONI F

La distribuzione F è solo un rapporto di due distribuzioni chi-quadrato separate, e viene utilizzata per confrontare la varianza di due campioni. Come risultato, ha due diversi gradi di libertà, uno per ciascun campione.

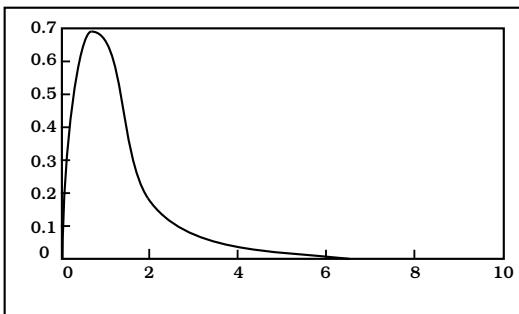
Ecco la funzione di densità di probabilità per una distribuzione F:

$$f(x) = \begin{cases} \frac{\left(\int_0^{\infty} x^{\frac{v_1+v_2}{2}-1} e^{-x} dx \right) \times (v_1)^{\frac{v_1}{2}} \times (v_2)^{\frac{v_2}{2}}}{\left(\int_0^{\infty} x^{\frac{v_1}{2}-1} e^{-x} dx \right) \times \left(\int_0^{\infty} x^{\frac{v_2}{2}-1} e^{-x} dx \right)} \times \frac{x^{\frac{v_1}{2}-1}}{(v_1 \times x + v_2)^{\frac{v_1+v_2}{2}}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Se la funzione di densità di probabilità di X è quella mostrata qui sopra, gli statistici dicono che "X segue una distribuzione F con gradi di libertà v_1 e v_2 ".

Invertendo il ruolo di v_1 e v_2 si ottengono curve leggermente diverse, come mostra la Figura 1-5 nel caso di $v_1 = 5$ e $v_2 = 10$ e di $v_1 = 10$ e $v_2 = 5$.

Se $v_1 = 5$ e $v_2 = 10$



Se $v_1 = 10$ e $v_2 = 5$

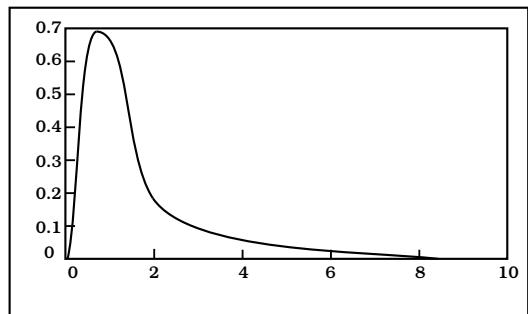


FIGURA 1-5: DENSITÀ DELLA DISTRIBUZIONE F PER 5 E 10 GRADI DI LIBERTÀ (A SINISTRA) O 10 E 5 GRADI DI LIBERTÀ (A DESTRA)

La Figura 1-6 riporta il grafico di una distribuzione F con gradi di libertà v_1 e v_2 . Il valore F è un punto sull'asse orizzontale, mentre l'area totale della zona tratteggiata a destra è la probabilità P che la variabile con distribuzione F assuma un valore maggiore del valore di F scelto.

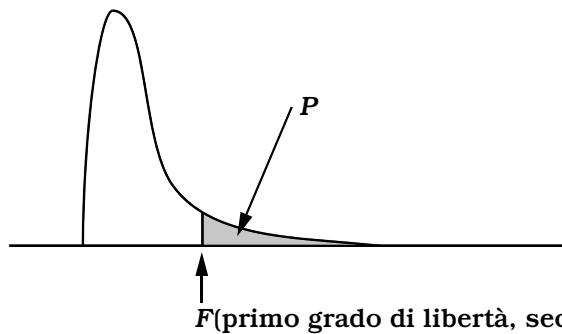


FIGURA 1-6: PROBABILITÀ CHE UN VALORE x SIA MAGGIORRE DEL VALORE CRITICO F

La Tabella 1-9 mostra la tavola di distribuzione della funzione F per $p = 0,05$.

TABELLA 1-9: TAVOLA DI DISTRIBUZIONE DELLA DENSITÀ F PER $p = 0,05$

v_1	1	2	3	4	5	6	7	8	9	10
v_2										
1	161,4	199,5	215,7	224,6	230,2	264,0	236,8	238,9	240,5	241,9
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4
3	10,1	9,6	9,3	9,1	9,0	8,9	8,9	8,8	8,8	8,8
4	7,7	6,9	6,6	6,4	6,3	6,2	6,1	6,0	6,0	6,0
5	6,6	5,8	5,4	5,2	5,1	5,0	4,9	4,8	4,8	4,7
6	6,0	5,1	4,8	4,5	4,4	4,3	4,2	4,1	4,1	4,1
7	5,6	4,7	4,3	4,1	4,0	3,9	3,8	3,7	3,7	3,6
8	5,3	4,5	4,1	3,8	3,7	3,6	3,5	3,4	3,4	3,3
9	5,1	4,3	3,9	3,6	3,5	3,4	3,3	3,2	3,2	3,1
10	5,0	4,1	3,7	3,5	3,3	3,2	3,1	3,1	3,0	3,0
11	4,8	4,0	3,6	3,4	3,2	3,1	3,1	2,9	2,9	2,9
12	4,7	3,9	3,5	3,3	3,1	3,0	2,9	2,8	2,8	2,8

Questa tavola si usa come quella della distribuzione chi-quadro, tranne che stavolta le colonne corrispondono ai gradi di libertà di un campione e le righe a quelli dell'altro. Si usa una tavola differente per ogni p -value di uso comune.

Se $v_1 = 1$ e $v_2 = 12$, la Tabella 1-7 riporta un valore critico pari a 4,7. Ciò significa che, verificando un'ipotesi, calcoliamo la statistica test e la confrontiamo con il valore critico di 4,7 ottenuto dalla tabella; se la statistica test calcolata è maggiore di 4,7, possiamo concludere che il nostro risultato è *statisticamente significativo*. Per qualsiasi statistica test maggiore dei valori di questa tabella, il p -value è minore di 0,05. Ciò significa che se $v_1 = 1$ e $v_2 = 12$, la probabilità che F sia uguale o maggiore a 4,7 quando l'ipotesi nulla è vera è del 5%, quindi c'è solo un 5% di probabilità di rifiutare l'ipotesi nulla quando in realtà è vera.

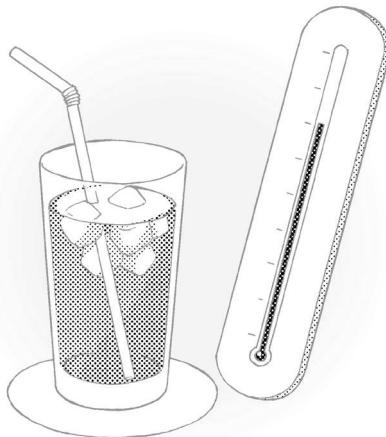
Consideriamo un altro esempio. La Tabella 1-10 mostra la distribuzione F per $p = 0,01$.

TABELLA 1-10: TAVOLA DI DISTRIBUZIONE DELLA DENSITÀ F PER $p = 0,01$

$v_1 \backslash v_2$	1	2	3	4	5	6	7	8	9	10
1	4052,2	4999,3	5403,5	5624,3	5764,0	5859,0	5928,3	5981,0	6022,4	6055,9
2	98,5	99,0	99,2	99,3	99,3	99,3	99,4	99,4	99,4	99,4
3	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2
4	21,2	18,8	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,5
5	16,3	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1
6	13,7	10,9	9,8	9,1	8,7	8,5	8,3	8,1	8,0	7,9
7	12,2	9,5	8,5	7,8	7,5	7,2	7,0	6,8	6,7	6,6
8	11,3	8,6	7,6	7,0	6,6	6,4	6,2	6,0	5,9	5,8
9	10,6	8,0	7,0	6,4	6,1	5,8	5,6	5,5	5,4	5,6
10	10,0	7,6	6,6	6,0	5,6	5,4	5,2	5,1	4,9	4,8
11	9,6	7,2	6,2	5,7	5,3	5,1	4,9	4,7	4,6	4,5
12	9,3	6,9	6,0	5,4	5,1	4,8	4,6	4,5	4,4	4,3

Se $v_1 = 1$ e $v_2 = 12$, stavolta il valore critico è 9,3. La probabilità che una statistica campione sia uguale o maggiore di 9,3 se l'ipotesi nulla è vera è soltanto dello 0,01. Quindi è molto improbabile sbagliare rifiutando l'ipotesi nulla. Osservate che il valore critico per $p = 0,01$ è maggiore che per $p = 0,05$: mantenendo costanti v_1 e v_2 , il valore critico aumenta al diminuire del p -value.

Z
ANALISI
DI REGRESSIONE
SEMPLICE



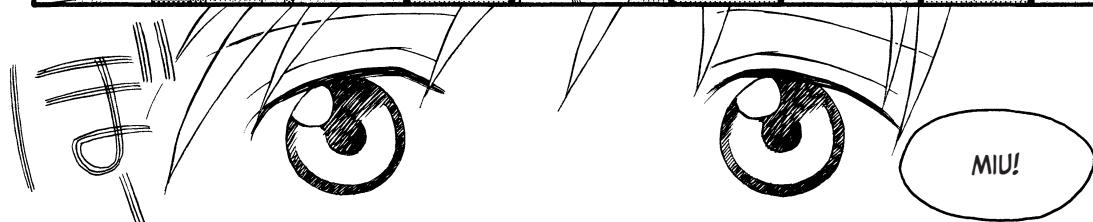
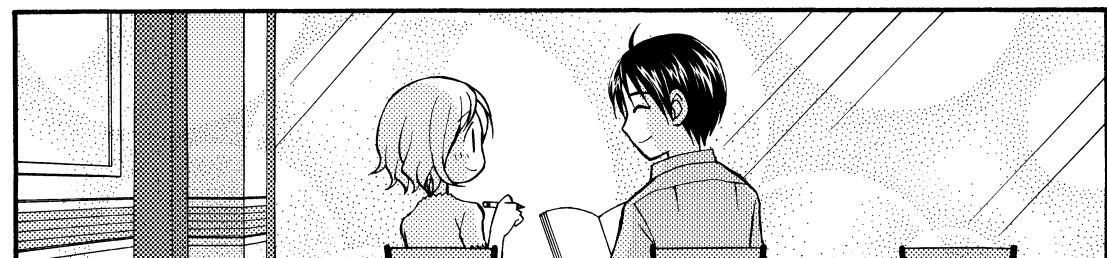
PRIMI PASSI

E QUINDI...

C'È UN LEGAME TRA I DUE,
GIUSTO?

ESATTO!

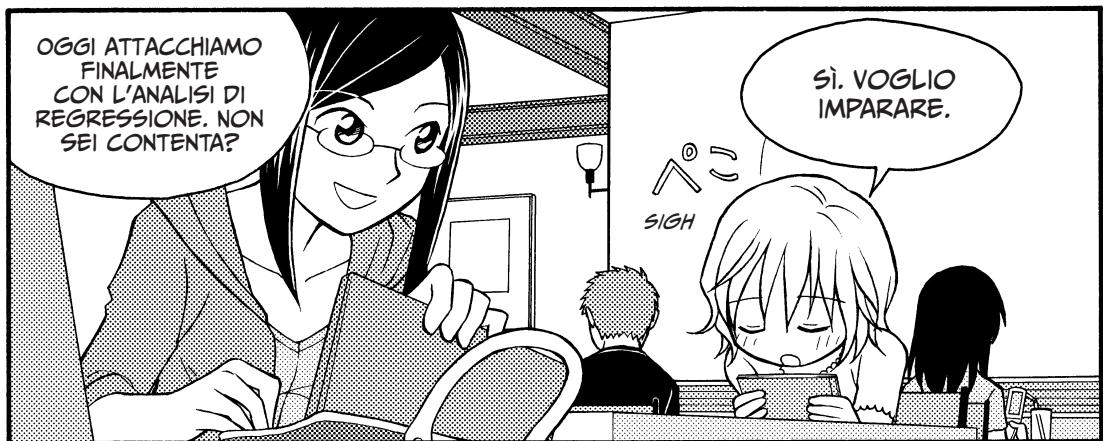
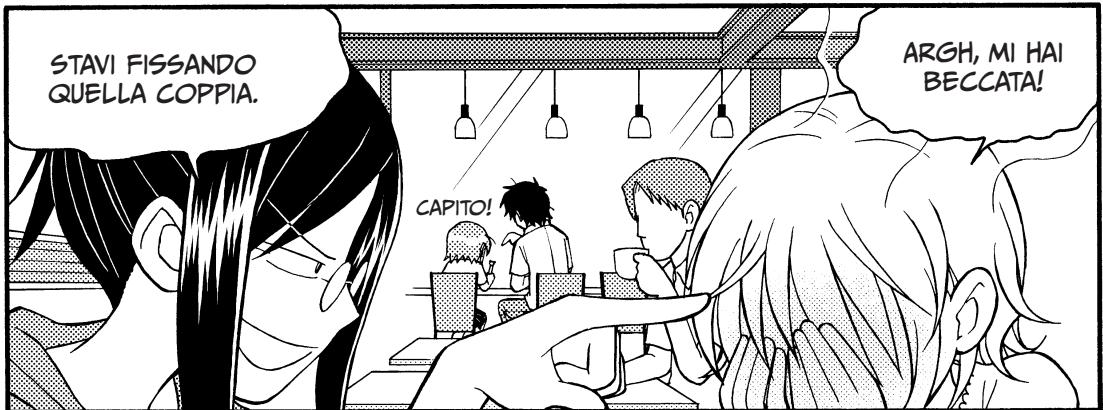
COME FAI A
SAPERE TANTE
COSE SULL'ANALISI DI
REGRESSIONE, MIU?



は

SGUARDO
PERSO

TERRA
CHIAMA MIU! MI
RICEVI?



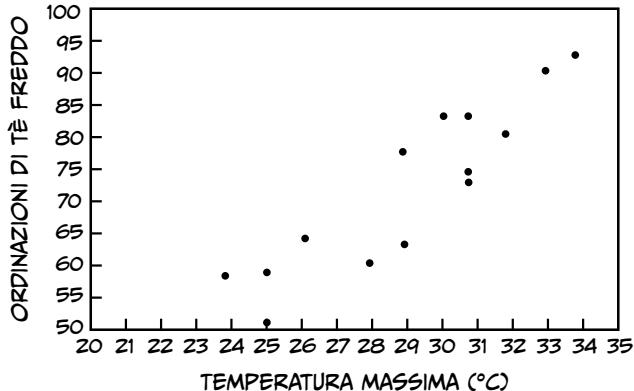
BENE, ALLORA
INIZIAMO! QUESTA
TABELLA RIPORTA LE
TEMPERATURE MASSIME
E LE ORDINAZIONI DI TÈ
FREDDO NEL CORSO DI
DUE SETTIMANE.

Data	Temperatura massima (°C)	Ordinazioni di tè freddo
Lun 22	29	77
Mar 23	28	62
Mer 24	34	93
Gio 25	31	84
Ven 26	25	59
Sab 27	29	64
Dom 28	32	80
Lun 29	31	75
Mar 30	24	58
Mer 31	33	91
Gio 1	25	51
Ven 2	31	73
Sab 3	26	65
Dom 4	30	84

GRAFICO DEI DATI

ORA...

...PRIMA DI TUTTO
NE RICAVIAMO IL
GRAFICO DI DI-
SPERSIONE...



...COSÌ.

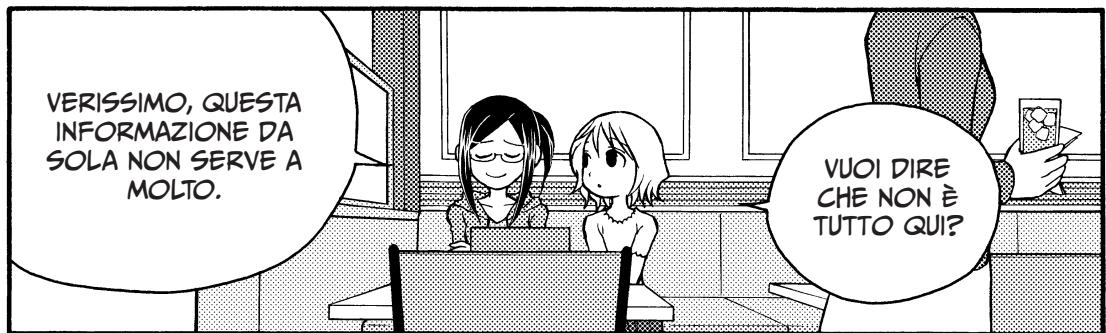
D'AC-
CORDO.

VEDI CHE I PUNTI SONO
GROSSO MODO ALLINEATI?
QUESTO FA PENSARE CHE LE
VARIABILI SIANO CORRELATE.
IL COEFFICIENTE DI CORRE-
LAZIONE, R , INDICA LA FORZA
DELLA CORRELAZIONE.

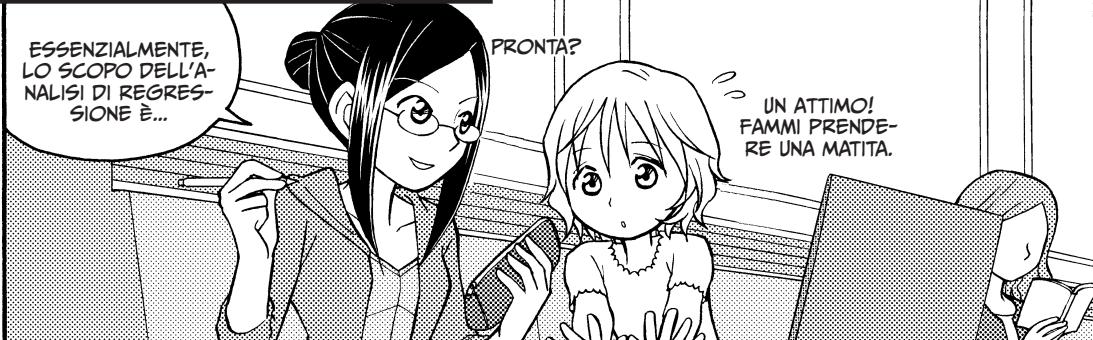
$$R = 0,9069$$

R VA DA +1 A -1,
E PIÙ SI ALLONTANA DA
ZERO, PIÙ È FORTE LA
CORRELAZIONE*. A PAGI-
NA 78 VEDREMO COME
RICAVARE IL COEFFI-
CIEN-
TE DI CORRELAZIONE.

*SE R È POSITIVO LA
CORRELAZIONE È POSITIVA, CIOÈ
 y AUMENTA ALL'AUMENTARE DI x .
SE R È NEGATIVO, y DIMINUISCE
ALL'AUMENTARE DI x .

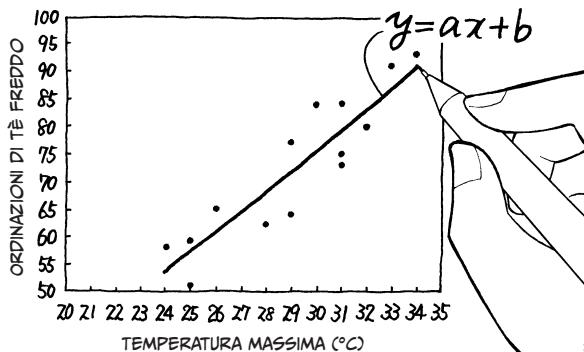
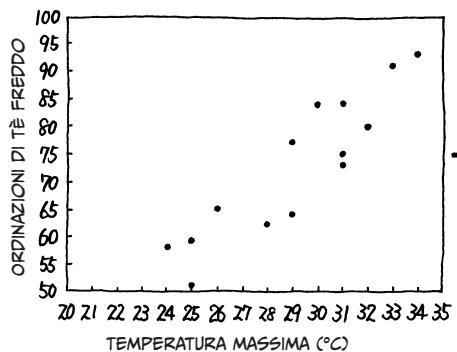


EQUAZIONE DI REGRESSIONE



...OTTENERE L'EQUAZIONE
DI REGRESSIONE...

...NELLA FORMA
 $y = ax + b$.



E QUESTO
A CHE CI SERVE?

SE METTI UN VALORE DELLA
TEMPERATURA MASSIMA AL
POSTO DI x ...

$y = ax + b$

SCRIB SCRIB

...PUOI
PREVEDERE
IL NUMERO DI
ORDINAZIONI DI
TÈ FREDDO (y).



CAPISCO!
L'ANALISI DI
REGRESSIONE
NON SEMBRA
TANTO DIFFICILE.

COME DICEVO PRIMA, y È LA VARIABILE DIPENDENTE O VARIABILE RESPONSO, E x È LA VARIABILE INDEPENDENTE O PREDITTORE.

$$y = ax + b$$

VARIABILE DIPENDENTE

VARIABILE INDIPENDENTE

a È IL COEFFICIENTE DI REGRESSIONE, CHE CI DÀ LA PENDENZA DELLA RETTA NEL GRAFICO.



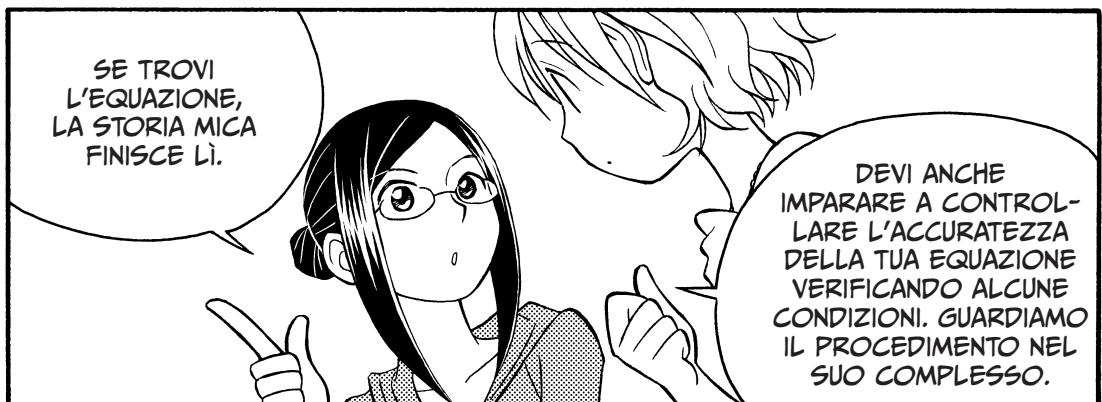
RIMANE D, L'INTER-
CETTA, CHE CI DICE
IL PUNTO IN CUI LA
RETTA INCONTRA
L'ASSE y.

OK, CI SONO.



ALLORA COME OTTENGO L'EQUAZIONE DI REGRESSIONE?

CALMA, MIU.

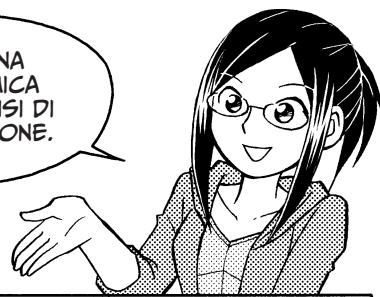


SE TROVI
L'EQUAZIONE,
LA STORIA MICA
FINISCE LÌ.

DEVI ANCHE
IMPARARE A CONTROL-
LARE L'ACCURATEZZA
DELLA TUA EQUAZIONE
VERIFICANDO ALCUNE
CONDIZIONI. GUARDIAMO
IL PROCEDIMENTO NEL
SUO COMPLESSO.

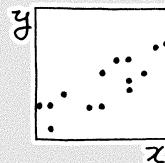
PROCEDIMENTO GENERALE PER L'ANALISI DI REGRESSIONE

ECCO UNA PANORAMICA DELL'ANALISI DI REGRESSIONE.



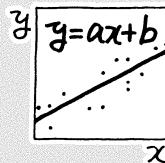
PASSO 1

TRACCIARE IL GRAFICO DI DISPERSIONE DELLA VARIABILE DIPENDENTE IN FUNZIONE DI QUELLA INDIPENDENTE. SE I PUNTI SI ALLINEANO, POTREBBERE ESSERCI CORRELAZIONE TRA LE VARIABILI



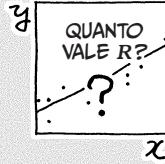
PASSO 2

CALCOLO DELL'EQUAZIONE DI REGRESSIONE



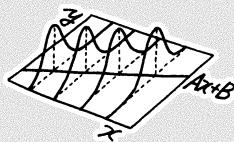
PASSO 3

CALCOLO DEL COEFFICIENTE DI CORRELAZIONE (r) E VALUTAZIONE DELLA POPOLAZIONE E DELLE IPOTESI



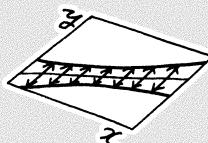
PASSO 4

ANALISI DELLA VARIANZA



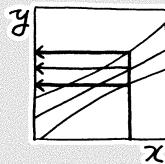
PASSO 5

CALCOLO DEGLI INTERVALLI DI CONFIDENZA

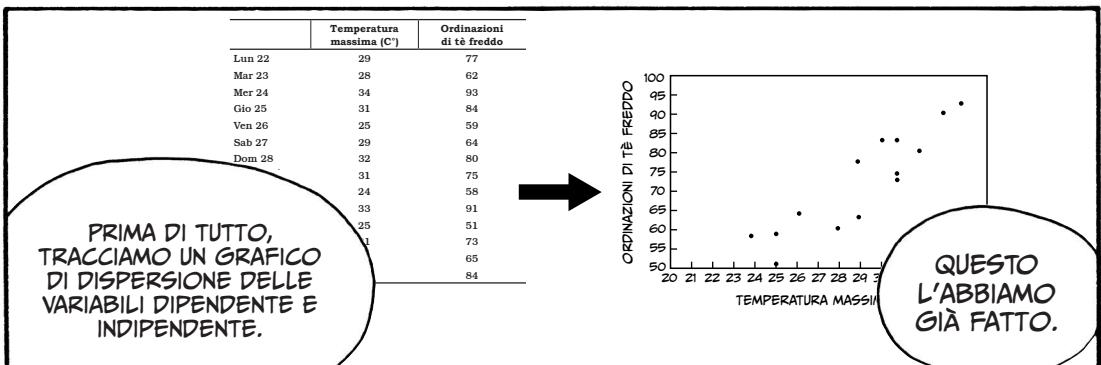
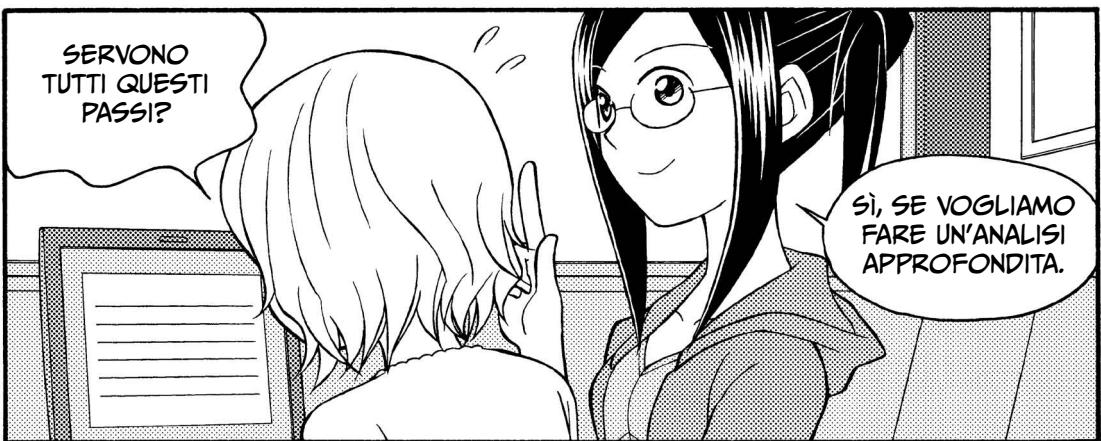


PASSO 6

FACCIAMO UNA PREVISIONE!



DIAGNOSTICA DI REGRESSIONE

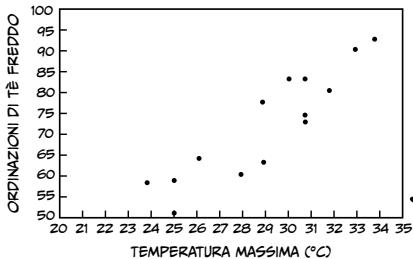


TRACCIANDO UN GRAFICO DELLE ORDINAZIONI DI TÈ FREDDO RISPETTO ALLE MASSIME GIORNALIERE, I PUNTI SEMBRANO ALLINEARSI.

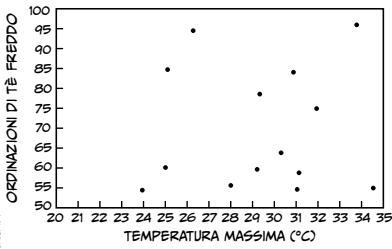
E SAPPIAMO GIÀ CHE R È 0,9069, ABBASTANZA ALTO.

SEMPRA CHE QUESTE VARIABILI SIANO CORRELATE.

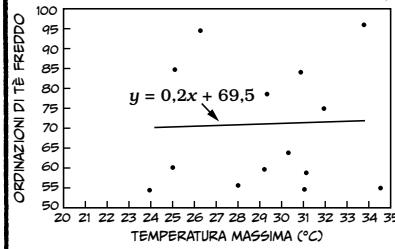
MA TUTTI QUESTI PUNTI SERVONO DAVVERO A QUALCOSA? PERCHÉ NON CALCOLIAMO R E BASTA?



LA FORMA DEI DATI È IMPORTANTE!

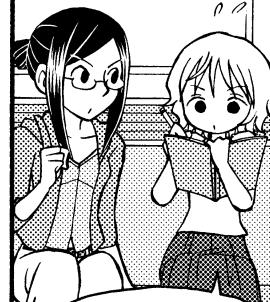


GUARDA QUESTO GRAFICO. INVECE DI DISPORSI LUNGO UNA RETTA, I PUNTI SONO SPARPAGLIATI A CASACCIO.



L'EQUAZIONE DI REGRESSIONE LA PUOI COMUNQUE RICAVARE, MA NON SIGNIFICA NULLA. IL BASSO VALORE DI R LO CONFERMA, MA IL GRAFIKO DI DISPERSIONE TE NE FA ACCORGERE A OCCHIO NUDO.

TRACCI SEMPRE IL GRAFICO PER PRIMA COSA, PER FARTI UN'IDEA DELLA DISPOSIZIONE DEI DATI.



PASSO 2: CALCOLO DELL'EQUAZIONE DI REGRESSIONE



ORA RICAVEREMO L'EQUAZIONE DI REGRESSIONE!

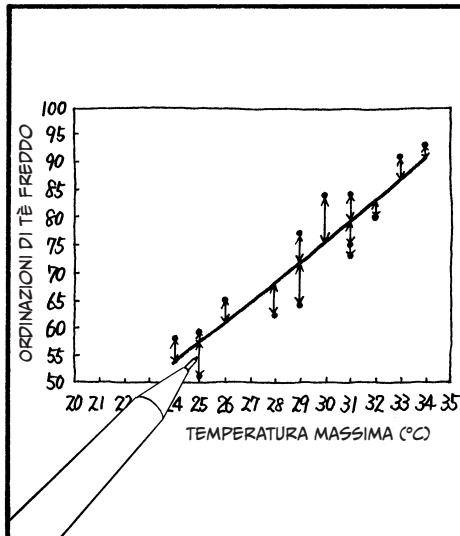
TROVIAMO A E B!

$$y = ax + b$$

FINALMENTE, È VENUTO IL MOMENTO.



TRACCIAMO UNA RETTA SEGUENDO L'ANDAMENTO DEI DATI PER QUANTO POSSIBILE.



LE FRECCETTE RAPPRESENTANO LA DISTANZA FRA I PUNTI DELLA RETTA, CHE CORRISPONDONO AI VALORI STIMATI, E I PUNTI REALMENTE OSSERVATI. QUESTE DISTANZE SONO DETTE RESIDUI; VOGLIAMO TROVARE LA RETTA CHE LI MINIMIZZA.

QUESTA È DETTA REGRESSIONE LINEARE DEI MINIMI QUADRATI.



ELEVIAMO I RESIDUI AL QUADRATO PER CALCOLARE LA SOMMA DEI QUADRATI, CHE CI SERVIRÀ PER TROVARE L'EQUAZIONE DI REGRESSIONE.

Passo 1

Calcolare S_{xx} (somma dei quadrati degli scarti in x), S_{yy} (somma dei quadrati degli scarti in y) e S_{xy} (somma dei prodotti degli scarti in x e y).

Passo 2

Calcolare S_e (somma dei residui al quadrato).

Passo 3

Calcolare le derivate di S_e rispetto ad a e b ed egualiarle a 0.

Passo 4

Raccogliere i termini in a e b .

Passo 5

Isolare la componente a .

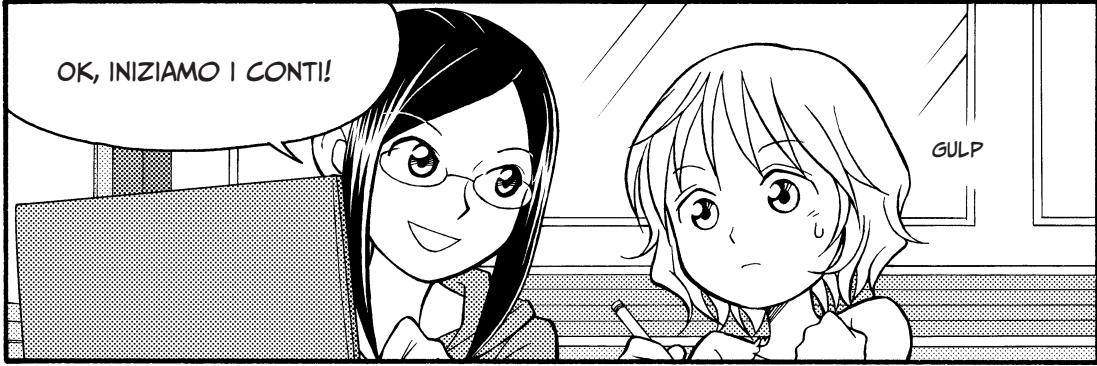
Passo 6

Trovare l'equazione di regressione.



ORA MI SEGNO TUTTO QUANTO.





OK, INIZIAMO I CONTI!

Passo 1

Determinare:

- La somma dei quadrati degli scarti in x , S_{xx} : $\sum (x - \bar{x})^2$
- La somma dei quadrati degli scarti in y , S_{yy} : $\sum (y - \bar{y})^2$
- La somma dei prodotti degli scarti in x E y , S_{xy} : $\sum (x - \bar{x})(y - \bar{y})$

Nota: la barra sopra la variabile (come in \bar{x}) è una notazione che indica la *media*; si legge “ x barrato”).

	Temperatura massima (°C)	Ordinazioni di tè freddo	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
Lun 22	29	77	-0,1	4,4	0,0	19,6	-0,6		
Mar 23	28	62	-1,1	-10,6	1,3	111,8	12,1		
Mer 24	34	93	4,9	20,4	23,6	417,3	99,2		
Gio 25	31	84	1,9	11,4	3,4	130,6	21,2		
Ven 26	25	59	-4,1	-13,6	17,2	184,2	56,2		
Sab 27	29	64	-0,1	-8,6	0,0	73,5	1,2		
Dom 28	32	80	2,9	7,4	8,2	55,2	21,2		
Lun 29	31	75	1,9	2,4	3,4	5,9	4,5		
Mar 30	24	58	-5,1	-14,6	26,4	212,3	74,9		
Mer 31	33	91	3,9	18,4	14,9	339,6	71,1		
Gio 1	25	51	-4,1	-21,6	17,2	465,3	89,4		
Ven 2	31	73	1,9	0,4	3,4	0,2	0,8		
Sab 3	26	65	-3,1	-7,6	9,9	57,8	23,8		
Dom 4	30	84	0,9	11,4	0,7	130,6	9,8		
Somma	408	1016	0	0	129,7	2203,4	484,9		
Media	29,1	72,6							
	\downarrow	\downarrow							
	\bar{x}	\bar{y}							
					\downarrow	\downarrow	\downarrow		
					S_{xx}	S_{yy}	S_{xy}		

*ALCUNE CIFRE DI QUESTO CAPITOLO SONO ARROTONDATE PER MAGGIOR LEGGIBILITÀ, MA SALVO INDICAZIONE CONTRARIA I CALCOLI SONO SEMPRE SVOLTI USANDO I VALORI NON ARROTONDATI.

Passo 2

Determinare la somma dei residui al quadrato, S_e :

- y è il valore osservato;
- \hat{y} è il valore stimato a partire dall'equazione di regressione;
- $y - \hat{y}$ è detto **residuo** e si indica con e .

Nota: il simbolo sopra la \hat{y} è familiarmente detto **cappello**, e il parametro stimato prende il nome di y "cappello".

	Temperatura massima (°C) x	Ordinazioni reali di tè freddo y	Ordinazioni stimate di tè freddo $\hat{y} = ax + b$	Residui (e) $y - \hat{y}$	Residui al quadrato $(y - \hat{y})^2$
Lun 22	29	77	$a \times 29 + b$	$77 - (a \times 29 + b)$	$[77 - (a \times 29 + b)]^2$
Mar 23	28	62	$a \times 28 + b$	$62 - (a \times 28 + b)$	$[62 - (a \times 28 + b)]^2$
Mer 24	34	93	$a \times 34 + b$	$93 - (a \times 34 + b)$	$[93 - (a \times 34 + b)]^2$
Gio 25	31	84	$a \times 31 + b$	$84 - (a \times 31 + b)$	$[84 - (a \times 31 + b)]^2$
Ven 26	25	59	$a \times 25 + b$	$59 - (a \times 25 + b)$	$[59 - (a \times 25 + b)]^2$
Sab 27	29	64	$a \times 29 + b$	$64 - (a \times 29 + b)$	$[64 - (a \times 29 + b)]^2$
Dom 28	32	80	$a \times 32 + b$	$80 - (a \times 32 + b)$	$[80 - (a \times 32 + b)]^2$
Lun 29	31	75	$a \times 31 + b$	$75 - (a \times 31 + b)$	$[75 - (a \times 31 + b)]^2$
Mar 30	24	58	$a \times 24 + b$	$58 - (a \times 24 + b)$	$[58 - (a \times 24 + b)]^2$
Mer 31	33	91	$a \times 33 + b$	$91 - (a \times 33 + b)$	$[91 - (a \times 33 + b)]^2$
Gio 1	25	51	$a \times 25 + b$	$51 - (a \times 25 + b)$	$[51 - (a \times 25 + b)]^2$
Ven 2	31	73	$a \times 31 + b$	$73 - (a \times 31 + b)$	$[73 - (a \times 31 + b)]^2$
Sab 3	26	65	$a \times 26 + b$	$65 - (a \times 26 + b)$	$[65 - (a \times 26 + b)]^2$
Dom 4	30	84	$a \times 30 + b$	$84 - (a \times 30 + b)$	$[84 - (a \times 30 + b)]^2$
Somma	408	1016	$408a + 14b$	$1016 - (408a + 14b)$	S_e

$$\begin{array}{cccccc}
\text{Media} & 29,1 & 72,6 & 29,1a + b & 72,6 - (29,1a + b) & = \frac{S_e}{14} \\
& & & = \bar{x}a + b & = \bar{y} - (\bar{x}a + b) & \\
\downarrow & \downarrow & & & & \\
\bar{x} & \bar{y} & & & &
\end{array}$$

$$S_e = [77 - (a \times 29 + b)]^2 + \dots + [84 - (a \times 30 + b)]^2$$

S_e , indicato anche con RSS, è la somma dei quadrati dei residui.



Passo 3

Derivare S_e rispetto ad a e a b , e uguagliare a zero le due derivate. Derivando $y = (ax + b)^n$ rispetto a x , si ottiene $\frac{dy}{dx} = n(ax + b)^{n-1} \times a$.

- Derivare rispetto ad a :

$$\frac{dS_e}{da} = 2[77 - (29a + b)] \times (-29) + \dots + 2[84 - (30a + b)] \times (-30) = 0 \quad \textcircled{1}$$

- Derivare rispetto a b :

$$\frac{dS_e}{db} = 2[77 - (29a + b)] \times (-1) + \dots + 2[84 - (30a + b)] \times (-1) = 0 \quad \textcircled{2}$$

Passo 4

Semplificare le espressioni **1** e **2** del passo precedente.

Semplificare **1**:

$$\begin{aligned} & 2[77 - (29a + b)] \times (-29) + \dots + 2[84 - (30a + b)] \times (-30) = 0 \\ & [77 - (29a + b)] \times (-29) + \dots + [84 - (30a + b)] \times (-30) = 0 \quad \text{DIVIDERE PER 2 ENTRAMBI I MEMBRI.} \\ & 29[(29a + b) - 77] + \dots + 30[(30a + b) - 84] = 0 \quad \text{MOLTIPLICARE PER -1.} \\ & (29 \times 29a + 29 \times b - 29 \times 77) + \dots + (30 \times 30a + 30 \times b - 30 \times 84) = 0 \quad \text{MOLTIPLICARE.} \\ \textcircled{3} \quad & (29^2 + \dots + 30^2)a + (29 + \dots + 30)b - (29 \times 77 + \dots + 30 \times 84) = 0 \quad \text{RACCOGLIERE I TERMINI IN } a \text{ E } b. \end{aligned}$$

Semplificare **2**:

$$\begin{aligned} & 2[77 - (29a + b)] \times (-1) + \dots + 2[84 - (30a + b)] \times (-1) = 0 \\ & [77 - (29a + b)] \times (-1) + \dots + [84 - (30a + b)] \times (-1) = 0 \quad \text{DIVIDERE PER 2 ENTRAMBI I MEMBRI.} \\ & [(29a + b) - 77] + \dots + [(30a + b) - 84] = 0 \quad \text{MOLTIPLICARE PER -1.} \\ & (29 + \dots + 30)a + \underbrace{b + \dots + b}_{14} - (77 + \dots + 84) = 0 \quad \text{RACCOGLIERE I TERMINI IN } a \text{ E } b. \\ & (29 + \dots + 30)a + 14b - (77 + \dots + 84) = 0 \\ & 14b = (77 + \dots + 84) - (29 + \dots + 30)a \quad \text{SOTTRARRE } 14b \text{ A ENTRAMBI I MEMBRI E MOLTIPLICARE PER -1.} \\ \textcircled{4} \quad & b = \frac{77 + \dots + 84}{14} - \frac{29 + \dots + 30}{14}a \quad \text{RISOLVERE PER } b. \\ \textcircled{5} \quad & b = \bar{y} - \bar{x}a \quad \text{I COEFFICIENTI DI } \textcircled{3} \text{ SONO LE MEDIE DI } y \text{ E } x. \end{aligned}$$

Passo 5

Sostituire il valore di b trovato alla riga ④ nella riga ③ (③ e ④ sono i risultati del Passo 4).

$$③ (29^2 + \dots + 30^2)a + (29 + \dots + 30) \left(\frac{77 + \dots + 84}{14} - \frac{29 + \dots + 30}{14}a \right) - (29 \times 77 + \dots + 30 \times 84) = 0$$

ORA È RIMASTA SOLO LA VARIABILE a .

$$(29^2 + \dots + 30^2)a + \frac{(29 + \dots + 30)(77 + \dots + 84)}{14} - \frac{(29 + \dots + 30)^2}{14}a - (29 \times 77 + \dots + 30 \times 84) = 0$$

$$\left[(29^2 + \dots + 30^2) - \frac{(29 + \dots + 30)^2}{14} \right] a + \frac{(29 + \dots + 30)(77 + \dots + 84)}{14} - (29 \times 77 + \dots + 30 \times 84) = 0$$

RACCOGLIERE I TERMINI IN a .

$$\left[(29^2 + \dots + 30^2) - \frac{(29 + \dots + 30)^2}{14} \right] a = (29 \times 77 + \dots + 30 \times 84) - \frac{(29 + \dots + 30)(77 + \dots + 84)}{14}$$

ISOLARE I TERMINI IN a .

Semplificare il lato sinistro dell'equazione.

$$(29^2 + \dots + 30^2) - \frac{(29 + \dots + 30)^2}{14}$$

$$= (29^2 + \dots + 30^2) - 2 \times \frac{(29 + \dots + 30)^2}{14} + \frac{(29 + \dots + 30)^2}{14}$$

SOMMIAMO E SOTTRAIAMO $\frac{(29 + \dots + 30)^2}{14}$.

$$= (29^2 + \dots + 30^2) - 2 \times (29 + \dots + 30) \times \frac{29 + \dots + 30}{14} + \left(\frac{29 + \dots + 30}{14} \right)^2 \times 14$$

MOLTIPLICHIAMO L'ULTIMO TERMINE PER $\frac{14}{14}$.

$$= (29^2 + \dots + 30^2) - 2 \times (29 + \dots + 30) \times \bar{x} + (\bar{x})^2 \times 14$$

$$\bar{x} = \frac{29 + \dots + 30}{14}$$

$$= (29^2 + \dots + 30^2) - 2 \times (29 + \dots + 30) \times \bar{x} + \underbrace{(\bar{x})^2 + \dots + (\bar{x})^2}_{14}$$

$$= [29^2 - 2 \times 29 \times \bar{x} + (\bar{x})^2] + \dots + [30^2 - 2 \times 30 \times \bar{x} + (\bar{x})^2]$$

$$= (29 - \bar{x})^2 + \dots + (30 - \bar{x})^2$$

$$= S_{xx}$$

Semplificare il lato destro dell'equazione.

$$(29 \times 77 + \dots + 30 \times 84) - \frac{(29 + \dots + 30)(77 + \dots + 84)}{14}$$

$$= (29 \times 77 + \dots + 30 \times 84) - \frac{29 + \dots + 30}{14} \times \frac{77 + \dots + 84}{14} \times 14$$

SOMMIAMO E SOTTRAIAMO $\bar{x} \times \bar{y} \times 14$.

$$= (29 \times 77 + \dots + 30 \times 84) - \bar{x} \times \bar{y} \times 14$$

$$= (29 \times 77 + \dots + 30 \times 84) - \bar{x} \times \bar{y} \times 14 - \bar{x} \times \bar{y} \times 14 + \bar{x} \times \bar{y} \times 14$$

$$= (29 \times 77 + \dots + 30 \times 84) - \frac{29 + \dots + 30}{14} \times \bar{y} \times 14 - \bar{x} \times \frac{77 + \dots + 84}{14} \times 14 + \bar{x} \times \bar{y} \times 14$$

$$= (29 \times 77 + \dots + 30 \times 84) - (29 + \dots + 30) \bar{y} - \bar{x} (77 + \dots + 84) + \bar{x} \times \bar{y} \times 14$$

$$= (29 \times 77 + \dots + 30 \times 84) - (29 + \dots + 30) \bar{y} - (77 + \dots + 84) \bar{x} + \underbrace{\bar{x} \times \bar{y} + \dots + \bar{x} \times \bar{y}}_{14}$$

$$= (29 - \bar{x})(77 - \bar{y}) + \dots + (30 - \bar{x})(84 - \bar{y})$$

$$= S_{xy}$$

⑥

$$a = \frac{S_{xy}}{S_{xx}}$$

ISOLIAMO a A MEMBRO SINISTRO.

