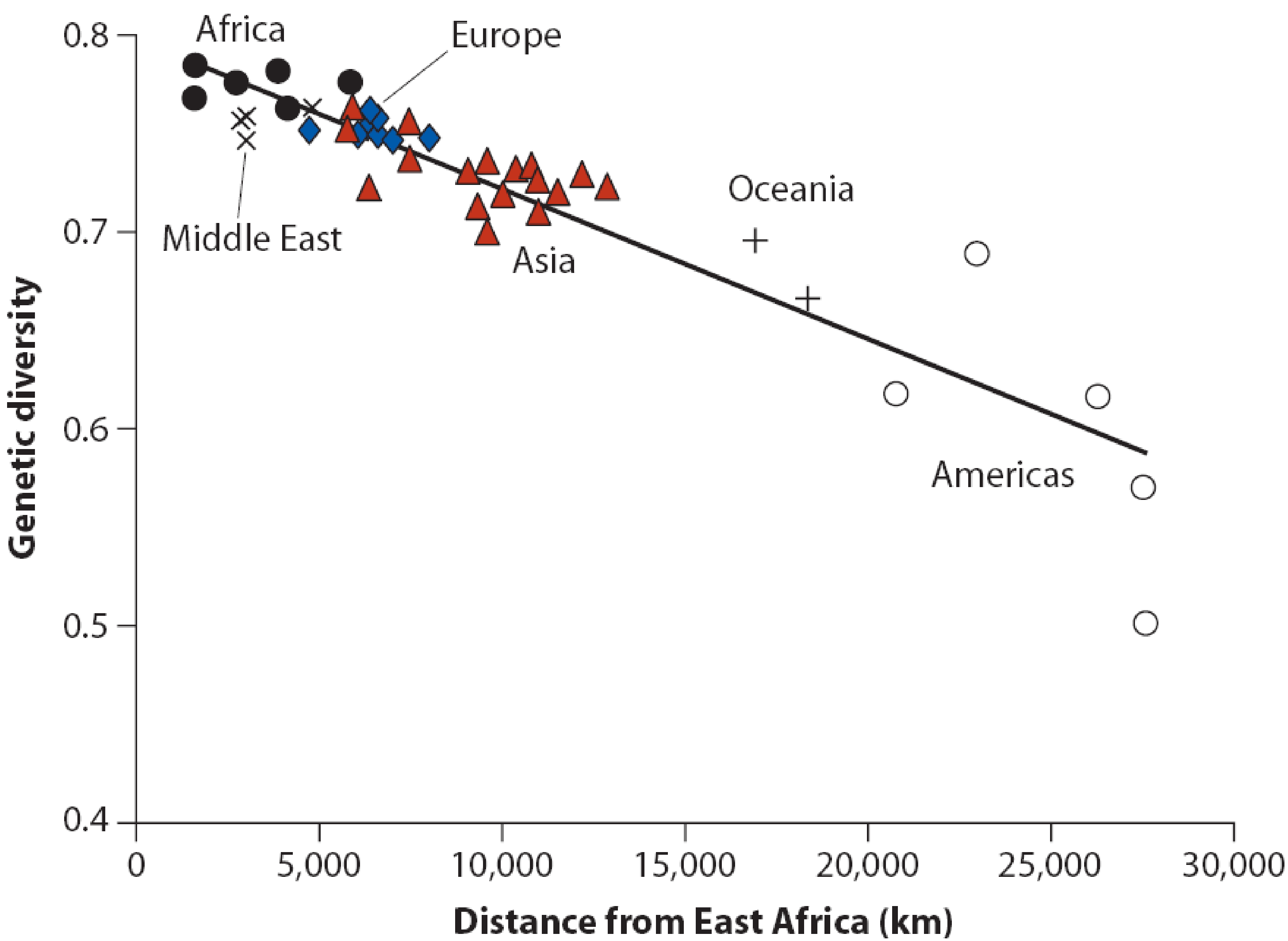
La regressione

Capitolo 17

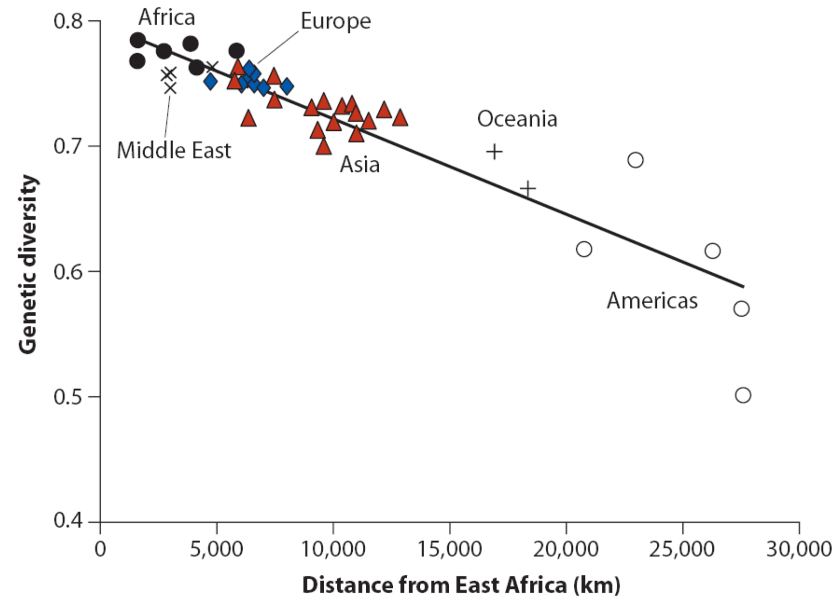
Analisi statistica dei dati biologici

E’ possibile prevedere il valore di una variabile?



La regressione lineare

La **regressione** è un metodo usato per prevedere il valore di una variabile numerica (variabile risposta) in base a quello di un'altra variabile numerica (variabile esplicativa).



Consiste nel trovare la retta di regressione, cioè quella retta che passa attraverso i punti del diagramma a dispersione e che descrive la relazione lineare tra le due variabili.

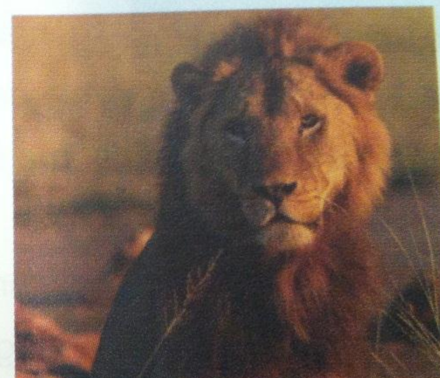
L'equazione della retta ci permette di:

- Predire il valore di Y (variabile risposta)
- Trovare il tasso di variazione di Y in funzione di X

Esempio

Esempio 17.1

Il naso del leone



La gestione della caccia al leone praticata in Africa è importante per la conservazione di questa specie. Un aspetto rilevante è la conoscenza dell'età degli esemplari maschi, dato che l'uccisione di maschi con più di 6 anni ha un impatto minore sull'organizzazione sociale rispetto alla perdita di individui più giovani. Whitman et al. (2004) hanno mostrato che la quantità di pigmentazione nera sul naso dei leoni maschi aumenta all'aumentare dell'età e quindi potrebbe essere usata per stimarne l'età. La relazione fra età e proporzione di pigmentazione nera sul naso di 323 leoni maschi di età nota in Tanzania è rappresentata nello scatter plot in Figura 17.1-1. I dati grezzi sono riportati

Questi dati possono essere utilizzati per stimare l'età di un leone in base alla proporzione di nero sul suo naso?

I dati grezzi

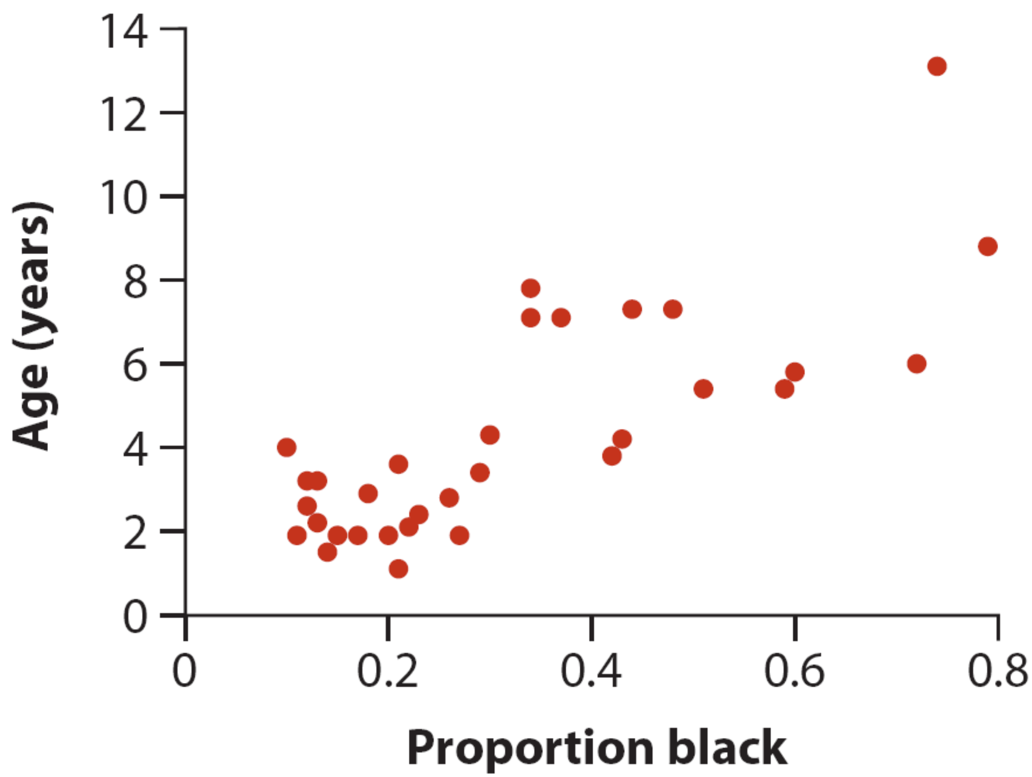
La proporzione di nero sul naso di 32 maschi di età nota

Età	Proporzione di nero
1.1	0.21
1.5	0.14
1.9	0.11
2.2	0.13
2.6	0.12
3.2	0.13
3.2	0.12
2.9	0.18
2.4	0.23
2.1	0.22
1.9	0.2
1.9	0.17
1.9	0.15
1.9	0.27
2.8	0.26
3.6	0.21

Età	Proporzione di nero
4.3	0.3
3.8	0.42
4.2	0.43
5.4	0.59
5.8	0.6
6	0.72
3.4	0.29
4	0.1
7.3	0.48
7.3	0.44
7.8	0.34
7.1	0.37
7.1	0.34
13.1	0.74
8.8	0.79
5.4	0.51

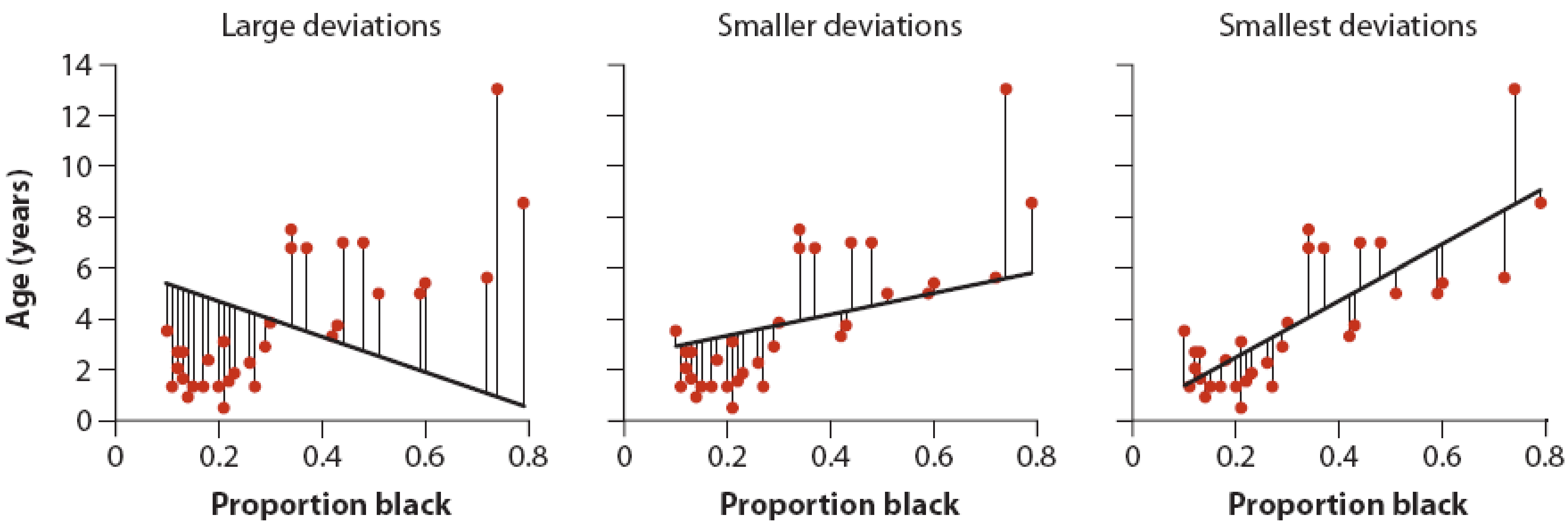
Diagramma a dispersione

L'età è la variabile risposta mentre la proporzione di nero è la variabile esplicativa. Infatti, siamo interessati a prevedere l'età in base alla proporzione di nero (e non il contrario!)

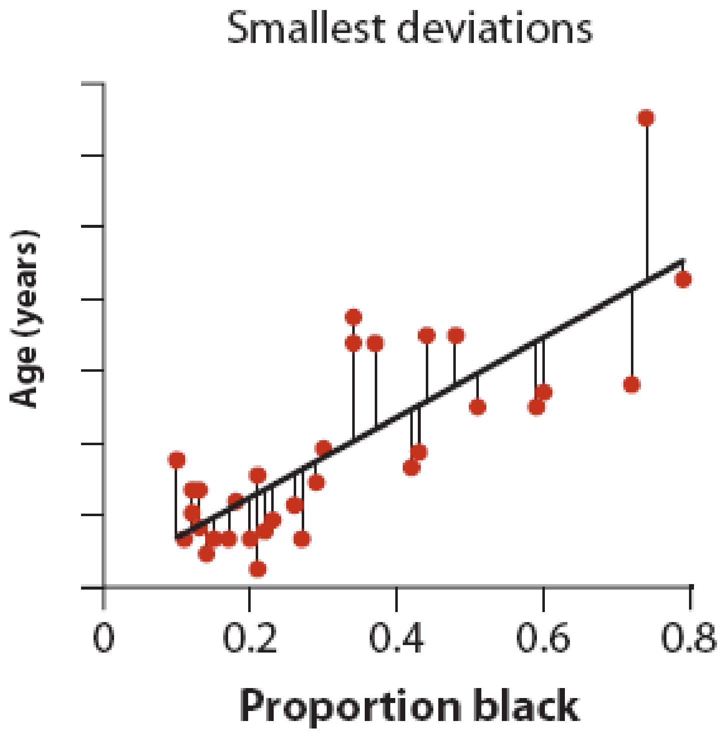


Trovare la retta di regressione: il metodo dei minimi quadrati

In un diagramma a dispersione possiamo tracciare molte rette: come facciamo a scegliere la «migliore»?



Minimi quadrati



Trovare la retta che renda più piccole possibili le distanze in Y (scarti verticali) tra i punti dei dati e la retta di regressione.

↓

Trovare la retta per la quale sia **minima la somma di tutti i quadrati** di queste distanze.

L’equazione della retta

Una retta di regressione è descritta matematicamente dalla seguente equazione:

$$Y = a + bX$$

Y è la **variabile risposta** (asse delle ordinate)

X è la **variabile esplicativa** (asse delle ascisse)

a è l’**intercetta** sull’asse Y

b è la **pendenza** della retta di regressione

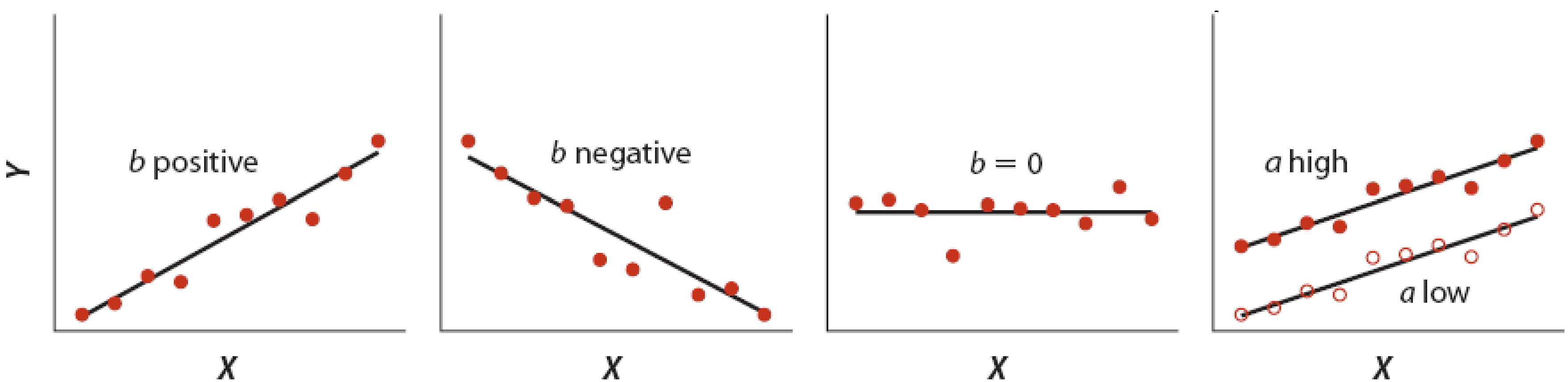
a e b

a

- l'intercetta sull'asse delle ordinate.
- Matematicamente è **il valore di Y quando X è uguale a 0**.
- La sua unità di misura è uguale a quella della variabile Y.

b

- Pendenza della retta di regressione.
- **Misura la variazione di Y quando X varia di una unità**.
- L'unità di misura della pendenza è il rapporto tra l'unità di misura di Y e quella di X.
- Se b è positivo, valori di X maggiori prevedono Y maggiori.
- Se b è negativo, valori di X



Calcolo di pendenza e intercetta della retta di regressione

La pendenza della retta di regressione dei minimi quadrati si calcola come:

$$b = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

X e Y sono le misure delle due variabili negli individui

$\bar{X}$ e $\bar{Y}$ sono le medie campionarie delle due variabili

Calcolo di pendenza e intercetta della retta di regressione

Una volta calcolato b, trovare l'intercetta (a) è semplice, perché la retta deve passare per il punto $(\bar{X}, \bar{Y})$. Quindi:

$$\bar{Y} = a + b\bar{X}$$

da cui otteniamo:

$$a = \bar{Y} - b\bar{X}$$

La retta di regressione nell'esempio dei leoni

Nell'esempio avremo:

$$\bar{X} = 0{,}3222 \qquad \bar{Y} = 4{,}3094$$

$$\sum (X - \bar{X})^2 = 1{,}2221 \qquad \sum (Y - \bar{Y})^2 = 222{,}0872$$

$$\sum (X - \bar{X})(Y - \bar{Y}) = 13{,}0123$$

la pendenza è data da:

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{13{,}0123}{1{,}2221} = 10{,}647$$

La pendenza b stima di quanto varia l'età dei leoni maschi quando la proporzione di pigmentazione nera sul naso varia di un'unità (anni per unità di proporzione).

Calcolo di pendenza e intercetta della retta di regressione

L'intercetta, misurata in anni, è:

$$a = \bar{Y} - b\bar{X} = 4{,}3094 - 10{,}647(0{,}3222) = 0{,}879$$

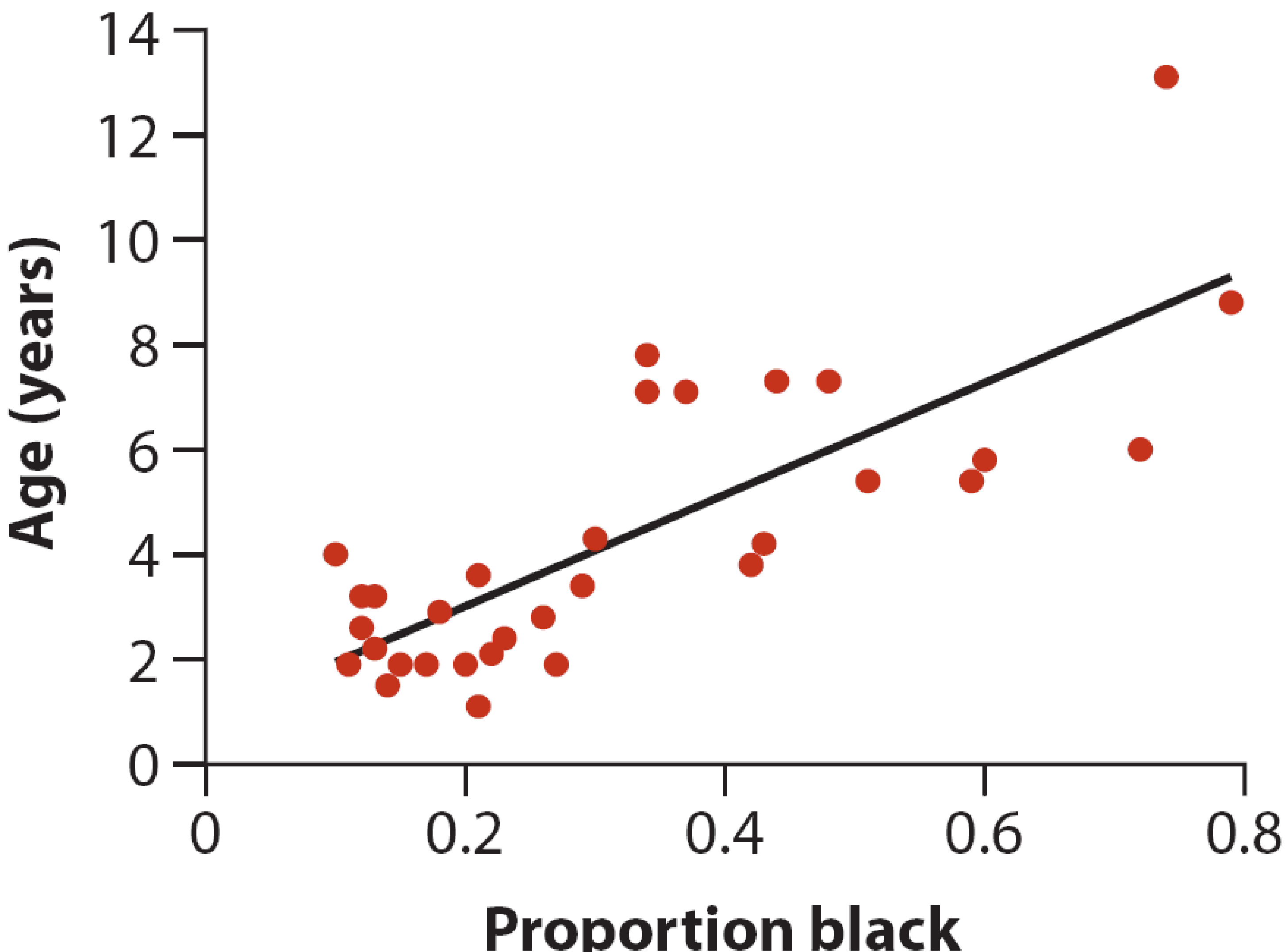
L'equazione della retta di regressione diventa perciò:

$$Y = 0{,}88 + 10{,}65X$$

e può essere scritta anche nella formula:

$$et\text{à} = 0{,}88 + 10{,}65(proporzione\ di\ nero)$$

La retta nel diagramma a dispersione



Risoluzione esempio con R (I)

```
#caricare i dati dell'esempio presenti nel file «regressione_esercizio_loni.txt»
dati<-read.table(choose.files(),header=T)
#assegno la variabile esplicativa a x e la variabile risposta a y
x<-dati$black_proportion
y<-dati$age
#visualizzo il diagramma a dispersione
matplot(x,y,pch=16,col="red")
#calcolo Xmedio e Ymedio
xmed<-mean(x)
ymed<-mean(y)
#calcolo la pendenza
bnum<-sum((x-xmed)*(y-ymed))
bden<-sum((x-xmed)^2)
b<-bnum/bden
#calcolo l'intercetta all'asse y
a<-ymed-b*xmed
```

Risoluzione esempio con R (II)

#utilizzo la funzione lm() per effettuare la regressione

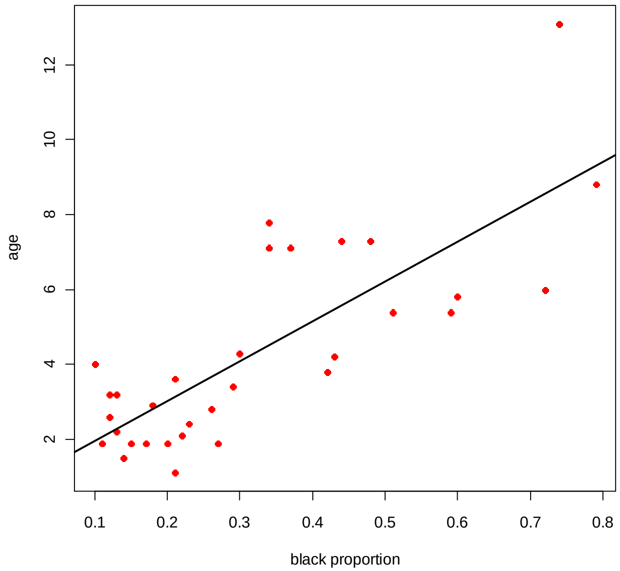
lmfit<-lm(age~black_proportion,data=dati)

#visualizzo il diagramma a dispersione

matplot(dati$black_proportion,dati$age,pch=16,col="red",ylab= "age", xlab="black proportion")

#aggiungo al grafico la retta di regressione

abline(lmfit,lwd=2)



Popolazioni e campioni

La retta di regressione serve per stimare la regressione «vera» di Y su X nella popolazione.