Passo 6 Calcolare l'equazione di regressione.

Da **6** nel Passo 5, $a = \frac{S_{xy}}{S_{xx}}$. Da **3** nel Passo 4, $b = \bar{y} - \bar{x}a$.

Sostituendo i valori calcolati nel Passo 1,

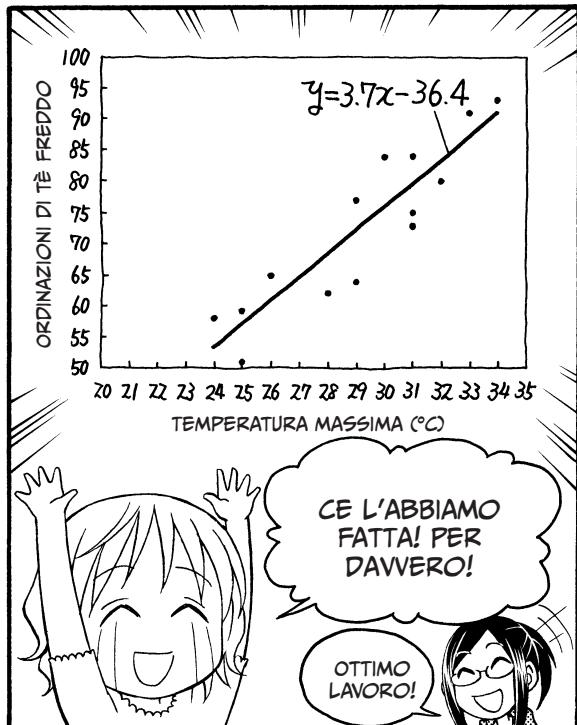
$$\begin{cases} a = \frac{S_{xx}}{S_{xy}} = \frac{484,9}{129,7} = 3,7 \\ b = \bar{y} - \bar{x}a = 72,6 - 29,1 \times 3,7 = -36,4 \end{cases}$$

si ottiene l'equazione di regressione:

$$y = 3,7x - 36,4.$$

Semplicissimo!

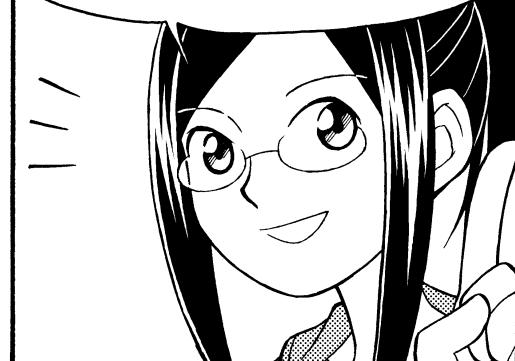
Nota: Questi valori sono stati arrotondati per leggibilità, ma il risultato (3,7, -36,4) si ottiene svolgendo i calcoli con i valori completi, non arrotondati.

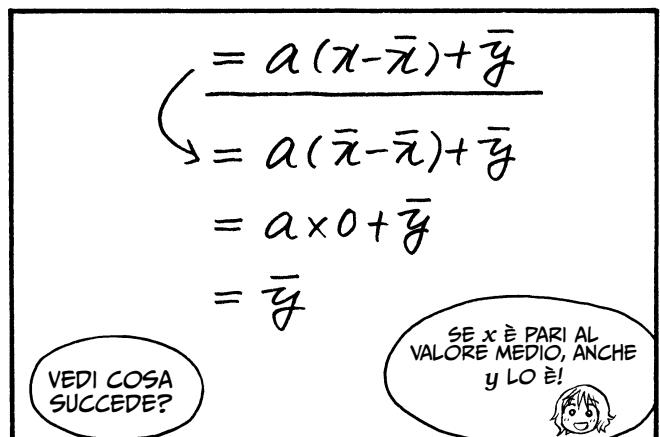
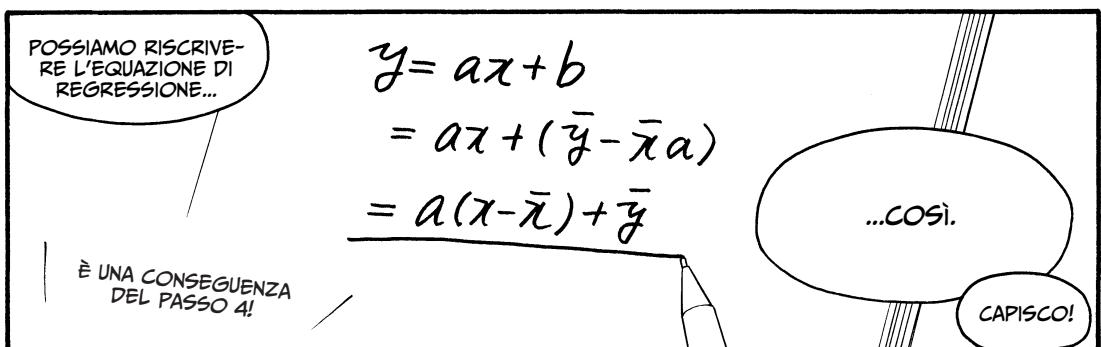
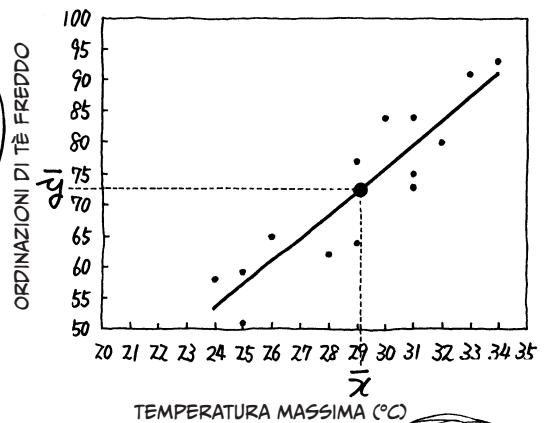


IL RAPPORTO TRA I RESIDUI, LA PENDENZA a E L'INTERCETTA b È SEMPRE

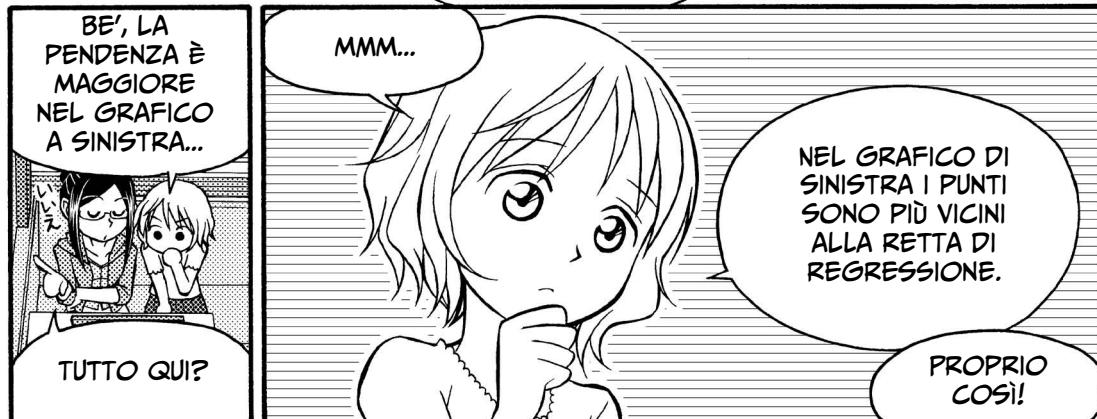
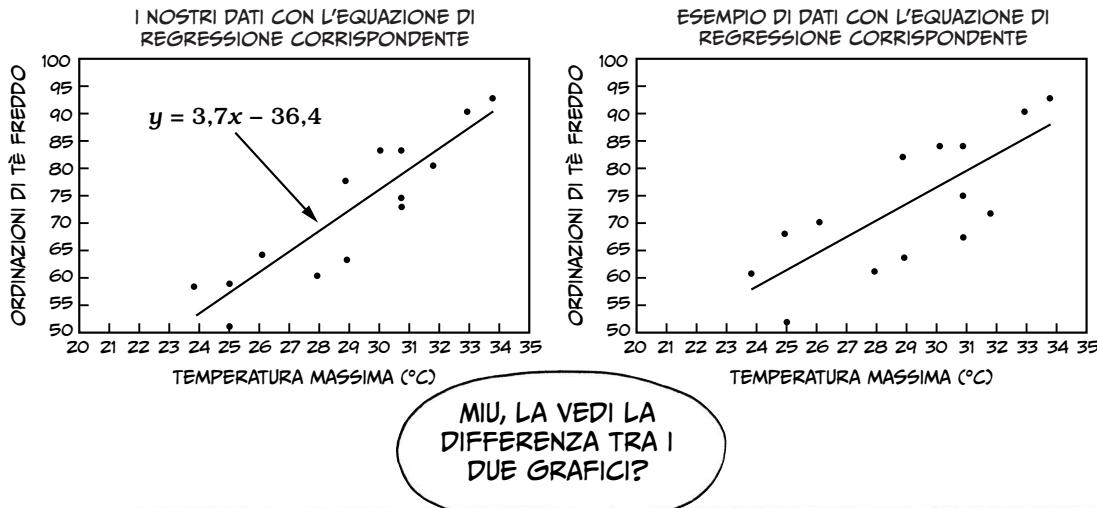
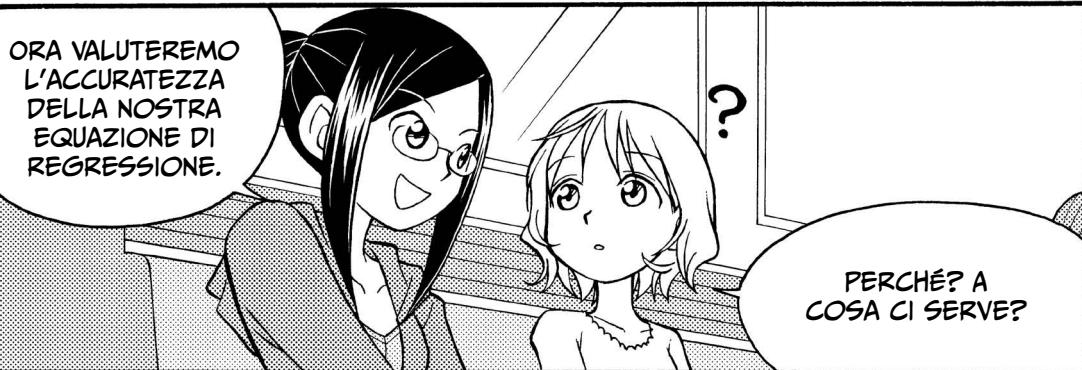
$$a = \frac{\text{somma dei prodotti di } x \text{ e } y}{\text{somma dei quadrati di } x} = \frac{S_{xy}}{S_{xx}}$$
$$b = \bar{y} - \bar{x}a$$

QUESTO VALE
PER TUTTE LE
REGRESSIONI LINEARI.





PASSO 3: CALCOLO DEL COEFFICIENTE DI CORRELAZIONE (R) E VALUTAZIONE DELLA POPOLAZIONE E DELLE IPOTESI



L'EQUAZIONE DI REGRESSIONE È PIÙ ACCURATA SE I VALORI STIMATI (I PUNTI DELLA RETTA) SONO PIÙ VICINI AI VALORI OSSERVATI (I PUNTI DEL GRAFICO).

CIOÈ ACCURATA SIGNIFICA REALISTICA?

INFATTI, L'ACCURATEZZA È IMPORTANTE, MA È TROPPO SOGGETTIVO VALUTARLA SOLTANTO GUARDANDO IL GRAFICO.

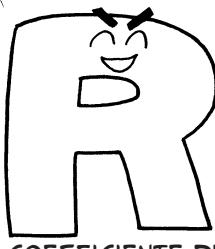
I PUNTI SONO VICINI.

I PUNTI SONO UN PO' LONTANNUCCI.

SÌ, È VERO.

ECCO PERCHÉ CI SERVE R!

TA-DA!



COEFFICIENTE DI CORRELAZIONE

È IL COEFFICIENTE DI CORRELAZIONE CHE ERA SPUNTATO FUORI PRIMA, NO?



ESATTO! R È UN INDICE CHE MISURA L'ACCURATEZZA DELL'EQUAZIONE DI REGRESSIONE. QUESTO INDICE CONFRONTA I DATI E LE PREVISIONI, CIOÈ I VALORI MISURATI x E y E QUELLI STIMATI \hat{x} E \hat{y} .

R È ANCHE DETTO COEFFICIENTE DI CORRELAZIONE DI PEARSON, IN ONORE DEL MATEMATICO KARL PEARSON.

OK!

ECCO LA FORMULA. I CONTI SOMIGLIANO A QUELLI DI PRIMA PER S_{xx} E S_{xy} .

$$R = \frac{\text{somma dei prodotti degli scarti di } y \text{ e } \hat{y}}{\sqrt{\text{somma dei quadrati degli scarti in } y \times \text{somma dei quadrati degli scarti in } \hat{y}}} = \frac{S_{y\hat{y}}}{\sqrt{S_{yy} \times S_{\hat{y}\hat{y}}}}$$

$$= \frac{1812,3}{\sqrt{2203,4 \times 1812,3}} = 0,9069$$



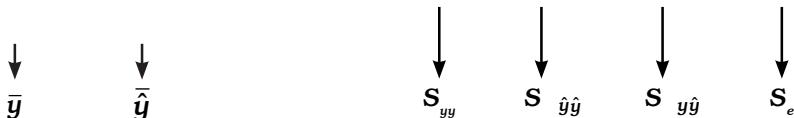
NIENTE DI TERRIBILE!



QUESTO MI RICORDA QUALCOSA

FUNZIONE DI REGRESSIONE!

	Valori osservati	Valori stimati $\hat{y} = 3,7x - 36,4$	$y - \bar{y}$	$\hat{y} - \bar{\hat{y}}$	$(y - \bar{y})^2$	$(\hat{y} - \bar{\hat{y}})^2$	$(y - \bar{y})(\hat{y} - \bar{\hat{y}})$	$(y - \hat{y})^2$
Lun 22	77	72,0	4,4	-0,5	19,6	0,3	-2,4	24,6
Mar 23	62	68,3	-10,6	-4,3	111,8	18,2	45,2	39,7
Mer 24	93	90,7	20,4	18,2	417,3	329,6	370,9	5,2
Gio 25	84	79,5	11,4	6,9	130,6	48,2	79,3	20,1
Ven 26	59	57,1	-13,6	-15,5	184,2	239,8	210,2	3,7
Sab 27	64	72,0	-8,6	-0,5	73,5	0,3	4,6	64,6
Dom 28	80	83,3	7,4	10,7	55,2	114,1	79,3	10,6
Lun 29	75	79,5	2,4	6,9	5,9	48,2	16,9	20,4
Mar 30	58	53,3	-14,6	-19,2	212,3	369,5	280,1	21,6
Mer 31	91	87,0	18,4	14,4	339,6	207,9	265,7	16,1
Gio 1	51	57,1	-21,6	-15,5	465,3	239,8	334,0	37,0
Ven 2	73	79,5	0,4	6,9	0,2	48,2	3,0	42,4
Sab 3	65	60,8	-7,6	-11,7	57,3	138,0	88,9	17,4
Dom 4	84	75,8	11,4	3,2	130,6	10,3	36,6	67,6
Somma	1016	1016	0	0	2203,4	1812,3	1812,3	391,1
Media	72,6	72,6						



IL CALCOLO DI R NON RICHIEDE S_e , MA L'HO INCLUSO PERCHÉ POI CI SERVIRÀ.

ELEVANDO R AL QUADRATO OTTENIAMO IL COEFFICIENTE DI DETERMINAZIONE, INDICATO CON R^2 .

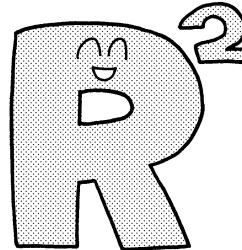
R^2 CI SEGNALA...

SONO UN COEFFICIENTE DI CORRELAZIONE.

ANCH'IO SONO UN COEFFICIENTE DI CORRELAZIONE.

$R \times R =$

...IN CHE MISURA LA NOSTRA EQUAZIONE DI REGRESSIONE RIPRODUCE LA VARIANZA DEI DATI.



SE R^2 È NULLO, LA VARIABILE PREDITTIVA NON RIESCE A PREVEDERE ACCURATAMENTE LA VARIABILE RESPONSO.

1

0

Più è accurata l'equazione di regressione, più R^2 si avvicina a 1, e viceversa.

Allora quanto deve essere grande R^2 per poter considerare accurata l'equazione di regressione?

Purtroppo non c'è un valore universalmente accettato.

MA IN GENERE VOGLIAMO CHE SIA ALMENO 0,5.



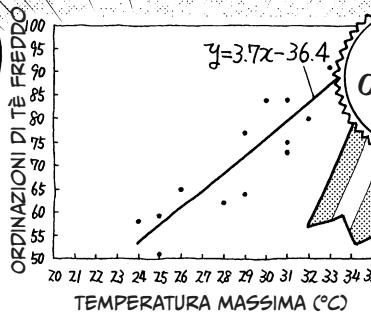
ORA PROVA A DETERMINARE IL VALORE DI R^2 .



$$R^2 = (0,9069)^2 \\ = 0,8225$$

VALE 0,8225.

IL VALORE DI R^2 È BEN MAGGIORI DI 0,5 PER LA NOSTRA EQUAZIONE, CHE QUINDI DOVREBBE STIMARE ABBASTANZA ACCURATAMENTE IL NUMERO DI ORDINAZIONI DI TÈ FREDDO.



$$R^2 = \left(\frac{\text{coefficiente di}}{\text{correlazione}} \right)^2 = \frac{a \times S_{xy}}{S_{yy}} = 1 - \frac{S_e}{S_{yy}}$$

SEGNATI QUESTA FORMULA. SI PUÒ CALCOLARE R^2 DIRETTAMENTE DA QUESTI VALORI. CON I DATI DELLA NOSTRA SALA DA TÈ
 $1 - (391,1 / 2203,4) = 8225!$

COMODO!

ABBIAMO COMPLETATO I PRIMI TRE PASSI.

EVVIVA!

CAMPIONI E POPOLAZIONI

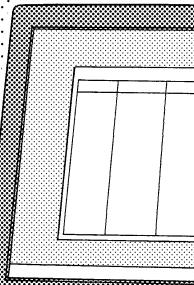
ORA VALUTIAMO LA POPOLAZIONE E VERIFICHIAMO CHE LE NOSTRE IPOTESI SIANO VALIDE!

OH...

VOLEVO GIUSTO CHIEDERTELLO. CHE POPOLAZIONE? I GIAPPONESI? TUTTI GLI ESSERI UMANI?

IN REALTÀ QUESTA POPOLAZIONE NON È FATTA DI PERSONE, MA DI DATI.

RIPRENDIAMO
I DATI DELLA
NOSTRA SALA
DA TÈ.



	Temperatura massima (°C)	Ordinazioni di tè freddo
Lun 22	29	77
Mar 23	28	62
Mer 24	34	93
Gio 25	31	84
Ven 26	25	59
Sab 27	29	64
Dom 28	32	80
Lun 29	31	75
Mar 30	24	58
Mer 31	33	91
Gio 1	25	51
Ven 2	31	73
Sab 3	26	65
Dom 4	30	84

IN QUANTI GIORNI LA MASSIMA È DI 31°C?



IL 25, IL 29 E
IL 2... QUINDI
SONO TRE.



ALLO-
RA...

POSSO
RAPPRESENTARE
LA TUA RISPOSTA
CON QUESTO
GRAFICO.

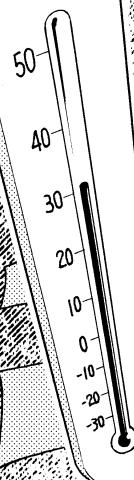
ORA,
RIFLETTI...

?

...QUESTI TRE
GIORNI NON SONO
GLI UNICI DELLA
STORIA CON UNA
MASSIMA DI 31°C,
GIUSTO?

CE NE SONO STATI
MOLTI ALTRI IN PASSATO
E CE NE SARANNO
MOLTI ALTRI NEL
FUTURO, NO?

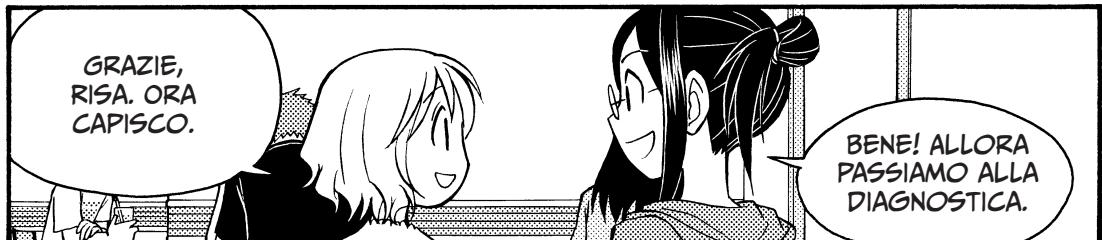
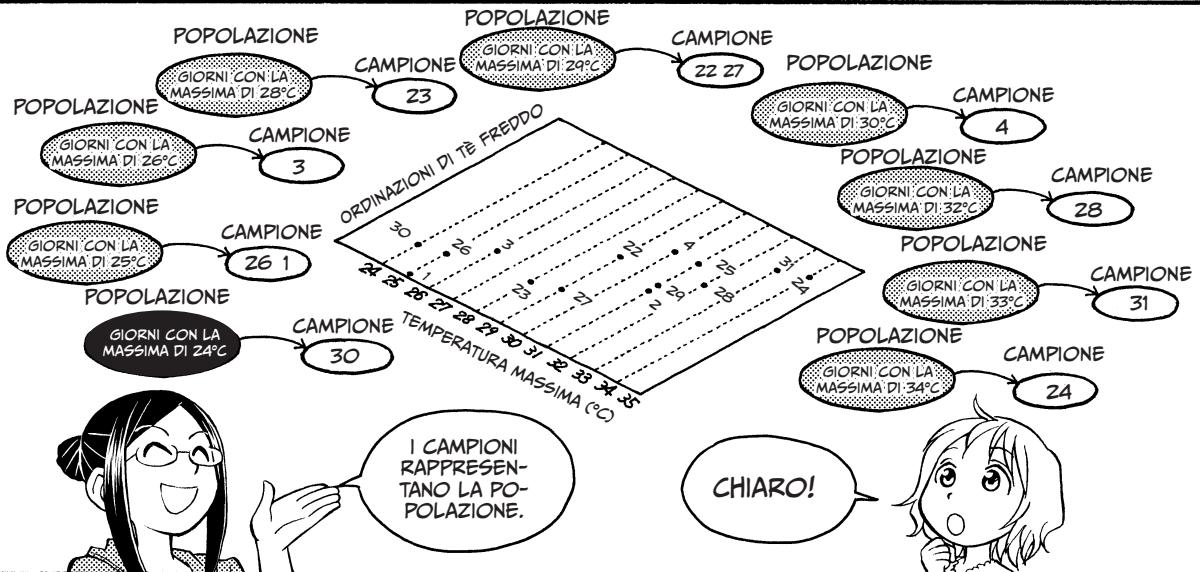
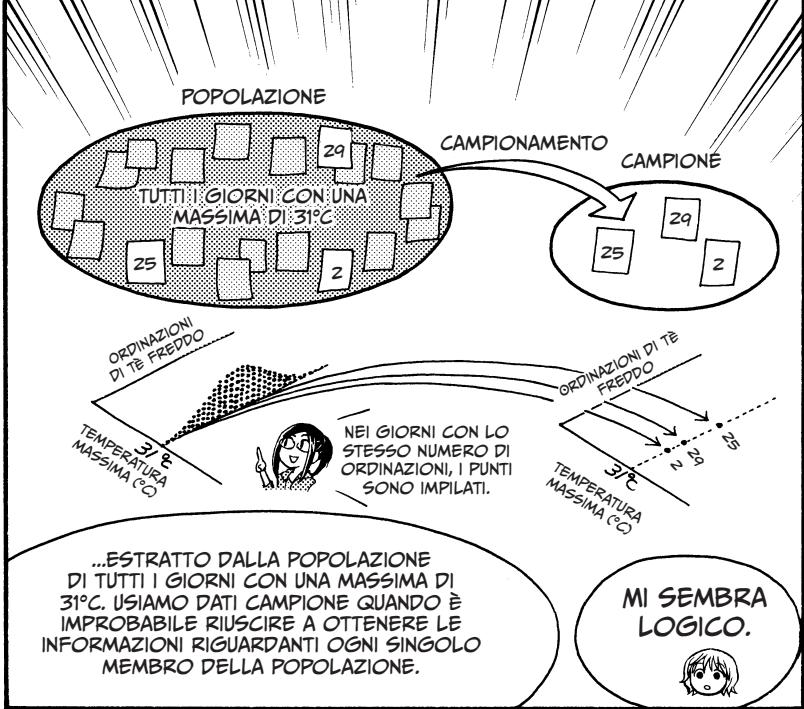
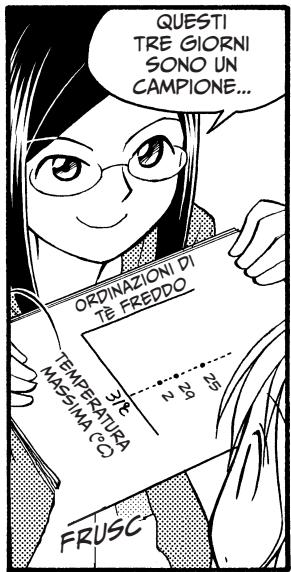
25



CERTO.



2



L'IPOTESI DI NORMALITÀ

L'EQUAZIONE DI REGRESSIONE HA SENSO SOLTANTO QUANDO È SODDISFATTA UNA CERTA IPOTESI.

QUALE?

ECCOLA:

IPOTESI ALTERNATIVA

IL NUMERO DI ORDINAZIONI DI TÈ FREDDO NEI GIORNI CON LA MASSIMA DI $x^{\circ}\text{C}$ SEGUE UNA DISTRIBUZIONE NORMALE CON MEDIA $Ax + B$ E DEVIAZIONE STANDARD σ (SIGMA).

ANDIAMO CON CALMA,
PRIMA DI TUTTO GUARDA
LE CURVE DI QUESTO GRAFICO.

QUESTE CURVE RAPPRESENTANO L'INTERA POPOLAZIONE DI ORDINAZIONI DI TÈ FREDDO PER OGNI MASSIMA DI TEMPERATURA. NON POTENDO CONOSCERE LA DISTRIBUZIONE ESATTA PER OGNI MASSIMA, DOBBIAMO IPOZZARE CHE SIA SEMPRE LA STESSA: UNA DISTRIBUZIONE NORMALE, RAPPRESENTATA DA UNA CURVA A CAMPANA.

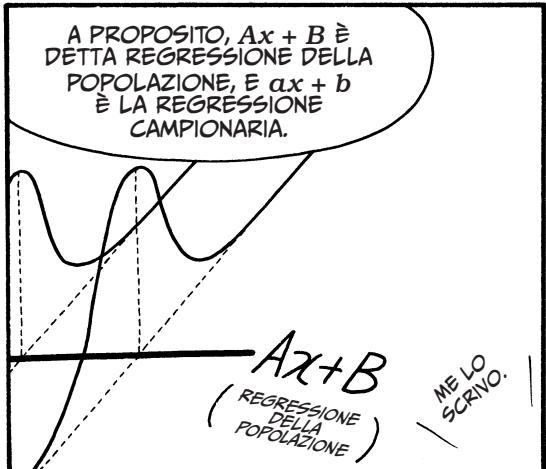
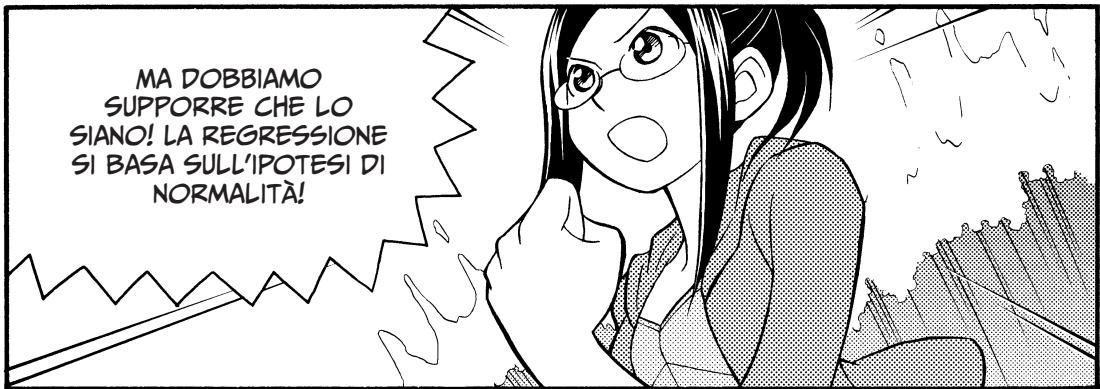
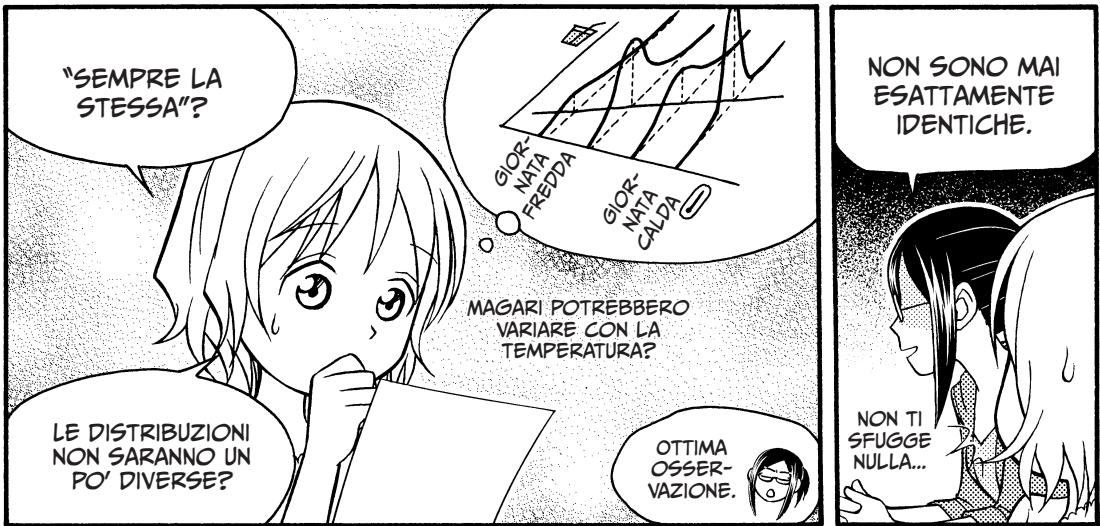
ORDINAZIONI
DI TÈ FREDDO

$Ax+B$

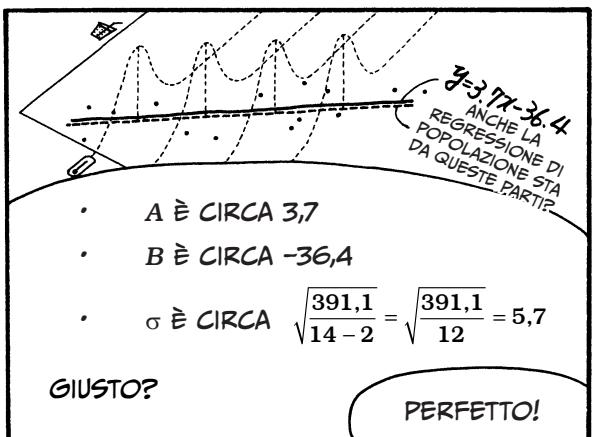
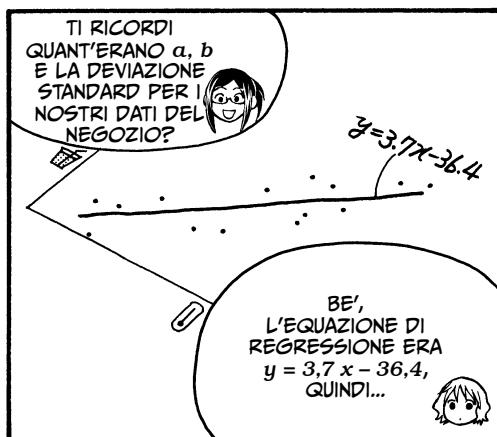
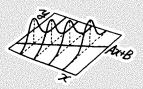
STESMO
ANDAMENTO

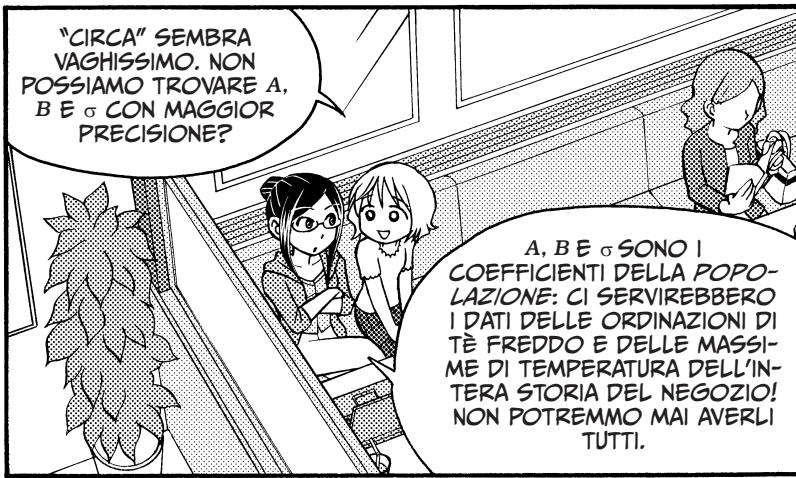
TEMPERATURA
MASSIMA ($^{\circ}\text{C}$)
26 28 30 32



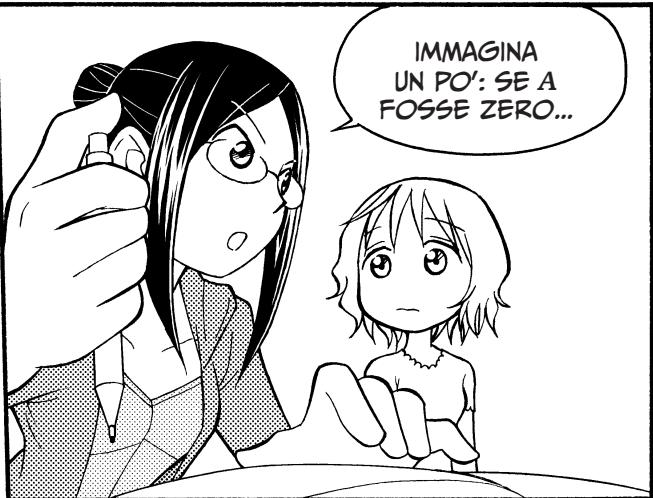
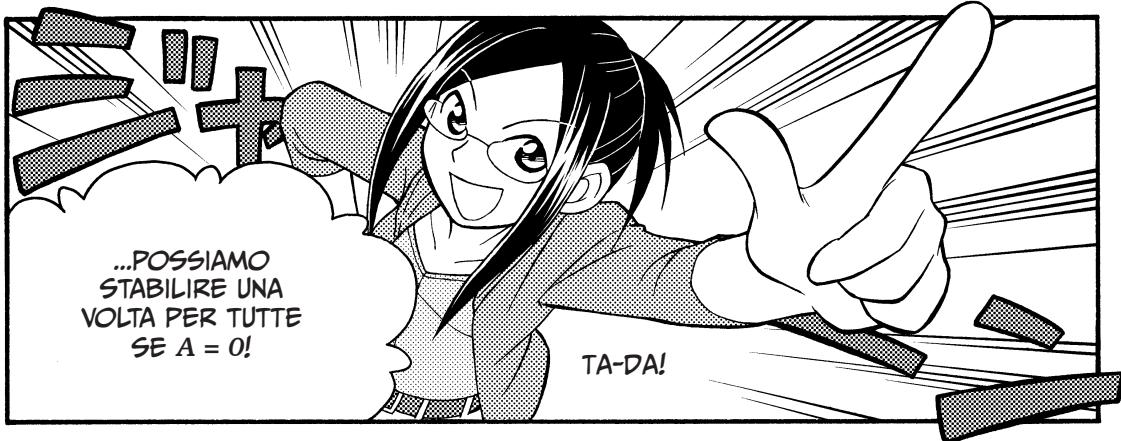


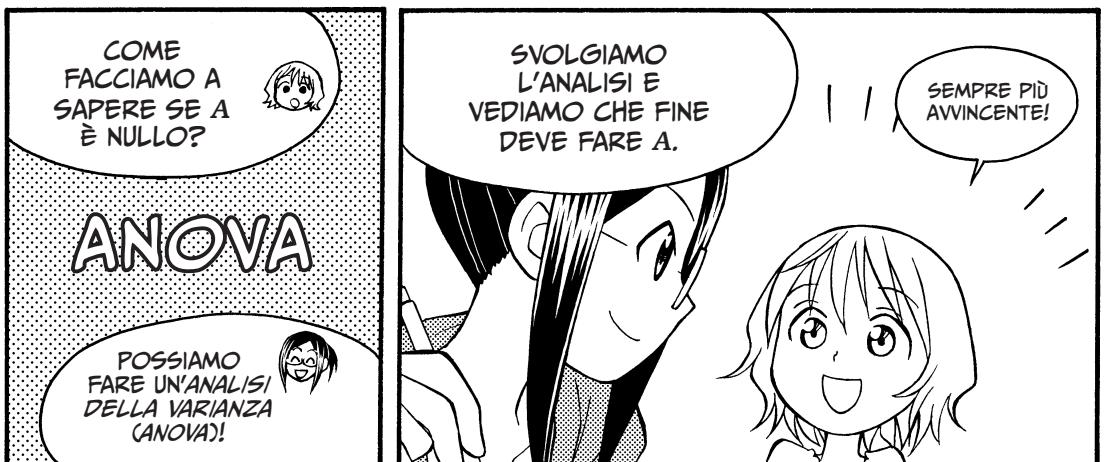
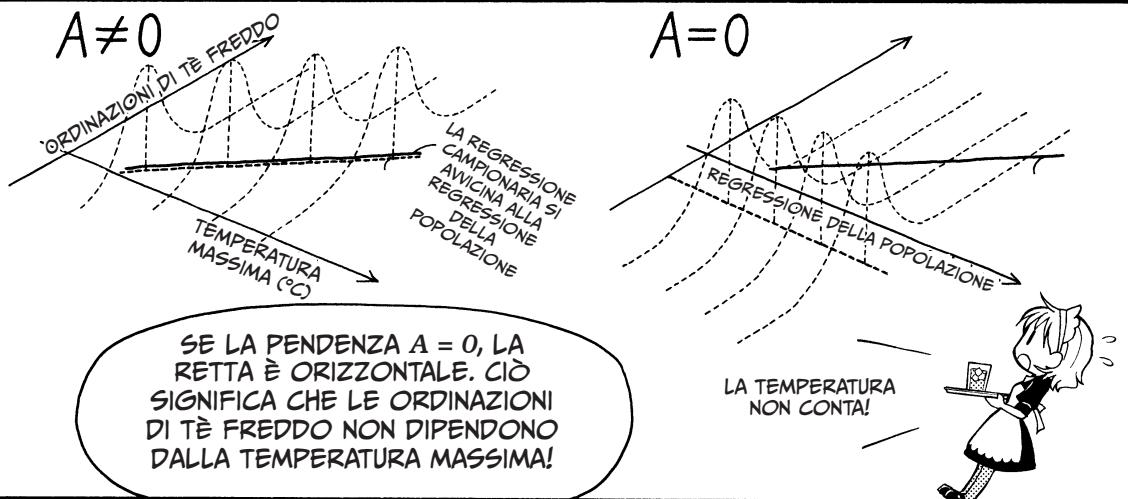
PASSO 4: ANALISI DELLA VARIANZA





A, B E C SONO I COEFFICIENTI DELLA POPOLAZIONE: CI SERVIREBBERO I DATI DELLE ORDINAZIONI DI TÈ FREDDO E DELLE MASSIME DI TEMPERATURA DELL'INTERA STORIA DEL NEGOZIO! NON POTREMMO MAI AVERLI TUTTI.





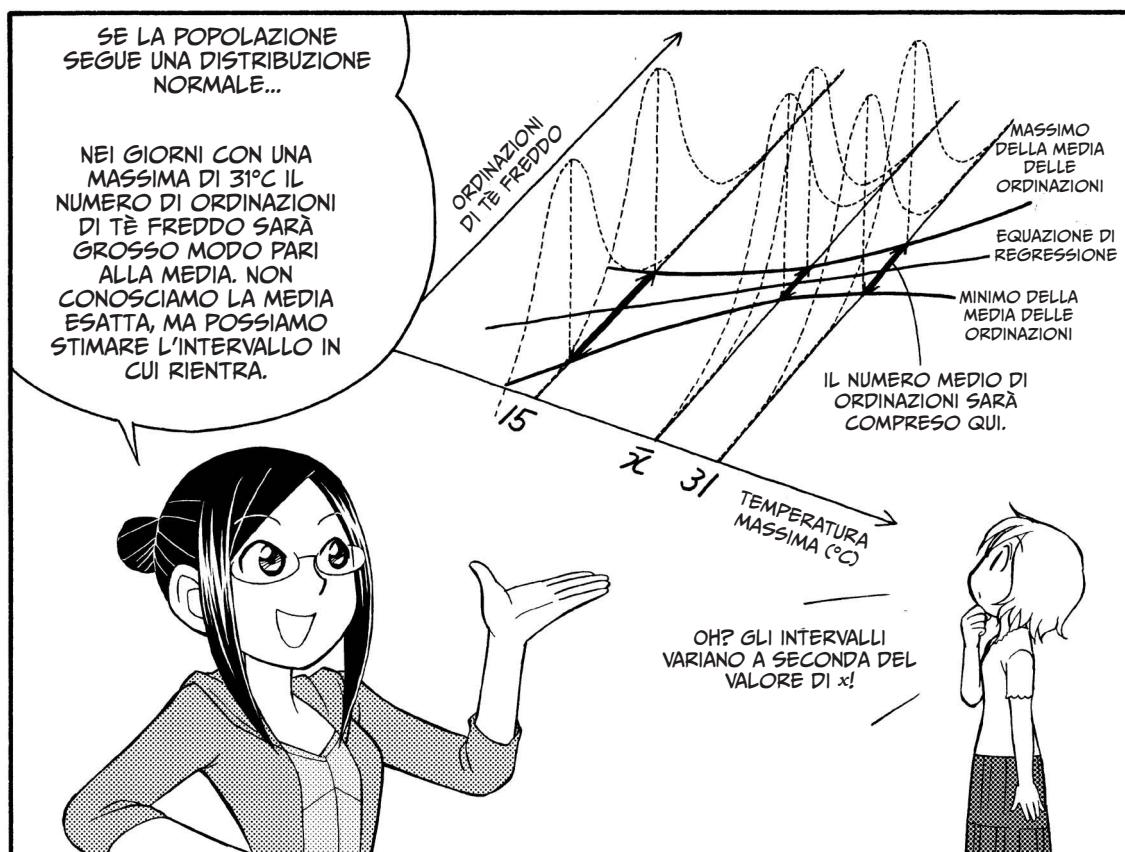
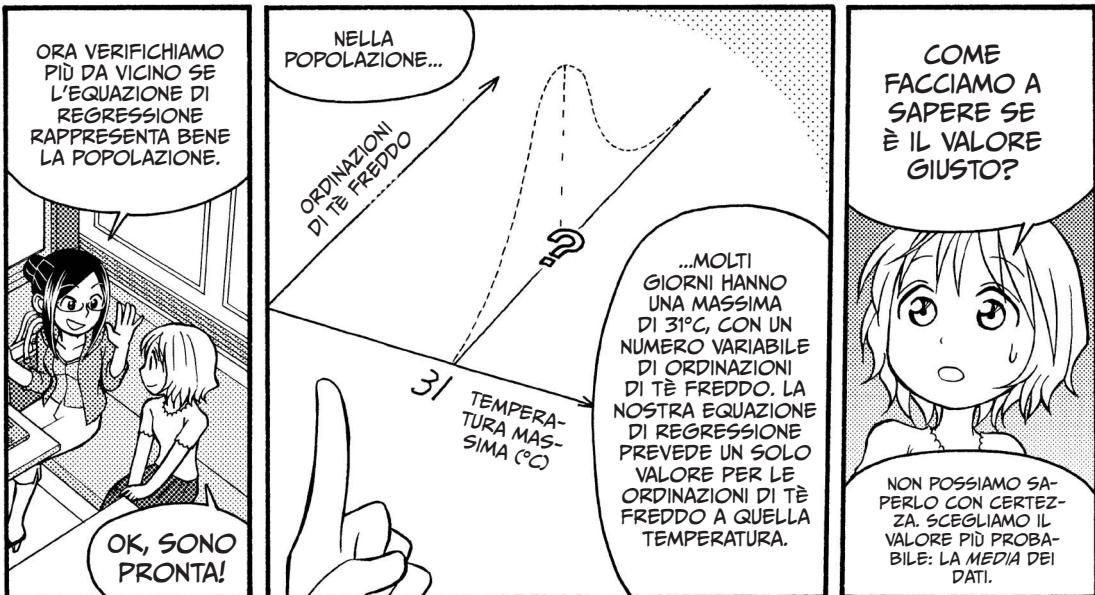
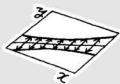
I PASSI DI ANOVA

Passo 1	Definire la popolazione.	La popolazione è "giorni con una temperatura massima di x gradi".
Passo 2	Formulare l'ipotesi nulla e l'ipotesi alternativa.	L'ipotesi nulla è $A = 0$. L'ipotesi alternativa è $A \neq 0$.
Passo 3	Scegliere il test di ipotesi da svolgere.	Useremo l'analisi della varianza univariata.
Passo 4	Scegliere il livello di significatività.	Porremo il livello di significatività a 0,05.
Passo 5	Calcolare la statistica test dai dati campione.	<p>La statistica test è:</p> $\frac{a^2}{\left(\frac{1}{S_{xx}}\right)} + \frac{S_e}{\text{numero di individui} - 2}$ <p>Sostituendo i valori dell'equazione di regressione campionaria:</p> $\frac{3,7^2}{\left(\frac{1}{129,7}\right)} + \frac{391,1}{14 - 2} = 55,6$ <p>La statistica test seguirà una distribuzione F con il primo grado di libertà pari a 1 e il secondo pari a 12 (numero di individui meno 2), se l'ipotesi nulla è vera.</p>
Passo 6	Determinare se il p -value per la statistica test derivante dal Passo 5 è minore del livello di significatività.	Al livello di significatività 0,05, con d_1 pari a 1 e d_2 pari a 12, il valore critico è 4,7472. La statistica test è 55,6.
Passo 7	Decidere se rifiutare l'ipotesi nulla.	Poiché la statistica test è maggiore del valore critico, scartiamo l'ipotesi nulla.