L'equazione della retta di regressione nella popolazione è:

$$Y = \alpha + \beta X$$

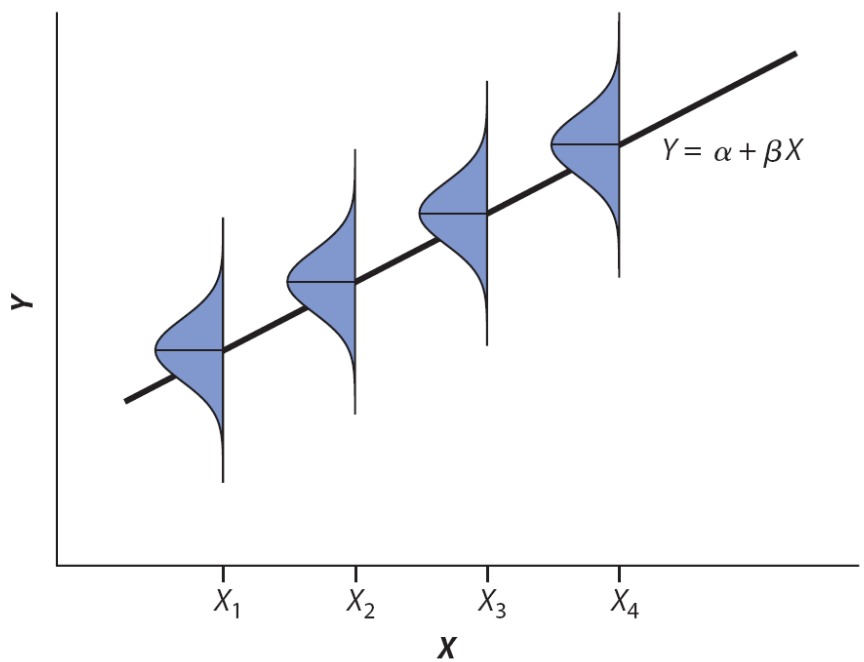
dove β è la pendenza nella popolazione e α è l'intercetta.

Le grandezze α e β sono parametri della popolazione, mentre a e b sono le loro stime campionarie.

Popolazioni e campioni

Nella regressione assumiamo che <u>esista una popolazione di possibili valori di Y</u> per ogni valore di X.

Il valore medio di Y per ogni valore di X giace sulla retta di regressione «vera».

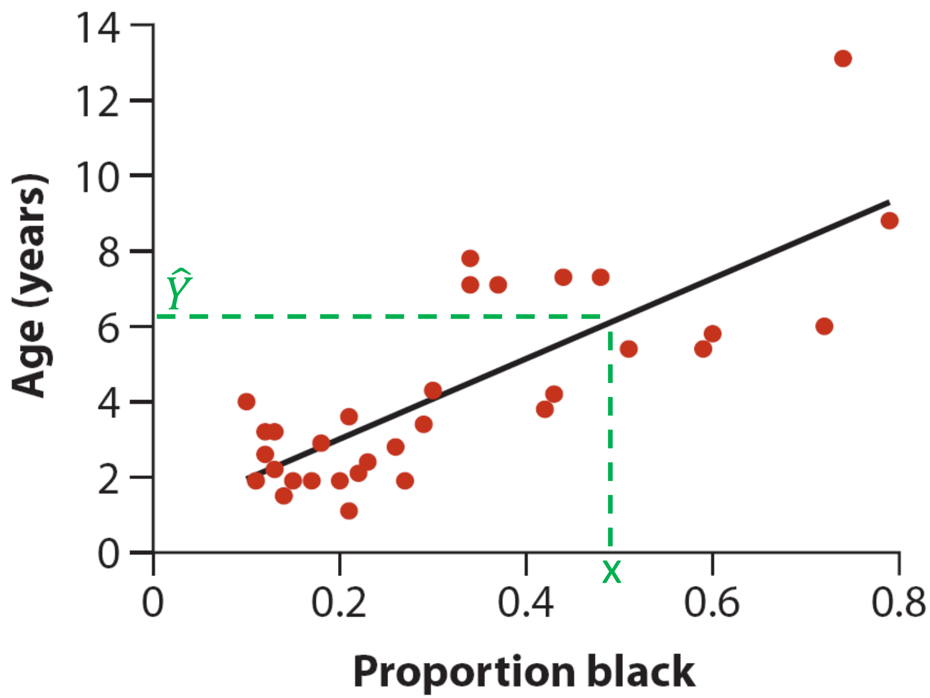


(la retta di regressione «vera» congiunge i valori medi di Y per ogni valore di X)

I valori previsti

Una volta in possesso della retta di regressione, si possono determinare i punti della retta che corrispondono a valori specificati di X.

Questi punti sono detti **previsioni** (o valori di Y previsti) e sono indicati con $\hat{Y}$.



I valori previsti

Il valore previsto di Y per un dato valore di X <u>stima la media di Y per l’intera popolazione di individui che hanno quel valore di X</u>.

Nell’esempio dei leoni, per prevedere l’età di un leone maschio corrispondente a un proporzione di nero di 0,50 avremo:

$$\hat{Y} = a + b(0{,}5) = 0{,}88 + 10{,}65(0{,}5) = 6{,}2$$

cioè che i leoni con una proporzione di nero del 50% abbiano un’età media di 6,2 anni.

<u>Attenzione: le previsioni sono affidabili solo per valori di X che cadono nell’intervallo dei valori osservati</u>

$\hat{Y}$ previsti con R

```
#caricare i dati dell’esempio presenti nel file «regressione_esercizio_loni.txt»
dati<-read.table(choose.files(),header=T)
#calcoliamo la retta di regressione
lmfit<-lm(age~black_proportion,data=dati)
#creiamo un vettore di valori di X per cui effettuare una previsione
x.new<-c(0.1,0.2,0.3,0.7)
#calcoliamo la previsione (primo modo) (y=a+bx)
y.hat<-lmfit$coefficients[1]+lmfit$coefficients[2]*x.new
#calcoliamo la previsione (secondo modo)
y.hat1<-predict(lmfit,newdata=data.frame(black_proportion=x.new))
#visualizziamo i punti
matplot(dati$black_proportion,dati$age,pch=16,col="red",ylab="age", xlab="black
proportion")
abline(lmfit,lwd=2)
points(x.new,y.hat,pch=16,col="green")
```

Esercizio sulle previsioni con R

```
#generare 50 osservazioni random da una distribuzione normale con media 5 e sd 2
x<-rnorm(50,5,2)
#scegliere  un valore per l'intercetta in Y
a.reale<-2
#scegliere un valore per la pendenza
b.reale<-4
#generare delle simulazioni da un modello lineare y=a+bX+err
y<-a.reale+b.reale*x+rnorm(50,5,5)
#visualizziamo il diagramma a dispersione
matplot(x,y,pch=16,col="red")
```

Esercizio sulle previsioni con R

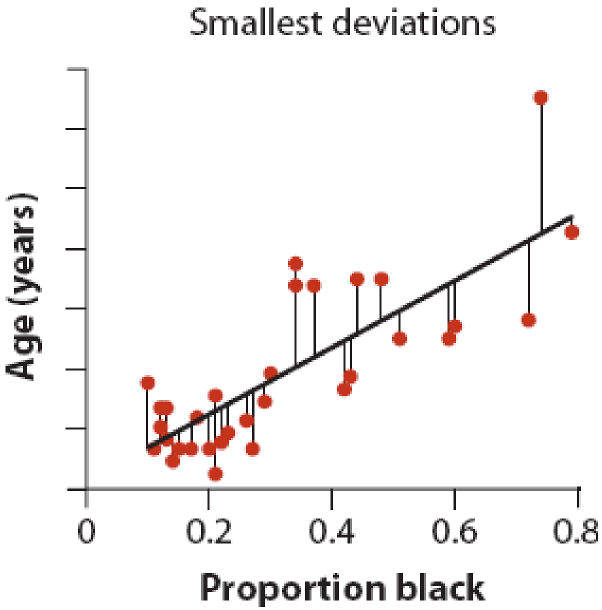
- Stimare i coefficienti della retta di regressione che descrive la relazione tra Y e X e confrontarli con i valori reali usati per generare i dati.

```
#calcolo della retta di regressione
fit<-lm(y~x)
#coefficienti stimati
fit
#grafico a dispersione con i punti generati
matplot(x,y,pch=16,col="red")
#aggiungo la retta di regressione stimata
abline(fit,col="green")
#aggiungo la retta con cui sono stati generati i dati originali
abline(a=a.reale,b=b.reale,col="black")
```

I residui

- Qual è il livello di adattamento della retta di regressione ai dati?

I **residui** misurano la dispersione dei punti al di sopra e al di sotto della retta di regressione.



Distanza di ogni osservazione Y dal suo valore previsto $\hat{Y}$.

Ad esempio, il 31-leone nel campione dell'esempio ha una proporzione di nero X=0,79. La corrispondente **età prevista** è:

$$\hat{Y} = 0{,}88 + 10{,}65(0{,}79) = 9{,}3$$

Il residuo è il valore osservato meno il valore previsto:

$$residuo = \left(Y - \hat{Y}\right) = (8{,}8 - 9{,}3) = -0{,}5$$

I residui

La **varianza dei residui** quantifica la dispersione dei punti al di sopra e al di sotto della retta di regressione.

$$MS_{residua} = \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2}$$

I residui

Una formula analoga, ma più rapida da calcolare:

$$MS_{residua} = \frac{\sum(Y_i - \bar{Y})^2 - b\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-2}$$

Applicata all'esempio dei leoni:

$$MS_{residua} = \frac{222{,}0872 - 10{,}647(13{,}0123)}{32-2} = 2{,}785$$

Errore standard della pendenza

Come per qualsiasi altra stima, alla stima campionaria di b della pendenza nella popolazione β, è associata un'incertezza.

L'incertezza è misurata dall'errore standard (deviazione standard della distribuzione campionaria di b)

Se sono soddisfatte le assunzioni della regressione lineare (che vedremo più avanti), allora la distribuzione campionaria di b è una distribuzione normale con media β ed errore standard pari a:

$$ES_b = \sqrt{\frac{MS_{residua}}{\sum(X_i - \bar{X})^2}}$$

Errore standard della pendenza

Per l'esempio dei leoni, l'errore standard associato alla stima di b è:

$$ES_b = \sqrt{\frac{MS_{residua}}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{2,785}{1,221}} = 1,510$$

Esercizio: calcolare l'errore standard di b con R

```
#calcolo la devianza di Y e X
devy<-sum((dati$age-mean(dati$age))^2)
devx<-sum((dati$black_proportion-mean(dati$black_proportion))^2)
#calcolo la somma dei prodotti
sp<-sum((dati$age-mean(dati$age))* (dati$black_proportion-mean(dati$black_proportion)))
#calcolo la pendenza della retta di regressione
b<-lm(age~black_proportion,data=dati)$coefficients[2]
#calcolo la varianza dei residui (MSresidua)
msr<-(devy-b*sp)/(length(dati$age)-2)
#calcolo l'errore standard di b
ESb<-sqrt(msr/devx)
```

Intervallo di confidenza della pendenza

- L'intervallo di confidenza del parametro B è dato da:

$$b - t_{\alpha/2,df} ES_b < \beta < b + t_{\alpha/2,df} ES_b$$

dove t è il valore critico a due code della distribuzione t con df=n-2 gradi di libertà

Per l'esempio dei leoni, t0.05,30=2,042:

$$10{,}647 - 2{,}042(1{,}510) < \beta < 10{,}647 + 2{,}042(1{,}510)$$

$$7{,}56 < \beta < 13{,}73$$

L'età media dei leoni aumenta tra 7,6 e 13,7 anni quando si ha un aumento unitario di pigmentazione (nb: 1u=100%)

Qualità delle previsioni

La retta di regressione contiene un errore associato alla stima dei parametri α e β a partire da un campione.

Perciò è importante conoscere il grado di precisione delle previsioni effettuate utilizzando la retta di regressione.

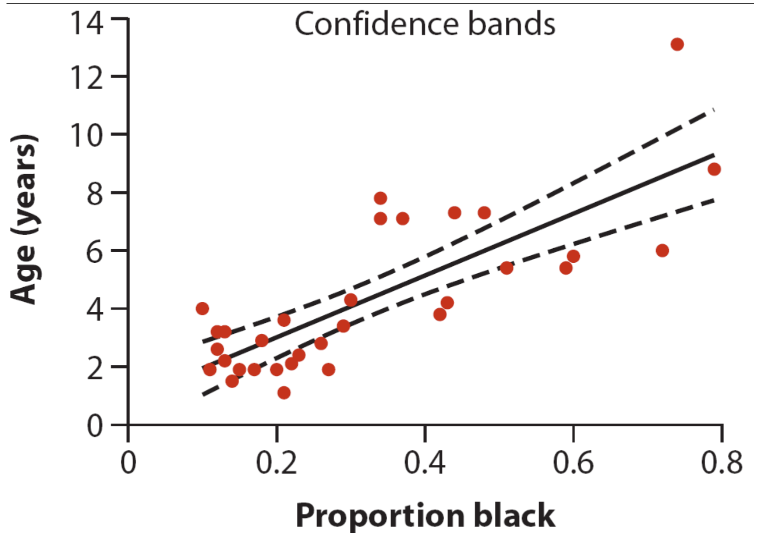
Esistono due tipi di intervalli di confidenza delle previsioni: **bande di confidenza** e **intervalli di previsione**

Bande di confidenza

Tipo di previsione: valore di $\hat{Y}$ medio nella popolazione per un certo valore di X.

Nell'esempio: età media di tutti i leoni maschi nella popolazione che hanno il naso nero al 60%.

Le bande di confidenza (per un certo valore di α) rappresentano gli intervalli di confidenza della media prevista di Y nell'intervallo di variazione di X.

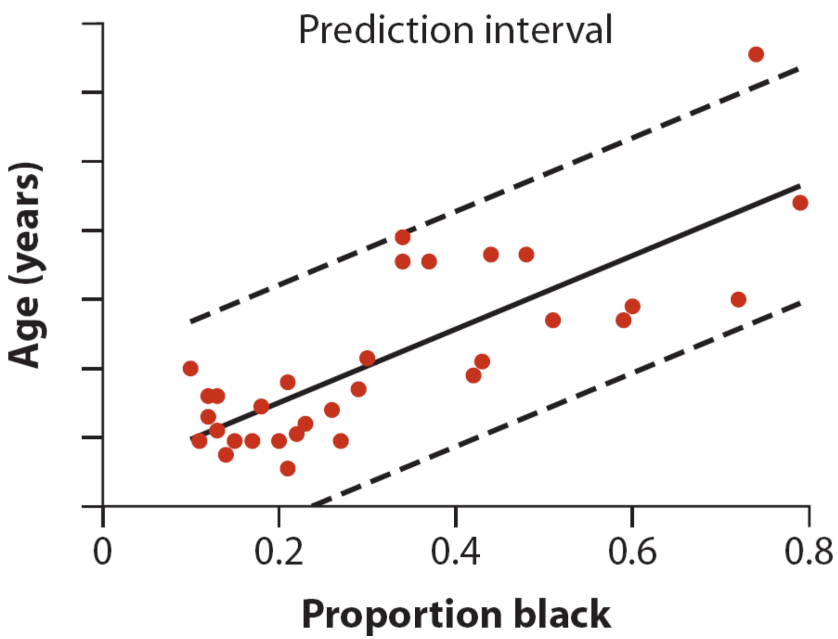


Intervalli di previsione

Tipo di previsione: valore di $\hat{Y}$ per un singolo valore di X.

Nell'esempio: età di uno specifico leone il cui naso è nero al 60%.

Gli intervalli di previsione (per un certo valore di α) rappresentano la precisione delle previsioni ottenibili per singoli valori X nell'intervallo di variazione.



Calcolo di bande di confidenza con R

```
#per calcolare le bande di confidenza si utilizza la funzione «predict.lm» di R
#assegno x e y
x<-dati$black_proportion
y<-dati$age
#calcolo i coefficienti della retta di regressione
lmfit<-lm(y~x)
#genero un intervallo di valori di x in cui predire y
xnew<-data.frame(x=seq(from=min(x),to=max(x),length.out=10))
#ottengo la predizione delle bande di confidenza al 95%
bc<-predict.lm(lmfit, xnew, interval="confidence", level=0.95)
#creo il diagramma a dispersione
matplot(x,y,pch=16,col="red")
#aggiungo la retta di regressione
abline(lmfit,lwd=2)
#aggiungo le bande di confidenza
matplot(xnew,bc[,2],type="l", lty="dashed", add=T)
matplot(xnew,bc[,3],type="l", lty="dashed", add=T)
```

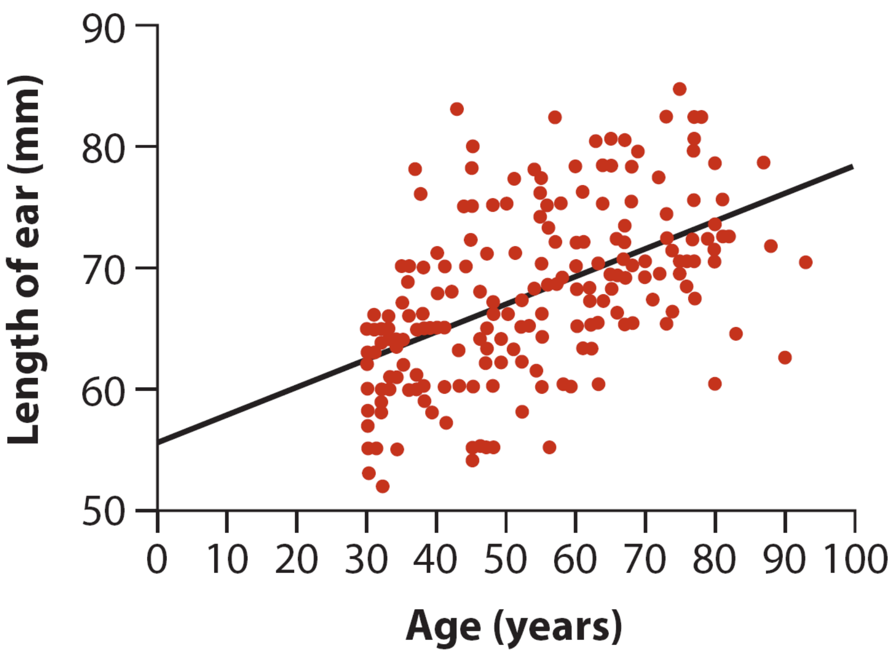
Calcolo degli intervalli di previsione con R

```
#per calcolare gli intervalli di previsione si utilizza ancora la funzione «predict.lm» di R
#assegno x e y
x<-dati$black_proportion
y<-dati$age
#calcolo i coefficienti della retta di regressione
lmfit<-lm(y~x)
#genero un intervallo di valori di x in cui predire y
xnew<-data.frame(x=seq(from=min(x),to=max(x),length.out=10))
#ottengo la predizione degli intervalli di previsione al 95%
ip<-predict.lm(lmfit, xnew, interval="prediction", level=0.95)
#creo il diagramma a dispersione
matplot(x,y,pch=16,col="red")
#aggiungo la retta di regressione
abline(lmfit,lwd=2)
#aggiungo le bande di confidenza
matplot(xnew, ip[,2],type="l", lty="dashed", add=T)
matplot(xnew, ip[,3],type="l", lty="dashed", add=T)
```

Estrapolazione

Le previsioni effettuate sulla base di una retta di regressione sono affidabili **SOLO** all'interno dell'intervallo di variazione di X.

L'**estrapolazione** è la previsione del valore di una variabile risposta all'esterno dell'intervallo di valori di X osservati nei dati.



lunghezza orecchio=55,9+0,22(età)

Attenzione: alla nascita un bambino dovrebbe avere un orecchio di 5,6 cm!

Le relazioni possono non essere lineari in tutto l'intervallo di valori in X.

Verifica di ipotesi sulla pendenza

Nella regressione, la verifica delle ipotesi viene usata per valutare se la pendenza nella popolazione sia uguale a un valore specificato dall'ipotesi nulla, β_0 (generalmente, ma non sempre, =0).

H_0: $\beta = \beta_0 = 0$
H_A: $\beta \neq \beta_0 = 0$

Statistica test

La statistica test t è:

$$t = \frac{b - \beta_0}{ES_b}$$

dove:

b è la stima della pendenza nel campione

ES_b è l'errore standard di b

Se è vera l'ipotesi nulla t segue una distribuzione t di student con n-2 gradi di libertà.

Esempio cince

Esempio 17.3

I richiami d'allarme delle cince

Gli animali variano i propri richiami sonori a seconda del contesto, e questo fatto suggerisce che i richiami trasmettano informazioni. Le cince bige americane, blackcapped chickadee in inglese, emettono il richiamo d'allarme «chick-a-dee-dee-dee» quando incontrano un predatore che non vola (ma emettono il suono «seet» completamente diverso quando il predatore vola). Templeton et al. (2005) hanno posto esemplari vivi, appollaiati, di 13 specie di uccelli predatori con masse corporee diverse nelle vicinanze di stormi di cince, e hanno registrato i richiami d'allarme delle cince stesse. Le 13 specie di uccelli predatori avevano dimensioni che andavano da quelle di piccoli e agili falchi e gufi, la cui dieta comprende spesso piccoli uccelli, a quelle di predatori grandi e meno agili, la cui dieta comprende pochi piccoli uccelli (Tabella 17.3-1). Il numero medio di suoni «dee» emessi durante ogni richiamo dalle cince minacciate dal predatore è riportato in Tabella 17.3-1. Un diagramma di dispersione di questi dati è presentato in Figura 17.3-1. Assumendo che le misure ottenute con differenti predatori siano indipendenti, il peso del predatore ci permette di prevedere il numero medio di suoni «dee» nei richiami d'allarme? ■

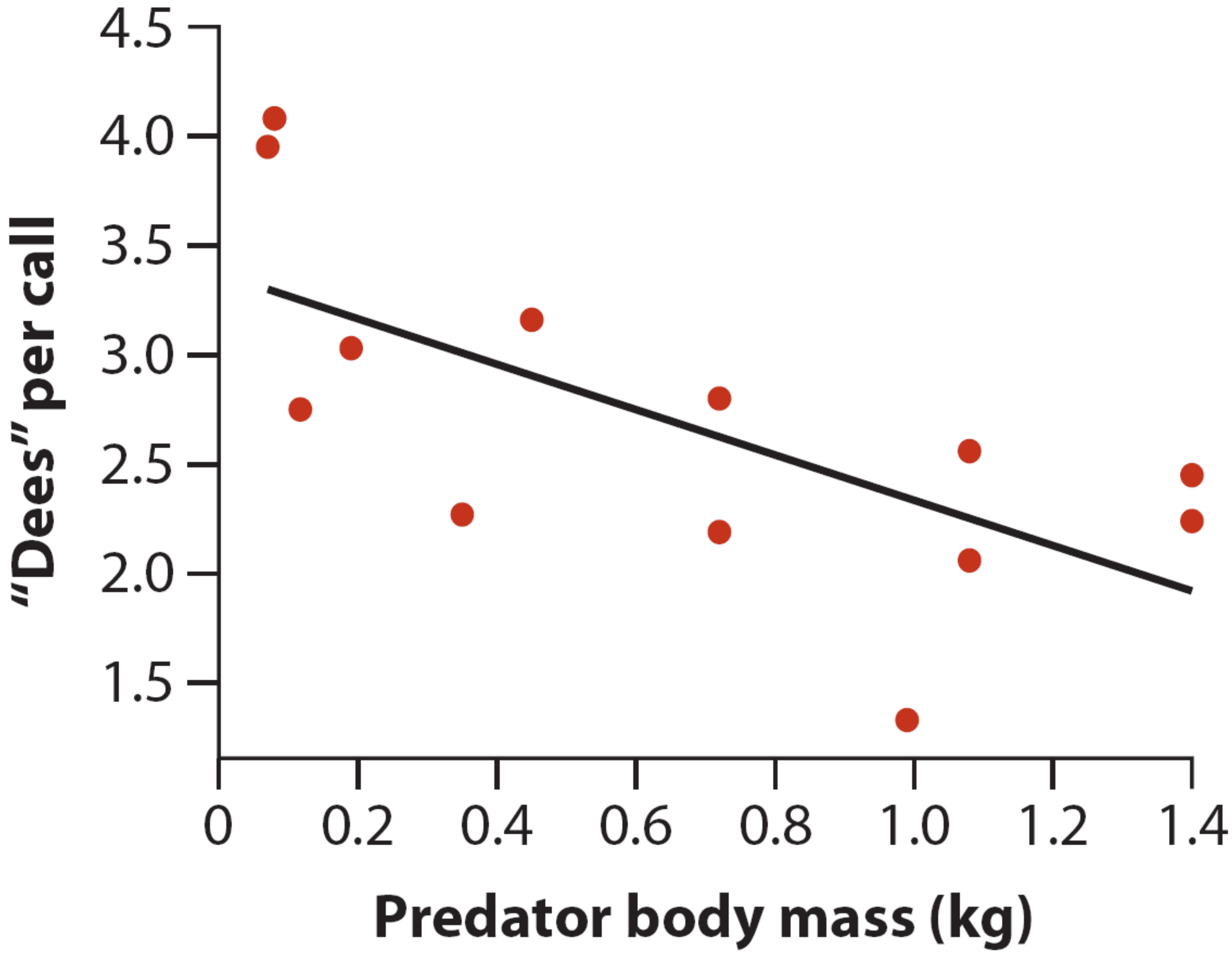
I dati grezzi

Tabella 17.3-1

Numero medio di suoni «dee» per ogni richiamo d'allarme emesso dalle cince Poecile atricapillus in presenza di un predatore vivo appollaiato.