LA STATISTICA F CI PERMETTE DI VERIFICARE LA PENDENZA DELLA RETTA STUDIANO LA VARIANZA. SE LA VARIABILITÀ ATTORNO ALLA RETTA È MOLTO MINORE DELLA VARIANZA TOTALE DI Y SIGNIFICA CHE LA RETTA RIPRODUCE LA VARIABILITÀ DI Y, E LA STATISTICA SARÀ UTILE. SE IL RAPPORTO È PICCOLO, LA RETTA NON RIPRODUCE BENE LA VARIABILITÀ DI Y, E PROBABILMENTE NON SERVE A NULLA!



PASSO 5: CALCOLO DEGLI INTERVALLI DI CONFIDENZA



CALCOLIAMO
L'INTERVALLO
PER CIASCUNA
TEMPERATURA.

COME HAI NOTATO,
L'AMPIEZZA VARIA.
È MINORE VICINO A \bar{x} ,
CHE È LA MEDIA DELLE
TEMPERATURE MASSIME.

NEANCHE QUESTO
INTERVALLO DÀ LA
GARANZIA ASSOLUTA
DI INCLUDERE LA
VERA MEDIA DELLA
POPOLAZIONE.
L'AFFIDABILITÀ È
DETERMINATA DAL
COEFFICIENTE DI
CONFIDENZA.

ORA,
QUESTO...

...NON È UN
COEFFICIENTE
QUALSIASI.

NON ESISTE
UN'EQUAZIONE
PER CALCOLARLO,
NÉ REGOLE
PRESTABILITE.

LO FISSI TU A
QUALUNQUE
PERCENTUALE
DESIDERI.

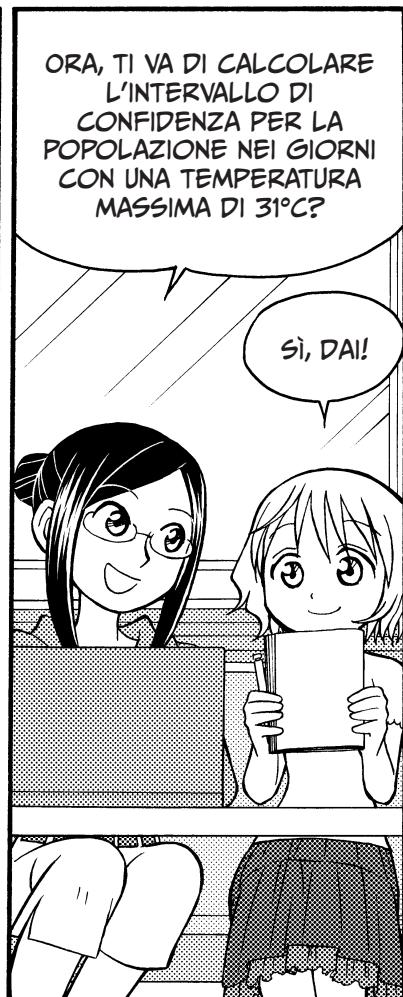
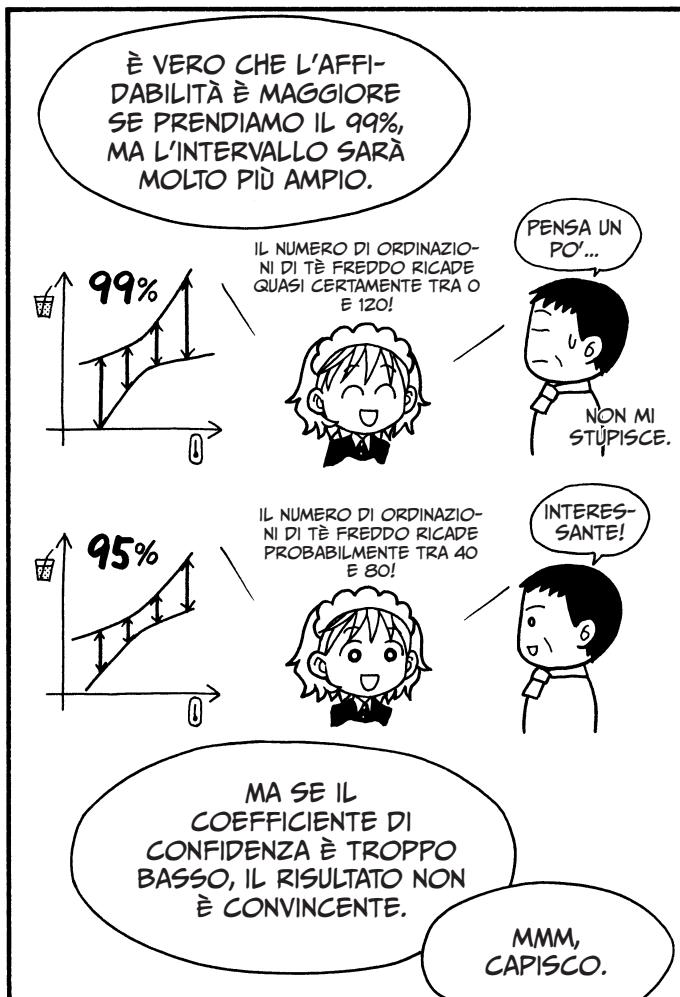
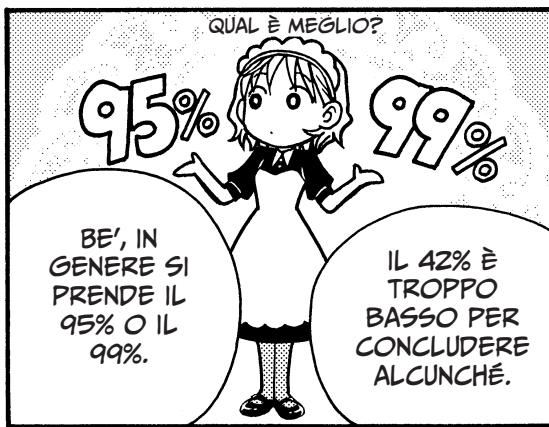
SCELGO IL 42%!



QUANDO CALCOLI UN INTERVALLO DI
CONFIDENZA, PRIMA DI TUTTO SCEGLI
IL COEFFICIENTE DI CONFIDENZA.

POI DICHI PER ESEMPIO CHE
"L'INTERVALLO DI CONFIDENZA DEL 42%
PER LE ORDINAZIONI DI TÈ FREDDO
ALLA MASSIMA TEMPERATURA DI 31°C
VA DA 30 A 35 ORDINAZIONI!"

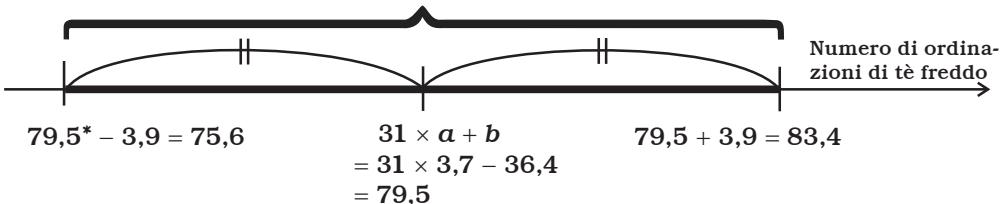
LO
SCELGO
IO?



ECCO COME SI CALCOLA UN INTERVALLO DI CONFIDENZA DEL 95% PER LE ORDINAZIONI DI TÈ FREDDO NEI GIORNI CON LA MASSIMA DI 31°C.



Questo è l'intervallo di confidenza



La distanza dalla media stimata è

$$\begin{aligned} & \sqrt{F(1, n - 2; 0,05) \times \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \times \frac{S_e}{n - 2}} \\ &= \sqrt{F(1, 14 - 2; 0,05) \times \left(\frac{1}{14} + \frac{(31 - 29,1)^2}{129,7} \right) \times \frac{391,1}{14 - 2}} \\ &= 3,9 \end{aligned}$$

dove n è il numero di dati del campione e F è il rapporto tra due distribuzioni chi-quadro, come descritto a pagina 57.

PER CALCOLARE UN INTERVALLO DI CONFIDENZA DEL 99% BASTA SOSTITUIRE

$$F(1, 14 - 2; 0,05) = 4,7$$

CON

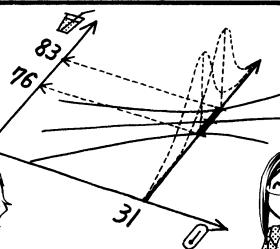
$$F(1, 14 - 2; 0,01) = 9,3$$



(A PAGINA 58 TROVATE LA SPIEGAZIONE DI $F(1, n - 2; 0,05) = 4,7$ E COSÌ VIA).

*IL VALORE 79,5 È STATO CALCOLATO CON CIFRE NON ARROTONDATE.

QUINDI SIAMO SICURE AL 95% CHE, CONSIDERANDO LA POPOLAZIONE DI GIORNI CON UNA MASSIMA DI 31°C, IL NUMERO MEDIO DI ORDINAZIONI DI TÈ FREDDO RICADA TRA 76 E 83.



PROPRIO COSÌ!

PASSO 6: FACCIAMO UNA PREVISIONE!



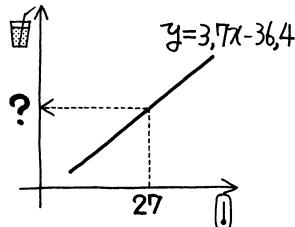
E INFINE,
FACCIAMO LA
PREVISIONE.

L'ULTIMO
PASSO!

SE DOMANI LA
MASSIMA SARÀ
DI 27°C...

...QUANTE SARANNO
LE ORDINAZIONI DI TÈ
FREDDO ALLA SALA
DA TÈ?

MMH, L'EQUA-
ZIONE DI RE-
GRESSIONE
È $y = 3,7x -$
 $36,4...$



$$\begin{aligned} y &= 3,7 \times 27 - 36,4 \\ &= 63,5 \\ &\approx 64 \end{aligned}$$

...QUINDI È
64*!

BINGO!

*QUESTO CALCOLO È STATO SVOLTO CON CIFRE ARROTONDATE.
RIPETENDOLO CON LE CIFRE COMPLETE, SENZA ARROTONDARE,
DOVRESTE OTTENERE 64,6.

MA GLI
ORDINI SARANNO
PROPRIO 64?

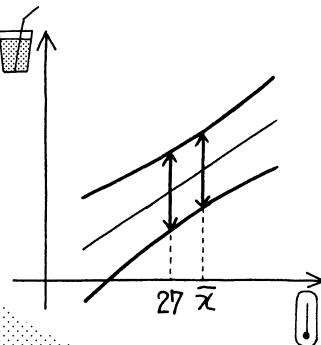
COME
POSSIAMO
ESSERNE
CERTE?

AH, SAPERLO.

LE ORDINAZIONI
DOVREBBERO
ESSERE VICINE
A 64 PERCHÉ R^2
È 0,8225, MA...
QUANTO VICINE?

TROVIAMO UN
INTERVALLO DI
PREVISIONE!

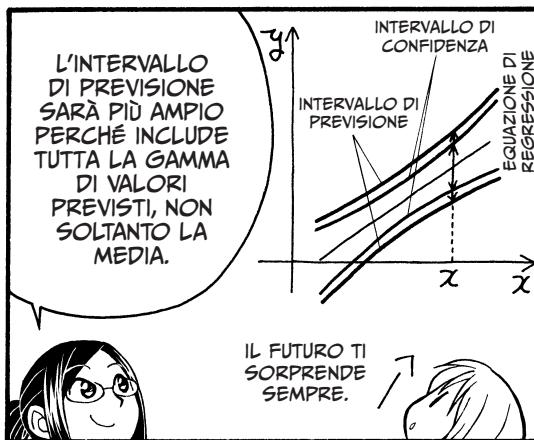
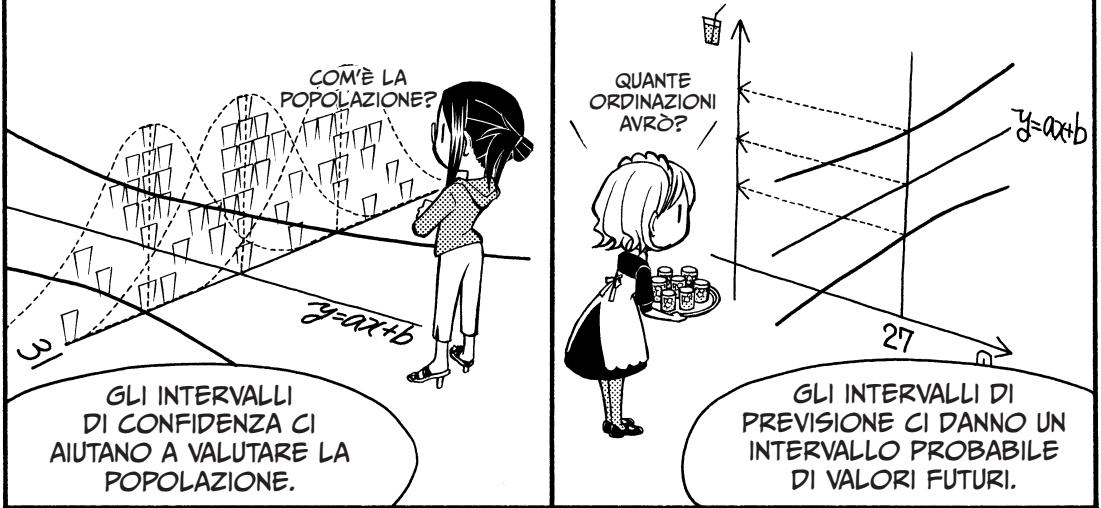
SCEGLIAMO UN
COEFFICIENTE E
POI CALCOLIAMO
UN INTERVALLO IN
CUI LE ORDINAZIONI
RICADRANNO CON
ALTA PROBABILITÀ.



MA NON È
QUELLO CHE
ABBiamo
APPENA
FATTO?

NON PROPRIAMENTE. PRIMA PRE-
VEDEVAMO IL NUMERO ME-
DIO DI ORDINAZIONI DI TÈ
FREDDO PER LA POPOLA-
ZIONE DI GIORNI CON UNA
CERTA TEMPERATURA MAS-
SIMA, ORA INVECE PREVE-
DIAMO IL NUMERO PROBA-
BILE DI ORDINAZIONI DI TÈ
FREDDO PER UN GIORNO
DATO CON UNA CERTA
TEMPERATURA.

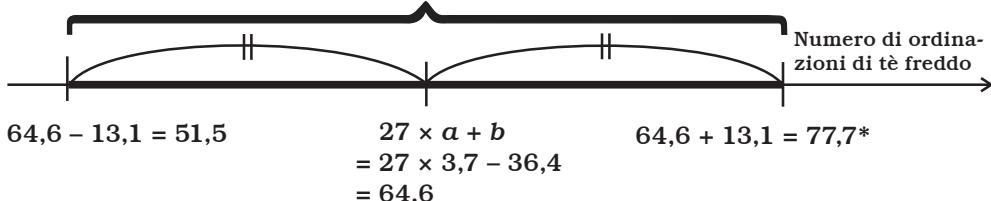
E CHE
DIFFERENZA
C'È?



ECCO COME SI CALCOLA L'INTERVALLO DI PREVISIONE DEL 95% PER LE VENDITE DI TÈ FREDDO DI DOMANI.



Questo è l'intervallo di previsione.



La distanza dal valore stimato è

$$\begin{aligned} & \sqrt{F(1, n-2; 0,05) \times \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \times \frac{S_e}{n-2}} \\ &= \sqrt{F(1, 14-2; 0,05) \times \left(1 + \frac{1}{14} + \frac{(27 - 29,1)^2}{129,7} \right) \times \frac{391,1}{14-2}} \\ &= 13,1 \end{aligned}$$

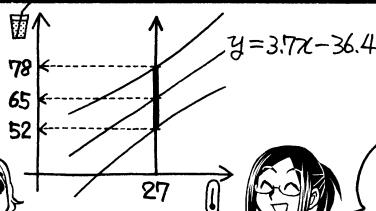
IL NUMERO STIMATO DI ORDINAZIONI DI TÈ FREDDO CALCOLATO PRIMA (A PAGINA 95) ERA ARROTONDATO, MA QUI ABBIAMO PRESO IL NUMERO DI ORDINAZIONI DI TÈ FREDDO STIMATO SENZA ARROTONDARE, 64,6.



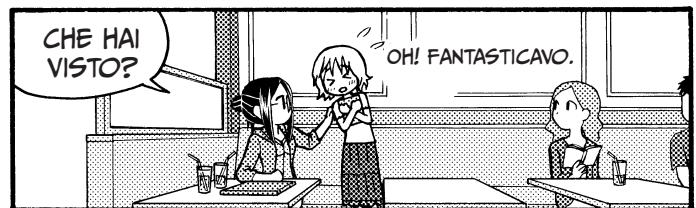
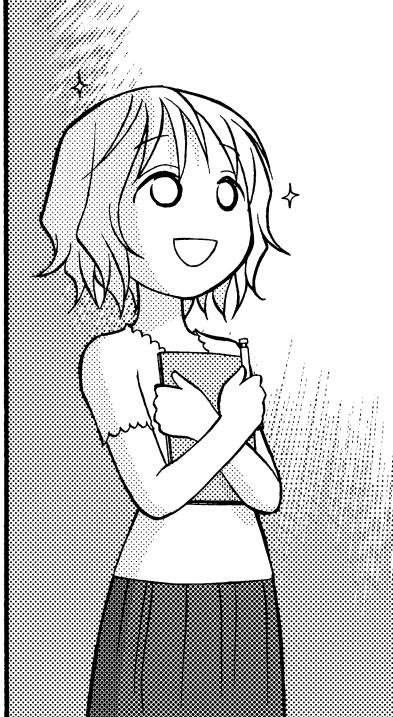
ABBIAMO USATO LA DISTRIBUZIONE F PER DETERMINARE L'INTERVALLO DI PREVISIONE E LA REGRESSIONE DI POPOLAZIONE. IN GENERE GLI STATISTICI ARRIVANO AGLI STESSI RISULTATI CON LA DISTRIBUZIONE T.

*QUESTO CALCOLO È STATO SVOLOTTO CON LE CIFRE ARROTONDATE MOSTRATE QUI. IL CALCOLO COMPLETO SENZA ARROTONDAMENTI DÀ 77,6.

QUINDI SIAMO CERTE AL 95% CHE IL NUMERO DI ORDINAZIONI DI TÈ FREDDO SARÀ COMPRESO TRA 52 E 78 QUANDO LA TEMPERATURA DEL GIORNO È DI 27°C.



L'IDEA È
QUELLA!



QUALI SONO I PASSI NECESSARI?

Ricordate il procedimento per l'analisi di regressione presentato a pagina 68?

1. Tracciare il grafico di dispersione della variabile dipendente in funzione di quella indipendente. Se i punti si allineano, potrebbe esserci correlazione tra le variabili.
2. Calcolare l'equazione di regressione.
3. Calcolare il coefficiente di correlazione (R) e valutare la popolazione e le ipotesi.
4. Svolgere l'analisi della varianza.
5. Calcolare gli intervalli di confidenza.
6. Fare una previsione!

In questo capitolo li abbiamo percorsi tutti, ma non è sempre necessario. Riprendiamo l'esempio dell'età e dell'altezza di Miu fatto a pagina 25.

- Dato 1: a questo mondo c'è una sola Miu.
- Dato 2: all'età di 10 anni Miu era alta 137,5 cm.

Ciò assodato, non ha senso dire che "l'altezza di Miu all'età di 10 anni segue una distribuzione normale con media $Ax + B$ e deviazione standard σ ". In altri termini, è insensato analizzare la popolazione delle altezze di Miu all'età di 10 anni: la sua altezza poteva assumere un solo valore, che ci è noto.

Nell'analisi di regressione, analizziamo l'intera popolazione o, assai più di frequente, analizziamo un campione di una popolazione più grande. Nel secondo caso bisogna svolgere tutti i passi. I Passi 4 e 5, tuttavia, valutano la rappresentatività del campione rispetto alla popolazione; sono superflui se i dati corrispondono all'intera popolazione e non soltanto a un campione.

NOTA *Usiamo il termine "statistica" per descrivere la misura di una caratteristica del campione, come la sua media, e "parametro" per descrivere una misura proveniente dalla popolazione, come la media o altri coefficienti della popolazione.*

RESIDUI STANDARDIZZATI

Ricordate che il *residuo* è la differenza tra il valore *misurato* e il valore *stimato* tramite l'equazione di regressione. Il *residuo standardizzato* è il residuo diviso per la sua deviazione standard

stimata; si usa per valutare se una certa misura si discosta dall'andamento in maniera significativa. Per esempio, immaginiamo che il giorno 4 un gruppo di sportivi accaldati faccia una pausa alla sala da tè, portando le ordinazioni reali di tè freddo a 84, benché le previsioni fossero circa 76 in base alla massima del giorno. Un evento simile farebbe crescere il residuo standardizzato.

I residui standardizzati si calcolano dividendo ogni residuo per una stima della sua deviazione standard, ricavata a sua volta dalla somma dei residui al quadrato. Poiché il calcolo è un po' complicato, e in genere i programmi informatici di statistica lo svolgono automaticamente, non lo ripercorriamo in dettaglio.

La Tabella 2-1 mostra i residui standardizzati per i dati della sala da tè usati in questo capitolo.

TABELLA 2-1: CALCOLO DEI RESIDUI STANDARDIZZATI

Temperatura massima x	Numero misurato di ordinazioni di tè freddo y	Numero stimato di ordinazioni di tè freddo $\hat{y} = 3,7x - 36,4$	Residuo	
			y - \hat{y}	Residuo standardizzato
Lun 22	29	77	72,0	5,0 0,9
Mar 23	28	62	68,3 -6,3 -1,2	
Mer 24	34	93	90,7 2,3 0,5	
Gio 25	31	84	79,5 4,5 0,8	
Ven 26	25	59	57,1 1,9 0,4	
Sab 27	29	64	72,0 -8,0 -1,5	
Dom 28	32	80	83,3 -3,3 -0,6	
Lun 29	31	75	79,5 -4,5 -0,8	
Mar 30	24	58	53,3 4,7 1,0	
Mer 31	33	91	87,0 4,0 0,8	
Gio 1	25	51	57,1 -6,1 -1,2	
Ven 2	31	73	79,5 -6,5 -1,2	
Sab 3	26	65	60,8 4,2 0,8	
Dom 4	30	84	75,8 8,2 1,5	

Come potete vedere, il residuo standardizzato per il giorno 4 vale 1,5. Se le ordinazioni di tè freddo fossero state 76 come previsto, il residuo standardizzato sarebbe stato nullo.

A volte un valore misurato è talmente lontano dall'andamento generale da compromettere l'analisi. Se il residuo standardizzato è maggiore di 3 o minore di -3, la misura è considerata *anomala*. Ci sono diversi rimedi possibili, tra cui escludere i valori anomali, assegnare loro un valore dato, o tenerli nell'analisi così come sono. Per scegliere come trattarli bisogna valutare quale ne sia la causa.

INTERPOLAZIONE ED ESTRAPOLAZIONE

Riguardando i valori x (temperatura massima) di pagina 64, si nota che il valore massimo è 34°C e il minimo è 24°C. Usando l'analisi di regressione si può *interpolare* il numero di ordinazioni di tè freddo nei giorni di massima compresa tra 24°C e 34°C ed *estrapolare* il numero di ordinazioni di tè freddo nei giorni di massima minore di 24°C o maggiore di 34°C. In altri termini, l'estrapolazione è la stima dei valori esterni all'intervallo dei dati osservati.

Poiché abbiamo osservato l'andamento soltanto tra i valori di 24°C e 34°C, non sappiamo se le vendite di tè freddo continuino a rispettarlo nei giorni freddissimi o caldissimi. L'estrapolazione è quindi meno affidabile dell'interpolazione, e alcuni specialisti la evitano del tutto.

Nell'uso quotidiano è accettabile estrapolare, finché si tiene presente che il risultato non è del tutto affidabile. Ma è meglio evitare estrapolazioni nel contesto accademico, o nella stima di valori lontani dall'intervallo dei dati misurati.

AUTOCORRELAZIONE

In questo capitolo abbiamo preso come variabile indipendente la temperatura massima, e l'abbiamo usata per prevedere le vendite di tè freddo. Nella maggior parte delle località è improbabile che un giorno la massima sia di 20°C e il giorno dopo schizzi a 30°C. Di solito la temperatura cresce o cala gradualmente nel corso di vari giorni, quindi se le due variabili sono correlate, anche il numero di ordinazioni di tè freddo dovrebbe aumentare o diminuire con gradualità. Abbiamo però fatto l'ipotesi che la deviazione (l'errore) assuma valori casuali. Da un giorno all'altro, quindi, i valori previsti non variano gradualmente come farebbero nella realtà.

Analizzando variabili che possono risentire del passaggio del tempo è bene controllare l'autocorrelazione. Questa si verifica quando l'errore è correlato nel corso del tempo, e può indicare la necessità di cambiare modello di regressione.

L'autocorrelazione è descritta dall'indice detto *statistica di Durbin-Watson*, da calcolare in questo modo:

$$d = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2}$$

Questa formula si legge come “la somma dei quadrati di ogni residuo meno il residuo precedente, divisa per la somma dei quadrati di tutti i residui”. Per l'esempio di questo capitolo, la statistica di Durbin-Watson vale:

$$\frac{(-6,3 - 5,0)^2 + (2,3 - (-6,3))^2 + \dots + (8,2 - 4,2)^2}{5,0^2 + (-6,3)^2 + \dots + 8,2^2} = 1,8$$

L'esatto valore critico della statistica di Durbin-Watson dipende da ogni singola analisi, ed esistono tabelle per determinarlo, ma in generale si fissa il limite a 1: valori più bassi possono indicare la presenza di autocorrelazione. Poiché questo risultato è vicino a 2, possiamo concludere che non c'è autocorrelazione nel nostro esempio.

REGRESSIONE NON LINEARE

A pagina 66 Risa ha detto:



LO SCOPO DELL'ANALISI DI REGRESSIONE È OTTENERE L'EQUAZIONE DI REGRESSIONE NELLA FORMA $y = ax + b$.

Questa è un'equazione lineare, ma non sempre le equazioni di regressione lo sono. Anche le seguenti sono possibili equazioni di regressione:

- $y = \frac{a}{x} + b$
- $y = a\sqrt{x} + b$
- $y = ax^2 + bx + c$
- $y = a \times \log x + b$

Nell'esempio dell'età e dell'altezza di Miu presentato a pagina 26, l'equazione di regressione era della forma $y = \frac{a}{x} + b$ e non $y = ax + b$.

A questo punto, svolgendo l'analisi di regressione sui propri dati, è naturale chiedersi come scegliere la forma dell'equazione. Aiutatevi con questo procedimento.

1. Tracciate un grafico di dispersione dei dati, con i valori della variabile indipendente sull'asse x e quelli della variabile dipendente sull'asse y . Esaminate la relazione tra le variabili suggerita dalla disposizione dei punti: seguono grosso modo una linea retta? O descrivono una curva? Nel secondo caso, qual è la forma della curva?
2. Fate un tentativo con l'equazione di regressione suggerita dalla forma del grafico tracciato in 1). Tracciate il grafico con i residui (o i residui standardizzati) sull'asse y e la variabile indipendente sull'asse x . I residui dovrebbero sembrare casuali: se invece appare una qualche regolarità, come una curva, è possibile che l'equazione di regressione non corrisponda alla relazione tra i dati.
3. Se il grafico dei residui di 2) mostra una qualche regolarità, cambiate l'equazione di regressione e ripetete 2). Provate con varie equazioni e scegliete quella che sembra meglio corrispondere ai dati. In genere è consigliabile scegliere la più semplice fra le equazioni che approssimano bene i dati.

TRASFORMAZIONE DELLE EQUAZIONI NON LINEARI IN EQUAZIONI LINEARI

Avendo a che fare con equazioni non lineari, possiamo in alternativa trasformarle in equazioni lineari. Consideriamo per esempio l'equazione per l'età e l'altezza di Miu (pagina 26):

$$y = -\frac{326,6}{x} + 173,3$$

Per trasformarla in un'equazione lineare, ricordate che:

$$\text{Se } \frac{1}{x} = X, \text{ abbiamo } \frac{1}{X} = x.$$

Definiamo quindi una nuova variabile X come $\frac{1}{x}$, e usiamo X nella solita equazione di regressione $y = aX + b$. Come mostrato a pagina 76, il valore di a e b in quest'equazione si può calcolare così:

$$\begin{cases} \alpha = \frac{S_{xy}}{S_{xx}} \\ b = \bar{y} - \bar{x}\alpha \end{cases}$$

Procediamo come al solito (riprendete la Tabella 2-2).

TABELLA 2-2: CALCOLO DELL'EQUAZIONE DI REGRESSIONE

Età <i>x</i>	$\frac{1}{\text{età}} = X$	Altezza <i>y</i>	$(X - \bar{X})$	$y - \bar{y}$	$(X - \bar{X})^2$	$(y - \bar{y})^2$	$(X - \bar{X})(y - \bar{y})$
4	0,2500	100,1	0,1428	-38,1625	0,0204	1456,3764	-5,4515
5	0,2000	107,2	0,0928	-31,0625	0,0086	964,8789	-2,8841
6	0,1667	114,1	0,0595	-24,1625	0,0035	583,8264	-1,4381
7	0,1429	121,7	0,0357	-16,5625	0,0013	274,3164	-0,5914
8	0,1250	126,8	0,0178	-11,4625	0,0003	131,3889	-0,2046
9	0,1111	130,9	0,0040	-7,3625	0,0000	54,2064	-0,0292
10	0,1000	137,5	-0,0072	-0,7625	0,0001	0,5814	-0,0055
11	0,0909	143,2	-0,0162	4,9375	0,0003	24,3789	-0,0802
12	0,0833	149,4	-0,0238	11,1375	0,0006	124,0439	-0,2653
13	0,0769	151,6	-0,0302	13,3375	0,0009	177,889	-0,4032
14	0,0714	154,0	-0,0357	15,7375	0,0013	247,6689	-0,5622
15	0,0667	154,6	-0,0405	16,3375	0,0016	266,9139	-0,6614
16	0,0625	155,0	-0,0447	16,7375	0,0020	280,1439	-0,7473
17	0,0588	155,1	-0,0483	16,8375	0,0023	283,5014	-0,8137
18	0,0556	155,3	-0,0516	17,0375	0,0027	290,2764	-0,8790
19	0,0526	155,7	-0,0545	17,4375	0,0030	304,0664	-0,9507
Somma	184	1.7144	2212,2	0,0000	0,0000	0,0489	5464,4575
Media	11,5	0,1072	138,3				-15,9563

Secondo la tabella:

$$\begin{cases} \alpha = \frac{S_{xy}}{S_{xx}} = \frac{-15,9563}{0,0489} = -326,6^* \\ b = \bar{y} - \bar{x}\alpha = 138,2625 - 0,1072 \times (-326,6) = 173,3 \end{cases}$$

*Se ottenete una cifra un po' diversa da 326,6, potrebbe essere colpa degli arrotondamenti. In tal caso la differenza dovrebbe essere piccola.

Quindi l'equazione di regressione è:

$$y = -326,6X + 173,3$$

↑ ↑
altezza \underline{1}
 età

Questo equivale a:

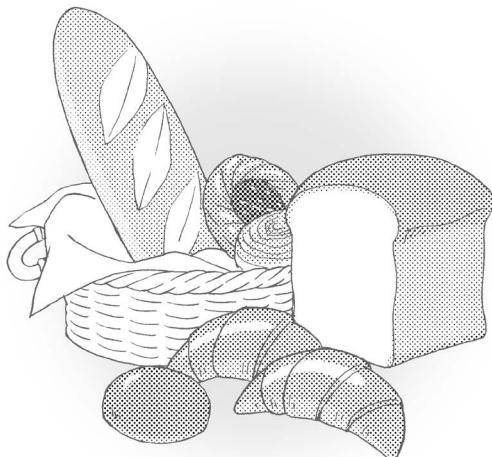
$$y = -\frac{326,6}{x} + 173,3$$

↑ ↑
altezza età

Abbiamo trasformato l'equazione non lineare di partenza in un'equazione lineare!

3

**ANALISI
DI REGRESSIONE
MULTIPLA**



PREVISIONI CON PIÙ DI UNA VARIABILE

GRAZIE PER
AVER PORTATO I
DATI.

DI NULLA.
SEI GENTILE AD
AIUTARE LA TUA
AMICA.

BE'...

...HO I MIEI
MOTIVI.



RISA...

PANT

PUFF

OH,
ECCOLA.

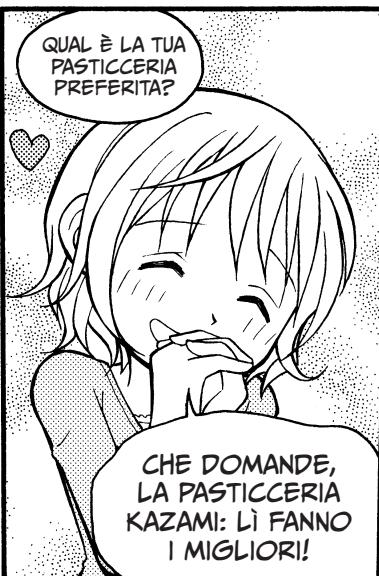
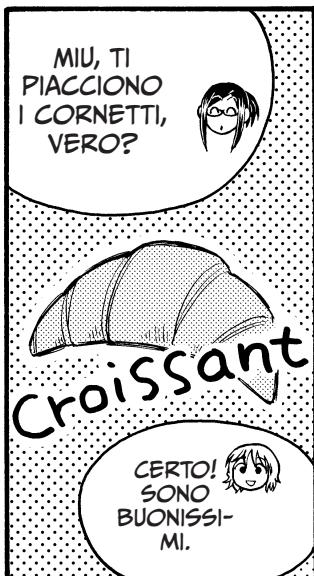
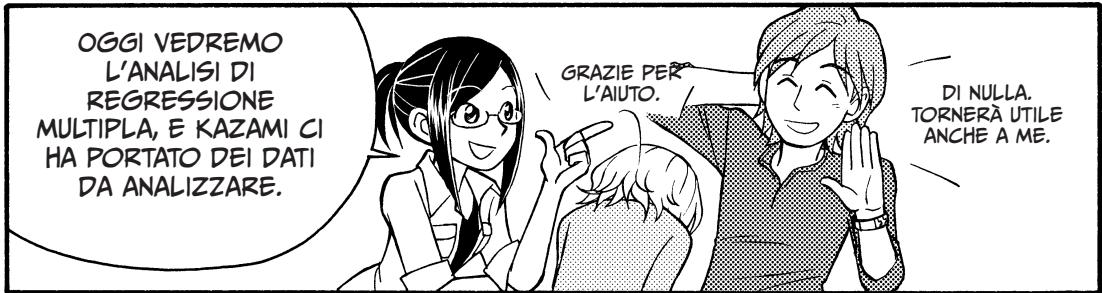
SIAMO QUI!

SCUSATE IL
RITARDO!

LA LEZIONE È
FINITA TARDI.
HO FATTO
UNA CORSA.

NON PREOC-
CUPARTI, SIA-
MO APPENA
ARRIVATI.





MA ALLORA TU
SEI... ?

L'EREDE
DELL'IMPERO
DELLE PASTIC-
CERIE KAZAMI!

RISA, NON
ESAGERARE.

È SOLO
UN'IMPRESA DI
FAMIGLIA.

SONO SOLO
DIECI, QUASI
TUTTE IN CEN-
TRO.

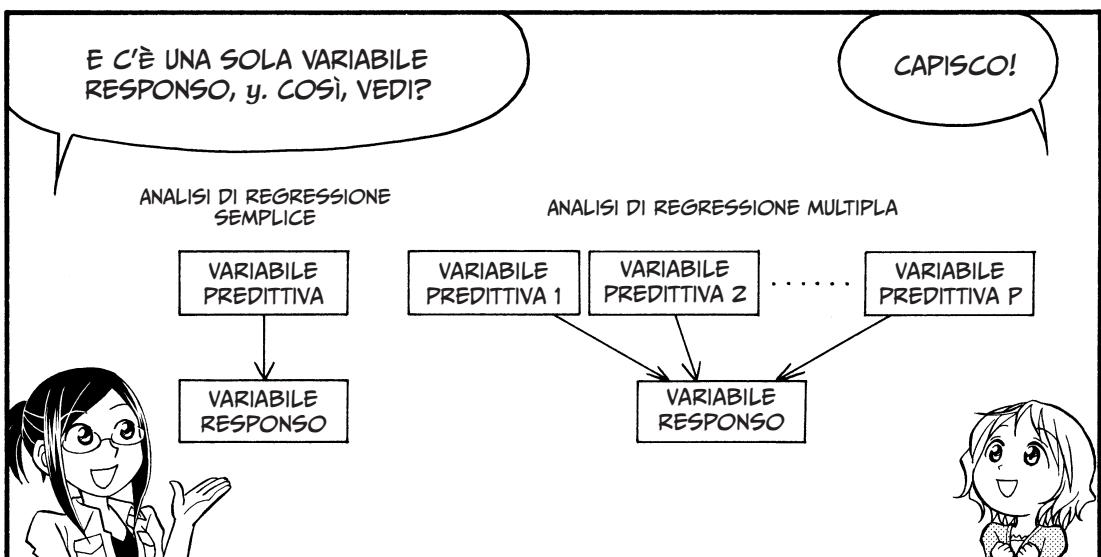
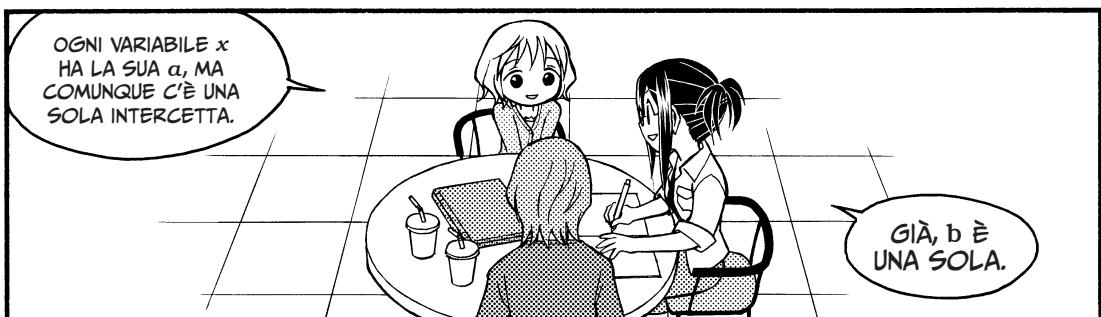
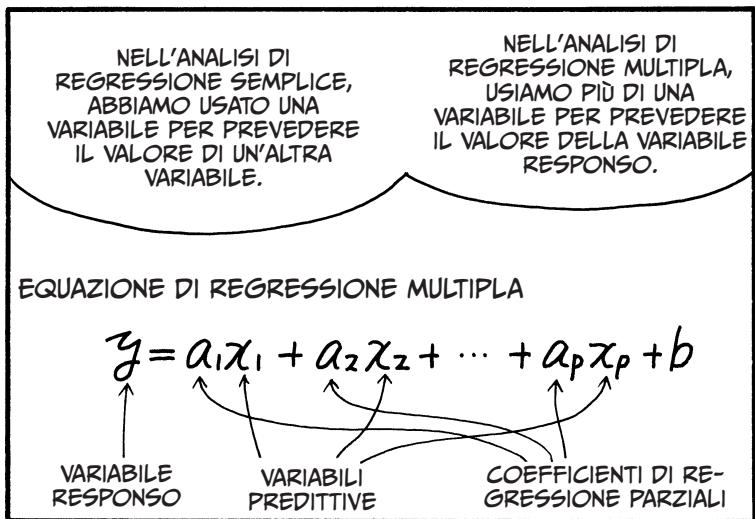
PERÒ PENSIAMO
DI APRIRE
UN'ALTRA A
BREVE.

MA HO VISTO
PASTICCERIE
KAZAMI IN TUTTA
LA CITTA!

QUINDI OGGI...

FAREMOS
PREVISIONI SULLE
VENDITE DEL
NUOVO NEGOZIO
CON L'ANALISI DI
REGRESSIONE
MULTIPLA.

WOW!



EQUAZIONE DI REGRESSIONE MULTIPLA

I PASSI SONO
GLI STESSI
DELL'EQUAZIONE
DI REGRESSIONE
SEMPLICE?

BE'...

SONO SIMILI, MA
NON PROPRIO
IDENTICI.

CLICK

PROCEDIMENTO PER L'ANALISI DI REGRESSIONE MULTIPLA

- PASSO 1 TRACCIARE DEI GRAFICI DI DISPERSIONE DI OGNI VARIABILE PREDITTIVA IN FUNZIONE DELLA VARIABILE RESPONSO, PER VEDERE SE SEMBRANO CORRELATE
- PASSO 2 CALCOLO DELL'EQUAZIONE DI REGRESSIONE MULTIPLA
- PASSO 3 VALUTAZIONE DELL'ACCURATEZZA DELL'EQUAZIONE DI REGRESSIONE MULTIPLA
- PASSO 4 SVOLGIMENTO DEL TEST DI ANALISI DELLA VARIANZA (ANOVA)
- PASSO 5 CALCOLO DEGLI INTERVALLI DI CONFIDENZA PER LA POPOLAZIONE
- PASSO 6 FACCIAMO UNA PREVISIONE!

DOBBIANO CONSIDERARE I PREDITTORI SINGOLARMENTE E ANCHE TUTTI INSIEME.



ME LO SCRIVO.