Specie di uccello predatore	Peso del predatore (kg)	Numero di suoni «dee» per ogni richiamo d'allarme
Glaucidium californicum (civetta nana della California)	0,07	3,95
Aegolius acadicus (civetta acadica)	0,08	4,08
Falco sparverius (gheppio americano)	0,12	2,75
Falco columbarius (smeriglio)	0,19	3,03
Asio flammeus (gufo di palude)	0,35	2,27
Accipiter cooperii (sparviere di Cooper)	0,45	3,16
Falco mexicanus (falco delle praterie)	0,72	2,19
Falco peregrinus (falco pellegrino)	0,72	2,80
Bubo virginianus (gufo della Virginia)	1,40	2,45
Buteo lagopus (poiana calzata)	0,99	1,33
Falco rusticolus (girifalco)	1,40	2,24
Buteo jamaicensis (falco coda rossa)	1,08	2,56
Strix nebulosa		

Il diagramma a dispersione per l'esempio delle cince



Il test t sulla pendenza della regressione lineare

Ipotesi nulla ed alternativa:

In questo esempio l'ipotesi nulla è che il numero medio di suoni «dee» in un richiamo d'allarme non possa essere previsto sulla base del peso del predatore (pendenza = 0), perciò:

H0: la pendenza della retta di regressione tra numero di suoni «dee» in ogni richiamo e il peso del predatore è uguale a 0 ($\beta = 0$)

HA: La pendenza della retta di regressione tra numero di suoni «dee» in ogni richiamo e il peso del predatore è diversa da 0 ($\beta \neq 0$)

Calcolo della retta di regressione

Nell'esempio avremo:

$$\bar{X} = 0{,}6654 \qquad \bar{Y} = 2{,}6823$$

$$\sum (X - \bar{X})^2 = 2{,}9009 \qquad \sum (Y - \bar{Y})^2 = 6{,}8210$$

$$\sum (X - \bar{X})(Y - \bar{Y}) = -3{,}0118$$

la pendenza stimata è data da:

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{-3{,}0118}{2{,}9009} = -1{,}0382$$

Calcolo della retta di regressione

L'intercetta nell'asse Y stimata (a):

$$a = \bar{Y} - b\bar{X} = 2{,}6823 + 1{,}0382(0{,}6654) = 3{,}37$$

Da qui otteniamo l'equazione della retta di regressione dei minimi quadrati:

Y = 3,37 - 1,04 X

che può essere espressa anche nella forma:

frequenza di «dee» = 3,37 - 1,04 (peso del predatore)

Calcolo dell'errore standard della pendenza

Per calcolare l'errore standard di b dobbiamo conoscere la media dei quadrati dei residui:

$$MS_{residua} = \frac{\sum(Y_i - \bar{Y})^2 - b\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-2}$$
$$= \frac{6{,}8210 - (-1{,}0382)(-3{,}0118)}{13-2}$$
$$= 0{,}3358$$

quindi l'errore standard di b è:

$$ES_b = \sqrt{\frac{MS_{residua}}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{0{,}3358}{2{,}9009}} = 0{,}3402$$

Calcolo di t

Ora disponiamo di tutti gli elementi per calcolare t:

$$t = \frac{b - \beta_0}{ES_b} = \frac{-1{,}038 - 0}{0{,}3402} = -3{,}051$$

Dobbiamo adesso confrontare questa statistica t con la distribuzione t con df=n-2=13-2=11 gradi di libertà:

$t_{0.05,11}$=2,201

La regione di accettazione è compresa nell'intervallo da -2,201 a +2,201.

Il t calcolato cade all'esterno della regione di accettazione quindi L'IPOTESI NULLA DEVE ESSERE RIFIUTATA.

Esercizio sulle cince con R (I)

```
#1°METODO MANUALE (effettuiamo tutti i calcoli SENZA formule ad hoc)
#carico il file «Regressione_esercizio_cince.txt»
dati<-read.table(choose.files(),header=T)
#assegno le variabili x e y
x<-dati$mass
y<-dati$dees
#visualizzo il diagramma a dispersione
matplot(x,y,pch=16,col="red")
#calcolo Xmedio e Ymedio
xmed<-mean(x)
ymed<-mean(y)
#calcolo la pendenza
bnum<-sum((x-xmed)*(y-ymed))
bden<-sum((x-xmed)^2)
b<-bnum/bden
#calcolo l'intercetta all'asse y
a<-ymed-b*xmed
#continua…
```

Esercizio sulle cince con R (II)

```
#continua…
#calcolo la devianza di Y e X
devy<-sum((y-ymed)^2)
devx<-sum((x-xmed)^2)
#calcolo la somma dei prodotti
sp<-sum((y-ymed)*(x-xmed))
#calcolo la varianza dei residui (MSresidua)
msr<-(devy-b*sp)/(length(x)-2)
#calcolo l’errore standard di b
ESb<-sqrt(msr/devx)
#EFFETTUO IL TEST SULLA PENDENZA
#calcolo t
tcalc<-(b-0)/Esb
#valore di t critico
qt(p=0.975,df=length(x)-2)
```

Esercizio sulle cince con R (III)

```
#2° METODO
#calcolo la retta di regressione
lmfit<-lm(dees~mass,data=dati)
#effettuo il test t sulla pendenza
out<-summary(lmfit)
#p-value t test
out
```

L’adattamento della retta di regressione ai dati: R^2

Si può misurare la froazione di varizione in Y «spiegata» da X:

$$R^2 = \frac{SS_{Regressione}}{SS_{Totale}}$$

dove $SS_{Regressione}$: $\sum\left(\hat{Y}_i - \bar{Y}\right)^2$

e SS_{Totale}: $\sum(Y_i - \bar{Y})^2$

L'adattamento della retta di regressione ai dati: R^2

- Varia da 0 a 1
- Se **R^2 è vicino a 1**: X prevede la maggior parte della variabilità di Y (le osservazioni Y stanno vicine alla retta di regressione)
- Se **R^2 è vicino a 0**: X non prevede molta della variabilità in Y (le osservazione Y saranno ampiamente dispersi intorno alla retta di regressione)

Nell'esempio delle cince:

$$R^2 = \frac{3{,}1268}{6{,}8210} = 0{,}46$$

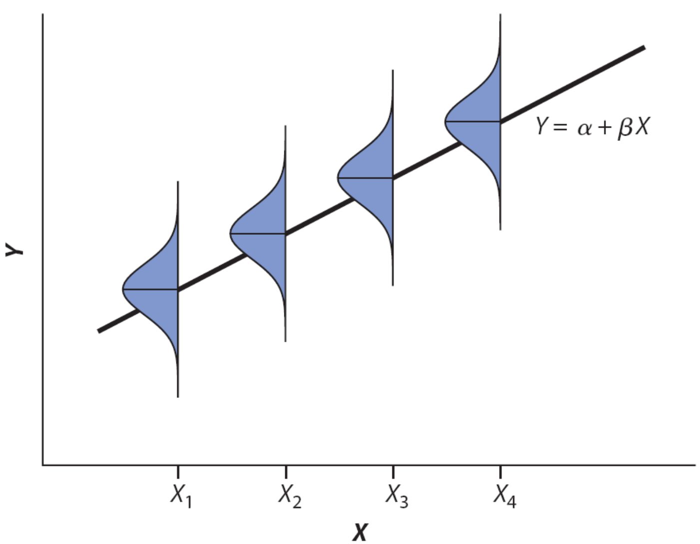
Calcolo di R2 con R

```
#carico il file «Regressione_esercizio_cince.txt»
dati<-read.table(choose.files(),header=T)
#assegno le variabili x e y
x<-dati$mass
y<-dati$dees
#calcolo Ymedio
ymed<-mean(y)
#calcolo la devianza di Y (SSTotale)
devy<-sum((y-ymed)^2)
#genero la retta di regressione
lmfit<-lm(dees~mass,data=dati)
#calcolo SSRegressione
SSreg<-sum((predict(lmfit)-ymed)^2)
#calcolo R2
R2<-SSreg/devy
#calcolo R2 alternativo
summary(lm(dees~mass,data=dati))$r.squared
```

Le assunzioni della regressione

Quando si usa la regressione lineare devono essere soddisfatte le seguenti assunzioni:

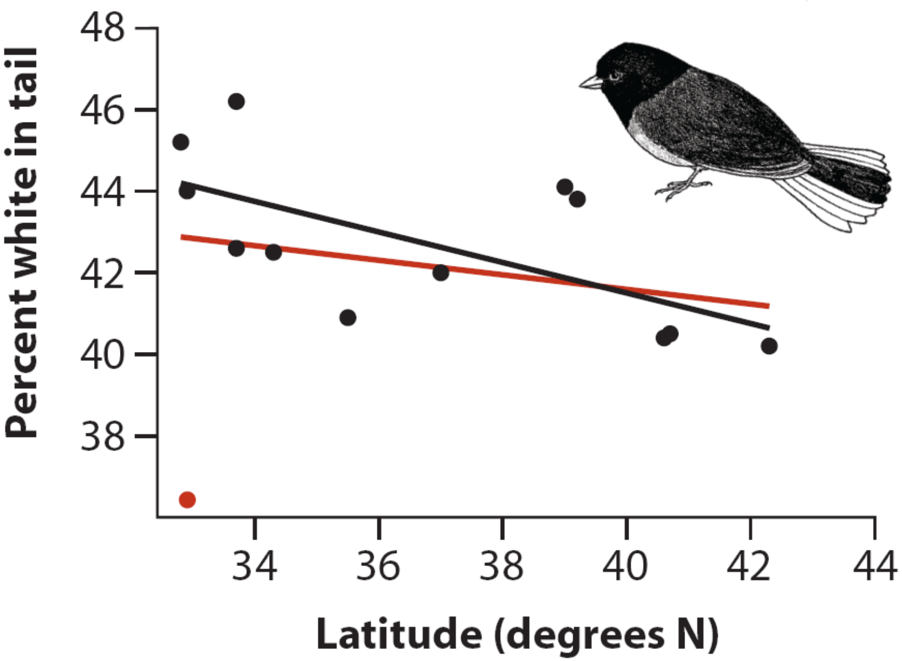
- Per ogni valore di X esiste una popolazione di possibili valori di Y la cui media giace sulla retta di regressione
- Per ogni valore di X la distribuzione dei possibili valori di Y è normale
- La varianza dei valori di Y è la stessa per tutti i valori di X
- Per ogni valore di X le misure di Y rappresentano un campione casuale estratto dalla popolazione dei possibili valori di Y



Attenzione: NON è necessario che X sia distribuita in maniera normale o sia campionata in maniera casuale!

Deviazioni dalle assunzioni della regressione lineare

- Outlier provocano:
  - Distribuzione non normale dei valori di Y
  - Varianze in Y non uguali
  - Influenzano la stima di α e β

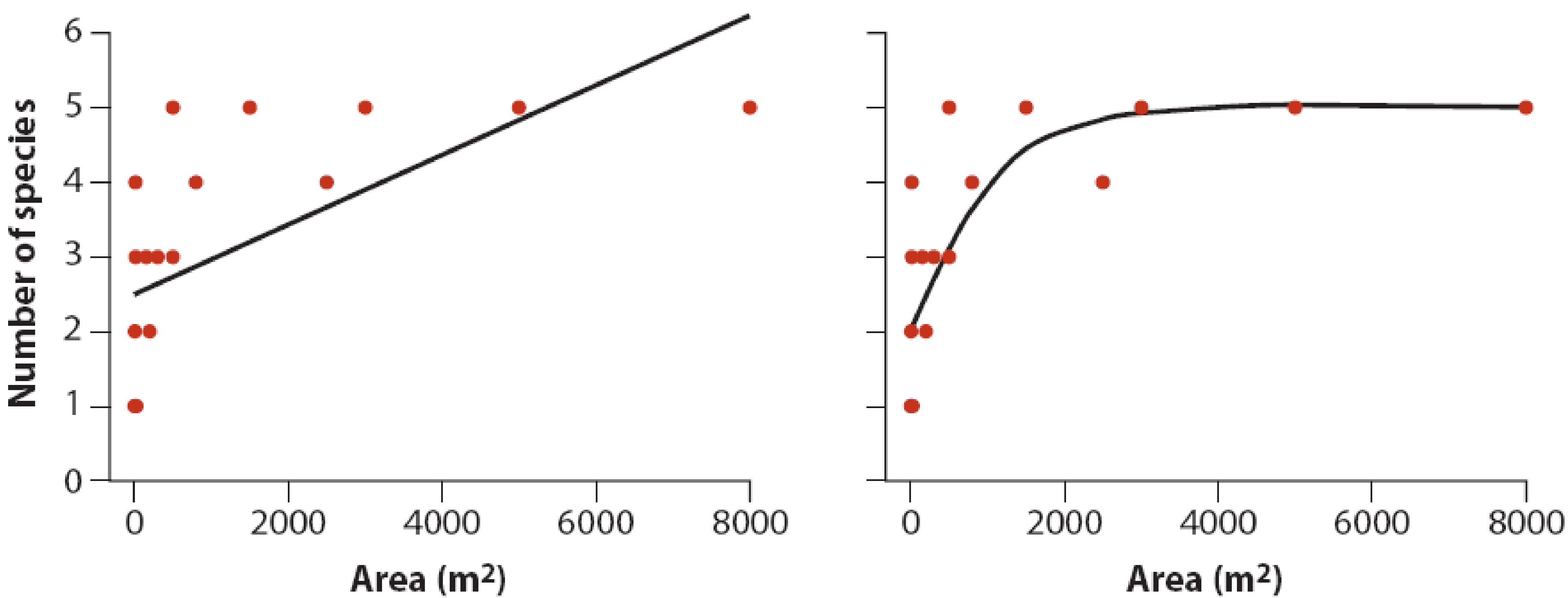


SOLUZIONI:

- Verificare i risultati senza outlier
- Trasformazione di X o Y
- Correlazione per ranghi

Deviazioni dalle assunzioni della regressione lineare

- Identificazione della non linearità:
  - Il diagramma a dispersione permette una diagnosi visiva della non linearità tra Y e X

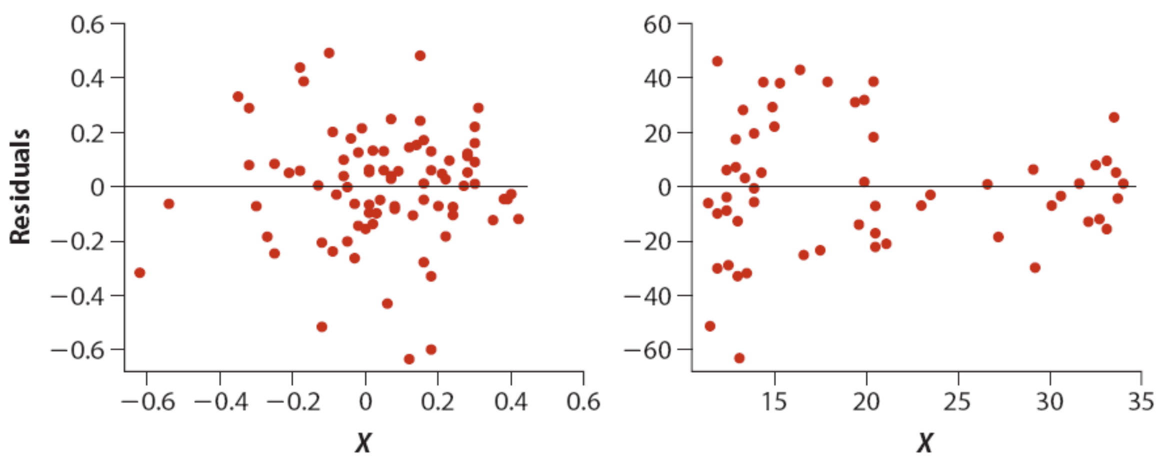


Deviazioni dalle assunzioni della regressione lineare

- Identificazione della non normalità e di varianze disuguali: il **grafico dei residui**:
  - Rappresentazione grafica di $(Y_i - \hat{Y}_i)$ in funzione di X

Se le assunzioni di normalità e di uguali varianze dei residui fossero soddisfatte:

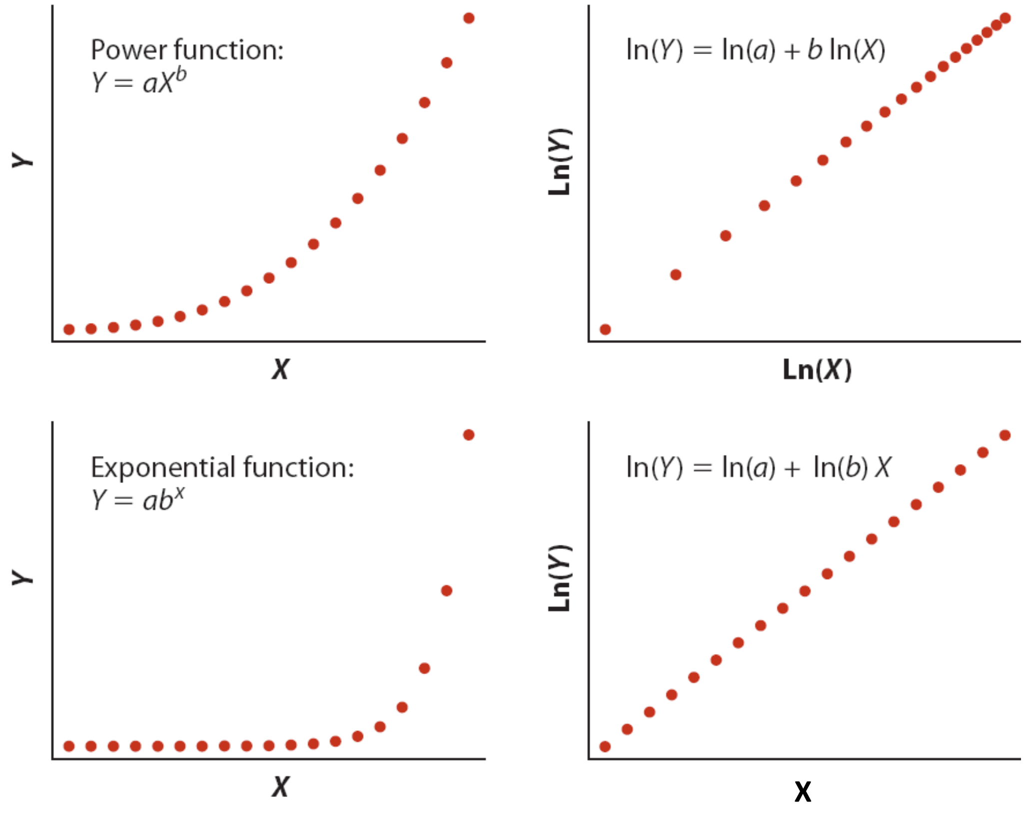
- **Nuvola di punti simmetrica sopra e sotto la retta** orizzontale corrispondente a Y=0, con una maggiore densità di punti vicino alla retta
- Una **curvatura poco rilevante o assente** spostandosi da sinistra a destra lungo X
- Una **dispersione circa uguale di punti sopra e sotto la retta Y=0** per tutti i valori di X



Le trasformazioni

Alcune relazioni non lineari (ma non tutte) possono essere rese tali con una trasformazione appropriata.

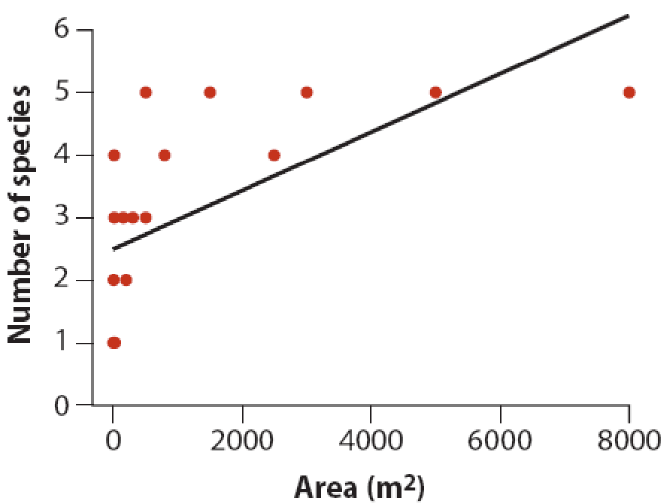
Una delle più utilizzate è la **trasformazione logaritmica**.



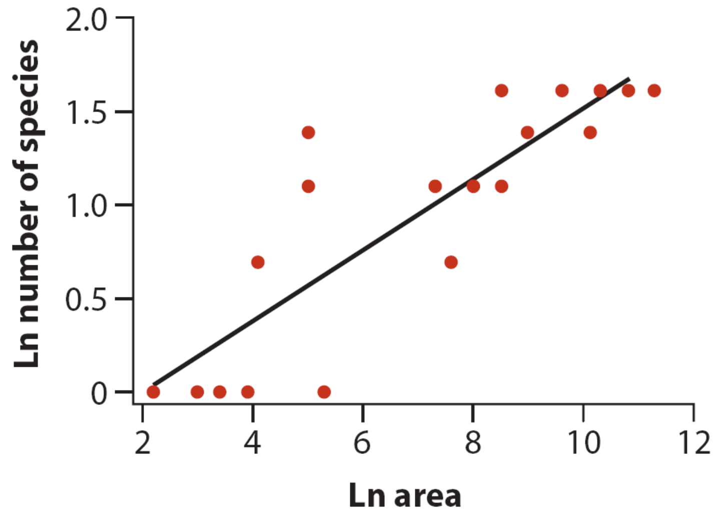
Le trasformazioni

In un caso reale si procede per tentativi: potrebbe essere necessario trasformare X, Y o entrambe.