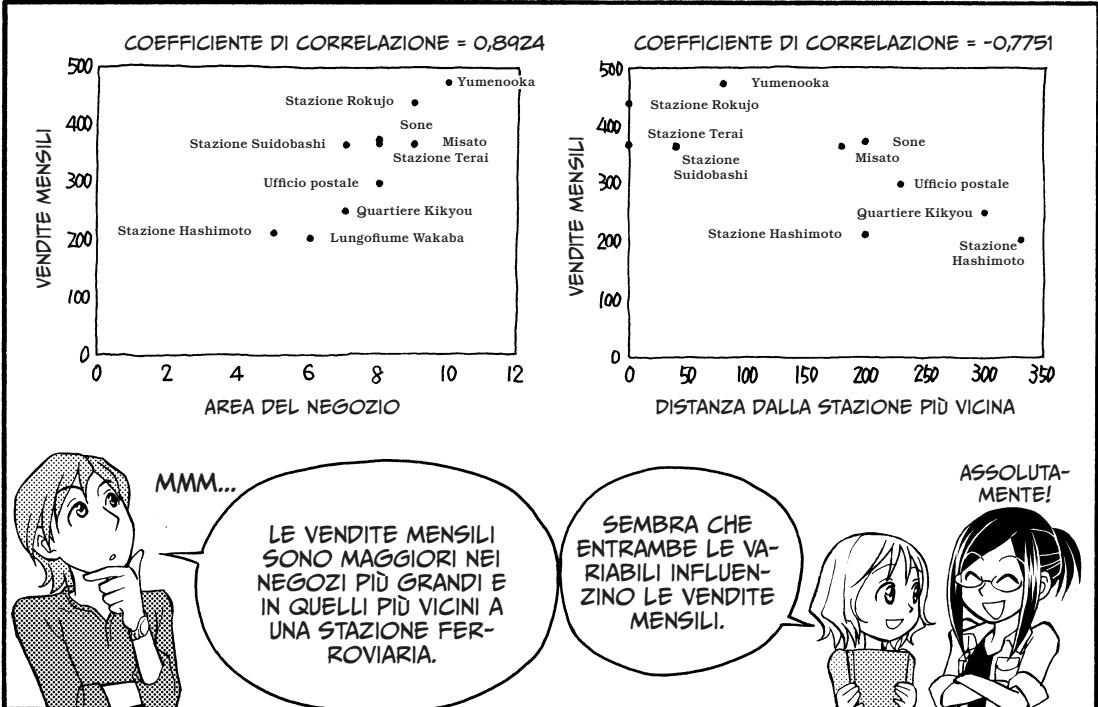
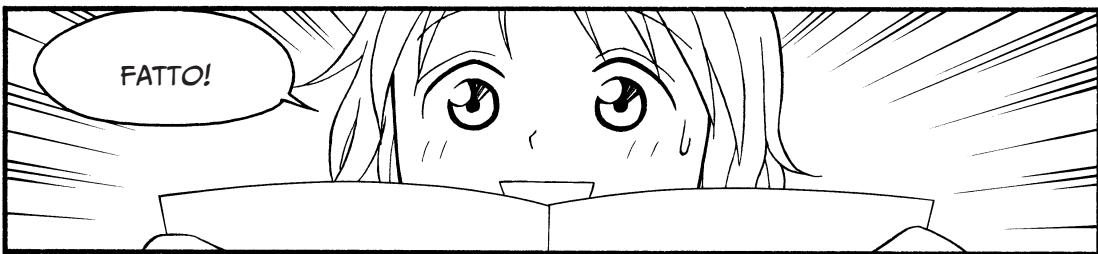
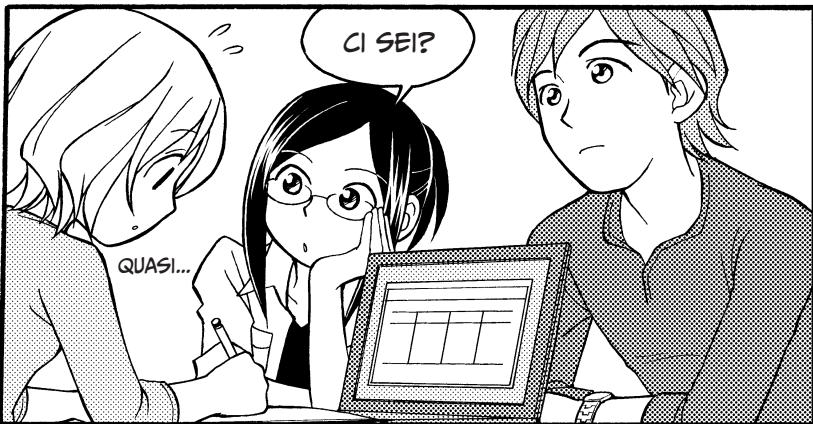
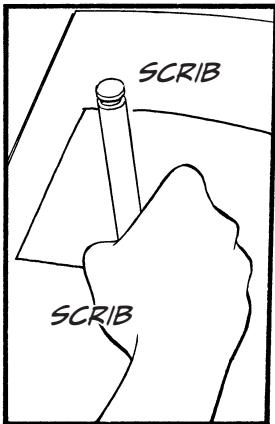
PASSO 1: TRACCIARE IL GRAFICO DI DISPERSIONE DI TUTTE LE VARIABILI PREDITTIVE IN FUNZIONE DELLA VARIABILE RESPONSO, PER VEDERE SE SEMBRANO CORRELATE



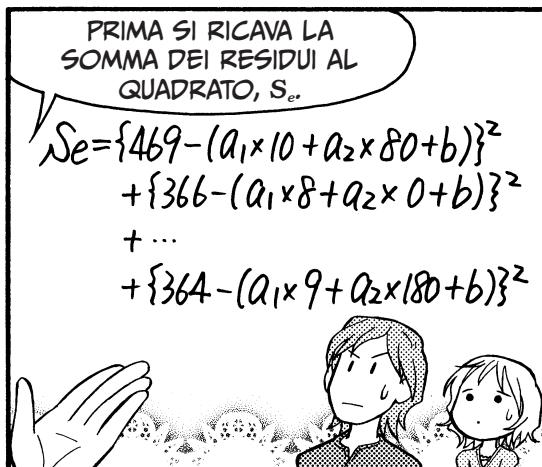
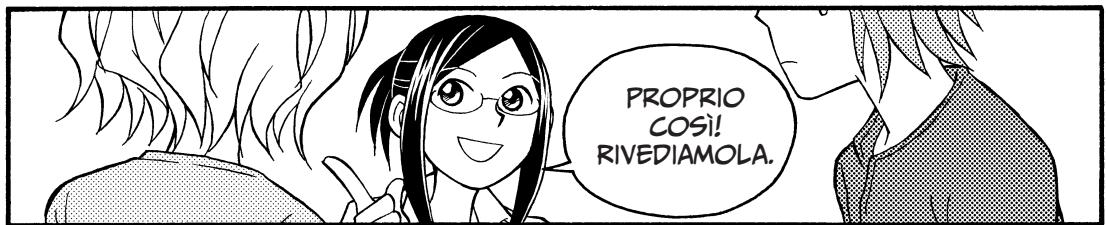
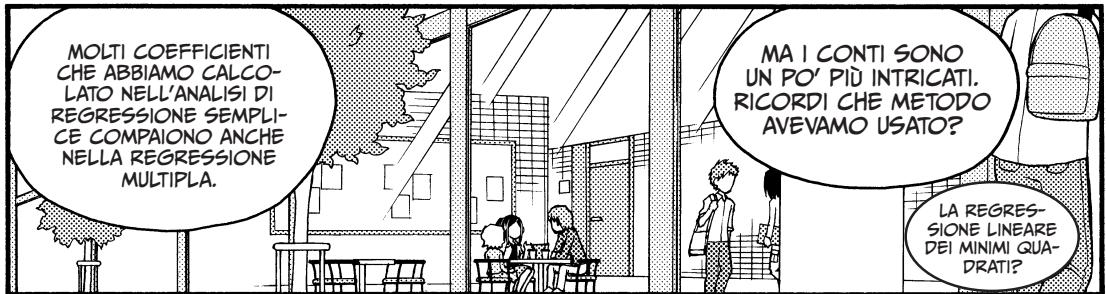
Pasticceria	Area del negozio (tsubo*)	Distanza dalla stazione più vicina (metri)	Vendita mensile (10.000 ¥)
Yumenooka	10	80	469
Stazione Terai	8	0	366
Sone	8	200	371
Stazione Hashimoto	5	200	208
Quartiere Kikyou	7	300	246
Ufficio postale	8	230	297
Stazione Suidobashi	7	40	363
Stazione Rokujo	9	0	436
Lungofiume Wakaba	6	330	198
Misato	9	180	364

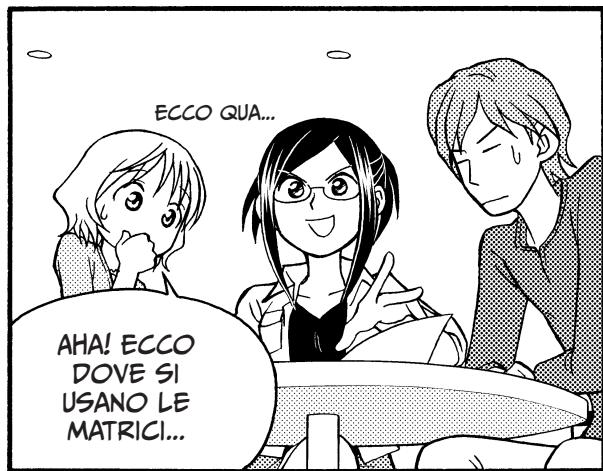
*1 tsubo equivale a circa 3,3 metri quadri.





PASSO 2: CALCOLO DELL'EQUAZIONE DI REGRESSIONE MULTIPLA



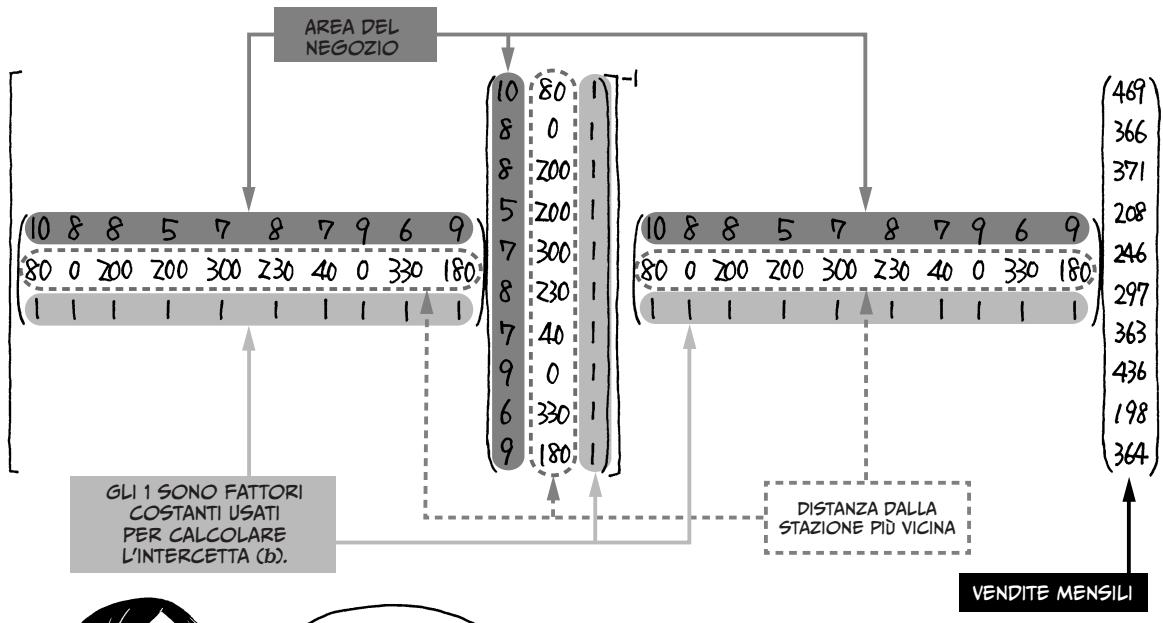


$$\begin{bmatrix} 10 & 80 & 1 \\ 8 & 0 & 1 \\ 8 & 200 & 1 \\ 5 & 200 & 1 \\ 7 & 300 & 1 \\ 8 & 230 & 1 \\ 7 & 40 & 1 \\ 9 & 0 & 1 \\ 6 & 330 & 1 \\ 9 & 180 & 1 \end{bmatrix}^{-1}$$
$$\begin{bmatrix} 469 \\ 366 \\ 371 \\ 208 \\ 246 \\ 297 \\ 363 \\ 436 \\ 198 \\ 364 \end{bmatrix}$$

QUESTO MALLOPPO È
UGUALE AL COEFFICIENTE
DI REGRESSIONE
PARZIALE!

...FACENDO
QUESTO
CALCOLO!

CHE DIAVOLO
È 'STA
ROBA?



E VA BENE.
FARÒ IO I
CONTI...

Variabile predittiva	Coefficienti di regressione parziali
Area del negozio (tsubo)	$a_1 = 41,5$
Distanza dalla stazione più vicina (metri)	$a_2 = -0,3$
Intercetta	$b = 65,3$

EVVIVA!!

*A PAGINA 209 TROVATE IL CALCOLO COMPLETO.



LA RETTA DESCRITTA
DALL'EQUAZIONE DI
REGRESSIONE MULTIPLA

$y = \alpha_1x_1 + \alpha_2x_2 + \dots + \alpha_px_p + b$
PASSERÀ SEMPRE PER IL
PUNTO $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p, \bar{y})$, DOVE \bar{x}_i
È LA MEDIA DI x_i .

ORA MI VA
IN PAPPA IL
CERVELLO.

MI RI-
CORDA
QUAL-
COSA...

RIFLETTI,
RIFLETTI. DOV'È
CHE L'HO GIÀ
VISTA?

IN ALTRI TERMINI, LA RETTA DESCRITTA DALLA NOSTRA EQUAZIONE

$y = 41,5x_1 - 0,3x_2 + 65,3$ PASSERÀ SEMPRE PER I PUNTI
CORRISPONDENTI AI VALORI MEDI DELL'AREA DEL NEGOZIO, DELLA
DISTANZA DALLA STAZIONE PIÙ VICINA E DELLE VENDITE MENSILI.

AH, ECCO!
QUANDO TRACCIAVI
IL GRAFICO DELL'E-
QUAZIONE, LA RETTA
PASSA PER I VALORI
MEDI.

PASSO 3: VALUTAZIONE DELL'ACCURATEZZA DELL'EQUAZIONE DI REGRESSIONE MULTIPLA

OK, ORA ABBIAMO
UN'EQUAZIONE, MA
POSSIAMO FARE
DAVVERO PREVI-
SIONI AFFIDABILI
SULLE VENDITE
DEL NUOVO NE-
GOZIO?

LO SCOPRIREMO CON
LA DIAGNOSTICA DI
REGRESSIONE. DOVRE-
MO TROVARE R^2 , E SE
È VICINO A 1, LA NOSTRA
EQUAZIONE È ABBASTAN-
ZA PRECISA!

CHE MEMORIA!

PRIMA DI R^2 DOBBIAMO TROVARE IL BUON VECCHIO R , CHE IN QUESTO CASO SI CHIAMA COEFFICIENTE DI CORRELAZIONE MULTIPLA. RICORDA CHE R È UN MODO DI CONFRONTARE I VALORI MISURATI (y) CON QUELLI STIMATI (\hat{y})*.



Pasticceria	Valore reale y	Valore stimato $\hat{y} = 41x_1 - 0,3x_2 + 65,3$	$y - \bar{y}$	$\hat{y} - \bar{\hat{y}}$	$(y - \bar{y})^2$	$(\hat{y} - \bar{\hat{y}})^2$	$(y - \bar{y})(\hat{y} - \bar{\hat{y}})$	$(y - \hat{y})^2$
Yumenooka	469	453,2	137,2	121,4	18823,8	14735,1	16654,4	250,0
Stazione Terai	366	397,4	34,2	65,6	1169,6	4307,5	2244,6	988,0
Sone	371	329,3	39,2	-2,5	1536,6	6,5	-99,8	1742,6
Stazione Hashimoto	208	204,7	-123,8	-127,1	15326,4	16150,7	15733,2	10,8
Quartiere Kikyou	246	253,7	-85,8	-78,1	7361,6	6016,9	6705,0	58,6
Ufficio postale	297	319,0	-34,8	-12,8	1211,0	163,1	444,4	485,3
Suidobashi	363	342,3	31,2	10,5	973,4	109,9	327,1	429,2
Stazione Rokujo	436	438,9	104,2	107,1	10857,6	11480,1	11164,5	8,7
Lungofiume Wakaba	198	201,9	-133,8	-129,9	17902,4	16870,5	17378,8	15,3
Misato	364	377,6	32,2	45,8	1036,8	2096,4	1474,3	184,6
Totale	3318	3318	0	0	76199,6	72026,6	72026,6	4173,0
Media	331,8	331,8						

↓
 \bar{y}

↓
 $\bar{\hat{y}}$

↓
 S_{yy}

↓
 $S_{\hat{y}\hat{y}}$

↓
 $S_{y\hat{y}}$



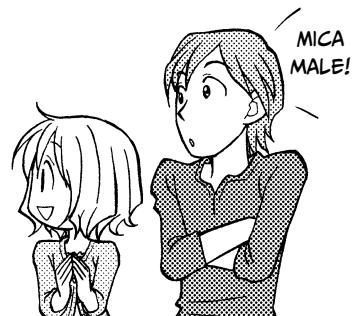
S_e PER ORA NON CI SERVE, MA PIÙ AVANTI LO USEREMO.

$$R = \frac{\text{somma di } (y - \bar{y})(\hat{y} - \bar{\hat{y}})}{\sqrt{\text{somma di } (y - \bar{y})^2 \times \text{somma di } (\hat{y} - \bar{\hat{y}})^2}} = \frac{S_{y\hat{y}}}{\sqrt{S_{yy} \times S_{\hat{y}\hat{y}}}} =$$

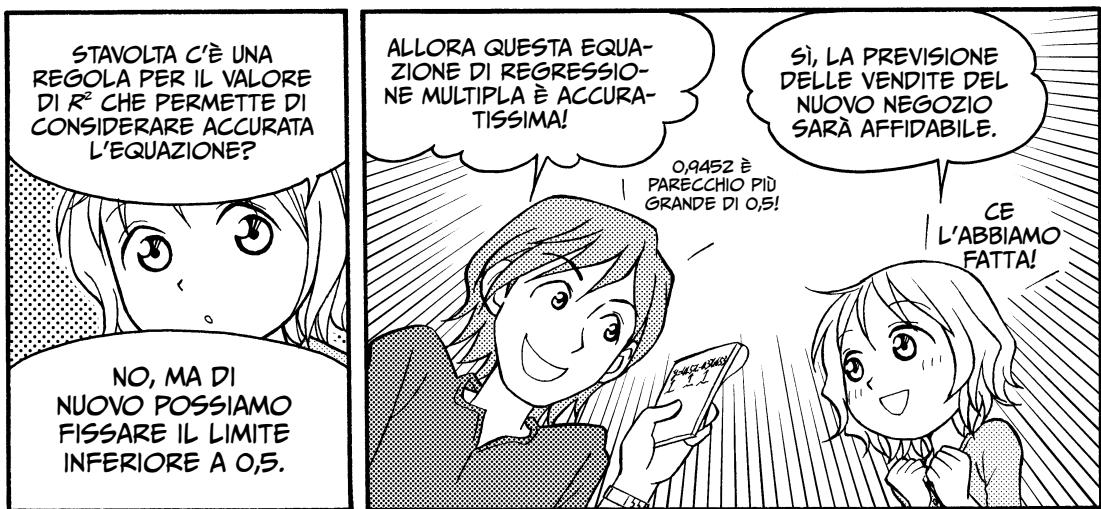
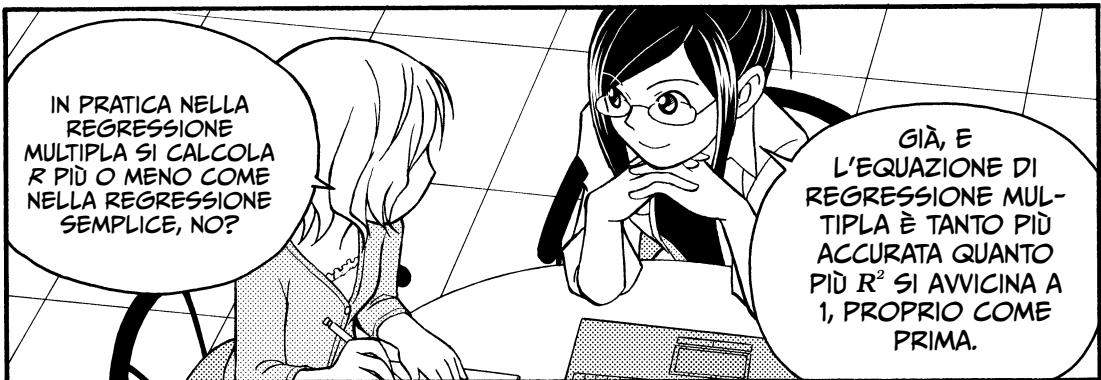
$$= \frac{72026,6}{\sqrt{76199,6 \times 72026,6}} = 0,9722$$

$$R^2 = (0,9722)^2 = 0,9452$$

R^2 VALE 0,9452.



*COME NEL CAPITOLO 2, ALCUNE CIFRE DI QUESTO CAPITOLO SONO ARROTONDATE PER MAGGIOR LEGGIBILITÀ, MA IN TUTTI I CALCOLI SONO STATE USATE LE CIFRE COMPLETE, NON ARROTONDATE, SALVO INDICAZIONE CONTRARIA.



*A PAGINA 144 TROVATE LA SPIEGAZIONE DI S_{1y}, S_{2y}, ..., S_{py}.

IL PROBLEMA DI R^2



Pasticceria	Area del negozio (tsubo)	Distanza dalla stazione più vicina (metri)	Età del negoziante (anni)	Vendite mensili (10.000 ¥)
Yumenooka	10	80	42	469
Stazione Terai	8	0	29	366
Sone	8	200	33	371
Stazione Hashimoto	5	200	41	208
Quartiere Kikyou	7	300	33	246
Ufficio postale	8	230	35	297
Suidobashi	7	40	40	363
Stazione Rokujo	9	0	46	436
Lungofiume Wakaba	6	330	44	198
Misato	9	180	34	364

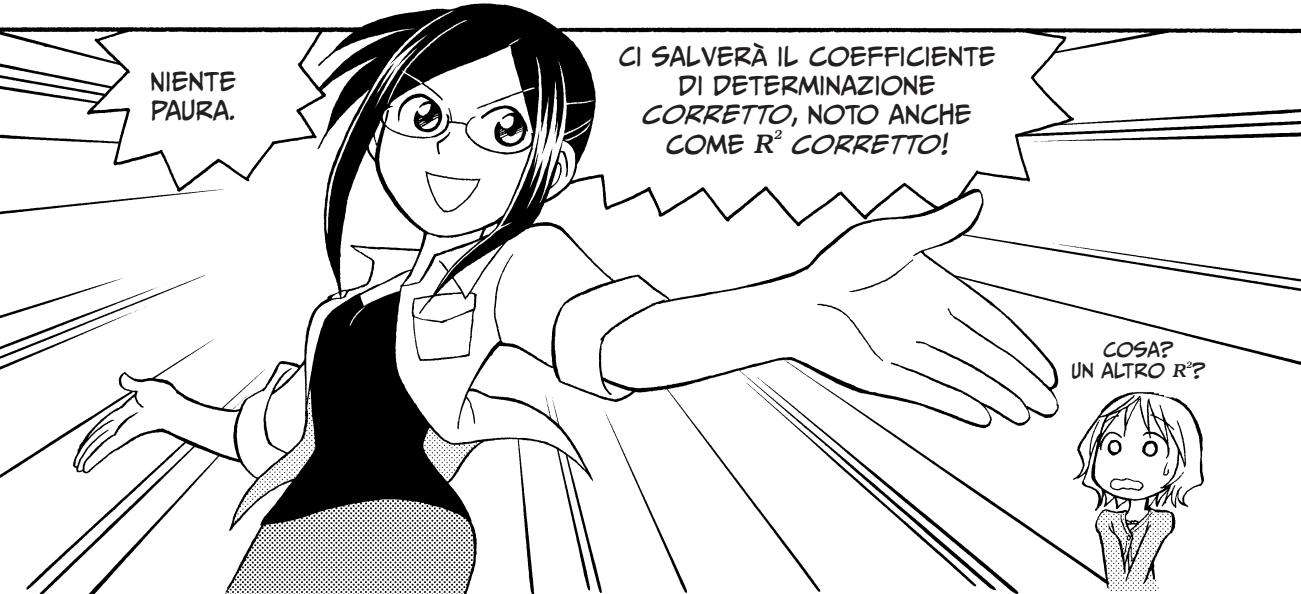
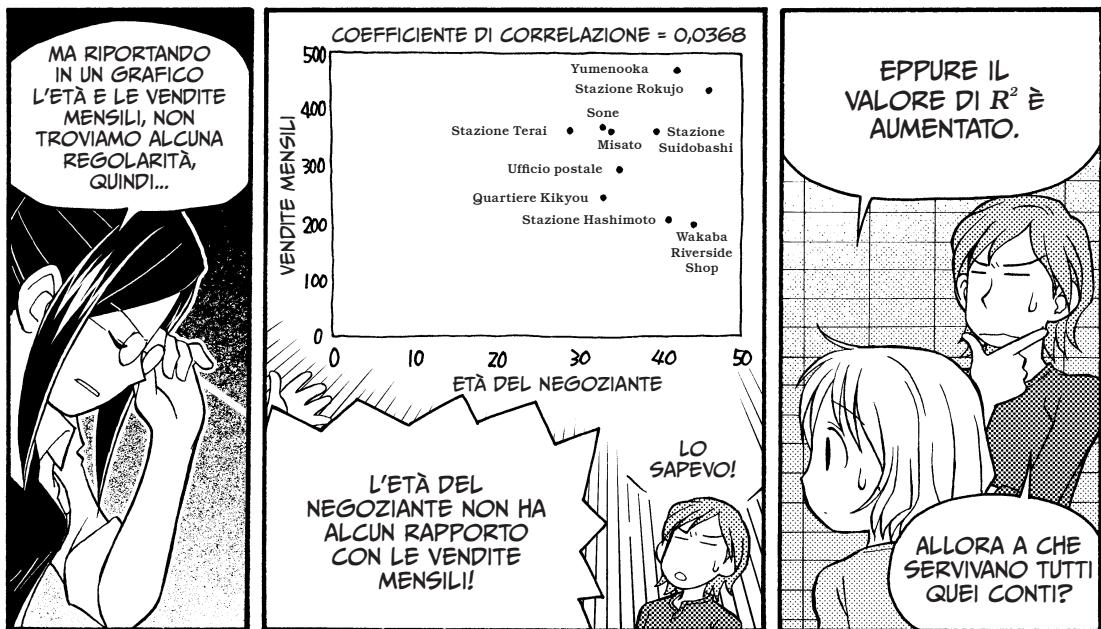
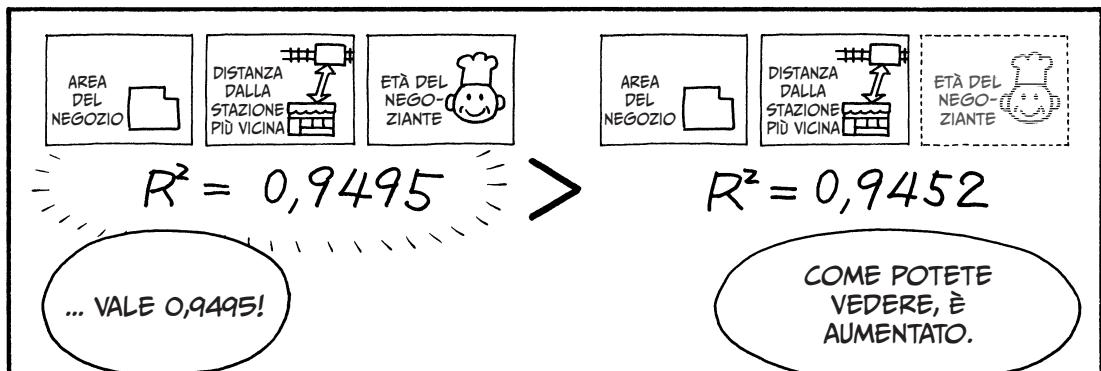
IMMAGINATE DI AGGIUNGERE AI DATI L'ETÀ DEL NEGOZIANTE.

ORA L'ETÀ È LA TERZA VARIABILE PREDITTIVA.

? MA CHE C'ENTRA L'ETÀ?

PRIMA DI AGGIUNGERE LA NUOVA VARIABILE, R^2 VALEVA 0,9452.

Dopo l'aggiunta di questa variabile...



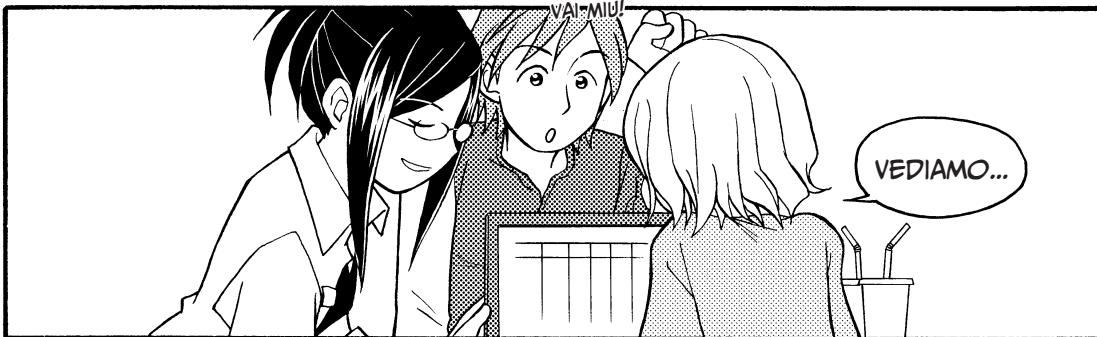
R² CORRETTO

IL VALORE DI R² CORRETTO (\bar{R}^2) SI OTTIENE CON QUESTA FORMULA.

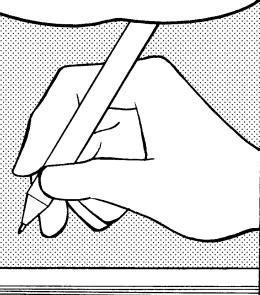
$$\bar{R}^2 = 1 - \left(\frac{\frac{S_e}{\text{dimensioni del campione} - \text{numero di variabili predittive} - 1}}{\left(\frac{S_{yy}}{\text{dimensioni del campione} - 1} \right)} \right)$$



MIU, SAPRESTI CALCOLARE IL VALORE DI R² CORRETTO CON O SENZA L'ETÀ DEL NEGOZIANTE?



QUANDO LE VARIABILI PREDITTIVE SONO SOLO L'AREA DEL NEGOZIO E LA DISTANZA...

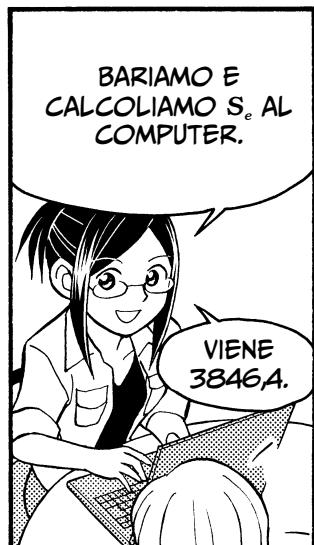
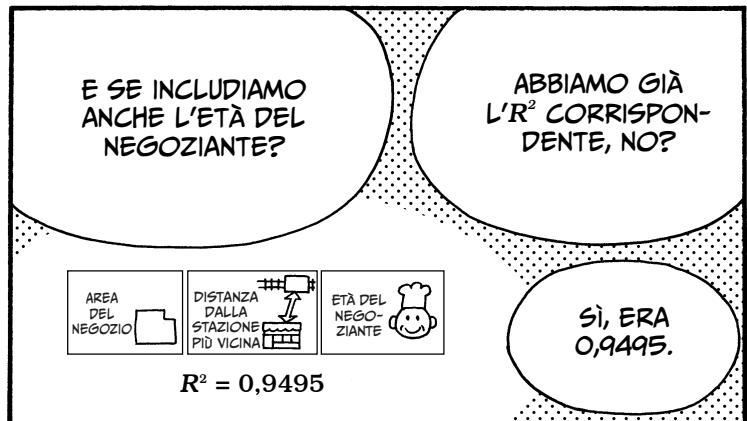
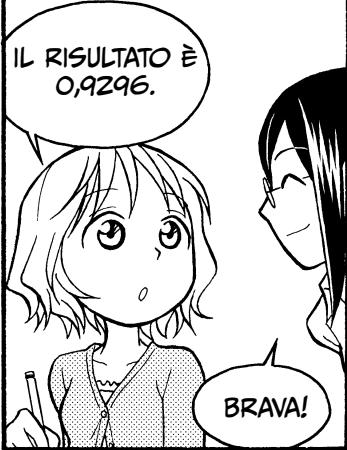


...R² VALE 0,9452.

QUINDI R² CORRETTO VALE:

$$1 - \left(\frac{\frac{S_e}{\text{dimensioni del campione} - \text{numero di variabili predittive} - 1}}{\left(\frac{S_{yy}}{\text{dimensioni del campione} - 1} \right)} \right) \\ = 1 - \frac{\left(\frac{4173,0}{10 - 2 - 1} \right)}{\left(\frac{76199,6}{10 - 1} \right)} = \underline{\underline{0,9296}}$$

L'HO TROVATO!



VARIABILI PREDITTIVE:

- AREA DEL NEGOZIO
- DISTANZA
- ETÀ DEL NEGOZIANTE

$$1 - \frac{\frac{S_e}{\text{dimensioni del campione} - \text{numero di variabili predittive} - 1}}{\left(\frac{S_{yy}}{\text{dimensioni del campione} - 1} \right)}$$

$$= 1 - \frac{\left(\frac{3846,4}{10 - 3 - 1} \right)}{\left(\frac{76199,6}{10 - 1} \right)} = \underline{\underline{0,9243}}$$

ASPETTA
UN PO'...

VARIABILI PREDITTIVE


GUARDA! R^2 CORRETTO È PIÙ GRANDE SE NON INCLUDIAMO L'ETÀ DEL NEGOZIANTE.

①
AREA DEL NEGOZIO E DISTANZA

②
AREA DEL NEGOZIO, DISTANZA ED ETÀ

R^2	0,9452 < 0,9495
\bar{R}^2	0,9296 > 0,9243


FUNZIONA!


VISTO? R^2 CORRETTO CI HA SALVATO!

EHI, GUARDATE UN PO' QUA.

R^2 CORRETTO È MINORE DI R^2 IN ENTRAMBI I CASI. È SEMPRE COSÌ?

	① AREA DEL NEGOZIO E DISTANZA	② AREA DEL NEGOZIO, DISTANZA ED ETÀ
--	----------------------------------	--

R^2	0,9452	0,9495
\bar{R}^2	0,9296	0,9243

OCCHIO DI LINCE! SÌ, È SEMPRE MINORE.

È UN BUON SEGNO?

SIGNIFICA CHE R^2 CORRETTO VALUTA L'ACCURATEZZA IN MODO PIÙ RIGOROSO, RENDENDO COSÌ PIÙ AFFIDABILE LA NOSTRA EQUAZIONE DI REGRESSIONE MULTIPLA.

R^2 CORRETTO È UNA FORZA.

VERIFICA DI IPOTESI CON LA REGRESSIONE MULTIPLA

ORA...

VISTO CHE R^2 CORRETTO CI STA BENE, VERIFicheremo le nostre ipotesi sulla popolazione.

\bar{R}^2

FAREMOS LA VERIFICA DI IPOTESI E CONTROLLEREMO IL COEFFICIENTE DI REGRESSIONE, GIUSTO?

SI, MA ADESSO NELL'ANALISI DI REGRESSIONE MULTIPLA CI SONO I COEFFICIENTI DI REGRESSIONE PARZIALI.

TI RICORDI COME FACEVAMO PRIMA LA VERIFICA DI IPOTESI?

MI SA DI SI. CONTROLLAVAMO CHE LA POPOLAZIONE CORRISPONDESSE ALL'EQUAZIONE E CHE A NON SI ANNULLASSE.

GIUSTO! NELLA REGRESSIONE MULTIPLA È PIÙ O MENO LA STESSA STORIA.

~ IPOTESI ALTERNATIVA ~

SE L'AREA DEL NEGOZIO È PARI A x_1 , TSUBO E LA STAZIONE PIÙ VICINA DISTA x_2 METRI, LE VENDITE MENSILI SEGUONO UNA DISTRIBUZIONE NORMALE CON MEDIA PARI A $A_1x_1 + A_2x_2 + B$ E DEVIAZIONE STANDARD σ .

ORA ABBIAMO PIÙ DI UNA x E PIÙ DI UN PARAMETRO A. QUESTE A NON DEVONO ESSERE TUTTE NULLE.

CHIARO!

PASSO 4: FARE IL TEST DI ANALISI DELLA VARIANZA (ANOVA)

ECCO LE NOSTRE IPOTESI
SUI COEFFICIENTI DI
REGRESSIONE PARZIALI.
 a_1 , a_2 E b SONO I
COEFFICIENTI DELL'INTERA
POPOLAZIONE.



SE OTTENIAMO LA SEGUENTE EQUAZIONE DI REGRESSIONE

$$y = a_1x_1 + a_2x_2 + b$$

- A_1 È CIRCA a_1 ;
- A_2 È CIRCA a_2 ;
- B È CIRCA b ;

$$\sigma = \sqrt{\frac{S_e}{\text{dimensioni del campione} - \text{numero di variabili predittive} - 1}}$$

SAPRESTI APPLICARE QUESTA FORMULA AI DATI DELLE PASTICCERIE KAZAMI?

CERTO.

L'EQUAZIONE DI REGRESSIONE MULTIPLA
È $y = 41,5x_1 - 0,3x_2 + 65,3$, QUINDI...

ECCO LE NOSTRE IPOTESI.

- A_1 È CIRCA 41,5;
 - A_2 È CIRCA -0,3;
 - B È CIRCA 65,3;
- $$\cdot \sigma = \sqrt{\frac{4173,0}{10-2-1}} = 24,4.$$

FANTASTICO!



ORA DOBBIAMO
VERIFICARE IL
NOSTRO MODELLO
CON UN TEST F.

CE NE SONO
DUE TIPI.

IL PRIMO CONTROLLA
I COEFFICIENTI DI
REGRESSIONE PARZIALI
TUTTI IN UNA VOLTA.

IL SECONDO TIPO DI
TEST CONTROLLA SEPARA-
MENTE I COEFFICIENTI DI RE-
GRESSIONE PARZIALI.

IPOTESI NULLA	$A_1 = 0 \text{ E } A_2 = 0$
IPOTESI ALTERNATIVA	NON VALE $A_1 = A_2 = 0$
IN ALTRI TERMINI, SIAMO IN UNO DI QUESTI CASI:	
• $A_1 \neq 0 \text{ E } A_2 \neq 0$	
• $A_1 \neq 0 \text{ E } A_2 = 0$	
• $A_1 = 0 \text{ E } A_2 \neq 0$	

IPOTESI NULLA	$A_i = 0$
IPOTESI ALTERNATIVA	$A_i \neq 0$



FISSIAMO IL
LIVELLO DI
SIGNIFICATIVITÀ
A 0,05. TE LA
SENTI DI FARE
QUESTI TEST?

SÌ, DAI!

INIZIAMO ESAMINANDO TUTTI I COEFFICIENTI DI REGRESSIONE PARZIALI ALLO STESSO TEMPO.



I PASSI DI ANOVA

Passo 1	Definire la popolazione.	La popolazione è composta da tutte le pasticcerie Kazami.
Passo 2	Formulare l'ipotesi nulla e l'ipotesi alternativa.	L'ipotesi nulla è $A_1 = 0$ e $A_2 = 0$. L'ipotesi alternativa è che A_1 e A_2 non si annullino entrambi.
Passo 3	Scegliere il test di ipotesi.	Effettuiamo un test F.
Passo 4	Fissare il livello di significatività.	Poniamo il livello di significatività a 0,05.
Passo 5	Calcolare la statistica test dai dati campione.	La statistica test è: $\frac{S_{yy} - S_e}{\text{numero di variabili predittive}} \div \frac{S_e}{\text{dimensioni del campione} - \text{numero di variabili predittive} - 1} =$ $\frac{76199,6 - 4173,0}{2} \div \frac{4173,0}{10 - 2 - 1} = 60,4$ La statistica test, 60,4, segue una distribuzione F con il primo grado di libertà pari a 2 (numero delle variabili predittive) e il secondo pari a 7 (dimensioni del campione - numero di variabili predittive - 1), se l'ipotesi nulla è vera.
Passo 6	Confrontare il p-value della statistica ricavato al Passo 5 con il livello di significatività.	Al livello di significatività 0,05, con $d_1 = 2$ e $d_2 = 7$ ($10 - 2 - 1$), il valore critico è 4,7374. La statistica test è 60,4.
Passo 7	Accettare o scartare l'ipotesi nulla.	Poiché la statistica test è maggiore del valore critico, scartiamo l'ipotesi nulla.

ORA INVECE CONTROLLIAMO I COEFFICIENTI DI REGRESSIONE PARZIALI UNO PER UNO. VEDIAMO IN DETTAGLIO IL CASO DI A_1 .



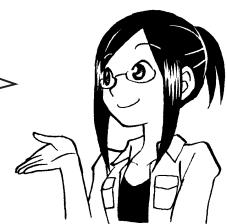
I PASSI DI ANOVA

Passo 1	Definire la popolazione.	La popolazione è composta da tutte le pasticcerie Kazami.
Passo 2	Formulare l'ipotesi nulla e l'ipotesi alternativa.	L'ipotesi nulla è $A_1 = 0$. L'ipotesi alternativa è $A_1 \neq 0$.
Passo 3	Scegliere il test di ipotesi.	Effettuiamo un test F.
Passo 4	Fissare il livello di significatività.	Poniamo il livello di significatività a 0,05.
Passo 5	Calcolare la statistica test dai dati campione.	La statistica test è: $\frac{a_1^2}{S_{11}} \div \frac{S_e}{\text{dimensioni del campione} - \text{numero di variabili predittive} - 1} =$ $\frac{41,5^2}{0,0657} \div \frac{4173,0}{10 - 2 - 1} = 44$ <p>La statistica test segue una distribuzione F con il primo grado di libertà pari a 1 e il secondo pari a 7 (dimensioni del campione – numero di variabili predittive – 1), se l'ipotesi nulla è vera (il valore di S_{11} è spiegato alla pagina seguente).</p>
Passo 6	Confrontare il p-value della statistica ricavato al Passo 5 con il livello di significatività.	Al livello di significatività 0,05, con $d_1 = 1$ e $d_2 = 7$ ($10 - 2 - 1$), il valore critico è 5,5914. La statistica test è 44.
Passo 7	Accettare o scartare l'ipotesi nulla.	Poiché la statistica test è maggiore del valore critico, scartiamo l'ipotesi nulla.

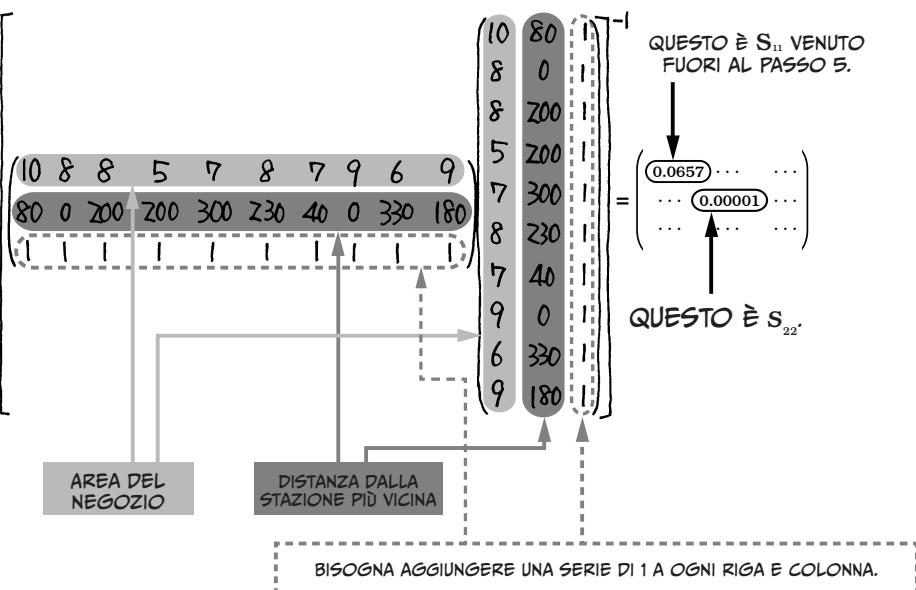
A PRESCINDERE DAL RISULTATO DEL PASSO 7,
SE IL VALORE DELLA STATISTICA TEST

$$\frac{a_1^2}{S_{11}} \div \frac{S_e}{\text{dimensioni del campione} - \text{numero di variabili predittive} - 1}$$

È MAGGIORE O UGUALE A Z, LA VARIABILE PREDITTIVA CORRISPONDENTE A QUEL COEFFICIENTE DI REGRESSIONE PARZIALE È COMUNQUE RITENUTA UTILE PER PREVEDERE LA VARIABILE RESPONSO.



DETERMINAZIONE DI
 S_{11} E S_{22}



SI USA UNA MATRICE PER DETERMINARE S_{11} E S_{22} . ALLA PAGINA PRECEDENTE S_{11} CI È SERVITO PER CALCOLARE LA STATISTICA TEST; ORA USIAMO S_{22} ALLO STESSO MODO, PER VERIFICARE INDIPENDENTEMENTE IL SECONDO COEFFICIENTE*.

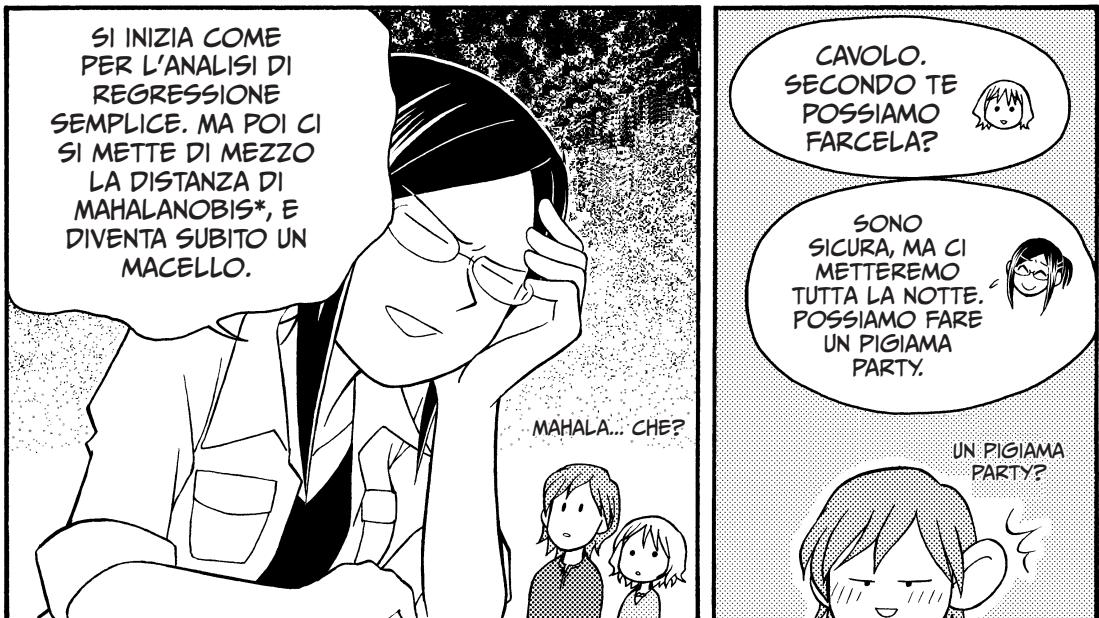
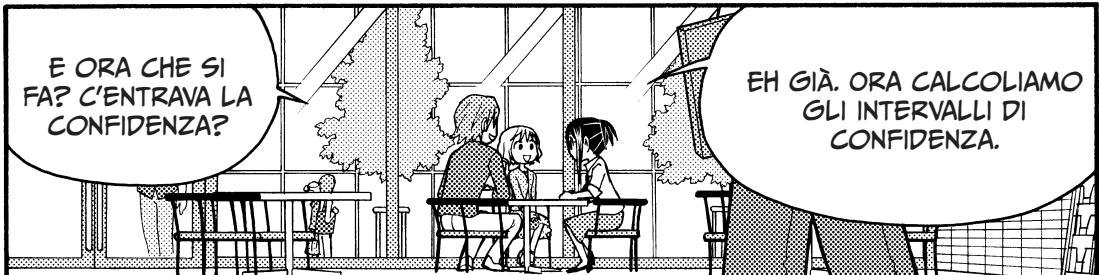


ALLORA A₁ NON È ZERO! POSSIAMO RIFIUTARE L'IPOTESI NULLA.