- **Effetto sulle osservazioni:**



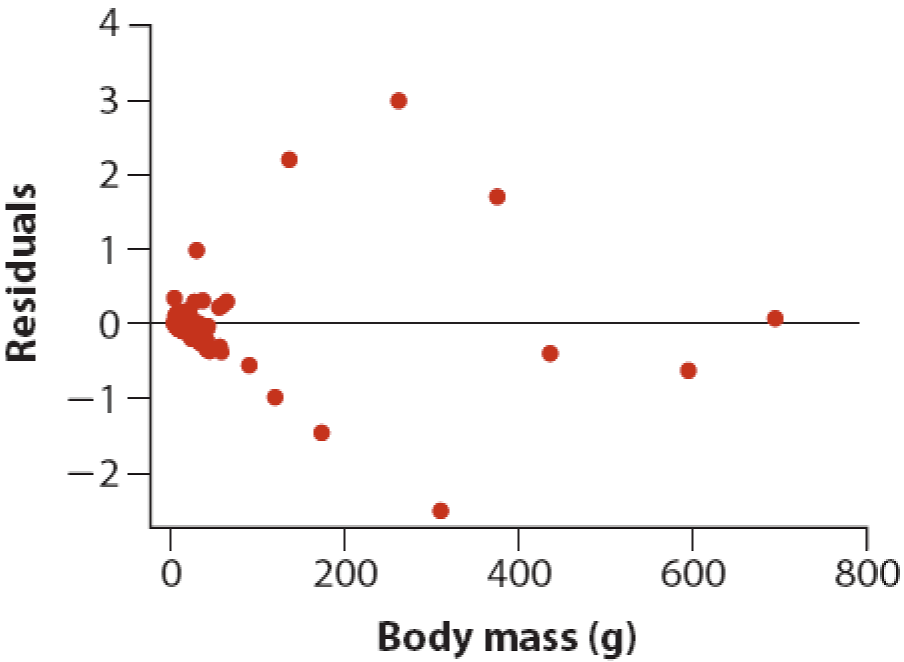
Variabili nella scala originale

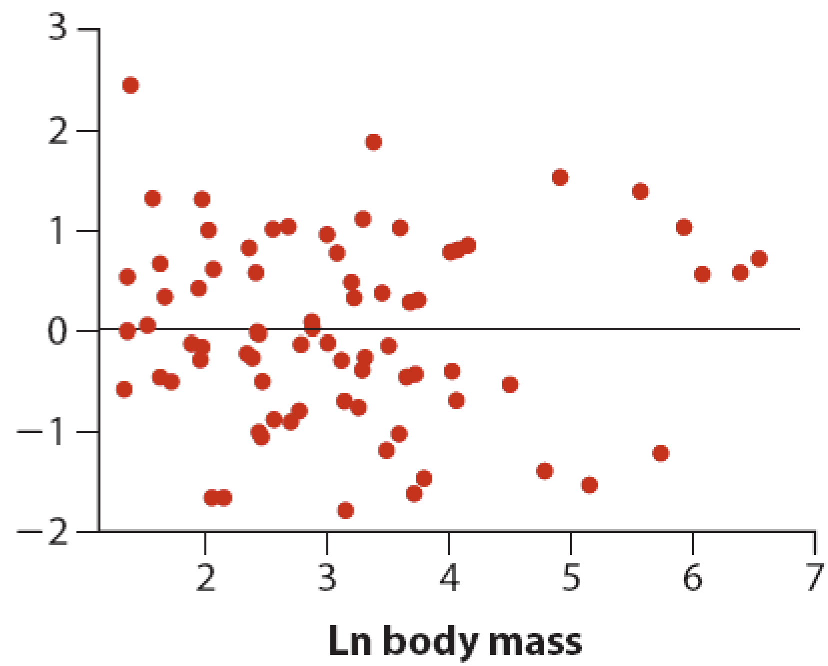


Variabili nella scala trasformata

Le trasformazioni

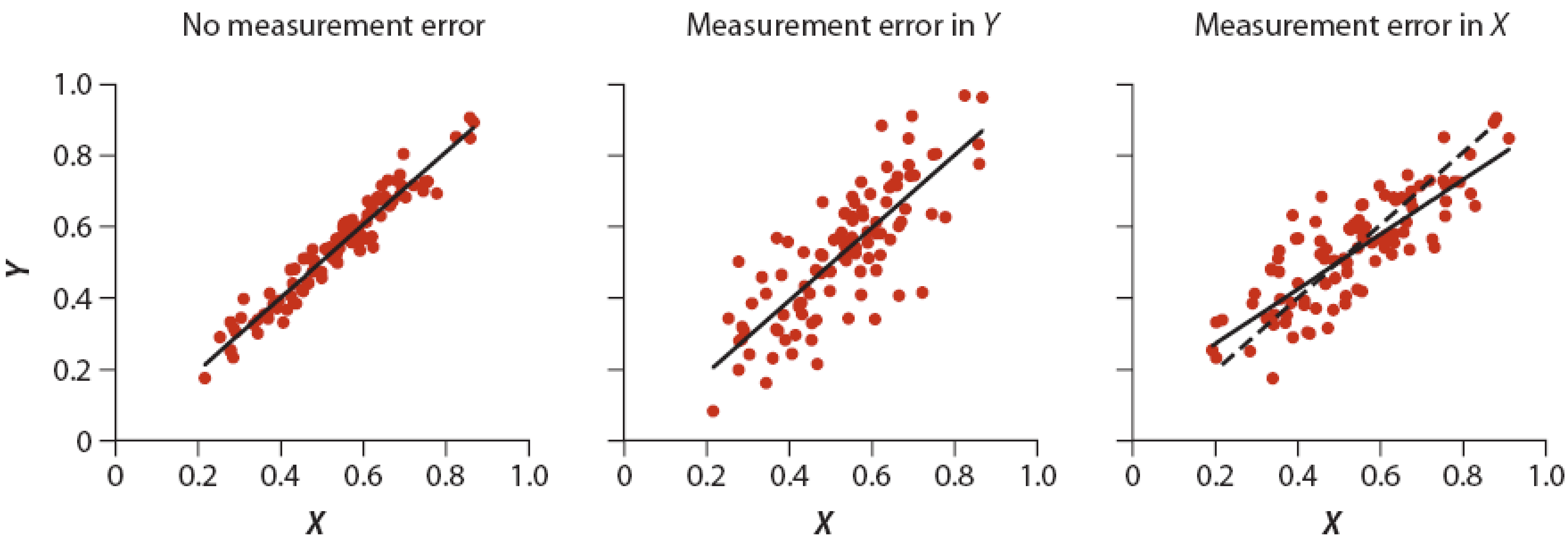
– **Effetto sui residui**





Gli effetti dell'errore di misura sulla regressione

- Errore introdotto quando una variabile (X o Y) non è misurata con la dovuta accuratezza



Measurement error in Y:

- Aumenta la varianza dei residui
- Aumenta l'Errore standard della pendenza (e delle previsioni)

Measurement error in X:

- Aumenta la varianza dei residui
- Distorsione nella stima della pendenza ($b \rightarrow 0$)

La regressione non lineare

- Non sempre le trasformazioni riescono a convertire una relazione non lineare in una relazione lineare.

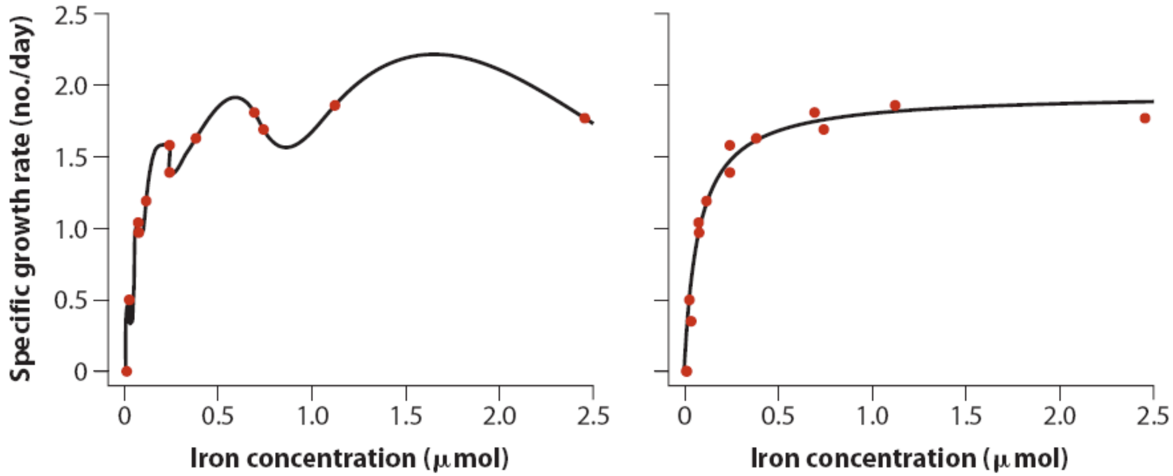
- Nella regressione non lineare assumiamo che la relazione vera tra X e Y NON sia lineare, ma si adatti ad un’altra funzione.

Principali tipi di relazione:

- Le curve con asintoto
- Le curve quadratiche
- Adattamento di una curva senza formula
- Regressione logistica (variabile risposta binaria)

La regressione non lineare

- Le curve con asintoto



- Pessima capacità di predire una nuova osservazione
- Difficile giustificazione biologica della curva

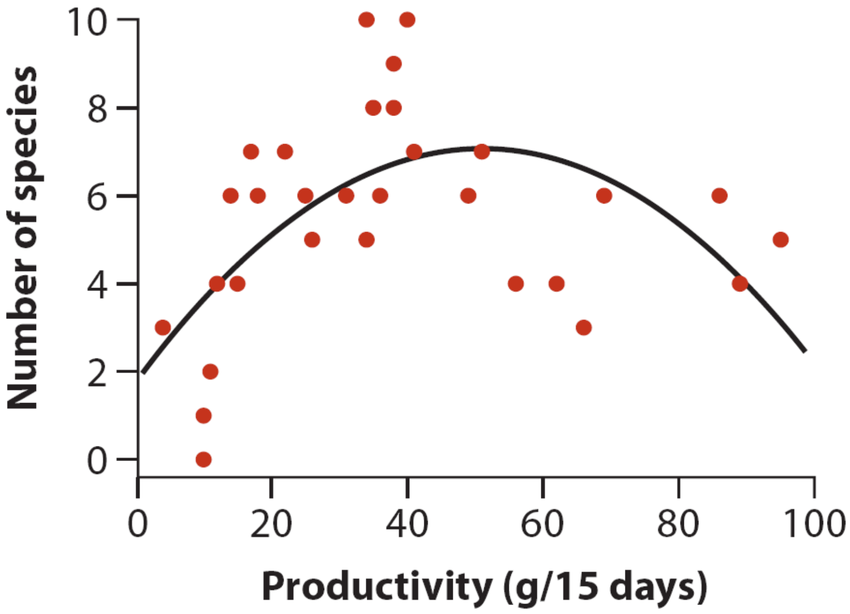
Utilizzando la funzione

$$Y = \frac{aX}{b + X}$$

risolviamo entrambi i problemi

La regressione non lineare

- Le curve quadratiche



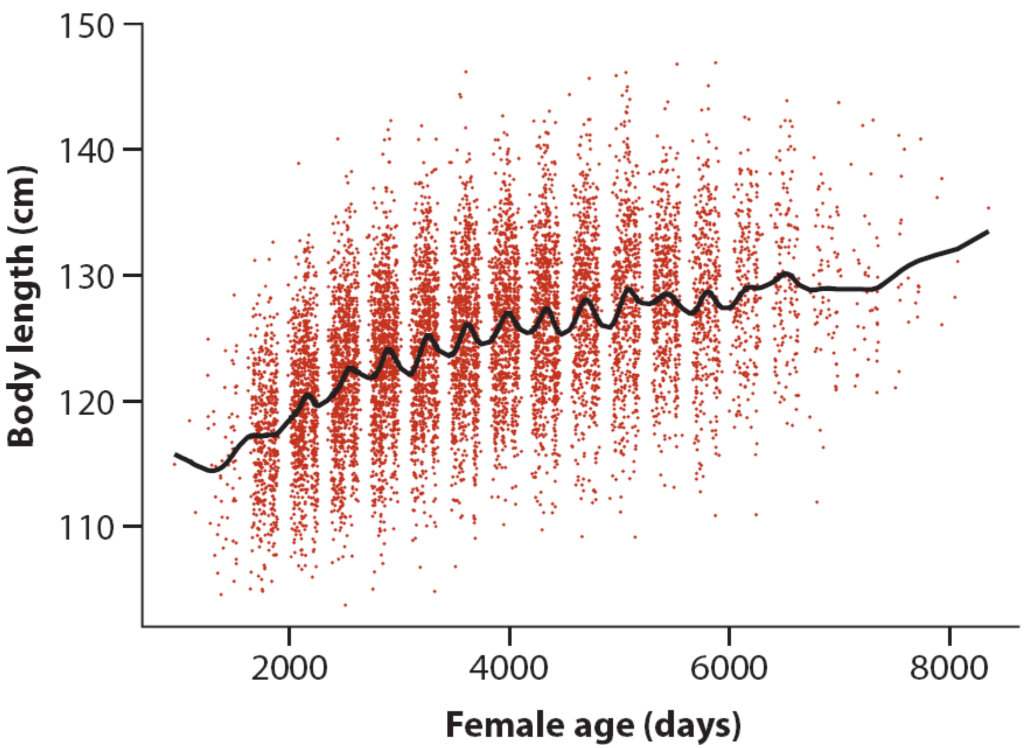
- Usata per adattare una «curva a gobba» (parabola)
- Quando c è negativo la concavità è verso il basso
- Quando c è positivo la concavità è verso l'alto

Equazione:

$$Y = a + bX + cX^2$$

La regressione non lineare

- Adattamento di una curva senza formula (**smoothing**)



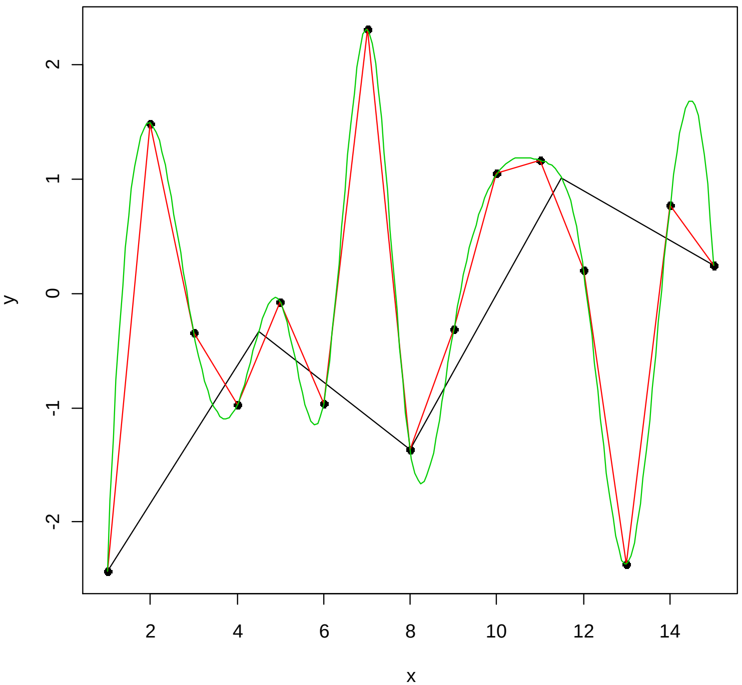
- Indipendente dalla specifica di un’equazione per la curva
- Si cerca di stimare come varia la media di Y la crescere dei valori di X
- Esistono vari metodi, tra i più usati:
    - Kernel smoothing
    - Spline smoothing
    - LOESS smoothing

Attenzione: la curva stimata è molto dipendente dal «coefficiente di smoothing»

La regressione non lineare

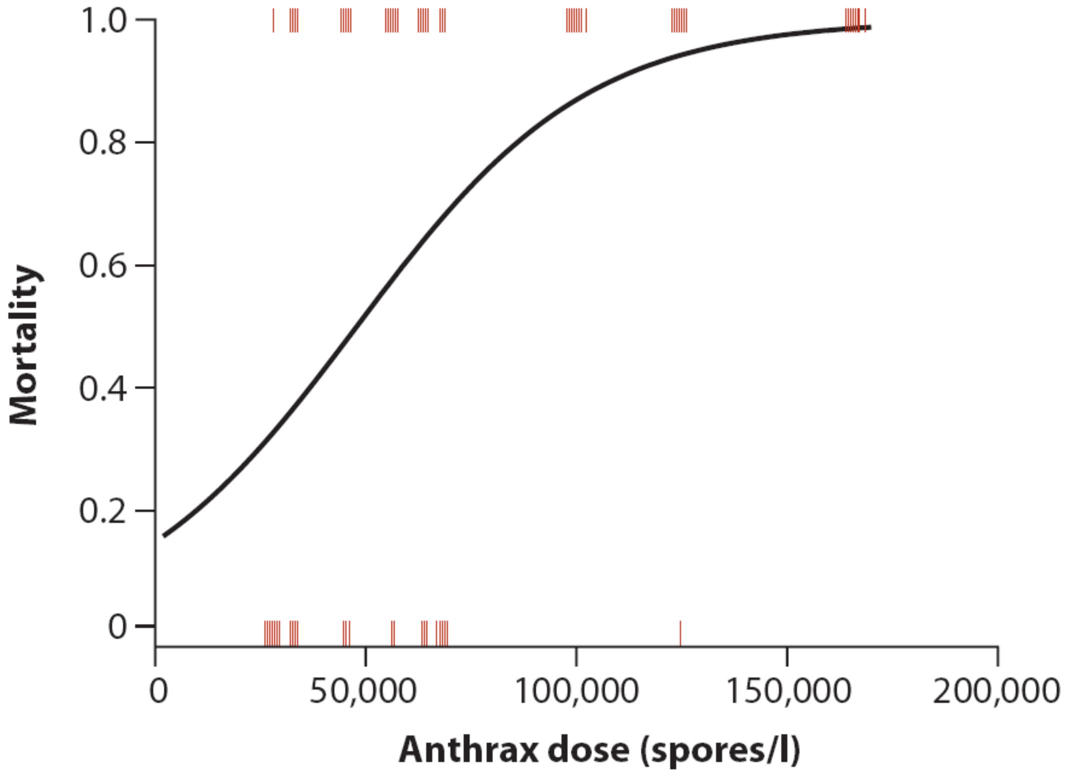
- Adattamento di una curva senza formula (**smoothing**) con R

```
#genero n osservazioni casuali
#da una distribuzione normale standard
n <- 15
x <- 1:n
y <- rnorm(n)
#creo il grafico con i punti
matplot(x, y,pch=16)
#adatto tre diverse curve spline
lines(spline(x, y, n = 5), col = "black")
lines(spline(x, y, n = 15), col = "red")
lines(spline(x, y, n = 200), col = "green")
```



La regressione non lineare

- Regressione logistica



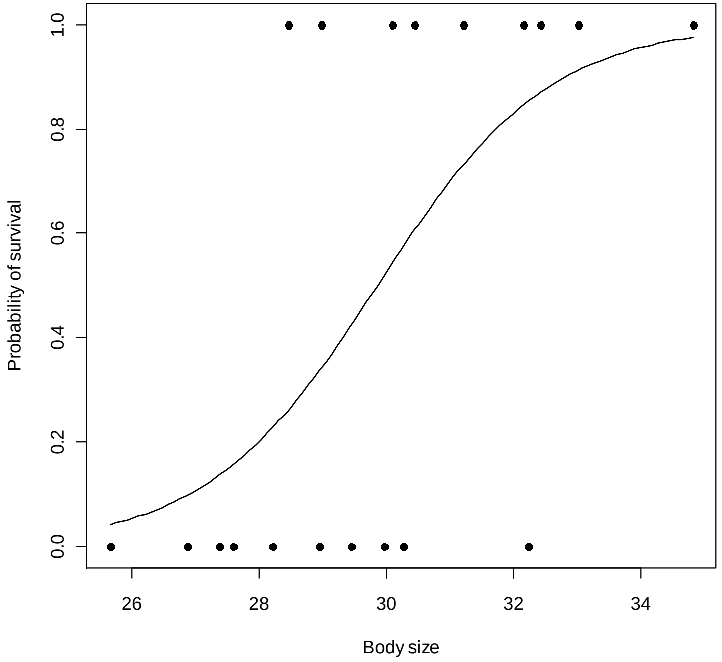
- Applicabile solo per variabili risposta di tipo binario (es: si/no o 0/1)
- Si basa sull'adattamento della seguente equazione:

$$log - odds(y) = a + bX$$

dove $odds = \frac{p}{1-p}$

La regressione non lineare

- Regressione logistica con R




```
#Creiamo un dataset fittizio di
#20 individui con differenti taglie corporee
bodysize<-rnorm(20,30,2)
bodysize<-sort(bodysize)
#assegnamo un indice di sopravvivenza
survive=c(0,0,0,0,0,1,0,1,0,0,1,1,0,1,1,1,0,1,1,1)
#trasformiamo i due vettori in un dataframe
dat<-as.data.frame(cbind(bodysize,survive))
#creiamo il grafico dei punti
matplot(bodysize,survive,xlab="Body size",ylab="Probability of survival“,pch=16)
#adattiamo una curva logistica
g=glm(survive~bodysize,family=binomial,data=dat)
#aggiungiamo la curva al grafico
curve(predict(g,data.frame(bodysize=x),type="resp"),add=TRUE)
```

Esercizi di riepilogo – Es 1

Per verificare se il grasso sottocutaneo assicuri isolamento termico nell'uomo, Sloan e Keatinge (1973) hanno misurato la quantità di calore dispersa nell'unità di tempo da giovani che nuotavano fino a 40 min in acqua a 20,3 °C consumando circa 4,8 kcal/min. Il calore disperso nell'unità di tempo veniva misurato dalla diminuzione della temperatura corporea, registrata con un termometro inserito sotto la lingua, divisa per il tempo in minuti trascorso nuotando. Gli autori hanno misurato un indice di «magrezza» corporea di ciascun ragazzo, definito come il reciproco dello spessore delle pliche cutanee standardizzato per l'area della superficie cutanea totale (in m^2) e per la massa corporea (in kg). I dati sono riportati nella tabella seguente.

Magrezza corporea (m/kg)	Calore disperso nell'unità di tempo (°C/min)
7,0	0,103
7,0	0,097
6,2	0,090
5,0	0,091
4,4	0,071
3,3	0,024
3,6	0,014
2,8	0,041
2,4	0,031
2,1	0,010
2,1	0,006
1,7	0,002

Domande:

a) Costruite un diagramma a dispersione per questi dati, indicando il tipo di relazione tra quantità di calore dispersa nell'unità di tempo e magrezza corporea

b) La quantità di calore dispersa nell'unità di tempo dipende dalla magrezza corporea? Eseguite un test formale.

Esercizi di riepilogo – Soluzione ES1

```
#soluzione punto a
#carico il file «Regressione_esercizio_Es1.txt»
dati<-read.table(choose.files(),header=T)
#grafico a dispersione –attenzione a scegliere correttamente X e Y
matplot(dati$Magrezza,dati$Calore_disperso,pch=16,col="red")
#calcolo la retta di regressione e la aggiungo al grafico
lmfit<-lm(Calore_disperso~Magrezza,data=dati)
abline(lmfit,lwd=2)
```

Esercizi di riepilogo – Soluzione ES1

#soluzione punto b

#H0:la quantità di calore dispersa nell'unità di tempo non varia al variare della magrezza corporea ($\beta = 0$)

#HA: La quantità di calore dispersa nell'unità di tempo varia al variare della magrezza corporea ($\beta \neq 0$)

summary(lmfit)

#oppure controllo i limiti della regione di accettazione con alfa = 5%

qt(p=0.025,df=length(dati$Magrezza)-2)#limite sinistro

qt(p=0.975,df=length(dati$Magrezza)-2)#limite destro

#Risultato: rifiuto l'ipotesi nulla, concludiamo che la quantità di calore disperso aumenta all'aumentare della magrezza corporea

Esercizi di riepilogo – Es 2

Huesner (1991) ha raccolto i dati seguenti sul peso (in grammi) e il metabolismo basale (in watt) in 17 specie di primati.
Ricerche precedenti avevano indicato che il metabolismo basale (R) delle specie di mammiferi ha la seguente dipendenza dal peso (M): $R = \alpha M^{\beta}$, dove α e β sono costanti.

a) Usando la regressione lineare, stimate β per i primati. Denotate con «b» la vostra stima.
b) Rappresentate la retta e i dati in un diagramma a dispersione.
c) Calcolate l'Errore standard per b e l'intervallo di confidenza al 95% di β. Assumente che i dati sulle specie siano indipendenti.

Specie	Peso (g)	Metabolismo basale (W)
Alouatta palliata	4670,0	11,6
Aotus trivirgatus	1020,0	2,6
Arctocebus calabarensis	206,0	0,7
Callithrix jachus	190,0	0,9
Cebuella pygmaea	105,0	0,6
Cheirogaleus medius	300,0	1,1
Euoticus elegantulus	261,5	1,2
Galago crassicaudatus	1039,0	2,9
Galago demidovii	61,0	0,4
Galago elegantulus	261,5	1,2
Homo sapiens	70 000,0	82,8
Lemur fulvus	2330,0	4,2
Nycticebus coucang	1300,0	1,7
Papio anubis	9500,0	16,0
Perodicticus potto	1011,0	2,1
Saguinus geoffroyi	225,0	1,3
Saimiri sciureus	800,0	4,4

Esercizi di riepilogo – Soluzione ES2

```
#soluzione punto a
#carico il file «Regressione_esercizio_Es2.txt»
dati<-read.table(choose.files(),header=T)
#verifico se le assunzioni della regressione sono soddisfatte
matplot(dati$mass,dati$watts,pch=16,col="red")
fit<-lm(watts~mass,data=dati)
#verifico le assunzioni guardando anche il grafico dei residui
matplot(dati$watts,fit$residuals,pch=16,col="red")
#la funzione che lega y a x è di tipo potenza, linearizzo facendo il logaritmo naturale di
x e y
datilog<-log(dati)
#verifico la linerità tra ln(x) e ln(y)
matplot(datilog$mass,datilog$watts,pch=16,col="red")
#e i residui
fitlog<-lm(watts~mass,data=datilog)
matplot(datilog$watts,fitlog$residuals,pch=16,col="red")
#la stima di b
b<-fitlog$coefficients[2]
```

Esercizi di riepilogo – Soluzione ES2

```
#soluzione punto b
#rappresento retta Y = −4.05 + 0.74 X e i dati in scala ln
matplot(datilog$mass,datilog$watts,pch=16,col="red")
abline(fitlog,lwd=2)
#soluzione punto c
#calcolo ESb
Esb<-summary(fitlog)$coefficients[2,"Std. Error"]
#calcolo IC95%
IC95<-c( b-qt(p=0.975,df=length(datilog$mass)-2)*Esb,
b+qt(p=0.975,df=length(datilog$mass)-2)*Esb )
```

Esercizi di riepilogo – ES3

- Precedenti evidenze e alcuni modelli teorici prevedono che l'esponente β che descrive la relazione tra metabolismo basale e massa corporea sia uguale a ¾. Usando i dati forniti nel Es2, valutate se l'esponente differisca dal valore atteso di ¾.

Esercizi di riepilogo – Soluzione ES3

#H0: il cambiamento di ln(metabolismo basale) per unità di ln(massa corporea) è 0.75 ($\beta = 0.75$)

#H0: il cambiamento di ln(metabolismo basale) per unità di ln(massa corporea) è 0.75 ($\beta \neq 0.75$)

```
t<-(b-0.75)/Esb
```

#calcolo i limiti della regione di accettazione

```
c(qt(0.025,length(datilog$mass)-2),
qt(0.975,length(datilog$mass)-2))
```

#calcolo il p-value

```
pt(t,df=length(datilog$mass)-2)*2
```

#NON RIFIUTIAMO H0

#la pendenza della retta di regressione tra il ln(x) e ln(y) (che è l'esponente della funzione potenza) non è significativamente diversa da ¾