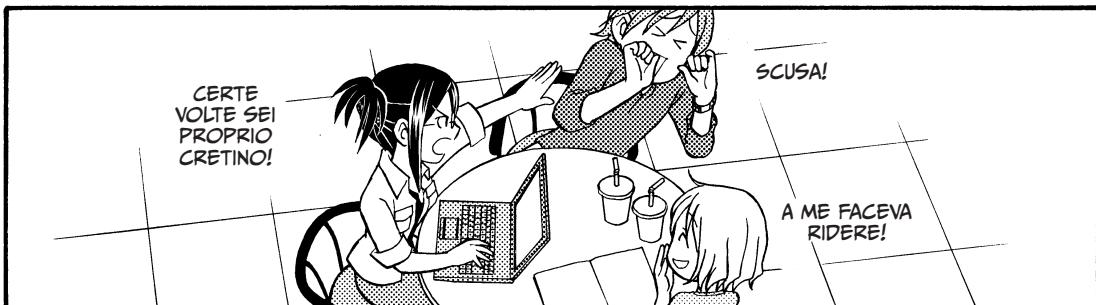
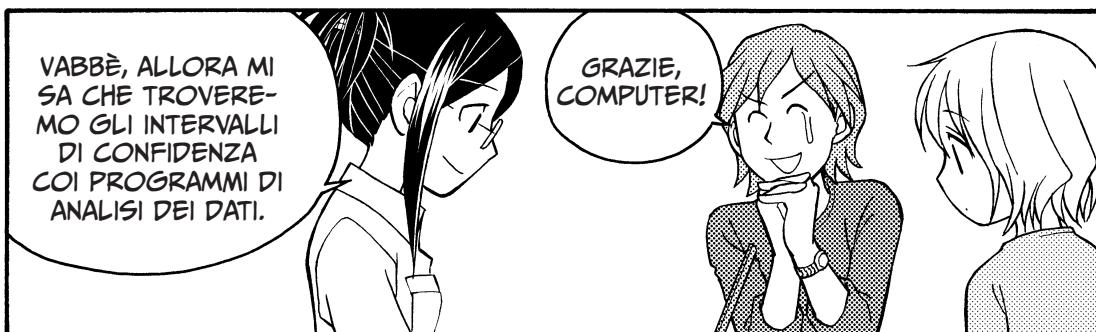
CE L'HA FATTA! SEI IL MIO MITO, MIU!

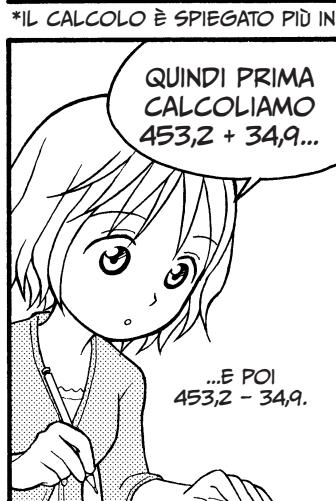
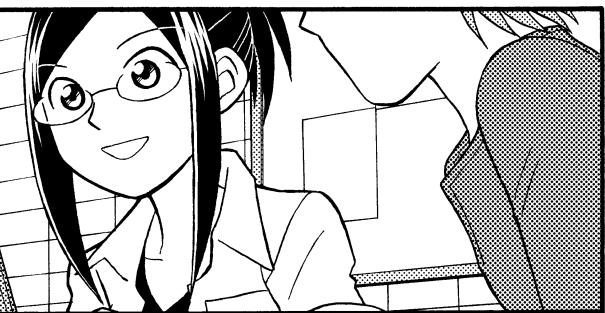
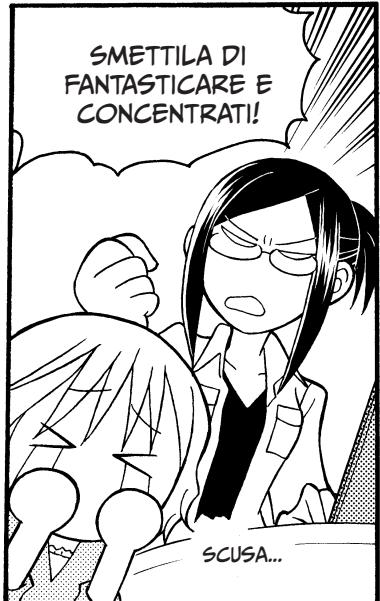
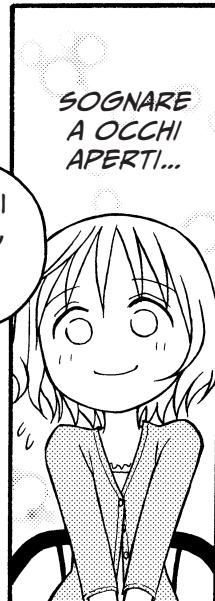
*INVECE DELLA DISTRIBUZIONE F, ALCUNI PREFERISCONO USARE LA DISTRIBUZIONE t PER SPIEGARE IL "TEST DEI COEFFICIENTI DI REGRESSIONE PARZIALI". I RISULTATI FINALI SARANNO COMUNQUE INDIPENDENTI DAL METODO SCELTO.

PASSO 5: CALCOLO DEGLI INTERVALLI DI CONFIDENZA PER LA POPOLAZIONE



*DOBBIAMO AL MATEMATICO P.C. MAHALANOBIS IL CONFRONTO DI POPOLAZIONI TRAMITE LE DISTANZE MULTIVARIATE.





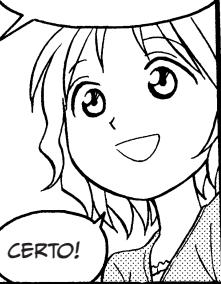
PASSO 6: FACCIAMO UNA PREVISIONE!

ECCO I DATI
PER IL NEGOZIO
DI PROSSIMA
APERTURA.

	Area del negozio (tsubo)	Distanza dalla stazione più vicina (metri)
Isebashi	10	110

UN NEGOZIO
A ISEBASHI? È
DIETRO CASA
MIA!

SAPRESTI
PREVEDERE
LE VENDITE,
MIU?



CERTO!

$$\begin{aligned}
 y &= 41,5x_1 - 0,3x_2 + 65,3 \\
 &= 41,5 \times 10 - 0,3 \times 110 + 65,3 \\
 &= \underline{447,3}^*
 \end{aligned}$$

4.473.000
YEN AL MESE!

*IL CALCOLO È STATO SVOLTO CON CIFRE ARROTONDATE. USANDO LE CIFRE COMPLETE, NON ARROTONDATE, IL RISULTATO È DI 442,96.

SEI UN GENIO, MIU!
BISOGNEREBBE
INTITOLARTI IL NUOVO
NEGOZIO.

MI SA CHE
DOVRESTI
INTITOLAR-
LO A RISA...



MA COME POSSIAMO
SAPERE LE VENDITE
ESATTE DI UN NEGOZIO
CHE ANCORA NON
ESISTE? NON DOVREMMO
CALCOLARE L'INTERVALLO
DI PREVISIONE?

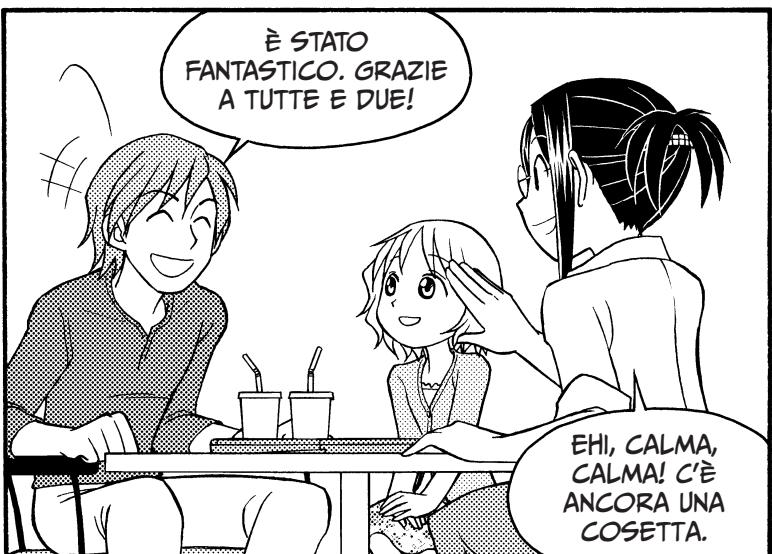
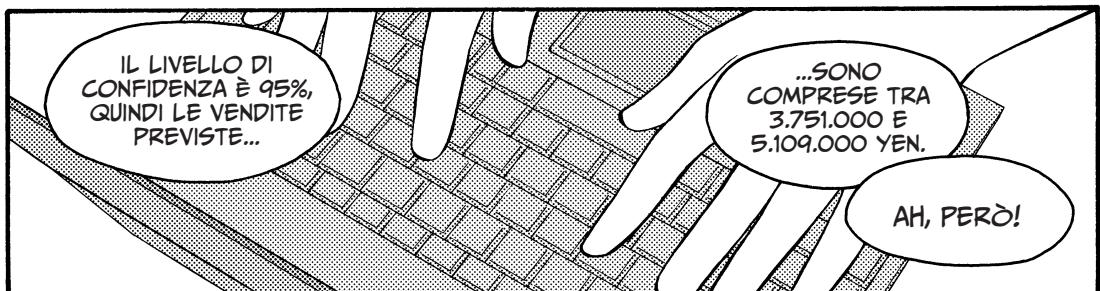
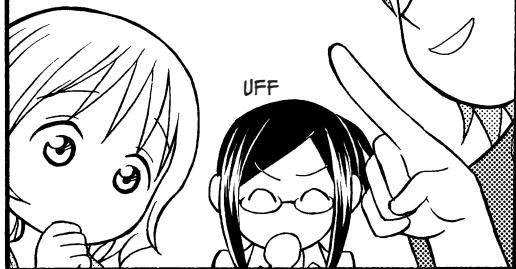
CERTAMENTE.

NELL'ANALISI DI REGRESSIONE
SEMPLICE, I METODI PER TROVARE
GLI INTERVALLI DI CONFIDENZA
E DI PREVISIONE ERANO SIMILI.
È COSÌ ANCHE PER L'ANALISI DI
REGRESSIONE MULTIPLA?

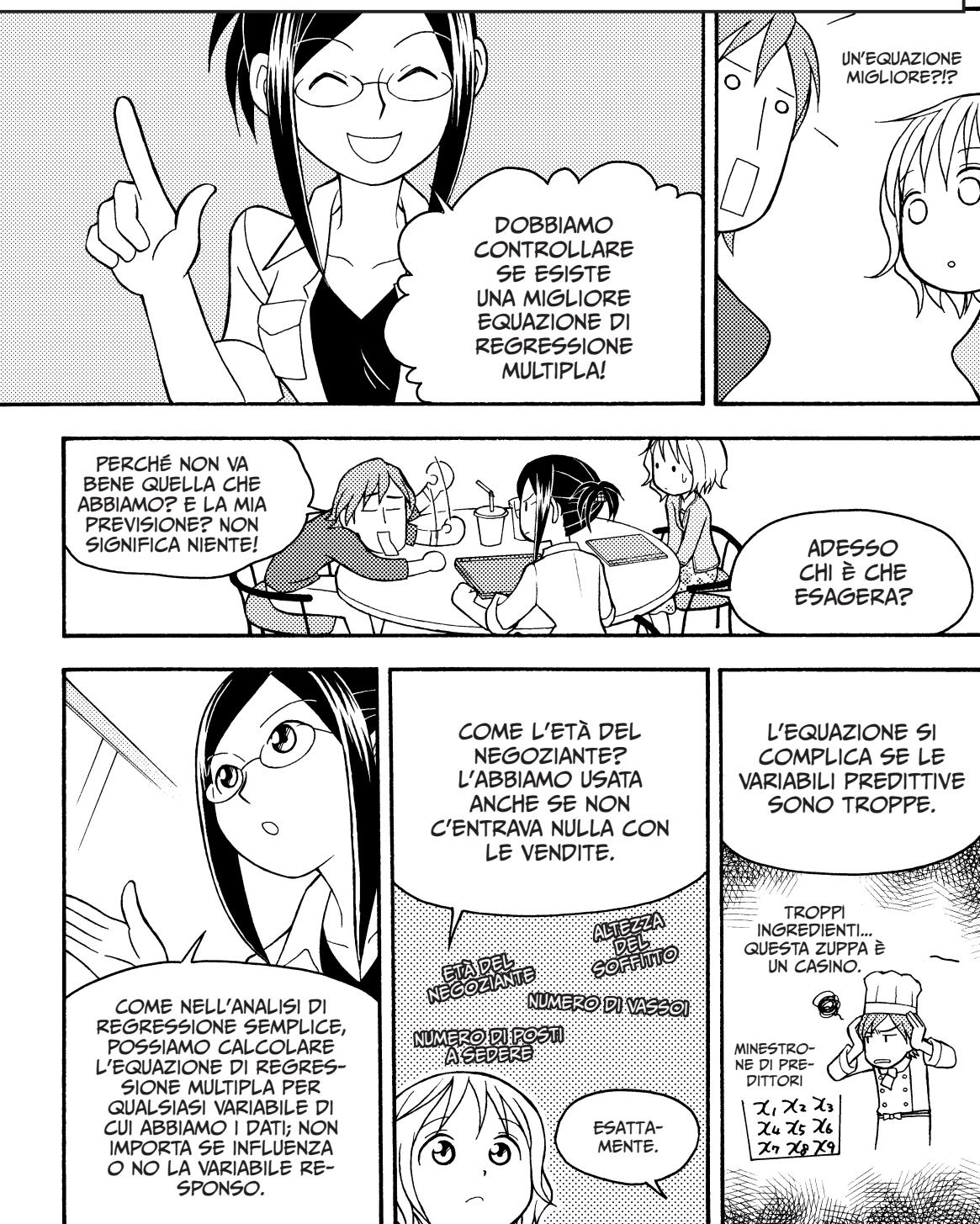
SÌ, C'È POCA
DIFERENZA.



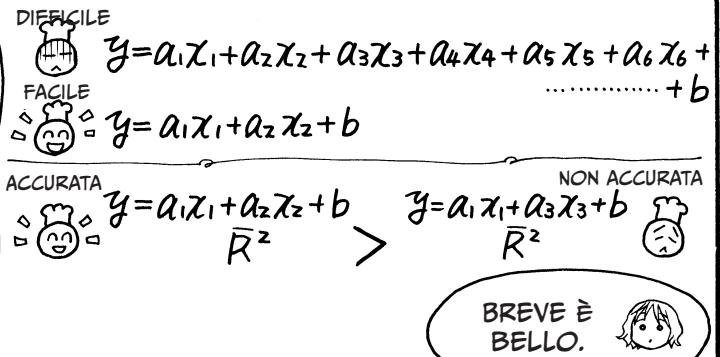
SÌ, LA DI-
STANZA CO-
MESI CHIAMA.



SCELTA DELLA MIGLIORE COMBINAZIONE DI VARIABILI PREDITTIVE



LA MIGLIORE EQUAZIONE DI REGRESSIONE RAPPRESENTA IL GIUSTO EQUILIBRIO FRA ACCURATEZZA E COMPLESSITÀ, INCLUDENDO SOLO LE VARIABILI PREDITTIVE NECESSARIE A FARE LA PREVISIONE MIGLIORE.



CI SONO VARI MODI PER INDIVIDUARE L'EQUAZIONE CHE SALVA CAPRA E CAVOLI.

- SELEZIONE IN AVANTI
- ELIMINAZIONE ALL'INDIETRO
- SELEZIONE AVANTI-INDIETRO PER PASSI
- INDIVIDUAZIONE DELLE VARIABILI PIÙ IMPORTANTI GRAZIE A UN ESPERTO DEL CAMPO

ECCO I PIÙ COMUNI.

OGGI USEREMO UN METODO ANCORA PIÙ SEMPLICE, CHE SI CHIAMA REGRESSIONE DEI SOTTOINSIEMI MIGLIORI, O ANCHE METODO ROUND ROBIN.

ROUND
ROBIN?

ROBIN
HOOD È
INGRAS-
SATO?

CHE CAVOLO
È?

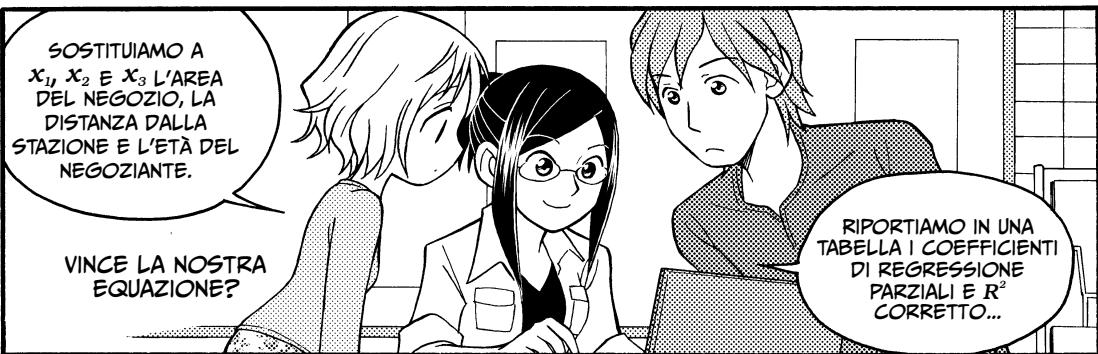
$x_1 \ x_2 \ x_3$

ORA TI SPIEGO.
IMMAGINA CHE
 x_1, x_2 E x_3 SIANO
POTENZIALI
VARIABILI
PREDITTIVE.

PRIMA DI TUTTO CALCOLIAMO
L'EQUAZIONE DI REGRESSIONE
MULTIPLA PER OGNI COMBINAZIONE DI
VARIABILI PREDITTIVE!

• x_1	• $x_1 x_2$	• $x_1 x_2 x_3$
• x_2	• $x_2 x_3$	
• x_3	• $x_1 x_3$	

AH, AH, AH!
UN SISTEMA
DAVERO
DIRETTO!

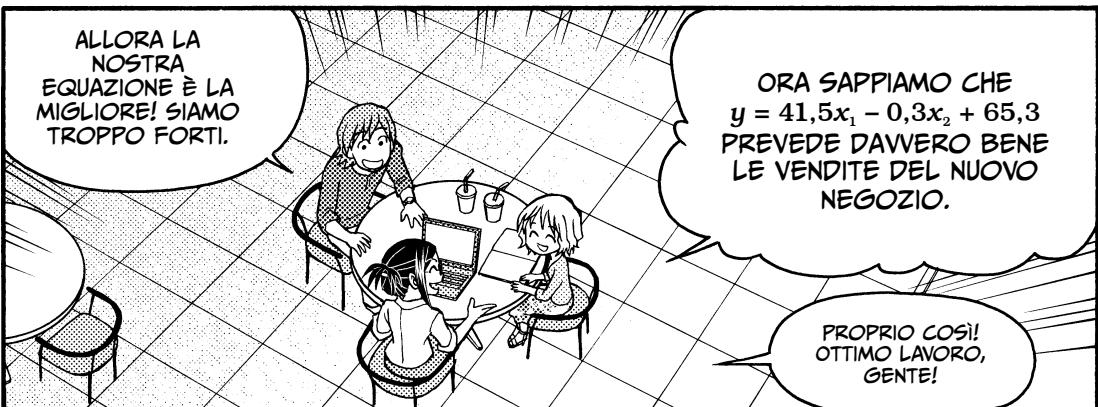


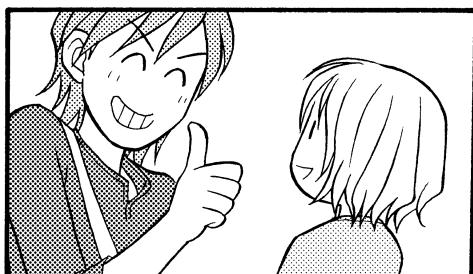
...COSÌ.

OPLÀ!

Variabili predittive	a_1	a_2	a_3	b	\bar{R}^2
1	54,9			-91,3	0,07709
2		-0,6		424,8	0,5508
3			0,6	309,1	0,0000
1 e 2	41,5	-0,3		65,3	0,9296
1 e 3	55,6		2,0	-170,1	0,7563
2 e 3		-0,6	-0,4	438,9	0,4873
1 e 2 e 3	42,2	-0,3	1,1	17,7	0,9243

1 È L'AREA DEL NEGOZIO, 2 È LA DISTANZA DALLA STAZIONE E 3 È L'ETÀ DEL NEGOZIANTE. R^2 CORRETTO È MASSIMO QUANDO SI USANO 1 E 2.





STIMA DI POPOLAZIONI CON L'ANALISI DI REGRESSIONE MULTIPLA

Rivediamo il procedimento per l'analisi di regressione multipla, illustrato a pagina 112.

1. Tracciare dei grafici di dispersione di ogni variabile predittiva e della variabile responso, per vedere se sembrano correlate.
2. Calcolo dell'equazione di regressione multipla.
3. Valutazione dell'accuratezza dell'equazione di regressione multipla.
4. Svolgimento del test di analisi della varianza (ANOVA).
5. Calcolo degli intervalli di confidenza per la popolazione.
6. Facciamo una previsione!

Come nel capitolo 2, abbiamo svolto i Passi da 1 a 6 come se fossero tutti obbligatori. In realtà, certi insiemi di dati permettono di saltare i Passi 4 e 5.

Al momento ci sono solo 10 pasticcerie Kazami, di cui una sola ha un'area di 10 tsubo¹ e dista 80 m dalla stazione più vicina. Risa però ha calcolato un intervallo di confidenza per la popolazione di negozi con l'area di 10 tsubo distanti 80 m da una stazione. Perché mai?

Be', è possibile che la famiglia Kazami decida di aprire un'altra pasticceria con l'area di 10 tsubo distante 80 m da una stazione ferroviaria. Se la catena continua a espandersi, decine di pasticcerie Kazami potrebbero avere queste caratteristiche. Svolgendo l'analisi, Risa ha ipotizzato l'eventuale apertura di altre pasticcerie con l'area di 10 tsubo distanti 80 m da una stazione.

L'utilità di quest'ipotesi è dubbia. Visto che la pasticceria a Yumenooka vende più di tutte le altre, la famiglia Kazami potrebbe decidere di aprire altri negozi simili. Il prossimo, a Isebashi, avrà un'area di 10 tsubo, ma disterà 110 m da una stazione. Anzi, forse non era necessario analizzare una popolazione di negozi tanto specifica. Risa avrebbe potuto passare subito dal calcolo di R^2 corretto alla previsione, ma da buona amica voleva mostrare a Miu tutti i passi.

1. Ricordate che 1 tsubo equivale a circa 3,3 metri quadri.

RESIDUI STANDARDIZZATI

Come nell'analisi di regressione semplice, in quella di regressione multipla si calcolano i residui standardizzati per valutare se l'equazione approssima bene i dati campione raccolti.

TABELLA 3-1: RESIDUI STANDARDIZZATI NELL'ESEMPIO DELLE PASTICCERIE KAZAMI

Pasticceria	Area del negozio x_1	Distanza dalla stazione più vicina x_2	Vendite mensili y	Vendite mensili $\hat{y} = 41,5x_1 - 0,3x_2 + 65,3$	Residuo $y - \hat{y}$	Residuo standardizzato
Yumenooka Shop	10	80	469	453,2	15,8	0,8
Stazione Terai	8	0	366	397,4	-31,4	-1,6
Sone	8	200	371	329,3	41,7	1,8
Stazione Hashimoto	5	200	208	204,7	3,3	0,2
Quartiere Kikyou	7	300	246	253,7	-7,7	-0,4
Ufficio postale	8	230	297	319,0	-22,0	1,0
Stazione Suidobashi	7	40	363	342,3	20,7	1,0
Stazione Rokujo	9	0	436	438,9	-2,9	-0,1
Lungofiume Wakaba	6	330	198	201,9	-3,9	-0,2
Misato	9	180	364	377,6	-13,6	-0,6

Se il residuo è positivo, la misura è maggiore di quanto preveda l'equazione: se è negativo, la misura è minore del previsto; se il residuo è nullo, la misura coincide con la previsione. Il valore assoluto del residuo indica se l'equazione riproduce bene la realtà: maggiore è il valore assoluto, maggiore è la differenza tra la misura e la previsione.

Se il valore assoluto del residuo standardizzato è maggiore di 3, possiamo concludere che il dato è *anomalo*. Le misure anomale sono quelle che non seguono l'andamento generale. In questo caso, il dato anomalo potrebbe corrispondere alla chiusura di una pasticceria, a lavori stradali nei dintorni, o a un evento speciale che si è tenuto nel negozio: qualunque cosa possa influenzare significativamente le vendite. Quando si individua un valore anomalo, bisogna studiarne le cause per capire se occorre rimuoverlo e ricalcolare l'equazione di regressione.

DISTANZA DI MAHALANOBIS

La *distanza di Mahalanobis* è stata definita nel 1936 dal matematico e scienziato P.C. Mahalanobis, fondatore dell'Indian Statistical Institute; è un concetto statistico utilissimo perché considera l'intero insieme dei dati, invece di valutare ogni misura singolarmente. Al contrario della definizione abituale di distanza, quella euclidea, la distanza di Mahalanobis tiene conto della correlazione tra misure per determinare la somiglianza di un campione a un certo insieme di dati. Poiché i calcoli riflettono un legame più complesso, le equazioni lineari non bastano. Si usano invece le matrici, che condensano una complessa griglia di informazioni in forma più maneggevole, utilizzabile per calcolare tutte queste distanze in un colpo solo.

A pagina 137, Risa ha usato il computer per ricavare l'intervallo di previsione dalla distanza di Mahalanobis. Rivediamo ora il calcolo per capire come ha ottenuto un intervallo di previsione compreso tra 3.751.000 e 5.109.000 yen, al livello di confidenza del 95%.

PASSO 1

Calcoliamo la matrice inversa di

$$\begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{pmatrix}, \text{ che è } \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{pmatrix}^{-1} = \begin{pmatrix} S^{11} & S^{12} & \cdots & S^{1p} \\ S^{21} & S^{22} & \cdots & S^{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S^{p1} & S^{p2} & \cdots & S^{pp} \end{pmatrix}.$$

La prima è la matrice di covarianza, calcolata a pagina 132. La diagonale di questa matrice (S_{11} , S_{22} eccetera) corrisponde alla varianza di una certa variabile.

L'inversa di questa matrice, cioè la terza matrice riportata, è anche detta *matrice di concentrazione* per le diverse variabili predittive: l'area del negozio e la distanza dalla stazione più vicina.

Per esempio, S_{22} è la varianza dei valori della distanza dalla stazione più vicina. S_{25} sarebbe la covarianza della distanza dalla stazione più vicina e di una quinta variabile predittiva.

A pagina 132 i valori di S_{ii} e S_{22} sono stati ottenuti in questa maniera. I valori di S_{ii} e S_{ij} in

$$\begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{pmatrix}^{-1}$$

sono sempre uguali ai valori di S_{ii} e S_{ij} ottenuti svolgendo test individuali sui coefficienti di regressione parziali. Cioè, i valori di S_{ii} e S_{ij} ottenuti con la regressione parziale saranno equivalenti a quelli trovati calcolando la matrice inversa.

PASSO 2

Ora calcoliamo il quadrato della distanza di Mahalanobis per un punto dato tramite l'equazione seguente:

$$D_M^2(x) = (x - \bar{x})^T (S^{-1})(x - \bar{x})$$

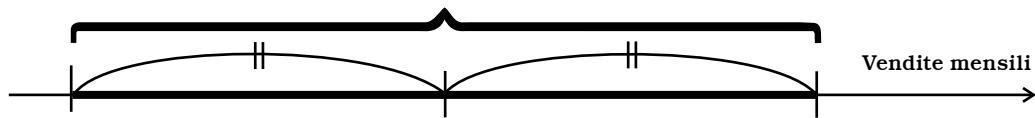
I valori di x vengono dai predittori, \bar{x} è la media di un dato insieme di predittori e S^{-1} è la matrice di concentrazione del Passo 1. Il quadrato della distanza di Mahalanobis per il negozio di Yumenooka è questa:

$$\begin{aligned} D^2 &= \left\{ \begin{array}{l} (x_1 - \bar{x}_1)(x_1 - \bar{x}_1)S^{11} + (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)S^{12} + \cdots + (x_1 - \bar{x}_1)(x_p - \bar{x}_p)S^{1p} \\ + (x_2 - \bar{x}_2)(x_1 - \bar{x}_1)S^{21} + (x_2 - \bar{x}_2)(x_2 - \bar{x}_2)S^{22} + \cdots + (x_2 - \bar{x}_2)(x_p - \bar{x}_p)S^{2p} \\ \dots \\ + (x_p - \bar{x}_p)(x_1 - \bar{x}_1)S^{p1} + (x_p - \bar{x}_p)(x_2 - \bar{x}_2)S^{p2} + \cdots + (x_p - \bar{x}_p)(x_p - \bar{x}_p)S^{pp} \end{array} \right\} (\text{numero di individui} - 1) \\ D^2 &= \left\{ \begin{array}{l} (x_1 - \bar{x}_1)(x_1 - \bar{x}_1)S^{11} + (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)S^{12} \\ + (x_2 - \bar{x}_2)(x_1 - \bar{x}_1)S^{21} + (x_2 - \bar{x}_2)(x_2 - \bar{x}_2)S^{22} \end{array} \right\} (\text{numero di individui} - 1) \\ &= \left\{ \begin{array}{l} (10 - 7,7)(10 - 7,7) \times 0,0657 + (10 - 7,7)(80 - 156) \times 0,0004 \\ + (80 - 156)(10 - 7,7) \times 0,0004 + (80 - 156)(80 - 156) \times 0,00001 \end{array} \right\} (10 - 1) \\ &= 2,4 \end{aligned}$$

PASSO 3

Ora calcoliamo l'intervallo di confidenza, come mostrato qui:

Questo è l'intervallo di confidenza.



$$453,2 - 35 = 418$$

$$\begin{aligned} & a_1 \times 10 + a_2 \times 80 + b \\ &= 41,5 \times 10 - 0,3 \times 80 + 65,3 \\ &= 453 \end{aligned}$$

$$453 + 35 = 488$$

Gli estremi dell'intervallo di confidenza hanno la stessa distanza dalla media; in altri termini, l'intervallo di confidenza “è centrato” sulla media. La distanza dalla media si calcola come mostrato qui sotto (D_2 indica il quadrato della distanza di Mahalanobis, e x rappresenta il numero totale di predittori, non un certo valore di uno di essi):

$$\begin{aligned} & \sqrt{F(1, \text{dimensioni del campione} - x - 1; 0,05) \times \left(\frac{1}{\text{dimensioni del campione}} + \frac{D^2}{\text{dimensioni del campione} - 1} \right) \times \frac{S_e}{\text{dimensioni del campione} - x - 1}} \\ &= \sqrt{F(1, 10 - 2 - 1; 0,05) \times \left(\frac{1}{10} + \frac{2,4}{10 - 1} \right) \times \frac{4173,0}{10 - 2 - 1}} \\ &= 35 \end{aligned}$$

Come nella semplice analisi di regressione, quando si calcola l'intervallo di previsione, si aggiunge 1 al secondo termine:

$$\sqrt{F(1, \text{dimensioni del campione} - x - 1; 0,05) \times \left(1 + \frac{1}{\text{dimensioni del campione}} + \frac{D^2}{\text{dimensioni del campione} - 1} \right) \times \frac{S_e}{\text{dimensioni del campione} - x - 1}}$$

Se il livello di confidenza è del 99%, basta sostituire 0,05 con 0,01:

$$\begin{aligned} F(1, \text{dimensioni del campione} - x - 1; 0,05) &= F(1, 10 - 2 - 1; 0,05) = 5,6 \\ F(1, \text{dimensioni del campione} - x - 1; 0,01) &= F(1, 10 - 2 - 1; 0,01) = 12,2 \end{aligned}$$

Potete notare che per avere una maggior sicurezza di includere il dato reale nell'intervallo di previsione dobbiamo ampliare questo intervallo.

USO DEI DATI CATEGORICI NELL'ANALISI DI REGRESSIONE MULTIPLA

Ricorderete dal capitolo 1 che i dati categorici sono quelli non misurabili tramite numeri. Per esempio, il colore degli occhi del negoziante è un dato categorico (e sarà anche una pessima variabile predittiva per le vendite mensili). Benché sia possibile *rappresentarle* in maniera numerica (1 = azzurro, 2 = verde), le variabili categoriche sono discrete: non esiste un “verde e mezzo”. Inoltre non possiamo dire che 2 (occhi verdi) è maggiore di 1 (occhi azzurri). Finora abbiamo usato dati numerici (che ammettono una rappresentazione significativa tramite dati numerici continui: una distanza di 110 m dalla stazione ferroviaria è maggiore di una distanza di 109,9 m), riportati anche a pagina 113.

TABELLA 3-2: DATI DELL'ESEMPIO DELLE PASTICCERIE KAZAMI

Pasticceria	Area del negozi (tsubo)	Distanza dalla stazione più vicina (metri)	Vendite mensili (10.000 ¥)
Yumenooka	10	80	469
Stazione Terai	8	0	366
Sone	8	200	371
Stazione Hashimoto	5	200	208
Quartiere Kikyou	7	300	246
Ufficio Postale	8	230	297
Stazione Suidobashi	7	40	363
Stazione Rokujo	9	0	436
Lungofiume Wakaba	6	330	198
Misato	9	180	364

La variabile predittiva *Area del negozio* si misura in tsubo, la *Distanza dalla stazione più vicina* in metri, e le *Vendite mensili* in yen. È ovvio che sono tutte grandezze misurabili numericamente. Nell'analisi di regressione multipla, la variabile responso deve essere misurabile, di tipo numerico, ma le variabili predittive possono essere

- tutte numeriche;
- alcune numeriche, altre categoriche;
- tutte categoriche.

Le Tabelle 3-3 e 3-4 mostrano entrambe un insieme di dati accettabile. La prima include variabili sia numeriche sia categoriche, mentre nella seconda tutte le variabili predittive sono categoriche.

TABELLA 3-3: COMBINAZIONE DI DATI NUMERICI E CATEGORICI

Pasticceria	Area del negozio (tsubo)	Distanza dalla stazione più vicina (metri)	Assaggi gratuiti	Vendite mensili (10.000 ¥)
Yumenooka	10	80	1	469
Stazione Terai	8	0	0	366
Sone	8	200	1	371
Stazione Hashimoto	5	200	0	208
Quartiere Kikyou	7	300	0	246
Ufficio postale	8	230	0	297
Stazione Suidobashi	7	40	0	363
Stazione Rokujo	9	0	1	436
Lungofiume Wakaba	6	330	0	198
Misato	9	180	1	364

Nella Tabella 3-3 abbiamo incluso una variabile predittiva categorica: *assaggi gratuiti*. Alcune pasticcerie Kazami mettono a disposizione un vassoio di assaggini gratuiti (1), e altre no (0). Includendo questi dati nell'analisi, otteniamo la seguente equazione di regressione multipla:

$$y = 30,6x_1 - 0,4x_2 + 39,5x_3 + 135,9$$

dove y rappresenta le vendite mensili, x_1 l'area del negozio, x_2 la distanza dalla stazione più vicina e x_3 gli assaggi gratuiti.

TABELLA 3-4: SOLO DATI PREDITTIVI CATEGORICI

Pasticceria	Area del negozio (tsubo)	Distanza dalla stazione più vicina (metri)	Assaggi gratuiti quotidiani	Assaggi gratuiti solo nel fine settimana	Vendite mensili (10.000 ¥)
Yumenooka	1	0	1	0	469
Stazione Terai	1	0	0	0	366
Sone	1	1	1	0	371
Stazione Hashimoto	0	1	0	0	208
Quartiere Kikyou	0	1	0	0	246
Ufficio postale	1	1	0	0	297
Stazione Suidobashi	0	0	0	0	363
Stazione Rokujo	1	0	1	1	436
Lungofiume Wakaba	0	1	0	0	198
Misato	1	0	1	1	364

↑
Minore di 8 tsubo = 0
8 o più tsubo = 1

↑
Minore di 200 m = 0
200 o più m = 1

↑
Non offre assaggi = 0
Offre assaggi = 1

Nella Tabella 3-4 abbiamo convertito alcuni dati numerici (area del negozio e distanza dalla stazione) in dati categorici, creando alcune categorie generali. Usando questi dati possiamo calcolare l'equazione di regressione multipla

$$y = 50,2x_1 - 110,1x_2 + 13,4x_3 + 75,1x_4 + 336,4$$

dove y rappresenta le vendite mensili, x_1 l'area del negozio, x_2 la distanza dalla stazione più vicina, x_3 gli assaggi gratuiti quotidiani e x_4 quelli offerti soltanto nel fine settimana.

MULTICOLLINEARITÀ

La multicollinearità si verifica quando due variabili predittive sono strettamente correlate. In questo caso è difficile distinguere l'effetto singolo di queste variabili sulla variabile responso, e l'analisi può risentirne in vari modi:

- stima meno accurata dell'effetto di una data variabile sulla variabile responso;
- errori standard dei coefficienti di regressione insolitamente grandi;
- mancato rifiuto dell'ipotesi nulla;
- *adattamento eccessivo*, cioè l'equazione di regressione descrive il legame della variabile responso con l'errore casuale, invece che con la variabile predittiva.

Si può rilevare la multicollinearità usando un indice come la tolleranza o il suo inverso, noto come *fattore di inflazione della varianza* (VIF, acronimo del nome inglese: *variance inflation factor*). In genere, si ritiene che la tolleranza minore di 0,1 o il VIF maggiore di 10 indichino notevole multicollinearità, ma a volte si usano limiti più rigidi.

Se siete alle prime armi con l'analisi di regressione multipla, non state a preoccuparvi troppo di tutto ciò. Ricordate solo che una notevole multicollinearità può essere problematica. Se alcune variabili predittive sono fortemente correlate tra loro, quindi, è forse consigliabile eliminarne una e rifare l'analisi dei dati.

DETERMINAZIONE DELL'INFLUENZA RELATIVA DELLE VARIABILI PREDITTIVE SULLA VARIABILE RESPONSO

A volte si usa l'analisi di regressione multipla per esaminare l'influenza relativa di ciascuna variabile predittiva sulla variabile responso. Si tratta di una strategia abbastanza diffusa e accettata, ma non sempre consigliabile.

Nell'esempio seguente, un ricercatore usa l'analisi di regressione multipla per valutare l'effetto relativo di vari fattori sul gradimento complessivo di un certo tipo di caramella.

Il signor Torikoshi si occupa di sviluppare nuovi prodotti in un'azienda dolciaria. Di recente ha creato una nuova caramella, Magic Fizz, che dà una sensazione frizzante al contatto con la lingua. Le vendite sono impressionanti. Per scoprire da cosa dipenda un tale successo, l'azienda regala campioncini di queste caramelle agli studenti dell'università locale, chiedendo loro di valutare il prodotto con il seguente questionario.

QUESTIONARIO SU MAGIC FIZZ

Esprimete il vostro parere su Magic Fizz rispondendo alle seguenti domande. Cerchiate la risposta più vicina alla vostra opinione.

Sapore	1. Insufficiente 2. Soddisfacente 3. Eccezionale
Consistenza	1. Insufficiente 2. Soddisfacente 3. Eccezionale
Sensazione di effervesienza	1. Insufficiente 2. Soddisfacente 3. Eccezionale
Confezione	1. Insufficiente 2. Soddisfacente 3. Eccezionale
Gradimento complessivo	1. Insufficiente 2. Soddisfacente 3. Eccezionale

Venti studenti completano il questionario; la Tabella 3-5 riporta i risultati. Osservate che al contrario dell'esempio delle pasticcerie Kazami, i valori della variabile responso – il gradimento complessivo – sono già noti. Nel caso delle pasticcerie, lo scopo era prevedere la variabile responso (*Profitto*) di un negozio non ancora aperto, sulla base dell'andamento di quelli esistenti. In questo caso, lo scopo dell'analisi è studiare l'effetto relativo delle diverse variabili predittive, per individuare quale (*Sapore*, *Consistenza*, *Sensazione*, *Presentazione*) influenzi di più il risultato (*Gradimento*).

TABELLA 3-5: RISULTATI DEL QUESTIONARIO SU MAGIC FIZZ

Intervistato	Sapore	Consisten- za	Sensazione di efferveszenza	Confezione	Gradimento complessivo
1	2	2	3	2	2
2	1	1	3	1	3
3	2	2	1	1	3
3	2	2	1	1	1
4	3	3	3	2	2
5	1	1	2	2	1
6	1	1	1	1	1
7	3	3	1	3	3
8	3	3	1	2	2
9	3	3	1	2	3
10	1	1	3	1	1
11	2	3	2	1	3
12	2	1	1	1	1
13	3	3	3	1	3
14	3	3	1	3	3
15	3	2	1	1	2
16	1	1	3	3	1
17	2	2	2	1	1
18	1	1	1	3	1
19	3	1	3	3	3
20	3	3	3	3	3

Prima di calcolare l'equazione di regressione multipla, ogni variabile è stata normalizzata. La normalizzazione riduce gli effetti dell'errore o della scala, permettendo confronti più accurati tra diverse variabili. L'equazione risultante è

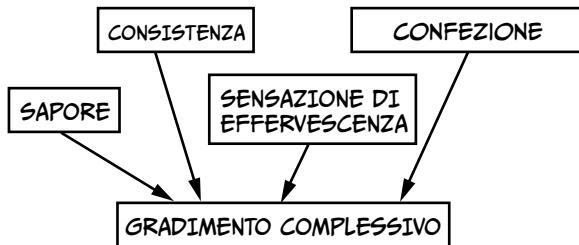
$$y = 0,41x_1 + 0,32x_2 + 0,26x_3 + 0,11x_4$$

dove y rappresenta il gradimento complessivo, x_1 il sapore, x_2 la consistenza, x_3 la sensazione di efferveszenza e x_4 la confezione.

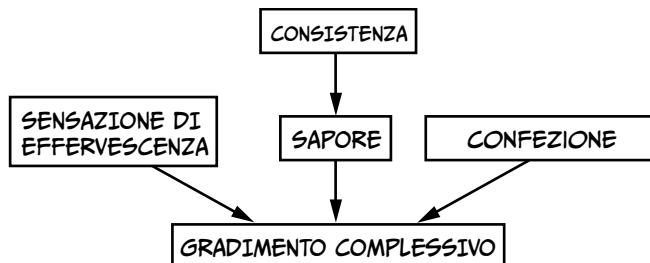
Confrontando i coefficienti di regressione parziali per le quattro variabili predittive, potete vedere che il più grande corrisponde al sapore. Su questa base, Torikoshi conclude che sia questo il fattore più importante nel gradimento complessivo.

Il suo ragionamento non regge. La variabile responso è pari alla somma delle variabili predittive moltiplicate per i coefficienti di regressione parziali; moltiplicandone una per un fattore più grande, il suo effetto sul risultato finale dovrebbe crescere, giusto? Be', a volte sì, ma le cose non sono sempre così semplici.

Esaminiamo meglio il ragionamento di Torikoshi con questo schema:



In altri termini, Torikoshi sta ipotizzando che tutte le variabili siano legate indipendentemente e direttamente al gradimento complessivo. Ma non è detto che le cose stiano così. La consistenza potrebbe influenzare il gradimento del sapore:



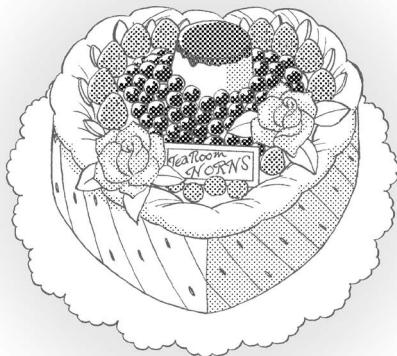
Un metodo migliore per confrontare l'effetto relativo di più variabili predittive sul risultato è il *modello di equazioni strutturali* (SEM, acronimo del nome inglese: *structural equation modeling*). Rispetto alla regressione lineare, esso formula ipotesi più flessibili, ed è applicabile anche a insiemi di dati che presentano multicollinearità. Non è però una panacea: si basa sull'ipotesi che i dati siano rilevanti per il problema di interesse.

SEM presuppone inoltre che il modello scelto per rappresentare i dati sia corretto. Va notato che le domande del questionario richiedono un'interpretazione soggettiva. Se Miu valutasse la caramella con due "soddisfacente" e due "eccezionale", starebbe a lei decidere se il suo gradimento complessivo corrisponde a "soddisfacente" o a "eccezionale". La scelta potrebbe dipendere dal suo umore del giorno!

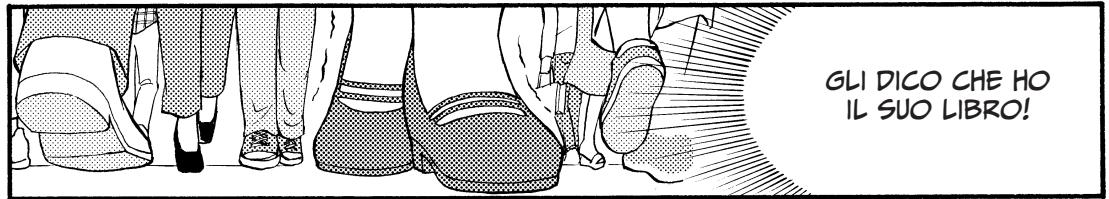
Risa potrebbe dare la stessa valutazione di Miu nelle quattro categorie primarie e una valutazione diversa del gradimento complessivo, pur rimanendo convinta di essere stata obiettiva. Poiché Miu e Risa non concordano sull'ultima categoria, è possibile che il modello non rappresenti i dati in maniera corretta. Il modello di equazioni strutturali può comunque dare risultati utili, segnalando l'influenza di alcune variabili sulle altre, piuttosto che sul risultato finale.

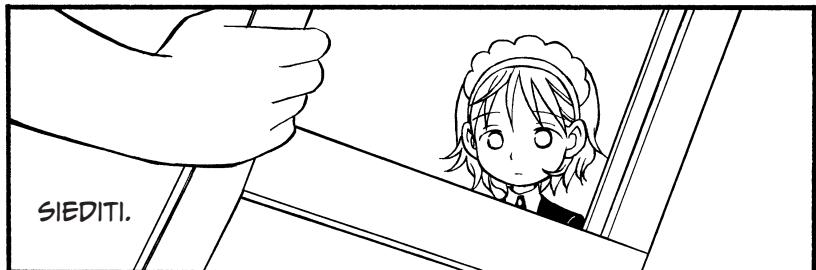
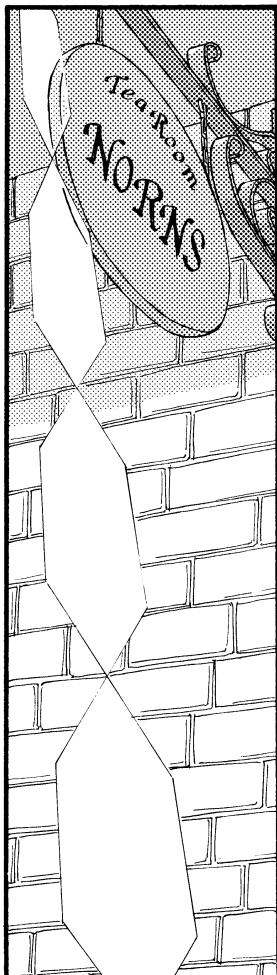
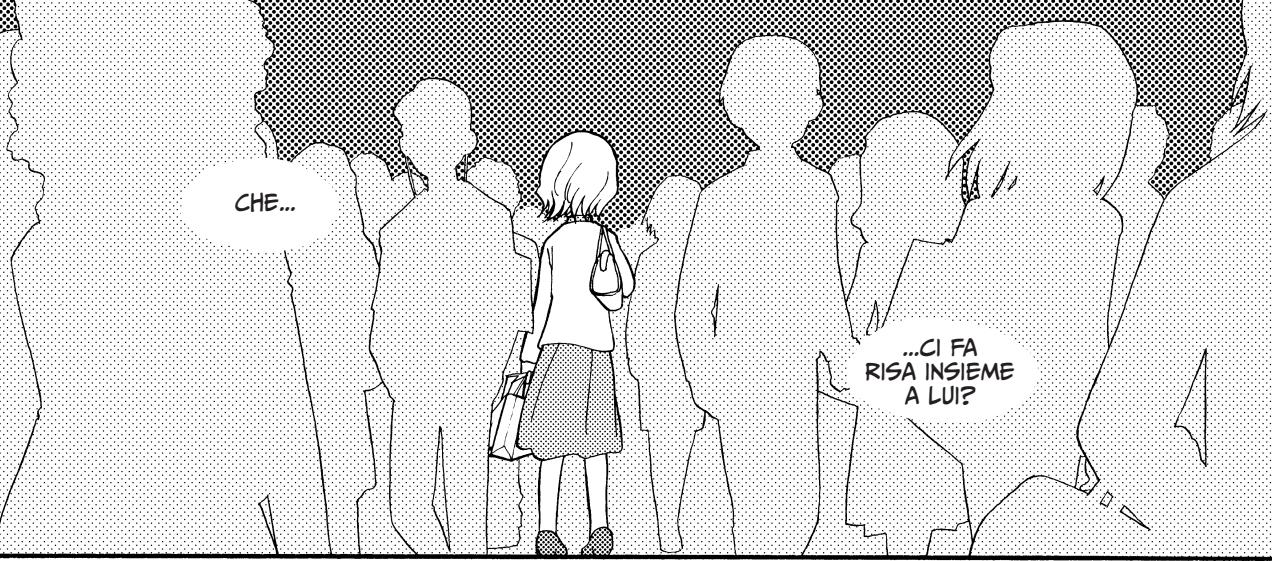
4

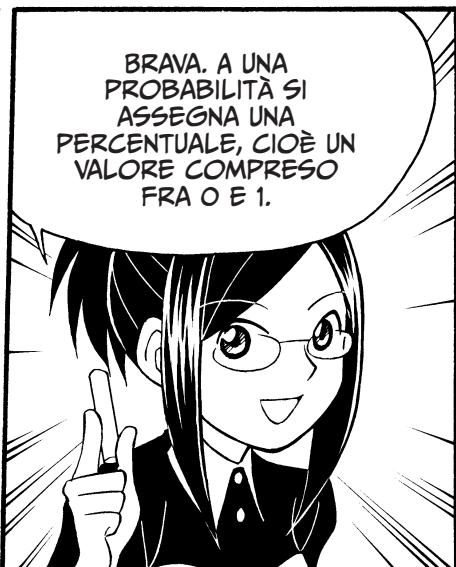
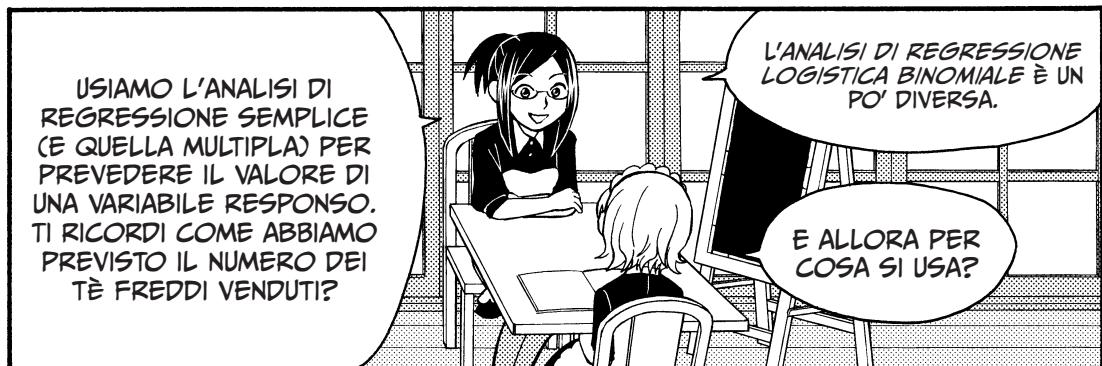
ANALISI DI REGRESSIONE LOGISTICA









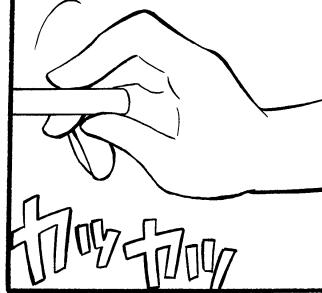


MA 70% È PIÙ
GRANDE DI 1, NO?

IN REALTÀ, 70% È
UGUALE A 0,7. QUANDO
SVOGLIAMO UN'ANALISI
DI REGRESSIONE
LOGISTICA, LA
RISPOSTA SARÀ
MINORE DI 1.

PER ESPRIMERLA
COME PERCENTUALE,
MOLTIPLICHIAMO
PER 100 E USIAMO IL
SIMBOLO %.

L'EQUAZIONE DI
REGRESSIONE
LOGISTICA...



...HA QUESTO
ASPECTO.

$$y = \frac{1}{1 + e^{-(a_1x_1 + a_2x_2 + \dots + a_px_p + b)}}$$

↑
VARIABILE
RESPONSO (y)

↑
VARIABILE
PREDITTIVA (x)

↑
COEFFICIENTE DI
INTERCETTA
REGRESSIONE

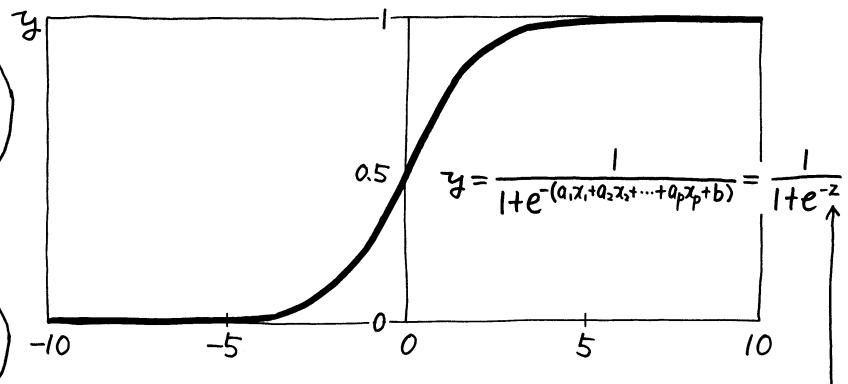
The equation is a logistic regression formula. It consists of a fraction where the numerator is 1 and the denominator is 1 plus an exponential term. The exponential term is the negative sum of coefficients (a1 through ap) multiplied by their respective variables (x1 through xp), plus a constant term b. Arrows point from labels to specific parts of the equation: 'VARIABLE RESPONSO (y)' points to the y variable, 'VARIABLE PREDITTIVA (x)' points to one of the x variables, and 'COEFFICIENTE DI INTERCETTA REGRESSIONE' points to the constant term b.

CHE ESPO-
NENTE GROSSO!
SEMBRA COM-
PLICATO...

NON TI PREOCCUPARE,
POI TI FACCIO VEDERE UN
MODO PIÙ SEMPLICE PER
SCRIVERLO. PROCEDIAMO
PASSO PASSO, COSÌ NON
SEMBRERÀ TANTO DIFFICILE.

IL GRAFICO
DELL'EQUAZIO-
NE HA QUESTO
ASPETTO:

HA UNA
FORMA A S.



HO RISCRITTO L'EQUAZIONE USANDO Z COME ESPONENTE. $f(z)$ È LA PROBABILITÀ DEL NOSTRO RISULTATO!

QUALUNQUE SIA Z,
IL VALORE DI Y NON
SARÀ MAI MAGGIORE
DI 1 O MINORE DI 0.

SÌ! A QUANTO
PARE LA S È
STATA APPIATTITA
APPosta.

ADESSO, PRIMA DI
ANDARE AVANTI CON
LA REGRESSIONE
LOGISTICA, DEVI
IMPARARE LA MASSIMA
VEROSIMIGLIANZA.

LA MASSIMA
VEROSIMIGLIANZA SI
USA PER STIMARE I
VALORI DEI PARAMETRI
DI UNA POPOLAZIONE
USANDO UN CAMPIONE
RAPPRESENTATIVO. LE
STIME SI FORMULANO
USANDO LA PROBABILITÀ.

MASSIMA
VEROSIMIGLIANZA?

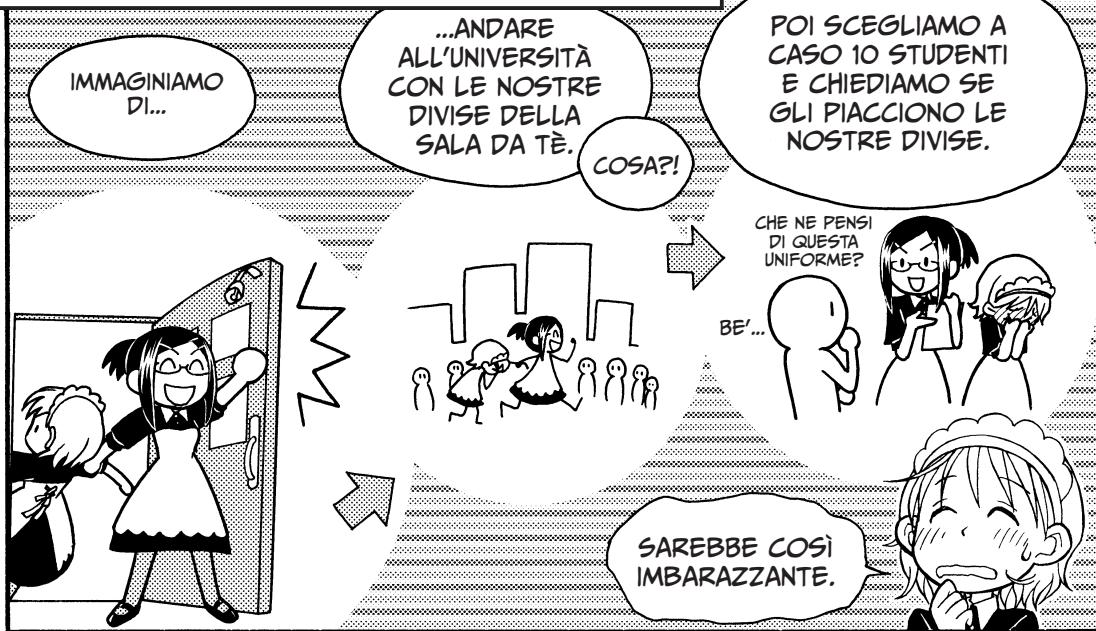
MASSIMA
VEROSIMIGLIANZA

ANCORA
PROBABILI-
TÀ!

PER SPIEGARTELLO,
USERÒ UNA SITUA-
ZIONE IPOTETICA IN
CUI SIAMO LE PRO-
TAGONISTE!

NON SO SE
SONO TAGLIATA
PER FARE LA
PROTAGONISTA.

IL METODO DELLA MASSIMA VEROSIMIGLIANZA



STUDENTI	TI PIACE LA DIVISA DELLA SALA DA TÈ NORNS?
A	SI
B	NO
C	SI
D	NO
E	SI
F	SI
G	SI
H	SI
I	NO
J	SI



...ALLORA LA PROBABILITÀ IN BASE AL RISULTATO IMMAGINARIO DEL SONDAGGIO È QUESTA:

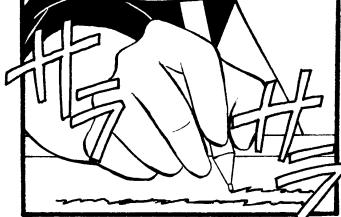
$$\begin{matrix} \text{SI} & \text{NO} & \text{SI} & \text{NO} & \text{SI} & \text{SI} & \text{SI} & \text{SI} & \text{NO} & \text{SI} \end{matrix}$$

$$P \times (1-P) \times P \times (1-P) \times P \times P \times P \times P \times (1-P) \times P$$

$$= P^7(1-P)^3$$

È UN'EQUAZIONE?

SÌ, E LA RISOLVIAMO TROVANDO IL VALORE PIÙ VEROSIMILE PER p .



$$P^7(1-P)^3$$

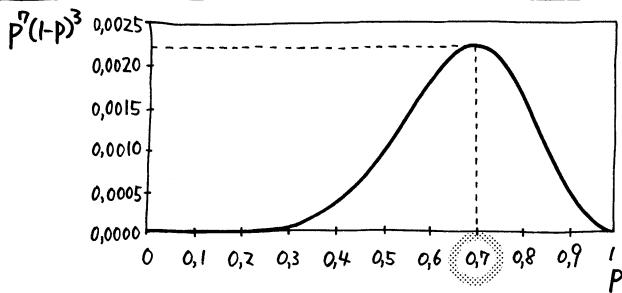
OPPURE

$$\log\{P^7(1-P)^3\}^*$$

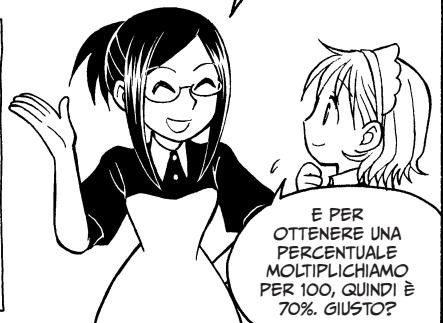
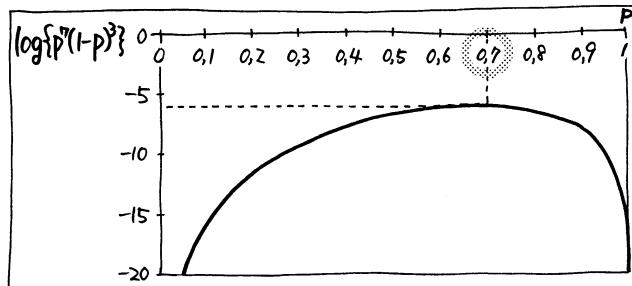
USIAMO UNA DI QUESTE FUNZIONI DI VEROSIMIGLIANZA.

IN ENTRAMBI I CASI IL RISULTATO È LO STESSO.

* PRENDERE IL LOGARITMO DELLA FUNZIONE PUÒ FACILITARE I CALCOLI SUCCESSIVI.



COME VEDI,
QUANDO TRACCIAMO I GRAFICI, HANNO ENTRAMBI UN MASSIMO IN 0,7. È IL VALORE PIÙ VEROSIMILE PER p !



GIUSTO. PRENDIAMO IL LOGARITMO PERCHÉ RENDE PIÙ FACILE CALCOLARE LA DERIVATA, CHE CI SERVIRÀ PER TROVARE LA MASSIMA VEROSIMIGLIANZA.

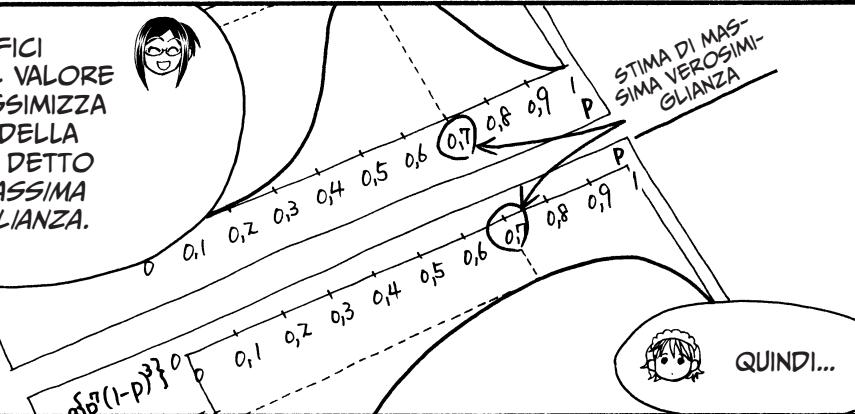
$$P^r(1-P)^3$$

→ FUNZIONE DI VEROSIMIGLIANZA

$$\log \{P^r(1-P)^3\}$$

→ FUNZIONE DI LOG-VEROSIMIGLIANZA

NEI GRAFICI CERCHIAMO IL VALORE DI P CHE MASSIMIZZA IL VALORE DELLA FUNZIONE. È DETTO STIMA DI MASSIMA VEROSIMIGLIANZA.



...DATO CHE LE FUNZIONI HANNO UN MASSIMO NELLO STESSO PUNTO, PUR AVENDO UNA FORMA DIVERSA, CI DANNO LO STESSO RISULTATO.



E ORA ANALIZZIAMO LA STIMA DI MASSIMA VEROSIMIGLIANZA PER LA POPOLARITÀ DELLE NOSTRE DIVISE.



TROVARE LA MASSIMA VEROLOGIANZA CON LA FUNZIONE DI VEROLOGIANZA

Passo 1

Trovare la funzione di verosimiglianza. Qui p sta per Sì, e $1 - p$ sta per No. Ci sono stati 7 Sì e 3 No.

$$\begin{aligned} & p \times (1-p) \times p \times (1-p) \times p \times p \times p \times p \times (1-p) \times p \\ & = p^7 (1-p)^3 \end{aligned}$$

Passo 2

Trovare la funzione di log-verosimiglianza e riscriverla.

$$\begin{aligned} L &= \log \left\{ p^7 (1-p)^3 \right\} \\ &= \log p^7 + \log (1-p)^3 \quad \leftarrow \text{Prendiamo il log di ogni addendo} \\ &= 7 \log p + 3 \log (1-p) \quad \leftarrow \text{Usiamo la regola dell'esponenziazione di pagina 22.} \end{aligned}$$

D'ORA IN POI USIAMO L PER INDICARE LA FUNZIONE DI LOG-VEROSIMIGLIANZA.



Passo 3

Deriviamo L rispetto a p e poniamo l'espressione uguale a 0. Ricordiamo che quando la variazione di una funzione è pari a 0, stiamo trovando i massimi.

$$\frac{dL}{dp} = 7 \times \frac{1}{p} + 3 \times \frac{1}{1-p} \times (-1) = 7 \times \frac{1}{p} - 3 \times \frac{1}{1-p} = 0$$

Passo 4

Esplicitiamo p dall'uguaglianza del Passo 3.

$$\begin{aligned} 7 \times \frac{1}{p} - 3 \times \frac{1}{1-p} &= 0 \\ \left(7 \times \frac{1}{p} - 3 \times \frac{1}{1-p} \right) \times p(1-p) &= 0 \times p(1-p) \quad \leftarrow \text{Moltiplichiamo ambo i membri per } p(1-p). \\ 7(1-p) - 3p &= 0 \end{aligned}$$

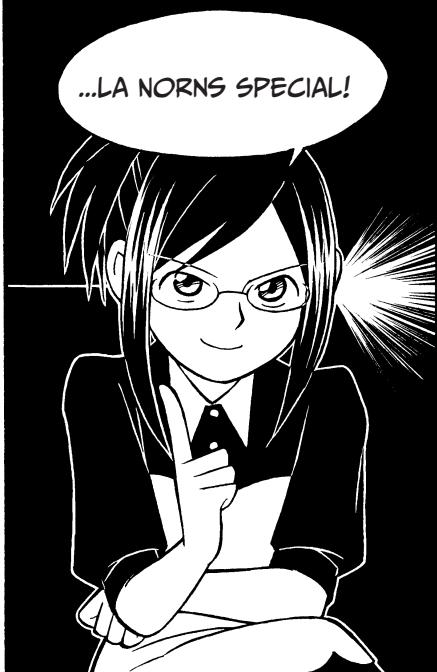
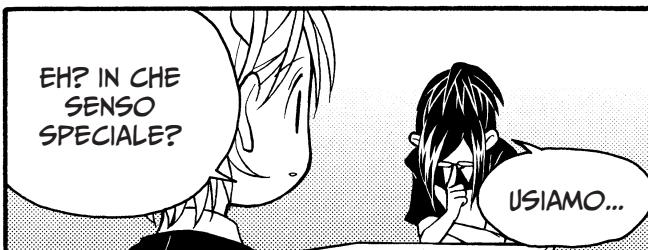
$$7 - 7p - 3p = 0$$

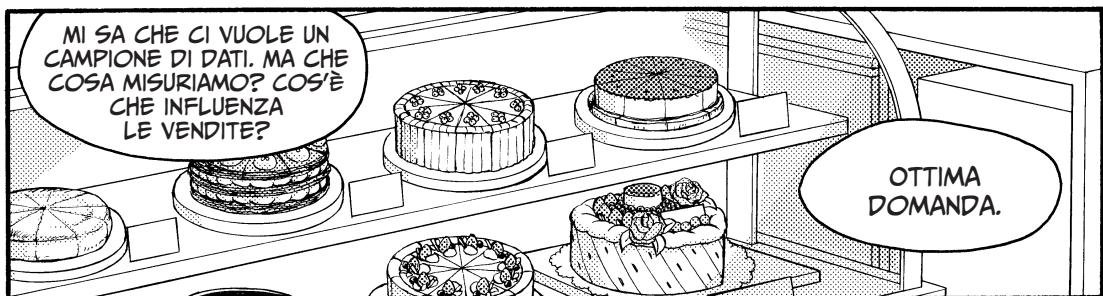
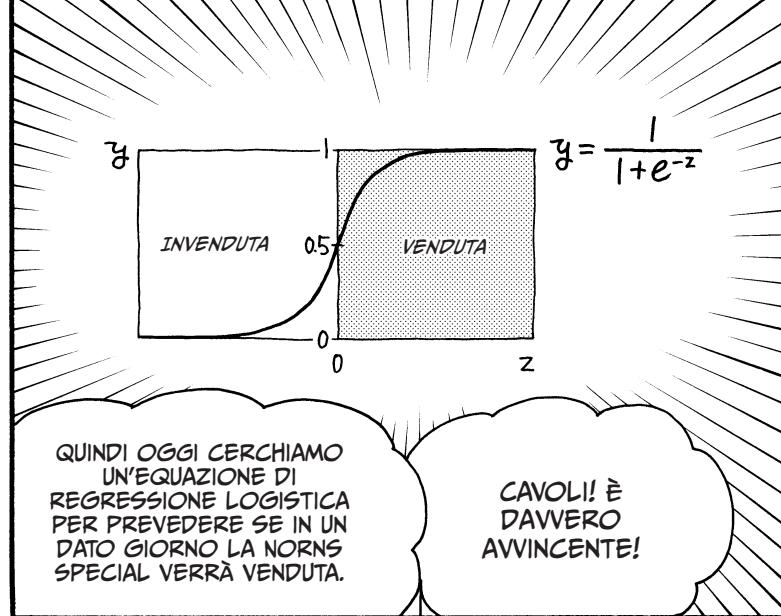
$$7 - 10p = 0$$

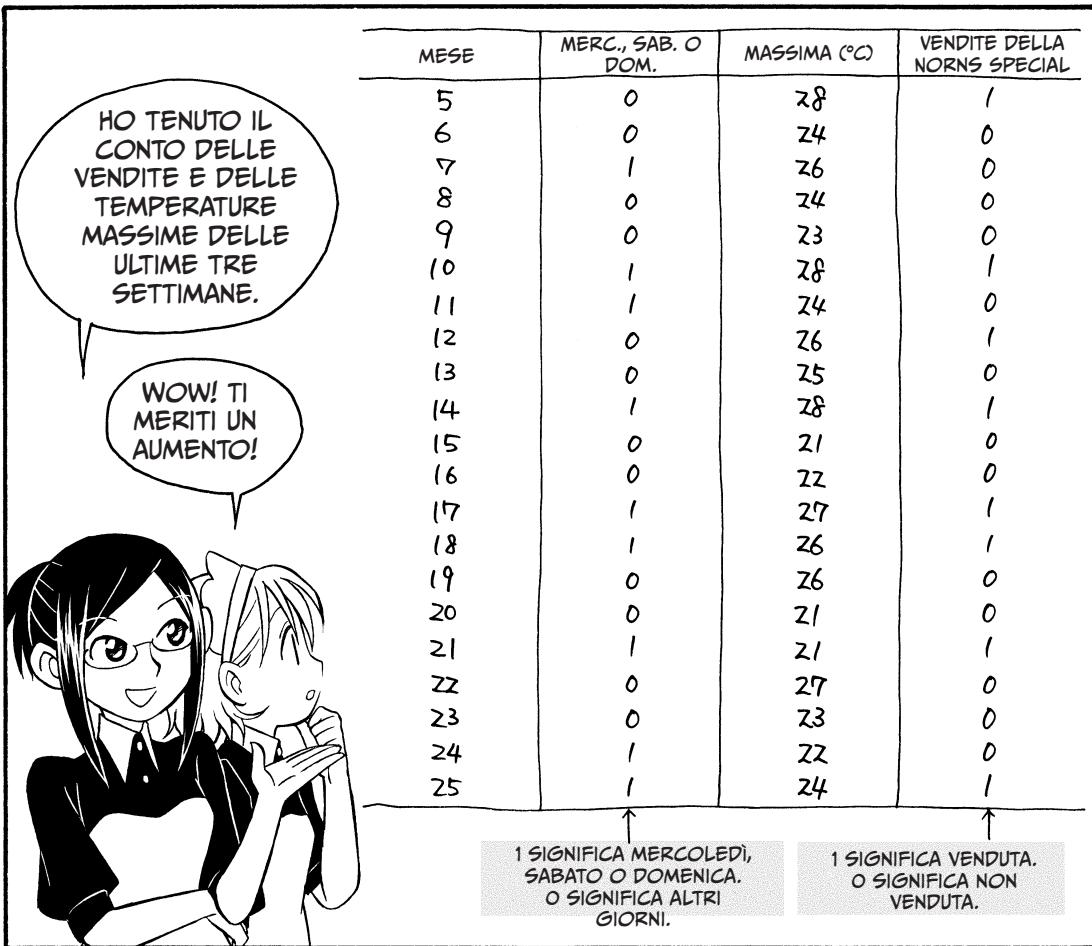
$$p = \frac{7}{10}$$



SCEGLIERE LE VARIABILI PREDITTIVE







ABBIANO USATO
1 PER DIRE
VENDUTA E 0 PER
INVENDUTA...

1 = VENDUTA

0 = INVENDUTA

...CHE È IL
MODO IN CUI
RAPPRESENTIA-
MO I DATI CATE-
GORICI COME
NUMERI, GIUSTO?

GIUSTO.

BE', NELLA REGRESSIONE LOGISTICA
QUESTI NUMERI NON SONO SOLO
ETICHETTE; MISURANO LA PROBABILITÀ
CHE LA TORTA SIA STATA VENDUTA. 1
SIGNIFICA 100% E 0 SIGNIFICA 0%.

OH! DATO CHE
SAPPIAMO CHE È
STATA VENDUTA, C'È
UNA PROBABILITÀ
DEL 100% CHE SIA
STATA VENDUTA.

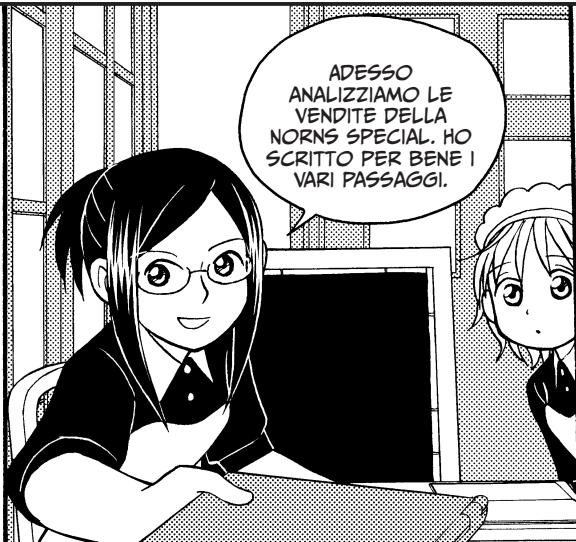
E SAPPIAMO
ANCHE PER
CERTO SE ERA
MERCOLEDÌ,
SABATO O
DOMENICA.

PROPRIO
COSÌ.

IN QUESTO CASO, LA
TEMPERATURA È UN DATO
MISURABILE, E QUINDI LA USIAMO,
COME NELLA REGRESSIONE
LINEARE. ANCHE I DATI CATEGORICI
FUNZIONANO PIÙ O MENO COME
NELLA REGRESSIONE LINEARE,
E POSSIAMO DI NUOVO USARE
QUALSIASI COMBINAZIONE DI DATI
CATEGORICI E NUMERICI.

MA I DATI
CATEGORICI HANNO
PROBABILITÀ
MISURABILI.

ANALISI DI REGRESSIONE LOGISTICA IN AZIONE!



ADESSO
ANALIZZIAMO LE
VENDITE DELLA
NORMS SPECIAL. HO
SCRITTO PER BENE I
VARI PASSAGGI.



CALCOLIAMO
L'EQUAZIONE E
TROVIAMO R^2 ?
E POI TROVIAMO
GLI INTERVALLI DI
CONFIDENZA E DI
PREVISIONE? AH,
E C'È LA COSA
DELL'IPOTESI?

SÌ, UNA COSA
DEL GENERE.

PROCEDURA PER L'ANALISI DI REGRESSIONE LOGISTICA

PASSO 1 TRACCIARE UN GRAFICO DI DISPERSIONE DELLE VARIABILI PREDITTIVE E DELLA VARIABILE RESPONSO PER VEDERE SE SEMBRANO CORRELATE.

PASSO 2 CALCOLARE L'EQUAZIONE DI REGRESSIONE LOGISTICA.

PASSO 3 VALUTARE L'ACCURATEZZA DELL'EQUAZIONE.

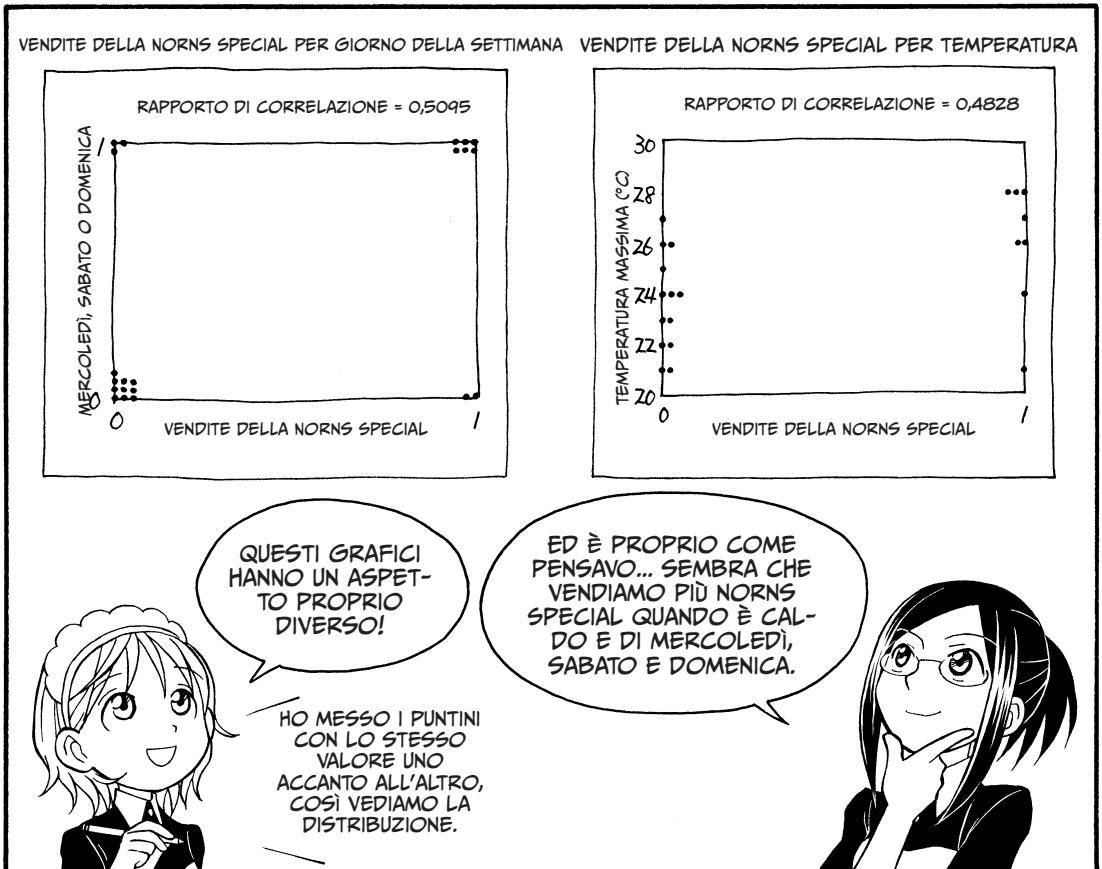
PASSO 4 SVOLGERE IL TEST D'IPOTESI.

PASSO 5 FACCIAMO UNA PREVISIONE!

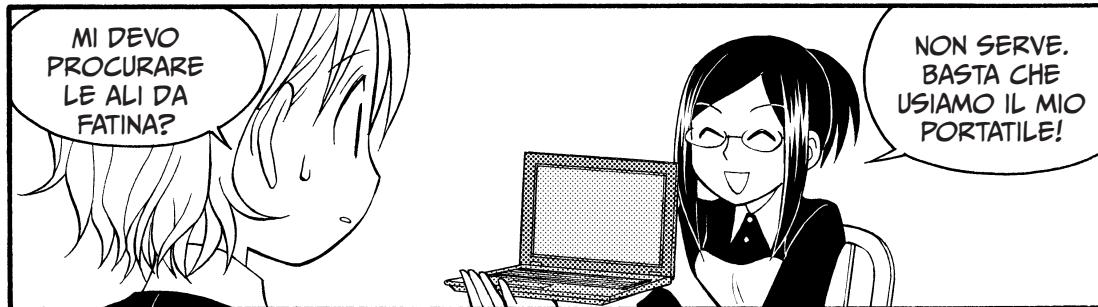
ECCO I CINQUE
PASSI FONDAMENTALI
DELL'ANALISI DI
REGRESSIONE
LOGISTICA.

NON È
TANTO
DIVERSO.

PASSO 1: TRACCIARE UN GRAFICO DI DISPERSIONE DELLE VARIABILI PREDITTIVE E DELLA VARIABILE RESPONSO PER VEDERE SE SEMBRANO CORRELATE



PASSO 2: CALCOLARE L'EQUAZIONE DI REGRESSIONE LOGISTICA



Passo 1

Determinare l'equazione logistica binomiale per ognuno dei campioni.

Mercoledì, sabato o domenica x_1	Temperatura massima x_2	Vendite della Norns Special y	Vendite della Norns Special $\hat{y} = \frac{1}{1 + e^{-(a_1x_1 + a_2x_2 + b)}}$
0	28	1	$\frac{1}{1 + e^{-(a_1 \times 0 + a_2 \times 28 + b)}}$
0	24	0	$\frac{1}{1 + e^{-(a_1 \times 0 + a_2 \times 24 + b)}}$
:	:	:	:
1	24	0	$\frac{1}{1 + e^{-(a_1 \times 1 + a_2 \times 24 + b)}}$

Passo 2

Trovare la funzione di verosimiglianza. La formula del Passo 1 rappresenta una torta venduta, e $(1 - \text{la formula})$ rappresenta una torta invenduta.

$$\frac{1}{1 + e^{-(a_1 \times 0 + a_2 \times 28 + b)}} \times \left(1 - \frac{1}{1 + e^{-(a_1 \times 0 + a_2 \times 24 + b)}}\right) \times \cdots \times \frac{1}{1 + e^{-(a_1 \times 1 + a_2 \times 24 + b)}}$$

Venduta

Invenduta

Venduta

Passo 3

Prendere il logaritmo naturale per trovare la funzione di log-verosimiglianza, L .

$$\begin{aligned} L &= \log_e \left[\frac{1}{1 + e^{-(a_1 \times 0 + a_2 \times 28 + b)}} \times \left(1 - \frac{1}{1 + e^{-(a_1 \times 0 + a_2 \times 24 + b)}}\right) \times \cdots \times \frac{1}{1 + e^{-(a_1 \times 1 + a_2 \times 24 + b)}} \right] \\ &= \log_e \left(\frac{1}{1 + e^{-(a_1 \times 0 + a_2 \times 28 + b)}} \right) + \log_e \left(1 - \frac{1}{1 + e^{-(a_1 \times 0 + a_2 \times 24 + b)}} \right) + \cdots + \log_e \left(\frac{1}{1 + e^{-(a_1 \times 1 + a_2 \times 24 + b)}} \right) \end{aligned}$$

Passo 4

Trovare i coefficienti di massima verosimiglianza. Questi coefficienti massimizzano la funzione di log-verosimiglianza, L .

I valori sono:*

$$\begin{cases} a_1 = 2,44 \\ a_2 = 0,54 \\ b = -15,20 \end{cases}$$



Possiamo inserire questi valori nella funzione di verosimiglianza per calcolare L , che useremo per calcolare R^2 .

$$L = \log_e \left(\frac{1}{1 + e^{-(2,44 \times 0 + 0,54 \times 28 - 15,20)}} \right) + \log_e \left(1 - \frac{1}{1 + e^{-(2,44 \times 0 + 0,54 \times 24 - 15,20)}} \right) + \dots + \log_e \left(\frac{1}{1 + e^{-(2,44 \times 1 + 0,54 \times 24 - 15,20)}} \right)$$
$$= -8,9$$

* Vedi pagina 210 per una spiegazione più dettagliata di questi calcoli.

Passo 5

Calcolare l'equazione di regressione logistica.

Inseriamo i coefficienti calcolati nel Passo 4 per ottenere la seguente equazione di regressione logistica:

$$y = \frac{1}{1 + e^{-(2,44x_1 + 0,54x_2 - 15,20)}}$$

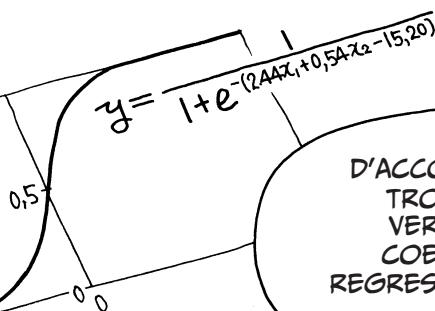
QUINDI È QUESTA L'EQUAZIONE CHE POSSIAMO USARE PER PREVEDERE SE VENDEREMO LA NORNS SPECIAL DI OGGI!

$$y = \frac{1}{1 + e^{-(2.44x_1 + 0.54x_2 - 15.20)}}$$

PROPRIO QUESTA.

PASSO 3 VALUTARE L'ACCURATEZZA DELL'EQUAZIONE

ADESSO DOBBIAMO CONTROLLARE CHE L'EQUAZIONE COMBACI BENE CON I NOSTRI DATI.



D'ACCORDO. ALLORA TROVIAMO R^2 E VERIFICHIAMO I COEFFICIENTI DI REGRESSIONE, GIUSTO?

È GIUSTO, MA LA REGRESSIONE LOGISTICA FUNZIONA IN MODO LIEVEMENTE DIVERSO.

UH? IN CHE SENSO?

NELL'ANALISI DI REGRESSIONE LOGISTICA, CALCOLIAMO UNO PSEUDO- R^2 .*

CIOÈ UN FALSO?

* IN QUESTO ESEMPIO USEREMO LA FORMULA DI MCFAFFDEN PER LO PSEUDO- R^2 .

ECCO L'EQUAZIONE CHE USIAMO PER CALCOLARE R^2 NELL'ANALISI DI REGRESSIONE LOGISTICA.

$$R^2 = 1 - \frac{\text{VALORE MASSIMO DELLA FUNZIONE DI LOG-VEROSIMIGLIANZA } L}{n_1 \log n_1 + n_0 \log n_0 - (n_1 + n_0) \log (n_1 + n_0)}$$

ARGH!!! È LUNGHIS- SIMA!



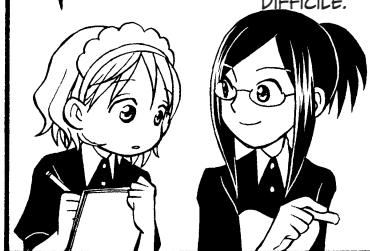
LE VARIABILI n TENGONO IL CONTO DELLE TORTE VENDUTE (n_1) E INVENDUTE (n_0).

n_1	NUMERO DI DATI IN CUI IL VALORE DELLA VARIABILE RESPONSO È 1
n_0	NUMERO DI DATI IN CUI IL VALORE DELLA VARIABILE RESPONSO È 0

ED ECCO UNA DEFINIZIONE PIÙ GENERALE.

NON SONO ANCORA SICURA DI COME USARE QUESTA EQUAZIONE CON I DATI SULLE NORNS SPECIAL.

NON TI PREOCUPARE, NON È DIFFICILE.



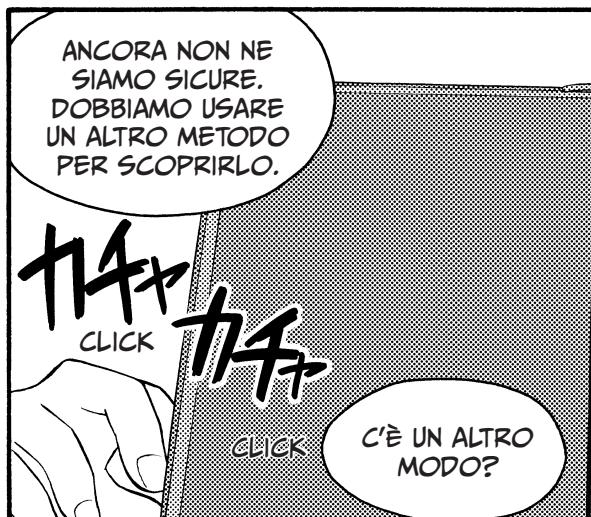
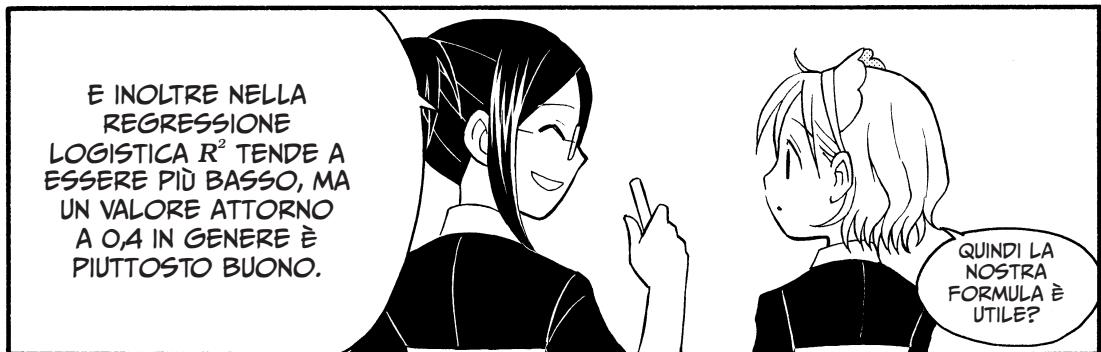
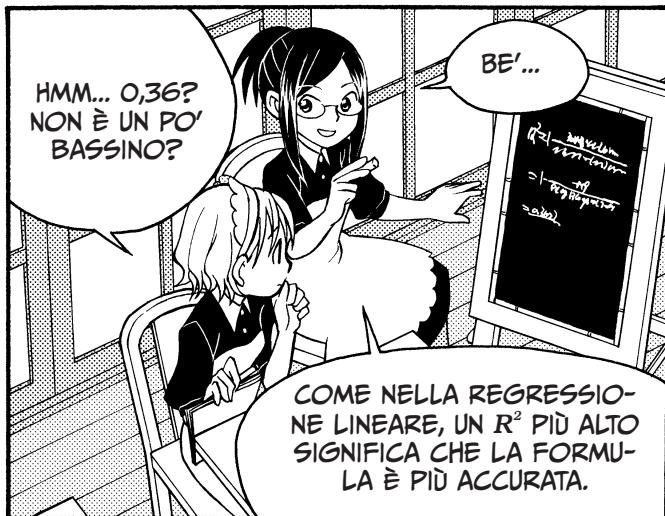
INSERIAMO I NUMERI RELATIVI ALLE NORNS SPECIAL...

$$R^2 = 1 - \frac{\text{VALORE MASSIMO DELLA FUNZIONE DI LOG-VEROSIMIGLIANZA } L}{n_1 \log n_1 + n_0 \log n_0 - (n_1 + n_0) \log (n_1 + n_0)}$$

$$= 1 - \frac{-8,9}{8 \log 8 + 13 \log 13 - (8+13) \log (8+13)}$$

$$= 0,3622$$

GULP! NON ME L'ASPETTAVO.



Gior-no	Mercoledì, sabato o domenica	Temperatu-ra massima (°C)	Vendite effettive	Vendite previste \hat{y}
	x_1	x_2	y	
5	0	28	1	0,51 venduta
6	0	24	0	0,11 invenduta
7	1	26	0	0,80 venduta
8	0	24	0	0,11 invenduta
9	0	23	0	0,06 invenduta
10	1	28	1	0,92 venduta
11	1	24	0	0,58 venduta
12	0	26	1	0,26 invenduta
13	0	25	0	0,17 invenduta
14	1	28	1	0,92 venduta
15	0	21	0	0,02 invenduta
16	0	22	0	0,04 invenduta
17	1	27	1	0,87 venduta
18	1	26	1	0,80 venduta
19	0	26	0	0,26 invenduta
20	0	21	0	0,02 invenduta
21	1	21	1	0,21 invenduta
22	0	27	0	0,38 invenduta
23	0	23	0	0,06 invenduta
24	1	22	0	0,31 invenduta
25	1	24	1	0,58 venduta

$$\frac{1}{1 + e^{-(2,44 \times 1 + 0,54 \times 24 - 15,20)}} = 0,58$$

↑

QUESTA TABELLA MOSTRA LE VENDITE EFFETTIVE DELLE NORMS SPECIAL E LE NOSTRE PREVISIONI. SE LA PREVISIONE È MAGGIORE DI 0,50, CORRISPONDE A UNA VENDITA.

MA LA TABELLA MOSTRA ANCHE ALTRE COSE. LE VEDI?



INTANTO, LA NORMS SPECIAL NON È STATA VENDUTA IL 7 E L'11, NONOSTANTE PREVEDESSIMO DI SÌ.



Gior-no	y	\hat{y}
7	0	0,80 venduta
11	0	0,58 venduta

OTTIMO!
CHE ALTRO?

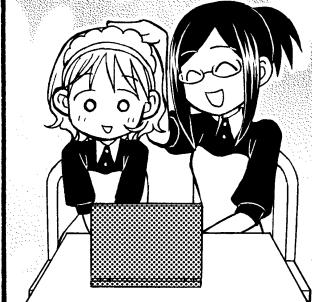


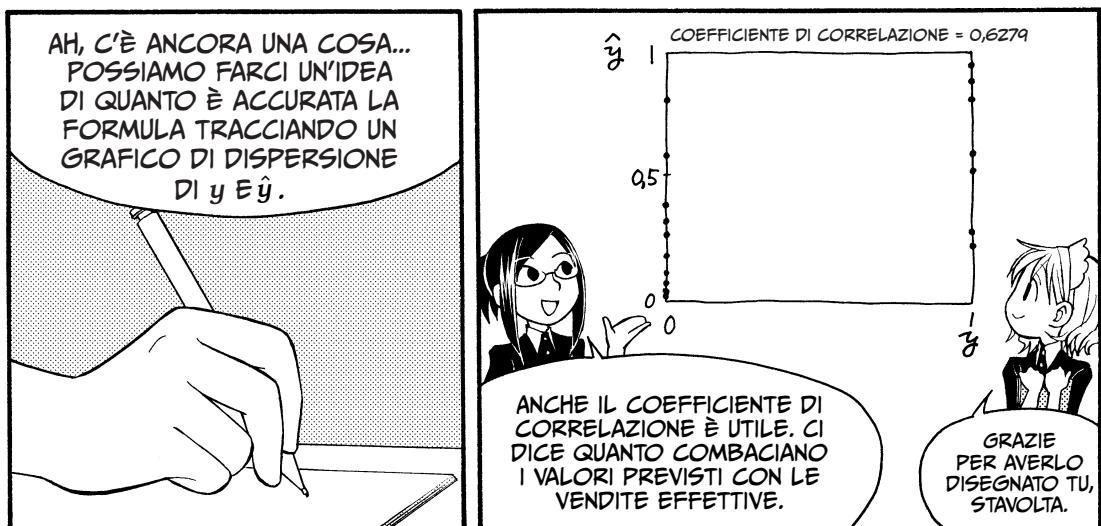
IL 12 E IL 21 AVEVAMO PREVISTO DI NON VENDERE, MA INVECE È STATA VENDUTA. È QUI CHE LA FORMULA HA SBAGLIATO.



PERFETTO!

LA MIGLIOR STUDENTESSA DI TUTTI I TEMPI!





PASSO 4: SVOLGERE IL TEST D'IPOTESI

COME PRIMA, DOBBIAMO VERIFICARE LE IPOTESI PER VEDERE SE I NOSTRI COEFFICIENTI DI REGRESSIONE SONO SIGNIFICATIVI.

E DATO CHE ABBIAMO DUE PREDITTORI POSSIAMO DI NUOVO PROVARE NEI DUE MODI!

TEST D'IPOTESI COMPLETO

IPOTESI NULLA	$A_1 = A_2 = 0$
IPOTESI ALTERNATIVA	UNA DELLE SEGUENTI È VERA: • $A_1 \neq 0$ E $A_2 \neq 0$ • $A_1 \neq 0$ E $A_2 = 0$ • $A_1 = 0$ E $A_2 \neq 0$

TEST D'IPOTESI PER UN SINGOLO COEFFICIENTE DI REGRESSIONE

IPOTESI NULLA	$A_i = 0$
IPOTESI ALTERNATIVA	$A_i \neq 0$

COSÌ.

GIUSTO.

USIAMO 0,05 COME LIVELLO DI SIGNIFICATIVITÀ.

OK.

SVOLGIAMO IL TEST DEL RAPPORTO DI VERO SIMIGLIANZA. CI PERMETTE DI ESAMINARE TUTTI I COEFFICIENTI INSIEME E DI VALUTARE LE RELAZIONI FRA I COEFFICIENTI.



I PASSI DEL TEST DEL RAPPORTO DI VERO SIMIGLIANZA

Passo 1	Definire le popolazioni.	Tutti i giorni in cui è stata venduta la Norns Special, confrontando i mercoledì, sabato e domenica rispetto agli altri giorni, a ogni temperatura massima.
Passo 2	Formulare un'ipotesi nulla e un'ipotesi alternativa.	L'ipotesi nulla è $A_1 = 0$ e $A_2 = 0$. L'ipotesi alternativa è $A_1 \neq 0$ oppure $A_2 \neq 0$.
Passo 3	Scegliere il test d'ipotesi da svolgere.	Svolgiamo il test del rapporto di verosimiglianza.
Passo 4	Scegliere il livello di significatività.	Useremo un livello di significatività di 0,05.
Passo 5	Calcolare la statistica test dai dati campione.	La statistica test è: $2[L - n_1 \log_e(n_1) - n_0 \log_e(n_0) + (n_1 + n_0) \log_e(n_1 + n_0)]$ Quando inseriamo i nostri dati troviamo: $2[-8.9010 - 8 \log_e 8 - 13 \log_e 13 + (8 + 13) \log_e (8 + 13)] = 10,1$ La statistica test segue una distribuzione chi-quadro con 2 gradi di libertà (il numero di variabili predittive), se l'ipotesi nulla è vera.
Passo 6	Determinare se il p -value per la statistica test ottenuto nel passo 5 è minore del livello di significatività.	Il livello di significatività è 0,05. Il valore della statistica test è 10,1, e quindi il p -value è 0,006. Infine, $0,006 < 0,05$.*
Passo 7	Decidere se possiamo scartare l'ipotesi nulla.	Dato che il p -value è minore del livello di significatività, scartiamo l'ipotesi nulla.

* A pagina 205 è spiegato come ottenere il p -value in una distribuzione chi-quadro.

ADESSO APPLICHIAMO IL TEST DI WALD PER VEDERE SE OGUNO DEI NOSTRI PREDITTORI HA UN EFFETTO SIGNIFICATIVO SULLA VENDITA DELLE NORNS SPECIAL. FACCIAMOLO CON I GIORNI DELLA SETTIMANA.



I PASSI DEL TEST DI WALD

Passo 1	Definire la popolazione.	Tutti i giorni in cui è stata venduta la Norns Special, confrontando i mercoledì, sabato e domenica rispetto agli altri giorni, a ogni temperatura massima.
Passo 2	Formulare un'ipotesi nulla e un'ipotesi alternativa.	L'ipotesi nulla è $A = 0$. L'ipotesi alternativa è $A \neq 0$.
Passo 3	Scegliere quale test d'ipotesi svolgere.	Svolgiamo il test di Wald.
Passo 4	Scegliere il livello di significatività.	Usiamo un livello di significatività di 0,05.
Passo 5	Calcolare la statistica test dai dati campione.	La statistica test per il test di Wald è $\frac{a_1^2}{S_{11}}$ In questo esempio il valore della statistica test è: $\frac{2,44^2}{1,5475} = 3,9$ La statistica test seguirà una distribuzione chi-quadro con 1 grado di libertà, se l'ipotesi nulla è vera.
Passo 6	Determinare se il p -value per la statistica test ottenuto nel passo 5 è minore del livello di significatività.	Il valore della statistica test è 3,9, e quindi il p -value è 0,0489. Osserviamo che $0,0489 < 0,05$, e quindi il p -value è inferiore.
Passo 7	Decidere se possiamo scartare l'ipotesi nulla.	Dato che il p -value è minore del livello di significatività, scartiamo l'ipotesi nulla.

IN ALCUNI TESTI QUESTO PROCEDIMENTO VIENE SPIEGATO USANDO LA DISTRIBUZIONE NORMALE ANZICHÉ IL CHI-QUADRO. IL RISULTATO FINALE SARÀ LO STESSO INDIPENDENTEMENTE DAL METODO USATO.



Ecco come calcoliamo la matrice degli errori standard. I valori di questa matrice sono usati per calcolare la statistica test di Wald nel passo 5 di pagina 180.

Temperatura massima

$$\begin{aligned}
 & \text{Mercoledì, sabato o domenica} \\
 & \left[\begin{array}{c|ccccc}
 & (\hat{y} \text{ il } 5) \times & 0 & \cdots & 0 \\
 & (1 - \hat{y} \text{ il } 5) & & & \\
 \hline
 0 & 0 & \cdots & 1 & \\
 28 & 24 & \cdots & 24 & \\
 \hline
 1 & 1 & \cdots & 1 &
 \end{array} \right]^{-1} \left[\begin{array}{c|ccccc}
 & (\hat{y} \text{ il } 6) \times & 0 & \cdots & 0 \\
 & (1 - \hat{y} \text{ il } 6) & & & \\
 \hline
 & & \vdots & & \vdots \\
 & & 0 & \cdots & (\hat{y} \text{ il } 25) \times \\
 & & & & (1 - \hat{y} \text{ il } 25)
 \end{array} \right]^{-1} \\
 & = \left[\begin{array}{c|ccccc}
 & 0,51 \times 0,49 & 0 & \cdots & 0 \\
 & 0 & 0,11 \times 0,89 & \cdots & 0 \\
 \hline
 0 & 28 & 24 & \cdots & 24 \\
 \hline
 1 & 1 & \cdots & 1 &
 \end{array} \right]^{-1} \left[\begin{array}{c|ccccc}
 & 0 & 28 & 1 \\
 & 0 & 24 & 1 \\
 \hline
 & & \vdots & & \vdots \\
 & & 1 & 24 & 1
 \end{array} \right]^{-1} \\
 & = \begin{pmatrix} 1,5388 & \cdots & \cdots \\ \cdots & 0,881 & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix}
 \end{aligned}$$

S_{11} del passo 5 è questo... ...e questo è S_{22} .

Questi 1 rappresentano una costante non misurabile. In altre parole, sono dei segnaposto.

QUINDI $A \neq 0$.
POSSIAMO SCAR-
TARE L'IPOTESI
NULLA!

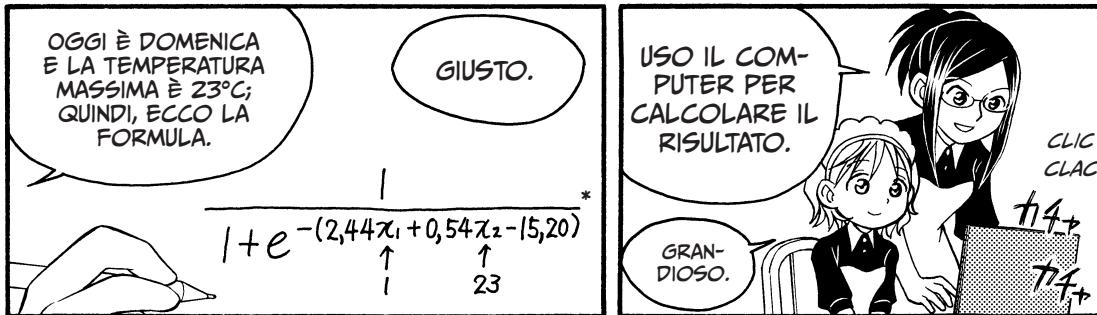
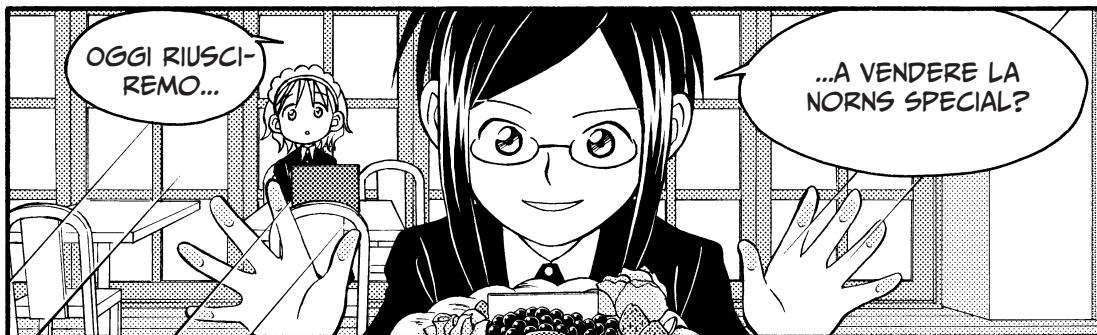
PROPRIO
COSÌ.

...LA PARTE PIÙ
IMPORTANTE.

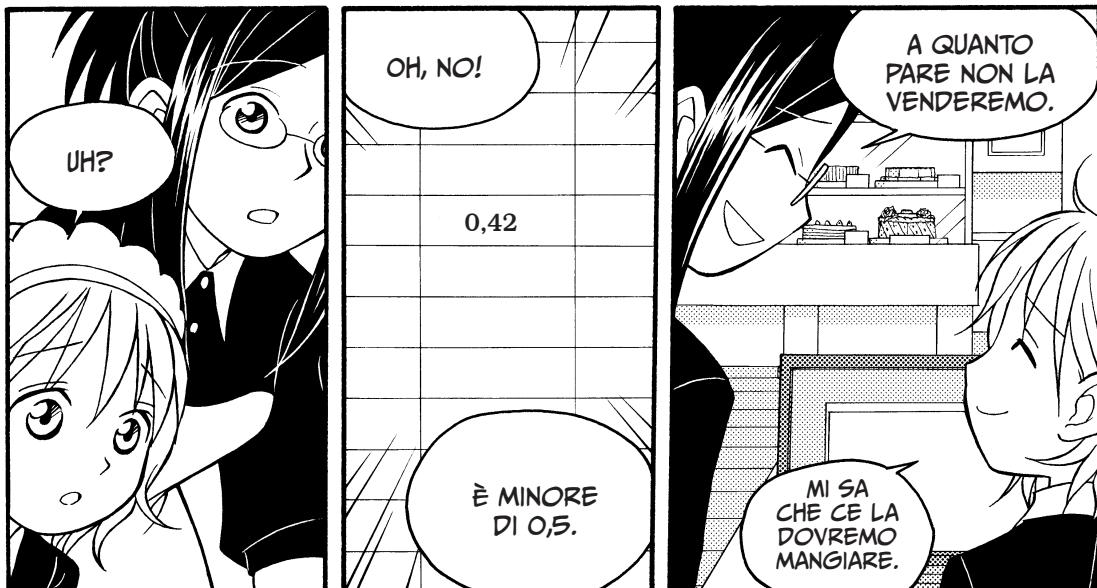
E
ADESSO...

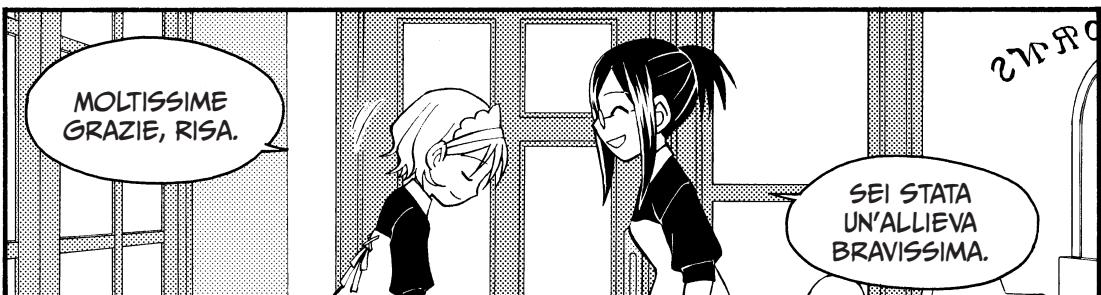
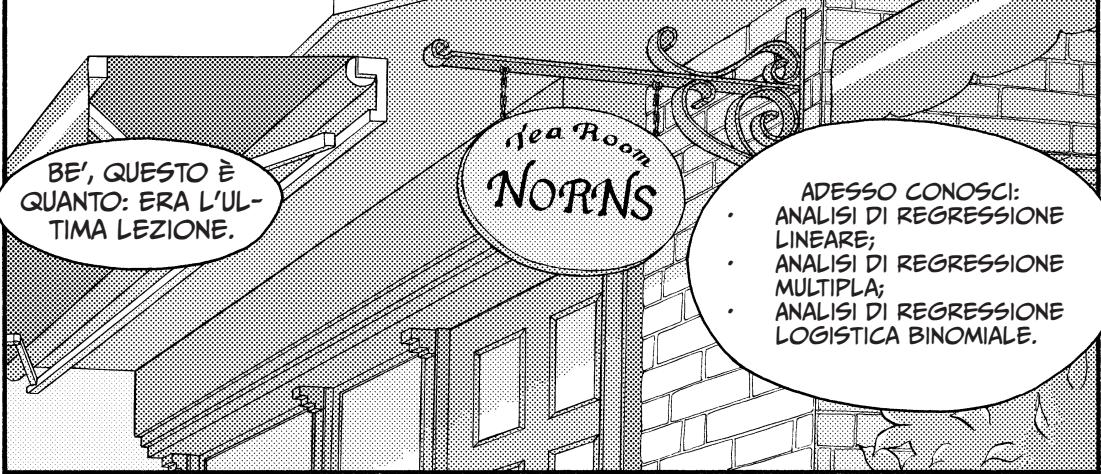
Frusc
A

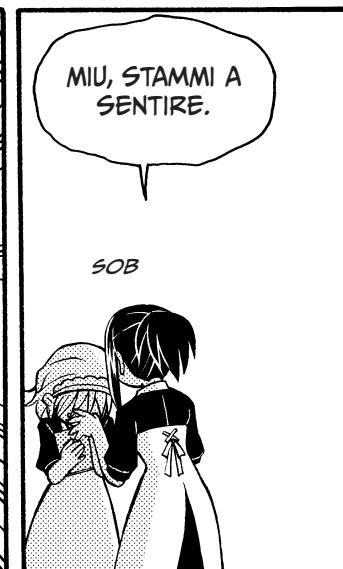
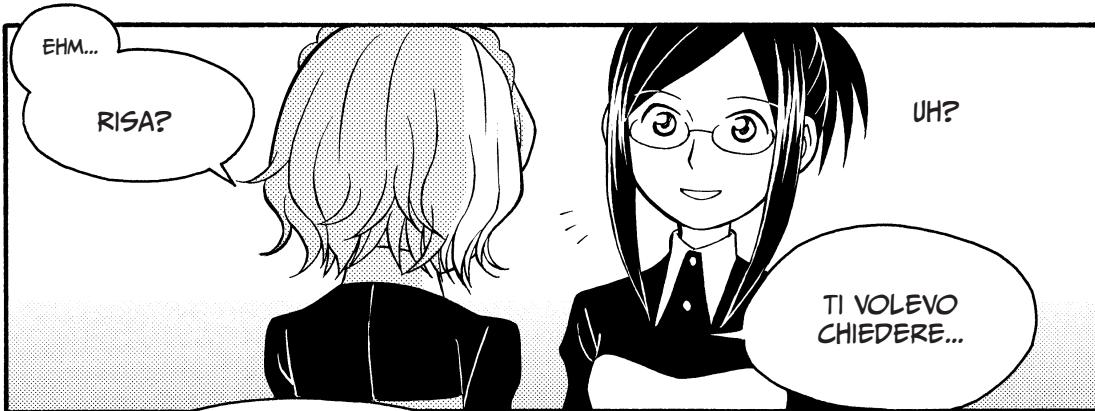
PASSO 5: PREVEDIAMO SE VERRÀ VENDUTA LA NORNS SPECIAL

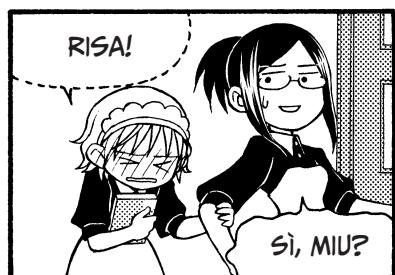
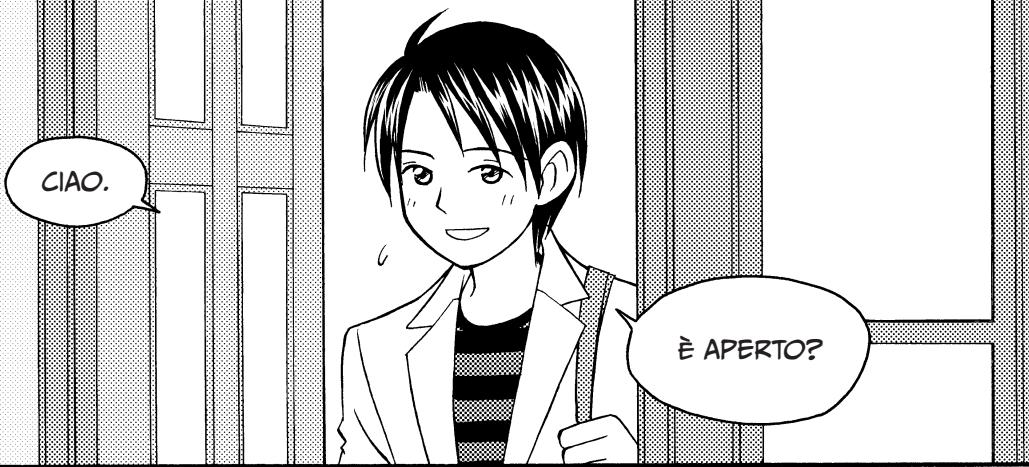


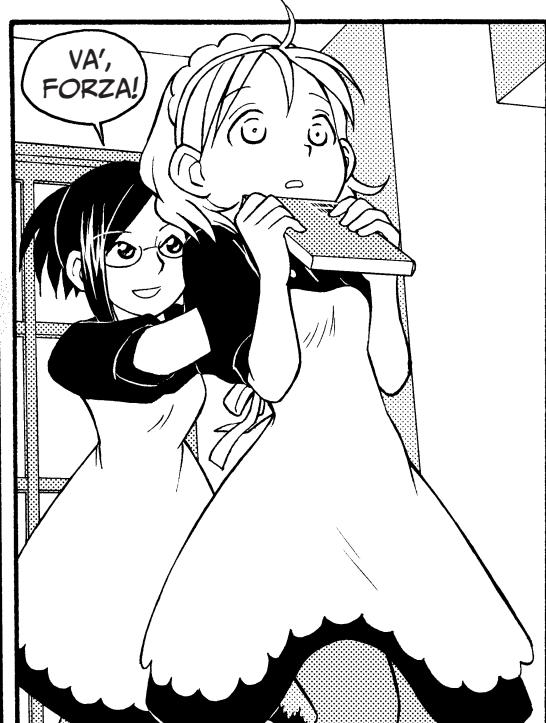
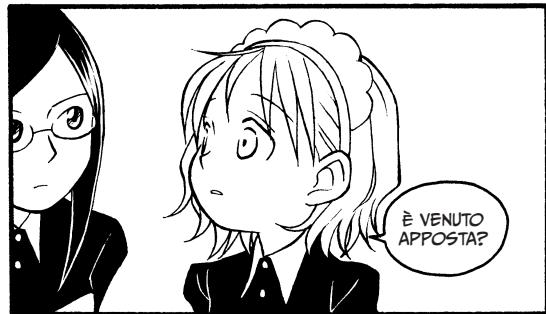
*QUESTO CALCOLO È STATO SVOLTO USANDO NUMERI ARROTONDATI. SE USATE QUELLI COMPLETI, OTTERRETE 0,44.

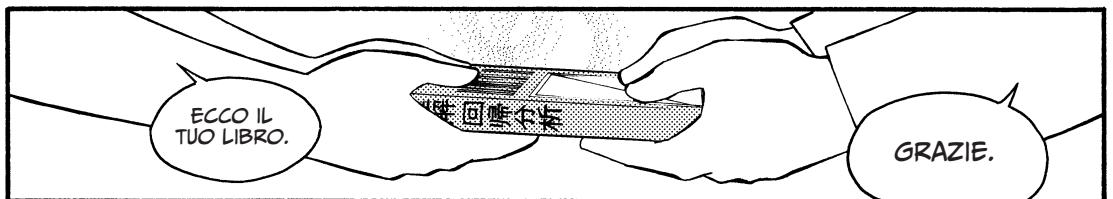


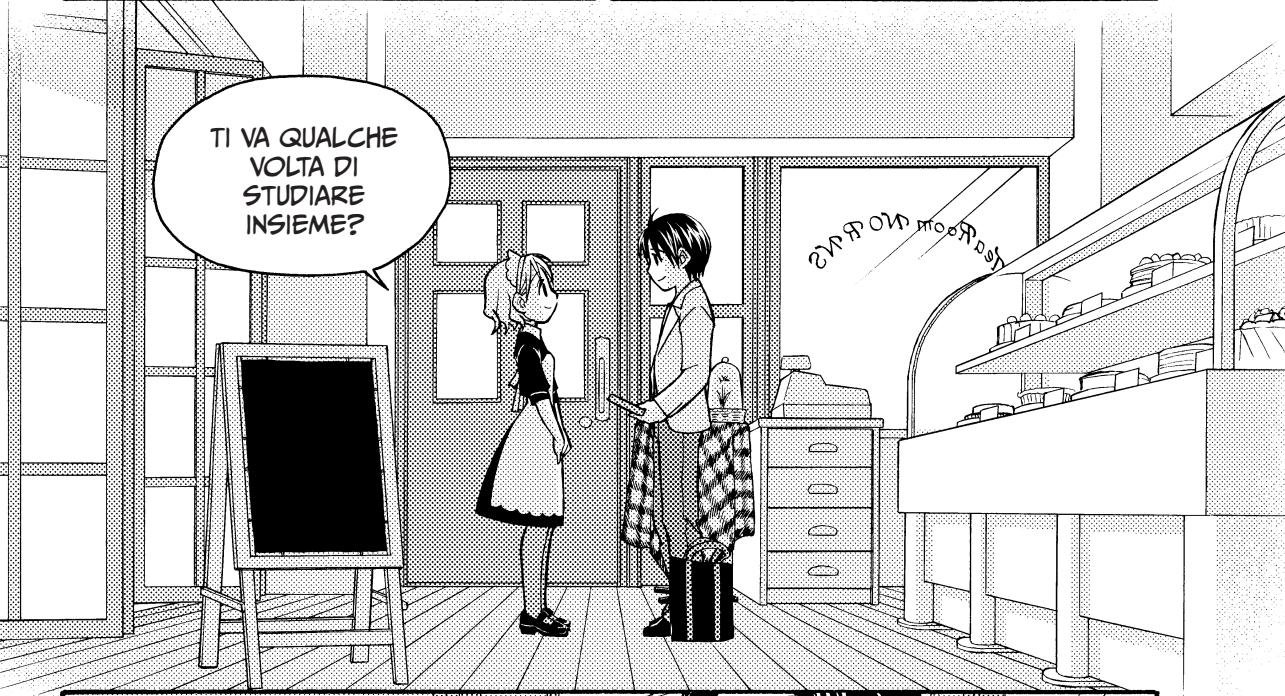


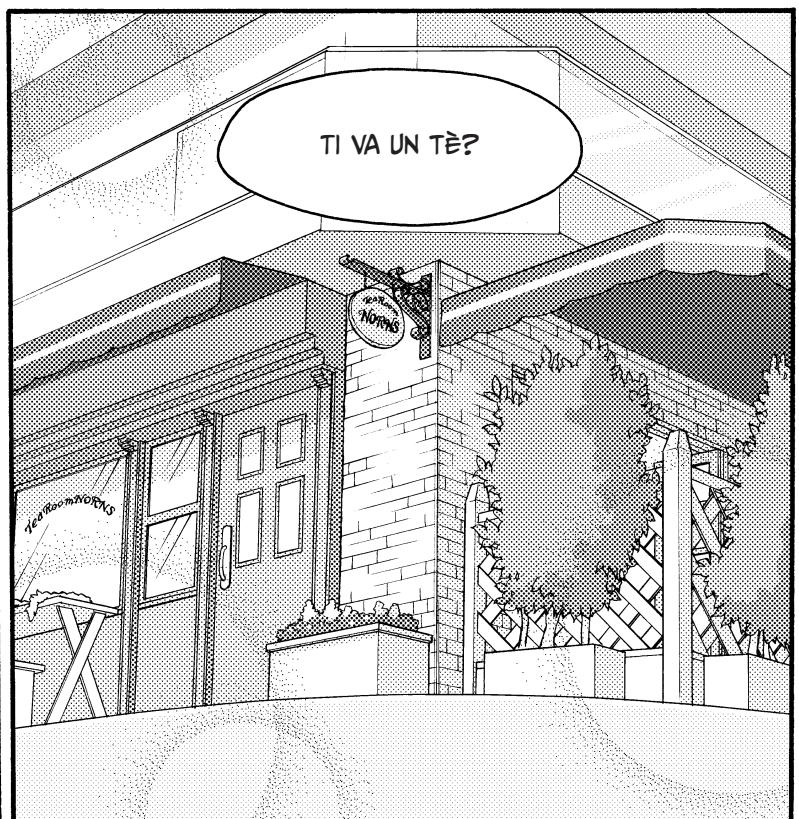
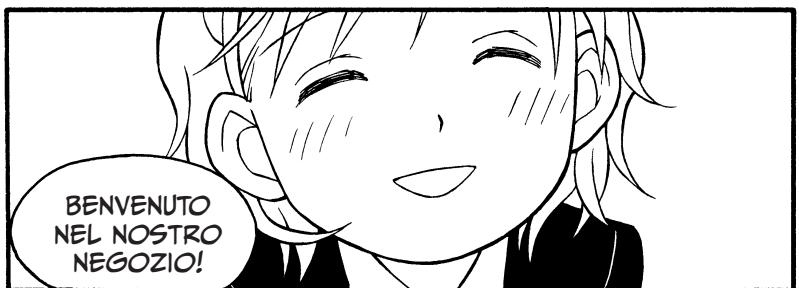
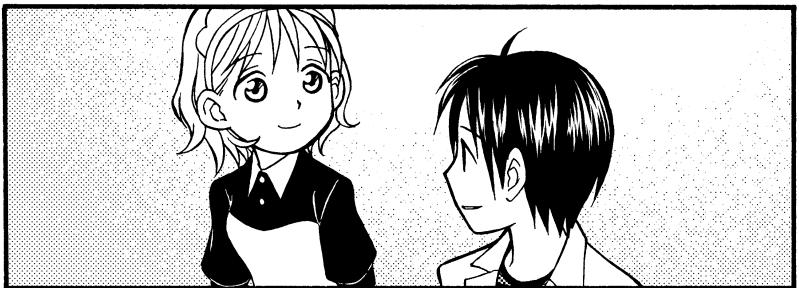












LA REGRESSIONE LOGISTICA NEL MONDO REALE

A pagina 68 Risa fa un elenco dei passi dell'analisi di regressione, ma poi abbiamo visto che non è sempre necessario svolgerli tutti. Per esempio, se analizziamo l'altezza di Miu nel corso degli anni, sappiamo che esiste una sola Miu e che a ogni età ha avuto una precisa e singola altezza. Non c'è una popolazione di altezze di Miu a 6 anni, e quindi non avrebbe senso analizzare la "popolazione".

Anche nel mondo reale succede abbastanza spesso di saltare il Passo 1, quello in cui si tracciano i grafici di dispersione, in particolare quando ci sono migliaia di punti da considerare. Per esempio, in un test clinico con molti partecipanti i ricercatori possono decidere di partire dal Passo 2 per risparmiare tempo, soprattutto se hanno un software che svolge rapidamente i calcoli al loro posto.

Inoltre, quando si applica la statistica al mondo reale, non ci si lancia ad applicare subito i test, bensì si pensa ai dati e all'obiettivo del test. Senza un contesto i numeri sono solo numeri e non significano niente.

LOGIT, ODDS RATIO E RISCHIO RELATIVO

Gli *odds* sono una grandezza che dà un'idea di quanto siano vicini un predittore e il responso. Sono definiti come il rapporto fra la probabilità che un evento si verifichi in una certa situazione (y) e la probabilità che non si verifichi ($1 - y$):

$$\frac{y}{1-y}$$

LOGIT

Il *logit* è il logaritmo degli odds. La funzione logistica è il suo inverso: prende dei log-odds e li trasforma in una probabilità. Il logit è imparentato matematicamente con i coefficienti di regressione: per ogni unità di incremento del predittore, il logit dell'evento aumenta del valore del coefficiente di regressione.

L'equazione della funzione logistica, che abbiamo già visto calcolando l'equazione di regressione logistica a pagina 170, è come segue:

$$y = \frac{1}{1 + e^{-z}}$$

dove z è il logit e y la probabilità.

Per trovare il logit, invertiamo l'equazione logistica così:

$$\log \frac{y}{1-y} = z.$$

Questa funzione inversa ci dà il logit basato sull'equazione di regressione logistica originaria. Il procedimento per trovare il logit è analogo al calcolo di qualsiasi altra inversa matematica:

$$\begin{aligned}
 y &= \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-z}} \times \frac{e^z}{e^z} = \frac{e^z}{e^z + 1} \\
 y \times (e^z + 1) &= \frac{e^z}{e^z + 1} \times (e^z + 1) \quad \text{MOLTIPLICHIAMO ENTRAMBI I MEMBRI DELL'EQUAZIONE PER } (e^z + 1) \\
 y \times e^z + y &= e^z \\
 y &= e^z - y \times e^z \quad \text{PASSIAMO DALL'ALTRA PARTE DELL'UGUALE.} \\
 y &= (1-y)e^z \\
 y \times \frac{1}{1-y} &= (1-y)e^z \times \frac{1}{1-y} \quad \text{MOLTIPLICHIAMO ENTRAMBI I MEMBRI DELL'EQUAZIONE PER } \frac{1}{1-y} \\
 \frac{y}{1-y} &= e^z \\
 \log \frac{y}{1-y} &= \log e^z = z
 \end{aligned}$$

Quindi l'equazione di regressione logistica per la vendita delle Norns Special (trovata a pagina 172),

$$y = \frac{1}{1 + e^{-(2,44x_1 + 0,54x_2 - 15,20)}},$$

si può riscrivere come

$$\log \frac{y}{1-y} = 2,44x_1 + 0,54x_2 - 15,20.$$

Quindi gli odds per la vendita della Norns Special un dato giorno a una data temperatura sono $e^{2,44x_1 + 0,54x_2 - 15,20}$, e il logit è $2,44x_1 + 0,54x_2 - 15,20$.

ODDS RATIO

Un altro modo per quantificare il nesso fra un predittore e un evento è l'*odds ratio* (OR). L'odds ratio confronta due insiemi di odds per condizioni diverse sulla stessa variabile.

Calcoliamo l'odds ratio della vendita della Norns Special di mercoledì, sabato o domenica rispetto agli altri giorni della settimana:

$$\frac{\left(\frac{\text{proporzione di vendite merc., sab. o dom.}}{1 - \text{proporzione di vendite merc., sab. o dom.}} \right)}{\left(\frac{\text{proporzione di vendite in giorni diversi da merc., sab. o dom.}}{1 - \text{proporzione di vendite merc., sab. o dom.}} \right)} = \frac{\left[\frac{(6/9)}{1 - (6/9)} \right]}{\left[\frac{(2/12)}{1 - (2/12)} \right]} = \frac{\left[\frac{(6/9)}{(3/9)} \right]}{\left[\frac{(2/12)}{(10/12)} \right]} =$$

$$\frac{(6/3)}{(2/10)} = \frac{6}{3} \div \frac{2}{10} = \frac{6}{3} \times \frac{10}{2} = 2 \times 5 = 10$$

Vediamo così che gli odds per la vendita della Norns Special in uno di questi tre giorni sono 10 volte più alti che negli altri giorni della settimana.

ODDS RATIO AGGIUSTATO

Finora abbiamo usato solo gli odds basati sul giorno della settimana. Se vogliamo la più fedele rappresentazione possibile dell'odds ratio, dobbiamo calcolare uno dopo l'altro l'odds ratio di ognuna delle variabili e poi mettere insieme i risultati. Questo è l'*odds ratio aggiustato*. Per i dati raccolti da Risa a pagina 176, ciò vuol dire trovare l'odds ratio per due variabili – giorno della settimana e temperatura – allo stesso tempo.

TABELLA 4-1: LE EQUAZIONI DI REGRESSIONE LOGISTICA E GLI ODDS PER I DATI DI PAGINA 176

Variabile predittiva	Equazione di regressione logistica	Odds
Solo "Merc., sab. o dom."	$y = \frac{1}{1 + e^{-(2,30x_1 - 1,61)}}$	$e^{(2,30x_1 - 1,61)}$
Solo "Temperatura massima"	$y = \frac{1}{1 + e^{-(0,52x_2 - 13,44)}}$	$e^{(0,52x_2 - 13,44)}$
"Merc., sab. o dom." e "Temperatura massima"	$y = \frac{1}{1 + e^{-(2,44x_1 + 0,54x_2 - 15,20)}}$	$e^{(2,44x_1 + 0,54x_2 - 15,20)}$

Gli odds per una vendita in base solo al giorno della settimana si calcolano come segue:

$$\frac{\text{odds di una vendita merc., sab. o dom.}}{\text{odds di una vendita un giorno diverso da merc., sab. o dom.}} = \frac{e^{2,30 \times 1 - 1,61}}{e^{2,30 \times 0 - 1,61}} =$$

$$e^{2,30 \times 1 - 1,61 - (2,30 \times 0 - 1,61)} = e^{2,30}$$

Questo è l'odds ratio non aggiustato per "mercoledì, sabato o domenica". Se lo valutiamo, troviamo $e^{2,30} = 10$, lo stesso valore che abbiamo ottenuto per l'odds ratio non aggiustato a pagina 192, come ci aspettavamo!

Per trovare gli odds per una vendita in base solo alla temperatura, consideriamo che effetto ha una variazione nella temperatura. Troviamo quindi che gli odds per la vendita di una torta con una differenza di 1 grado nella temperatura sono come segue:

$$\frac{\text{odds per una vendita con una massima di } (k+1) \text{ gradi}}{\text{odds per una vendita con una massima di } k \text{ gradi}} = \frac{e^{0,52 \times (k+1) - 13,44}}{e^{0,52 \times k - 13,44}} =$$

$$e^{0,52 \times (k+1) - 13,44 - (0,52 \times k - 13,44)} = e^{0,52}$$

Questo è l'odds ratio grezzo per un incremento di un grado nella temperatura.

L'equazione di regressione logistica che era stata calcolata con questi dati prendeva però in considerazione entrambe queste variabili insieme, e così i coefficienti di regressione (e quindi anche gli odds ratio) vanno modificati per tenere conto di più variabili.

In questo caso, con l'equazione di regressione calcolata usando sia il giorno della settimana che la temperatura, vediamo che sono cambiati sia gli esponenti che la costante. Per il giorno della settimana, il coefficiente è aumentato da 2,30 a 2,44, per la temperatura da 0,52 a 0,54 e la costante è adesso -15,20. Queste variazioni sono dovute alle *interazioni* fra variabili, quando un cambiamento di una variabile modifica gli effetti di un'altra: per esempio il fatto che il giorno sia sabato modifica gli effetti sulle vendite di un aumento della temperatura. Svolgiamo gli stessi calcoli con questi nuovi valori, variando prima il giorno della settimana:

$$\frac{e^{2,44 \times 1 + 0,54 \times k - 15,20}}{e^{2,44 \times 0 + 0,54 \times k - 15,20}} = e^{2,44 \times 1 + 0,54 \times k - 15,20 - (2,44 \times 0 + 0,54 \times k - 15,20)} = e^{2,44}$$

Questo è l'odds ratio aggiustato per "mercoledì, sabato o domenica". In altre parole, gli odds per il giorno della settimana sono stati modificati tenendo conto di eventuali effetti congiunti osservabili quando si considera anche della temperatura.

Analogamente, dopo aver aggiustato i coefficienti, si può ricalcare l'odds ratio per la temperatura:

$$\frac{e^{2,44 \times 1 + 0,54 \times (k+1) - 15,20}}{e^{2,44 \times 1 + 0,54 \times k - 15,20}} = \frac{e^{2,44 \times 0 + 0,54 \times (k+1) - 15,20}}{e^{2,44 \times 0 + 0,54 \times k - 15,20}} = e^{0,54 \times (k+1) - 15,20 - (0,54 \times k - 15,20)} = e^{0,54}$$

Questo è l'odds ratio aggiustato per "temperatura massima". In questo caso, l'odds ratio per la temperatura è stato modificato in modo da tenere conto di eventuali effetti del giorno della settimana.

VERIFICA D'IPOTESI CON GLI ODDS

Come ricorderete, nell'analisi di regressione lineare svolgiamo un test d'ipotesi chiedendoci se A è uguale a zero, così:

Ipotesi nulla	$A_i = 0$
Ipotesi alternativa	$A_i \neq 0$

Nell'analisi di regressione logistica, svolgiamo un test d'ipotesi valutando se il coefficiente A , come potenza di e , è uguale a e^0 :

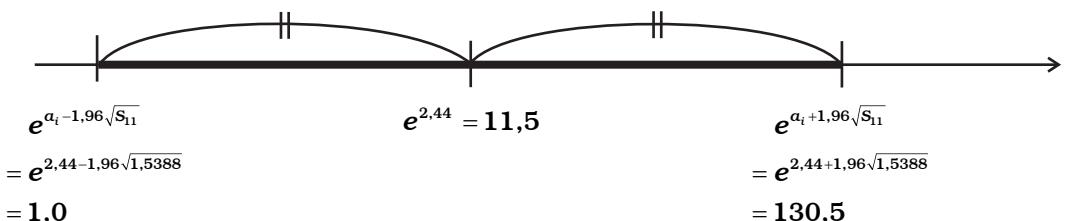
Ipotesi nulla	$e^{A_i} = e^0 = 1$
Ipotesi alternativa	$e^{A_i} \neq e^0 = 1$

Ricordiamo dalla Tabella 4-1 che $e^{(2,30x_1 - 1,61)}$ sono gli odds di vendere la Norns Special in base al giorno della settimana. Se invece trovassimo che gli odds sono $e^{0x_1 - 1,61}$, significherebbe che gli odds di venderla sono uguali ogni giorno della settimana e quindi l'ipotesi nulla sarebbe vera: il giorno della settimana non ha effetto sulle vendite. Controllare se $A_i = 0$ oppure $e^{A_i} = e^0 = 1$ è in realtà la stessa cosa, ma visto che l'analisi della regressione logistica lavora con odds e probabilità, ha più senso scrivere il test d'ipotesi in termini di odds.

INTERVALLO DI CONFIDENZA PER L'ODDS RATIO

Gli odds ratio si usano spesso negli studi clinici e li si presenta in genere insieme a un intervallo di confidenza. Per esempio, dei ricercatori, che stanno cercando di capire se lo zenzero aiuti a rimettersi in sesto dal mal di stomaco, possono dividere un gruppo di persone indisposte in due sottogruppi per somministrare a uno pillole di zenzero e all'altro un placebo. E si calcolerà il malessere delle persone dopo aver preso le pillole e si calcolerà un odds ratio; se questo mostrasse che le persone che hanno preso lo zenzero stanno meglio di quelle che hanno preso il placebo, i ricercatori potranno usare un intervallo di confidenza per farsi un'idea dell'errore standard e dell'accuratezza del risultato.

Possiamo anche calcolare un intervallo di confidenza per i dati delle Norns Special. Qui sotto calcoliamo l'intervallo con un livello di confidenza del 95%.



Se come popolazione prendiamo tutti i giorni in cui era in vendita la Norns Special, possiamo essere sicuri che l'odds ratio sia compreso fra 1 e 130,5. In altre parole, nella peggiore delle ipotesi non ci sono differenze nelle vendite in base ai giorni della settimana (quando l'odds ratio = 1) e nella migliore c'è una differenza molto spiccata in base al giorno. Se scegliamo un livello di confidenza del 99%, possiamo sostituire qui sopra 1,96 con 2,58, il che amplia l'intervallo: da 0,5 a 281,6. Come vedete, un livello di confidenza maggiore porta a un intervallo più ampio.

RISCHIO RELATIVO

Il *rischio relativo* (RR), un altro tipo di rapporto, confronta la probabilità che un evento si verifichi in un gruppo esposto a uno specifico fattore con la probabilità che lo stesso evento si verifichi in un gruppo non esposto. Questo rapporto si usa spesso in statistica quando si vogliono confrontare due possibili risultati e quello che ci interessa è relativamente raro. Per esempio, si usa spesso per studiare i fattori che possono portare a contrarre una malattia oppure gli effetti collaterali di un farmaco.

Possiamo usare il rischio relativo anche per studiare una questione meno seria (e meno rara): se il giorno della settimana aumenti le possibilità di vendere la Norns Special. Usiamo i dati di pagina 166.

Per prima cosa, costruiamo uno schema come la Tabella 4-2 con la condizione a sinistra e il risultato in alto. In questo caso la condizione è il giorno della settimana; la condizione deve essere binaria (sì o no) e quindi, visto che Risa ritiene che la Norns Special si venga di più di mercoledì, sabato e domenica consideriamo la condizione come presente ognuno di questi tre giorni e assente gli altri. Quanto al risultato, o la torta viene venduta o no.

TABELLA 4-2: INCROCI DEI DATI "MERCOLEDÌ, SABATO O DOMENICA"
E "VENDITE DELLA NORMS SPECIAL"

		Vendite della Norns Special		Somma
		Si	No	
Mer, Sab o Dom	Sì	6	3	9
	No	2	10	12
Somma		8	13	21

Per calcolare il rischio relativo, dobbiamo trovare il rapporto tra le Norns Special effettivamente vendute di mercoledì, sabato e domenica e quelle disponibili complessivamente in quei giorni. Nei dati del nostro campione, il numero di vendite è 6, e il numero di torte poste in vendita è 9 (3 non sono state vendute). Quindi il rapporto è 6:9.

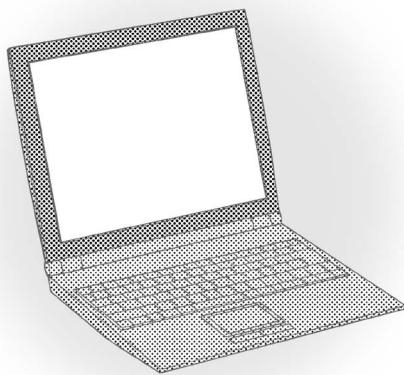
Ci serve poi il rapporto fra il numero di torte vendute negli altri giorni e quelle disponibili in questi altri giorni: questo rapporto è 2:12.

$$\frac{\text{rapporto delle vendite merc., sab. o dom.}}{\text{rapporto delle vendite in giorni diversi da merc., sab. o dom.}} = \frac{(6/9)}{(2/12)} = \frac{6}{9} \div \frac{2}{12} = \frac{6}{9} \times \frac{12}{2} = \frac{2}{3} \times 6 = 4$$

Quindi è quattro volte più probabile vendere la Norns Special di mercoledì, sabato e domenica. A quanto pare Risa aveva ragione!

È importante osservare che spesso i ricercatori riportano l'odds ratio anziché il rischio relativo perché il primo ha più strettamente a che fare con i risultati dell'analisi della regressione logistica e perché a volte non si riesce a calcolare il rischio relativo; per esempio, se non avessimo dati completi sulle vendite nei giorni diversi da mercoledì, sabato e domenica. Il rischio relativo, però, è più utile in alcune situazioni ed è spesso più facile da capire perché usa le probabilità e non gli odds.

APPENDICE
CALCOLI DI REGRESSIONE
CON UN FOGLIO ELETTRONICO



Questa appendice vi mostrerà come usare le funzioni di un foglio elettronico per calcolare:

- la costante di Eulero (e);
- le potenze;
- i logaritmi naturali;
- il prodotto di matrici;
- le matrici inverse;
- la statistica chi-quadro da un p -value;
- il p -value da una statistica chi-quadro;
- la statistica F da un p -value;
- il p -value da una statistica F ;
- il coefficiente di regressione parziale di un'analisi di regressione multipla;
- il coefficiente di regressione di un'equazione di regressione logistica.

Useremo un foglio elettronico che comprende già i dati degli esempi usati in questa appendice.

LA COSTANTE DI EULERO

La costante di Eulero (e), introdotta a pagina 19, è il numero usato come base per i logaritmi naturali. Questa funzione ci permetterà di elevare a qualsiasi potenza la costante di Eulero. In questo esempio calcoleremo e^1 .

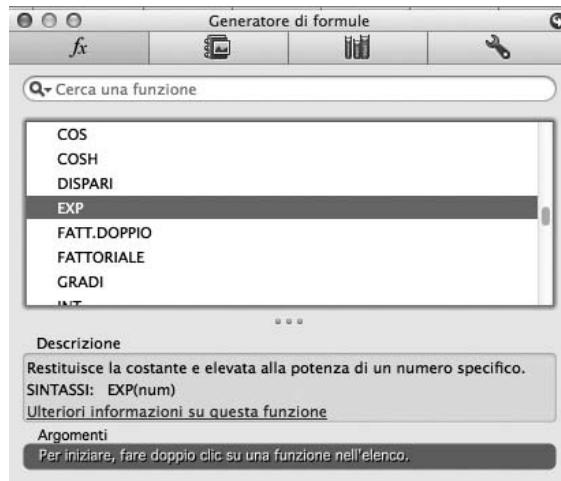
1. Andate al foglio **Costante di Eulero** del foglio elettronico.
2. Selezionate la cella **B1**.

B1	A	B	C	D
1	e^1			
2				
3				
4				

3. Selezionate dal menù principale **Generatore di Formule**.



4. Dal menù delle categorie, selezionate Matematiche e trigonometriche, poi selezionate la funzione EXP e premete OK.



5. A questo punto apparirà una finestra di dialogo in cui potrete inserire la potenza a cui elevare e . Inserite 1 e premete INVIO.



Dato che abbiamo calcolato la costante di Eulero alla potenza 1, otteniamo semplicemente il valore di e (con poche cifre decimali), ma con la funzione EXP possiamo elevare e a qualsiasi potenza.

	B1	\times	\checkmark	\leftarrow	\rightarrow	fx	=EXP(1)
1	A	B	C	D	E		
1	e^1	2,7182818					
2							
3							

NOTA

Invece di selezionare Formule / Generatore di Formule dal menù principale, potete scrivere direttamente “=EXP(x)” nella cella. Per esempio, inserire “=EXP(1)” restituisce il valore di e . Questo vale per qualsiasi funzione: dopo avere selezionato Formule / Generatore di Formule da menù, basta prendere nota del nome della formula da inserire direttamente nella cella.

POTENZE

Questa funzione si può usare per elevare qualsiasi numero a qualsiasi potenza. Usiamo la domanda dell'esempio a pagina 14: "Quanto fa 2 al cubo?".

1. Andate al foglio *Potenze* del foglio elettronico.
2. Scegliete la cella **B1** e digitate " $=2^3$ ". Premete INVIO.

	B1			
	A	B	C	D
1	2^3		8	
2				

Nei fogli elettronici (e anche più in generale) si usa il simbolo " $^$ " per indicare "alla potenza", cioè " 2^3 " significa 2^3 , e il risultato è 8. Ricordatevi di includere il segno di uguale ("=") all'inizio, altrimenti la formula non viene interpretata e il risultato non verrà calcolato.

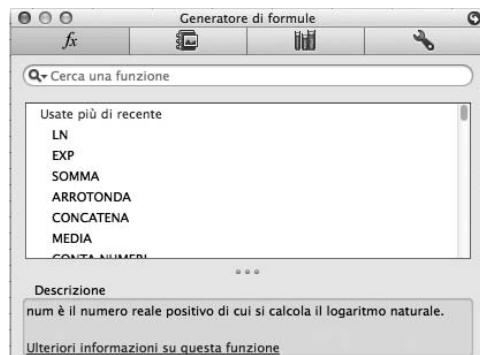
LOGARITMI NATURALI

Questa funzione calcola i logaritmi naturali (vedi pagina 20).

1. Andate al foglio *Logaritmi naturali* del foglio elettronico.
2. Cliccate la cella **B1**. Selezionate dal menù principale **Formule / Generatore di Formule**.
3. Dal menù delle categorie, scegliete **Matematiche e trigonometriche**. Scegliete la funzione **LN** e premete INVIO.



4. Inserite “ $\exp(3)$ ” e premete INVIO.

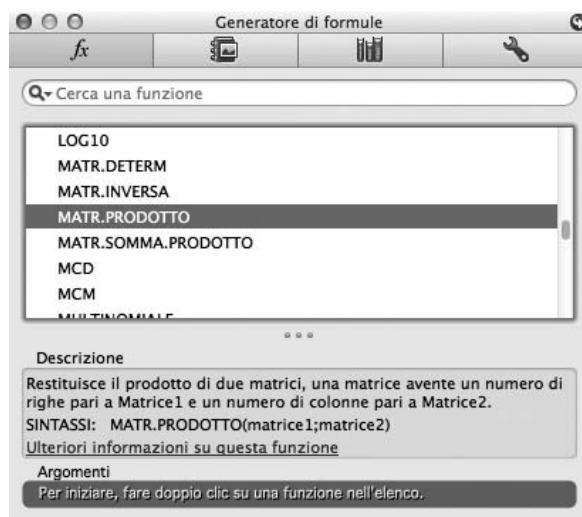


Il risultato sarà il logaritmo naturale di e^3 che, in base alla Regola 3 di pagina 22, sarà ovviamente 3. Potete inserire qualsiasi altro numero, come potenza di e oppure no, e trovarne il logaritmo naturale. Per esempio, inserendo “ $\exp(2)$ ” otterremo 2, mentre “2” darà come risultato (approssimato) 0,6931.

PRODOTTO DI MATRICI

Questa funzione si usa per moltiplicare matrici: faremo i calcoli relativi all'esempio mostrato nell'Esercizio risolto 1 di pagina 41.

1. Andate al foglio *Prodotto di matrici* del foglio elettronico.
2. Scegliete la cella G1. Selezionate dal menù principale **Formule / Generatore di Formule**.
3. Dal menù delle categorie, scegliete **Matematiche e trigonometriche**. Scegliete la funzione **MATR.PRODOTTO** e premete INVIO.



4. Cliccate il campo **matrice1** e inserite “A1:B2”. Poi cliccate nel secondo campo **matrice2** e inserite “D1:E2”. Premete INVIO.

5. A partire da **G1**, evidenziate una matrice di celle delle stesse dimensioni delle matrici che state moltiplicando: in questo esempio, da **G1** a **H2**. Poi cliccate nella barra delle funzioni.

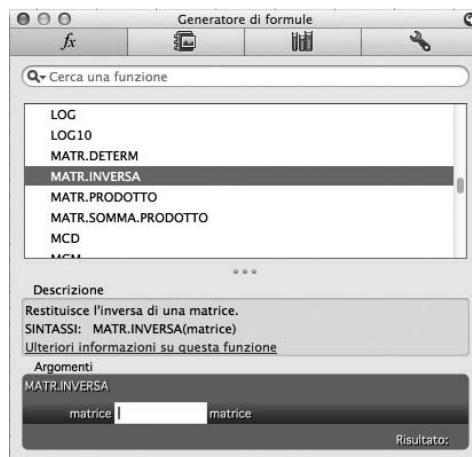
6. Premete **CTRL-MAIUSC-INVIO**. Nelle celle della matrice risultato appariranno i risultati corretti.

Otterrete gli stessi risultati trovati da Risa alla fine di pagina 41. Potete usare questo procedimento con qualsiasi coppia di matrici delle stesse dimensioni.

MATRICI INVERSE

Questa funzione calcola le inverse delle matrici; useremo l'esempio mostrato a pagina 44.

1. Andate al foglio *Inversa* di una matrice del foglio elettronico.
2. Cliccate sulla cella **D1**. Selezionate dal menù principale **Formule / Generatore di Formule**.
3. Dal menù delle categorie, scegliete **Matematiche e trigonometriche**. Selezionate la funzione **MATR.INVERSA** e premete INVIO.



4. Selezionate ed evidenziate la matrice nel foglio – cioè le celle da **A1** a **B2** – e premete INVIO.

5. A partire da **D1**, selezionate ed evidenziate una matrice di celle con le stesse dimensioni della prima matrice: in questo caso, da **D1** a **E2**. Poi cliccate nella barra delle funzioni.

	A	B	C	D	E	F	G
1	1	2		-2			
2	3	4					
3							

6. Premete **CTRL-MAIUSC-INVIO**: a partire da **D1** apparirà la matrice inversa di quella data.

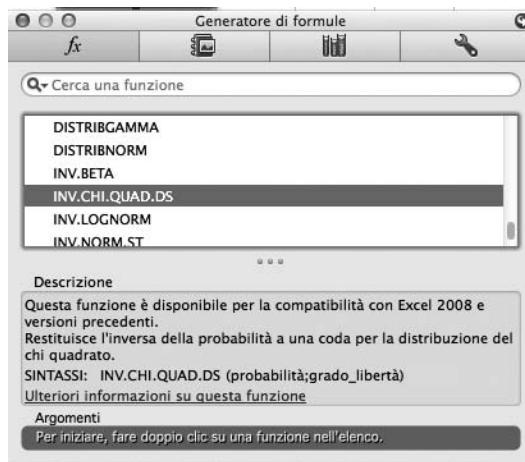
	A	B	C	D	E	F
1	1	2		-2	1	
2	3	4		1,5	-0,5	
3						

Si tratta degli stessi risultati ottenuti da Risa a pagina 44 e lo potete fare per qualsiasi matrice che volete invertire.

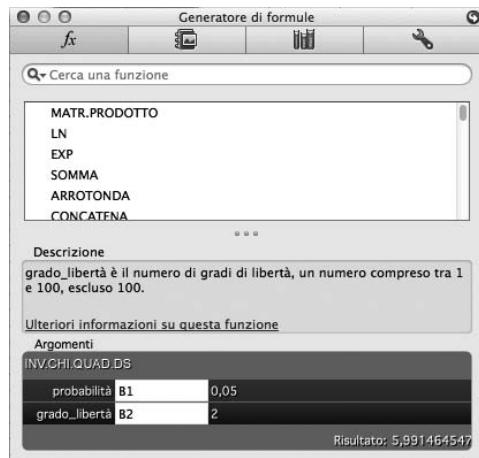
CALCOLARE UNA STATISTICA CHI-QUADRO DA UN P-VALUE

Questa funzione calcola una statistica test di una distribuzione chi-quadro, come abbiamo visto a pagina 54. Useremo un *p-value* di 0,05 e 2 gradi di libertà.

1. Andate al foglio *Chi-quadro da p-value* del foglio elettronico.
2. Scegliete la cella **B3**. Selezionate dal menù principale **Formule / Generatore di Formule**.
3. Dal menù delle categorie, scegliete **Statistiche**. Selezionate la funzione **INV.CHI.QUAD.DS**, poi premete **INVIO**.



4. Cliccate nel primo campo, **Probabilità**, e inserite **B1** per selezionare il valore di probabilità. Poi cliccate nel secondo campo, **Grado_libertà**, e inserite **B2** per selezionare il numero dei gradi di libertà. Quando nella cella **B3** appare (**B1, B2**), premete **INVIO**.

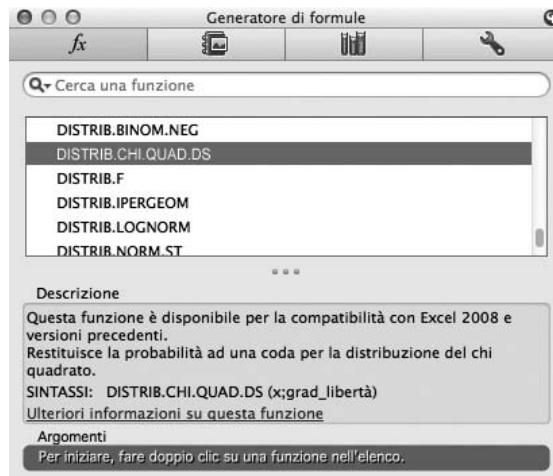


Potete confrontare questo risultato con i valori della Tabella 1-8 a pagina 56.

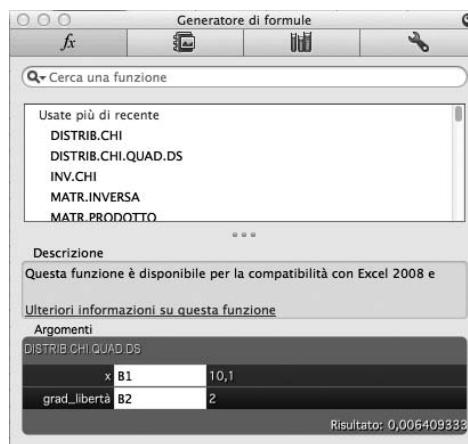
CALCOLARE UN *P*-VALUE DA UNA STATISTICA CHI-QUADRO

Questa funzione viene usata a pagina 179 nel test del rapporto di verosimiglianza per ottenere un *p*-value. Usiamo per la statistica test un valore pari a 10,1 e 2 gradi di libertà.

1. Andate al foglio *p*-value da chi-quadro del foglio elettronico.
2. Scegliete la cella B3. Selezionate dal menù principale **Formule / Generatore di Formule**.
3. Dal menu delle categorie, scegliete Statistiche. Selezionate la funzione **DISTRIB.CHI.QUAD.DS**, poi premete INVIO.



4. Inserite **B1** per selezionare il valore del chi-quadro. Poi inserite **B2** per i gradi di libertà. Quando nella cella **B3** appare (**B1, B2**), premete **INVIO**.



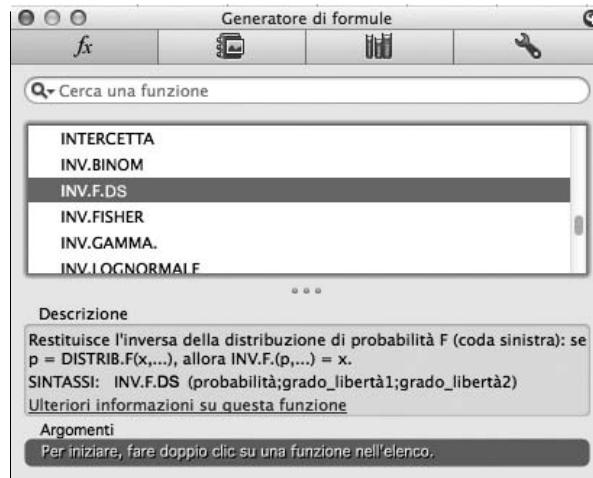
Troviamo 0,0064093, che a pagina 179 era stato arrotondato a 0,006.

	B3		=DISTRIB.CHI
	A	B	C
1	Chi-quadro	10,1	
2	Gradi di libertà	2	
3	Probabilità	0,0064093	
4			

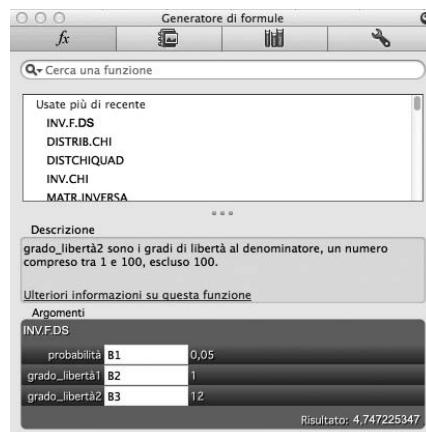
CALCOLARE UNA STATISTICA F DA UN P-VALUE

Questa funzione ci dà la statistica F che abbiamo calcolato a pagina 58.

1. Andate al foglio **Statistica F da p-value** del foglio elettronico.
2. Scegliete la cella **B4**. Selezionate dal menù principale **Formule / Generatore di Formule**.
3. Dal menù delle categorie, scegliete **Statistiche**. Selezionate la funzione **INV.F.DS**, poi premete **INVIO**.



4. Cliccate il primo campo, **probabilità**, e selezionate **B1** per inserire il valore di probabilità. Cliccate il secondo campo, **grad_libertà1**, e inserite **B2**, poi il terzo campo, **grad_libertà2**, e inserite **B3**. Quando nella cella **B3** appare (**B1, B2, B3**), premete **INVIO**.



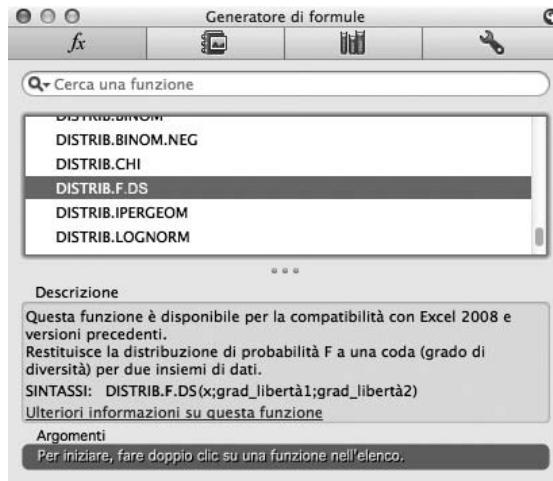
Troviamo 4,747225, che nella Tabella 1-7 di pagina 58 era stato arrotondato a 4,7.

J14		<input type="button" value="fx"/>	INV.F.DS(B1,B2,B3)	
	A	B	C	D
1	Probabilità	0,05		
2	I Grado di libertà	1		
3	II Grado di libertà	12		
4	F	4,747225		
5				

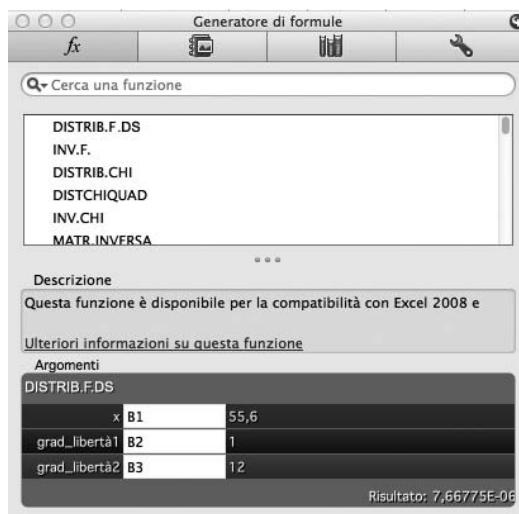
CALCOLARE UN *P*-VALUE DA UNA STATISTICA F

Questa funzione viene usata a pagina 90 per calcolare il *p*-value in un'ANOVA.

1. Andate al foglio *p-value per statistica F* del foglio elettronico.
2. Scegliete la cella B4. Selezionate dal menù principale **Formule / Generatore di Formule**.
3. Dal menù delle categorie, scegliete **Statistiche**. Selezionate la funzione **DISTRIB.F.DS** poi premete INVIO.



4. Cliccate B1 per selezionare il valore x da questa cella. Cliccate il secondo campo, **grad_libertà1** e inserite B2; poi cliccate il terzo campo, **grad_libertà2**, e inserite B3. Premete INVIO.



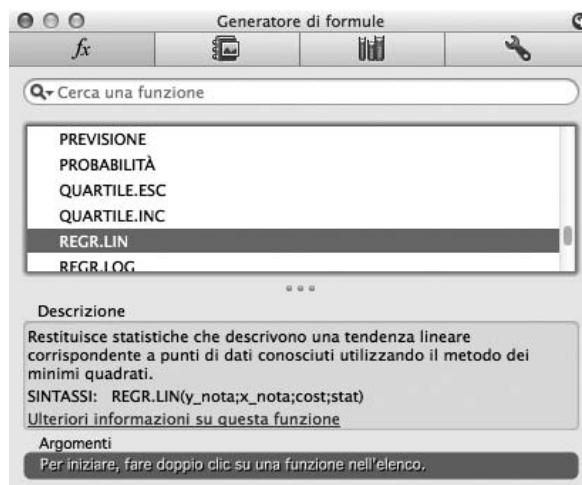
Il risultato è $7,66775 \times 10^{-6}$. Se stessimo svolgendo il test al livello $p = 0,05$, questo sarebbe un risultato significativo, perché è minore di 0,05.

	A	B	C	D	E
1	F	55,6			
2	I Grado di libertà	1			
3	II Grado di libertà	12			
4	Probabilità	7,66775E-06			
5					

COEFFICIENTE DI REGRESSIONE PARZIALE DI UN'ANALISI DI REGRESSIONE MULTIPLA

Questa funzione calcola il coefficiente di regressione parziale per i dati a pagina 113, dando i risultati che Risa ha trovato a pagina 118.

1. Andate al foglio *Coefficiente di regressione parziale* del foglio elettronico.
2. Scegliete la cella G2. Selezionate dal menù principale **Formule / Generatore di formule**.
3. Dal menù delle categorie, scegliete **Statistiche**. Selezionate la funzione **REGR.LIN**, poi premete INVIO.



4. Cliccate **y_nota** ed evidenziate le celle per le variabili **risponso**, nel nostro caso da **D2** a **D11**. Cliccate nel campo **x_nota** ed evidenziate le celle per le variabili predittive, nel nostro caso da **B2** a **C11**. Non servono valori per i campi **cost** e **stat**. Premete **INVIO**.



5. La funzione ci restituisce complessivamente tre valori. Selezionate quindi le celle da **G2** a **I2**, cliccate la barra delle funzioni e premete **CTRL-MAIUSC-INVIO**. Nei campi evidenziati compariranno i valori corretti.

	A	B	C	D	E	F	G	H	I
		Area del negozio (tsubo)	Distanza dalla stazione più vicina (metri)	Vendite mesili			Distanza dalla stazione più vicina (metri)	Area del negozi (tsubo)	Intercetta
1									
2	Yumenooka	10	80	469	Coefficiente di regressione parziale	-0,34088	41,5135	65,3239	
3	Stazione Terai	8	0	366					
4	Sone	8	200	371					
5	Stazione Hashimoto	5	200	208					
6	Quartiere Kikyō	7	300	246					
7	Ufficio postale	8	230	297					
8	Suidobashi Station Shop	7	40	363					
9	Stazione Rokujo	9	0	436					
10	Lungofiume Wakaba	6	330	198					
11	Misato	9	180	364					
12									

Come potete vedere, i risultati sono gli stessi di Risa a pagina 118 (nel testo sono stati arrotondati).

COEFFICIENTE DI REGRESSIONE DI UN'EQUAZIONE DI REGRESSIONE LOGISTICA

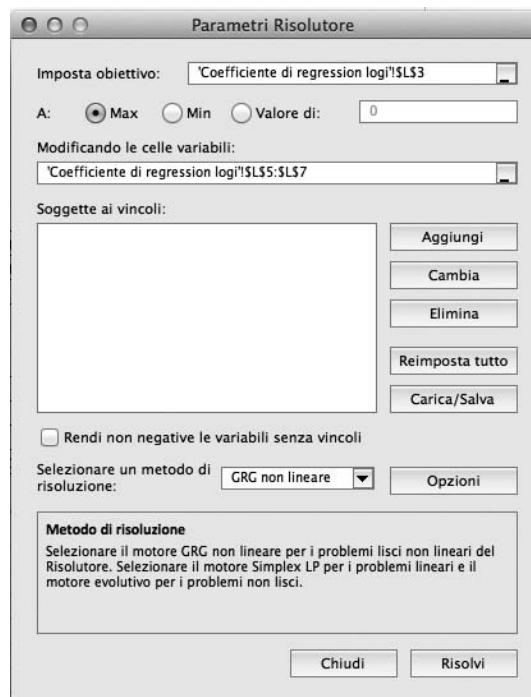
Non esiste una funzione che calcola il coefficiente di regressione logistica, ma possiamo usare lo strumento Risolutore del foglio di calcolo. In questo esempio calcoliamo i coefficienti di massima

verosimiglianza per l'equazione di regressione logistica usando i dati di pagina 166.

1. Andate al foglio *Coefficiente di regressione logistica* del foglio elettronico.
2. Selezionate dal menù principale **Dati / Risolutore**.



3. Cliccate sul pulsante **Risolutore**. Cliccate nel campo **Imposta obiettivo** e selezionate la cella L3 per scegliere il logaritmo dei dati di massima verosimiglianza. Cliccate il campo **Modificando le celle variabili** e selezionate le celle in cui volete che appaiano i vostri risultati; nel nostro caso, da L5 a L7. Cliccate **Risolvi**.



	A	B	C	D	E	F	G	H	I	J	K	L
1			Mercoledì , sabato e domenica	Temperatura massima	Vendita della Norns Specials	Valori previsti		Funzione di verosimiglianza				
2	5 giorni	Lunedì	0	28	1	0,51	0,51	Funzione di verosimiglianza		0,000136246		
3	6 giorni	Martedì	0	24	0	0,11	0,89	Logaritmo della Funzione di verosimiglianza		-8,901045158		
4	7 giorni	Mercoledì	1	26	0	0,80	0,20					
5	8 giorni	Giovedì	0	24	0	0,11	0,89	a1		2,442479882		
6	9 giorni	Venerdì	0	23	0	0,06	0,94	a2		0,544473597		
7	10 giorni	Sabato	1	28	1	0,92	0,92	b		-15,202633333		
8	11 giorni	Domenica	1	24	0	0,58	0,42					
9	12 giorni	Lunedì	0	26	1	0,26	0,26	Numero dei giorni in cui è stata venduta la Norns Special		8		
10	13 giorni	Martedì	0	25	0	0,17	0,83	Numero dei giorni in cui non è stata venduta la Norns Special		13		
11	14 giorni	Mercoledì	1	28	1	0,92	0,92	Pseudo R-quadro di McFadden		0,362165306		
12	15 giorni	Giovedì	0	21	0	0,02	0,98					
13	16 giorni	Venerdì	0	22	0	0,04	0,96	Le celle con un colore più chiaro contengono funzioni				
14	17 giorni	Sabato	1	27	1	0,87	0,87					
15	18 giorni	Domenica	1	26	1	0,80	0,80					
16	19 giorni	Lunedì	0	26	0	0,26	0,74					
17	20 giorni	Martedì	0	21	0	0,02	0,98					
18	21 giorni	Mercoledì	1	21	1	0,21	0,21					
19	22 giorni	Giovedì	0	27	0	0,38	0,62					
20	23 giorni	Venerdì	0	23	0	0,06	0,94					
21	24 giorni	Sabato	1	22	0	0,31	0,69					
22	25 giorni	Domenica	1	24	1	0,58	0,58					

Otterrete gli stessi risultati del Passo 4 di pagina 172 (nel testo sono stati arrotondati).

INDICE

0-9 e caratteri speciali

Δ (delta), 29

' (primo), 32

A

accuratezza. (v. anche coefficiente di determinazione)

- analisi di regressione logistica, equazione, 173–177

- regressione multipla, equazione, 119–126

adattamento eccessivo, 149

addizione di matrici, 39–40

alternativa, ipotesi (Ha), (v. ipotesi alternativa)

analisi della varianza (ANOVA), (v. varianza, analisi)

analisi di regressione, - lineare, (v. regressione lineare, analisi)

- logistica, (v. regressione logistica, analisi)

- multipla, (v. regressione multipla, analisi)

anomala, misura, 101, 144

autocorrelazione, 102–103

B

binomiale, analisi di regressione logistica, (v. regressione logistica, analisi)

binomiale, equazione logistica, 171

C

calcolo differenziale,

24–30

- derivazione, 31–36

campana, curva a, 53–54

campioni, 82–84

cancellazione degli esponenziali, regola, 22

chi-quadro (χ^2) distribuzione, 54–55, 56, 204–206

coefficiente,

- confidenza, (v. confidenza, coefficiente)

- correlazione (R), (v. correlazione, coefficiente)

- correlazione multipla, (v. correlazione multipla, coefficiente)

- determinazione (R_s), (v. determinazione, coefficiente)

- regressione parziale, (v. regressione parziale, coefficiente)

concentrazione, matrice, 145

confidenza,

- coefficiente, 92–93

- intervallo, (v. intervallo di confidenza, calcolo)

correlazione, coefficiente (R), 64–65, 70

- regressione, analisi, 78–82

- regressione multipla, analisi, 120

correlazione multipla, coefficiente,

- corretto, 124–126

- problema, 122–123

- regressione multipla, accuratezza equazione, 119–121

costante di Eulero, (v. Eulero, costante)

D

dati, (v. anche dati categorici), 46–47

- grafico, 64–65

- linearmente indipendenti, 47

- qualitativi, 46

- quantitativi, 46

dati categorici, 46

- conversione in dati numerici, 46–47

- in analisi di regressione logistica, 167

- in analisi di regressione multipla, 147–149

dati numerici, 46–47

delta (Δ), 29

densità di probabilità, funzioni, 52–53

- chi-quadro (χ^2), distribuzione, 54–55, 56, 204–206

- F, distribuzioni, 57–59, 206–209

- normale, distribuzione, 53–54

- tavole, 55–56

determinazione, coefficiente (R^2),

- corretto, 124–126

- regressione, analisi, 81–82

- regressione logistica, analisi, 173–177

- regressione multipla, analisi, 119–126

deviazione standard, 51–52

differenziale, calcolo, (v. calcolo differenziale)
dispersione, grafico,
- calcolo differenziale, 26
- analisi di regressione logistica, 169
- analisi di regressione multipla, 113–114
- tracciare i dati, 64–65
- analisi di regressione, 69–70
distribuzione normale, 53–54
distribuzioni *F*, 57–59, 206–209
Durbin-Watson, statistica, 102–103

E
equazione di regressione, (v. regressione, equazione)
equazioni lineari, trasformazioni, 104–106
equazioni strutturali, modello (SEM), 152
errore apparente, tasso, 177
esponenziazione, regola, 22
esponenti, 19–23, 200
- regola, 21
estrapolazione, 102
Eulero, costante, 19, 198–199

F
F (distribuzioni), (v. distribuzioni *F*)
F (test), (v. test *F*)
fattore di inflazione della varianza (VIF), 149
foglio di calcolo, funzioni, 198

funzioni, (v. anche densità di probabilità, funzioni)
- esponenziali, 19–23
- inverse, 14–18
- logaritmo, 19–23
- logaritmo naturale, 20, 200–201
- log-verosimiglianza, 161–163, 171–172
- verosimiglianza, 161–163, 171

G

gradi di libertà, 50–51
grafico, (v. anche dispersione, grafico)
- analisi di regressione logistica, equazione, 159
- dati, 64–65
- funzioni inverse, 17–18

H

H_0 (ipotesi nulla), 48
 H_a (ipotesi alternativa), 48

I

identità, matrice, (v. matrice identità)
interpolazione, 102
intervallo di confidenza, calcolo, 68
- analisi di regressione, 91–94
- analisi di regressione multipla, 133–135, 146
- odds ratio, 194–195
ipotesi,
- alternativa (H_a), 48
- di normalità, 85–86
- nulla (H_0), 48
ipotesi, verifica, 85–90
- analisi di regressione logistica, 178–181

- analisi di regressione multipla, 128–132
- con gli odds, 194

L

logaritmo, 19–23
- naturale, funzione, 20, 200–201
logit, 190–191
log-verosimiglianza, funzione, 161–163, 171–172

M

Mahalanobis, distanza, 133, 137, 144–146
massima verosimiglianza, stima, 162–163
matrici, 37–38
- addizione, 39–40
- colonne, 38
- elementi, 38
- identità, 44
- intervallo di previsione, calcolo, 144–146
- inverse, 44, 202–204
- moltiplicazione, 40–43, 201–202
- righe, 38
media, 49, 72
mediana, 49
minimi quadrati, regressione lineare, 71–76, 115
moltiplicazione
matriciale, (v. matrici, moltiplicazione)
multicollinearità, 149

N

normalità, ipotesi, 85–86
nulla, ipotesi, (H_0), 48
numero di Eulero, (v. Eulero, costante)

O

odds, 190

- test d'ipotesi, 194
- logit, 190–191
- odds ratio, 191–192
- aggiustato, 192–194
- intervallo di confidenza, calcolo, 194–195

P

- Pearson, coefficiente di correlazione, 79 (v. anche correlazione, coefficiente)
- popolazione, 82–84
 - intervallo di confidenza, calcolo, 133–135
 - media, 91
 - regressione, 86
- previsioni,
- regressione, analisi, 95–98
- regressione logistica, analisi, 182
- regressione multipla, analisi, 136–137
- predittore (variabile indipendente), 14, 67
- equazioni strutturali, modello, 152
- influenza relativa, 149–152
- miglior combinazione, scelta, 138–140
- multicollinearità, 149
- regressione logistica, analisi, 164–167
- primo ('), 32
- prodotto, regola, 23
- pseudo- R^2 , 173–177

Q

- quadrati degli scarti, somma, 50
- qualitativi, dati, (v. dati qualitativi)
- quantitativi, dati, (v. dati quantitativi)
- quoziante, regola, 21

R

- R (coefficiente di correlazione), (v. correlazione, coefficiente)
- R^2 (coefficiente di determinazione), (v. determinazione, coefficiente)
- regola degli esponenti, (v. esponenti, regola)
- regola del prodotto, (v. prodotto, regola)
- regola del quoziente, (v. quoziente, regola)
- regressione,
- campionaria, 86
- diagnostica, 119–121
- sottoinsiemi migliori, 139–140
- regressione, analisi,
- analisi della varianza, 87–90
- autocorrelazione, 102–103
- campioni e popolazione, 82–84
- confidenza, calcolo intervalli, 91–94
- correlazione, calcolo coefficiente, 78–82
- dispersione, grafico, 69–70
- equazione, 66–67
- equazione, calcolo, 71–77
- interpolazione ed estrapolazione, 102
- ipotesi di normalità, 85–86
- previsione, calcolo intervalli, 95–98
- procedimento generale, 68, 100
- regressione non lineare, 103–104
- residuo standardizzato, 100–101
- regressione, equazione, 66–67
- calcolo, 71–77
- trasformazioni, 104–106
- regressione lineare,
- analisi, 7
- minimi quadrati, (v. minimi quadrati, regressione lineare)
- regressione logistica, analisi, 8, 157
- accuratezza dell'equazione, valutazione, 173–177
- binomiale, 157
- equazione, calcolo, 158–159, 170–173
- grafico di dispersione, 169
- massima verosimiglianza, metodo, 159–163, 210–212
- odds ratio, 192–194
- odds ratio aggiustato, 192–194
- odds ratio, intervalli di confidenza, 194–195
- previsione con, 182
- procedura, 168, 190
- rischio relativo, 195–196
- test d'ipotesi, 178–181, 194
- variabili predittive, scelta, 164–167
- regressione multipla, analisi, 7–8, 111
- accuratezza dell'equazione, valutazione, 119–126
- dati categorici, uso, 147–149
- dispersione, grafico, 113–114

- equazione, calcolo, 115–119
 - intervalli di confidenza, calcolo, 133–135
 - intervallo di previsione, calcolo, 136–137
 - ipotesi, verifica, 127
 - Mahalanobis, distanza, 144–146
 - multicollinearità, 149
 - procedimento, 112, 142
 - residui standardizzati, 143–144
 - variabili predittive, scelta, 138–140
 - variabili predittive, determinazione influenza, 149–152
 - varianza, analisi, 128–132
 - regressione non lineare, 103–106
 - regressione parziale, coefficiente, 116–118
 - fogli elettronici, 209–210
 - ipotesi, verifica, 127, 129–131
 - residui, 71
 - quadrati, somma, 73–74
 - standardizzati, 100–101, 143–144
 - rischio relativo, 195–196
 - risultato statisticamente significativo, 58
 - round-robin, metodo, 139–140
- S**
- scarti, somma dei quadrati, 50
 - SEM (modello di equazioni strutturali), 152
- sottoinsiemi migliori, regressione, 139–140
 - spazio degli eventi, 53
 - statistica,
 - dati, tipi, 46–47
 - ipotesi, verifica, 48
 - variabilità, misura, 49–52
 - statisticamente significativo, risultato, 58
- T**
- test d'ipotesi, (v. ipotesi, verifica)
 - test F, 129–133
 - tolleranza, 149
- V**
- valore critico, 55
 - variabile responso, (v. responso, variabile)
 - variabili dipendenti, 14, 67, 149–152. (v. anche dispersione, grafico)
 - variabili indipendenti, (v. predittore)
 - varianza, 50–51
 - campionaria corretta, 50
 - varianza, analisi (ANOVA), 87–90, (v. anche ipotesi, verifica di)
 - regressione, 87–90
 - regressione logistica, 178–180
 - regressione multipla, 128–132
- verifica d'ipotesi, (v. ipotesi, verifica)
 - verosimiglianza,
 - funzioni di, 161–163, 171
 - test del rapporto, 179
 - VIF (fattore di inflazione della varianza), 149
- W**
- Wald, test, 180
- X**
- x barrato, 72
- Y**
- y cappello, 73

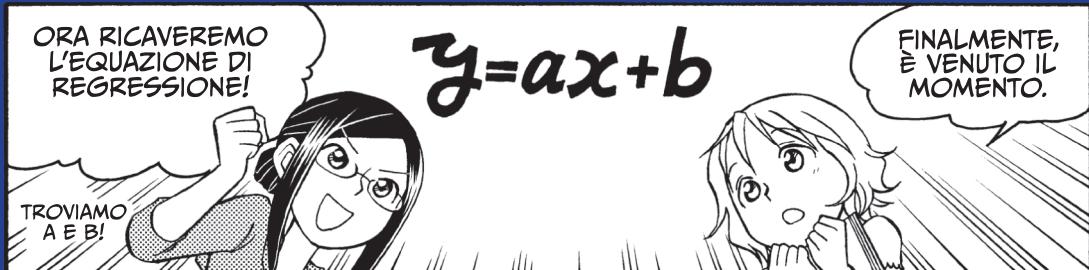


L'AUTORE

Shin Takahashi è nato nel 1972 a Niigata. Si è laureato al Kyushu Institute of Design (oggi Kyushu University). Ha lavorato come analista ed è ora autore di testi tecnici.

Homepage: <http://www.takahashishin.jp/>

UN'AFFASCINANTE GUIDA ALLA REGRESSIONE. A FUMETTI!



MIU HA SEMPRE AVUTO PROBLEMI CON LA REGRESSIONE STATISTICA MA ORA, FINALMENTE, CON UN PO' DI MOTIVAZIONE IN PIÙ E CON L'AUTO DELL'INTELLIGENTISSIMA RISA, LA SUA BRILLANTE COMPAGNA DI LAVORO IN CAFFETTERIA, È DECISA A VENIRNE A CAPO E AFFRONTARE QUESTA MATERIA.

NE **I MANGA DELLE SCIENZE - REGRESSIONE** SEGUIRETE INSIEME A NOI LE AVVENTURE DI MIU E RISA, MENTRE CALCOLANO L'EFFETTO DELLE MASSIME GIORNALIERE SUGLI ORDINI DI TÈ FREDDO, PREVEDONO GLI INCASSI DELLA PASTICCERIA E CALCOLANO LA PROBABILITÀ DELLE VENDITE DEI DOLCI CON L'AUTO DELLA REGRESSIONE STATISTICA SEMPLICE, MULTIPLA E LOGISTICA. DOPO UN RAPIDO RIPASSO DI CONCETTI DI BASE COME EQUAZIONI MATRICIALI, FUNZIONI INVERSE, LOGARITMI E DERIVATE, SARETE PRONTI A TUFFARVI IN ACQUE PIÙ PROFONDE, DOVE IMPARERETE A:

- » CALCOLARE L'EQUAZIONE DI REGRESSIONE;
- » VERIFICARE LA PRECISIONE DELLA VOstra EQUAZIONE USANDO IL COEFFICIENTE DI CORRELAZIONE;
- » ESEGUIRE VERIFCHE DI IPOTESI, ANALISI DELLA VARIANZA E CALCOLARE GLI INTERVALLI DI CONFIDENZA;
- » ESEGUIRE PREVISIONI USANDO GLI "ODDS RATIO" E GLI INTERVALLI DI PREVISIONE;
- » VERIFICARE LA VALIDITÀ DELL'ANALISI FACENDO USO DI TEST DIAGNOSTICI;
- » ESEGUIRE TEST CHI-QUADRO E TEST F PER VERIFICARE LA SIGNIFICATIVITÀ DELLA REGRESSIONE.

SIA CHE VI STIATE AVVICINANDO PER LA PRIMA VOLTA ALL'ANALISI DI REGRESSIONE SIA CHE CI ABBIATE PROVATO IN PASSATO MA NON SIETE SODDISFATTI, CON **I MANGA DELLE SCIENZE - REGRESSIONE** ARRIVERETE A PADRONEGGIARE QUESTA DISCIPLINA IN MODO IMMEDIATO E DIVERTENTE.



la Repubblica Le Scienze

600011
Pubblicazione settimanale da vendersi esclusivamente
in abbinamento a la Repubblica oppure a Le Scienze.
Supplemento al numero in edicola.

9,90 euro + il prezzo di Repubblica oppure di Le Scienze.