METODI STATISTICI PER IL MANAGEMENT

Lezione 18

Modelli statistici non lineari

Domenico Piccolo

Università degli Studi della Basilicata

domenico.piccolo@unibas.it

Modelli di regressione non lineari

In generale, un ***modello statistico non lineare*** è una riproduzione semplificata e per analogia che mira ad interpretare una *variabile dipendente* in funzione di una o più *variabili esplicative*, nonché di una ineliminabile *componente di errore*, avvalendosi di una relazione **non-lineare**.

- ***Tipologia dei modelli non lineari***
- ***Inferenza sui modelli non lineari***
- ***Applicazioni dei modelli non lineari***

Obiettivi: modelli lineari e non lineari

➤ É ingenuo assumere che la relazione tra i fenomeni reali debba essere solo di tipo lineare per cui il modello di regressione lineare va inteso come una prima approssimazione.

➤ D'altra parte la forma lineare può essere interpretata come la più semplice approssimazione geometrica, che vale quasi sempre localmente, come è dimostrabile in modo formale.

➤ Una funzione matematica *sufficientemente regolare* ammette sempre un'approssimazione di primo grado, grazie allo sviluppo in serie di Taylor.

➤ Infatti, per ogni $x = x_0$, si ha:

$$f(x) \simeq f(x_0) + f'(x_0)\,(x - x_0) = \left[f(x_0) - f'(x_0)\,x_0\right] + f'(x_0)\,x \;=\; \beta_0 + \beta_1\,x,$$

dove: $\beta_0 = f(x_0) - f'(x_0)\,x_0$ e $\beta_1 = f'(x_0)$.

➤ Con $f'(x_0)$ si indica la derivata prima della funzione $f(x)$ calcolata nel punto $x = x_0$.

➤ Si possono costruire modelli per funzioni polinomiali, inverse, trigonometriche, esponenziali, logaritmiche, etc. se gli argomenti di tali funzioni non contengono ulteriori parametri incogniti e purché la relazione tra funzioni e parametri sia sempre di tipo **lineare**.

➤ Il modello:

$$Y_i = \beta_0 + \beta_1 x_i^3 + \beta_2 log(x_i) + \beta_3 sen\left(\frac{2\pi}{12} x_i\right) + \epsilon_i , \qquad i = 1, 2, \ldots, n ,$$

ammette la possibilità di essere stimato con il metodo dei minimi quadrati ordinari, come una regressione lineare.

➤ Infatti, se si pone:

$$x_{i\,1} = x_i^3; \quad x_{i\,2} = log(x_i); \quad x_{i\,3} = sen\left(\frac{2\pi}{12} x_i\right); \qquad i = 1, 2, \ldots, n ,$$

si ottiene un modello lineare di regressione multipla:

$$Y_i = \beta_0 + \beta_1 x_{i\,1} + \beta_2 x_{i\,2} + \beta_3 x_{i\,3} + \epsilon_i , \qquad i = 1, 2, \ldots, n .$$

Modelli intrinsecamente non lineari

➤ In vari campi, gli studiosi individuano relazioni tra fenomeni che sono intrinsecamente non lineari sulla base dei principi basilari per la loro disciplina *(**TEORIA**)*.

➤ In tali circostanze, si utilizzano i dati campionari per verificare se i risultati empirici si adattano ad una specificazione non lineare che è stata ricavata sulla base di assunzioni teoriche sui fenomeni di cui si discute.

➤ Per agevolare la rappresentazione grafica, nel seguito, si considera il solo caso di un *modello di regressione non lineare semplice* (quindi, con una sola variabile esplicativa).

➤ Moltissime proposte sono presenti nella letteratura di settore.

➤ Nel seguito si introdurranno e, poi, si esemplificheranno solo due modelli non lineari scelti fra quelli più utilizzabili.

- ***Modello di Michaelis e Menten (1913)***.
- ***Modello della funzione logistica (1838)***.

Un modello generale di regressione non lineare

➤ Un *modello di regressione non lineare* è specificato dalle medesime ipotesi classiche del modello lineare, tranne la prima che è modificata nel modo seguente:

$$Y_i \;=\; h(x_i\,;\boldsymbol{\beta}) \;+\; \epsilon_i\,, \qquad i = 1, 2, \ldots, n\,.$$

ove $\boldsymbol{\beta}$ è un vettore di parametri incogniti e la funzione $h(.)$ è una generica *funzione non lineare* dei parametri.

➤ Poiché il valore medio delle v.c. errori ϵ_i è uguale a 0, la specificazione del modello di regressione non lineare implica che il valore medio della risposta Y_i sia:

$$\mathbb{E}\,(Y_i) \;=\; h(x_i\,;\boldsymbol{\beta})\,, \qquad i = 1, 2, \ldots, n\,,$$

➤ La variabile esplicativa è considerata *deterministica* (cioè fissa e nota senza errori) ovvero il modello è *condizionato ai dati disponibili* per la variabile esplicativa:

$$\mathbb{E}\,(Y_i \mid X = x_i) \;=\; h(x_i\,;\boldsymbol{\beta})\,, \qquad i = 1, 2, \ldots, n\,,$$

➤ Si desidera che il valore medio delle v.c. Y_i si adatti il meglio possibile alle osservazioni y_i, ovvero che le *deviazioni* $y_i - h(x_i\,;\boldsymbol{\beta})$ siano minime.

➤ *Per un prefissato vettore di parametri* $\boldsymbol{\beta}$, le differenze $y_i - h(x_i\,;\boldsymbol{\beta})$ sono detti ***scarti*** ed occorre trovare quel vettore $\boldsymbol{\beta}$ tale che tali scarti siano minimi.

➤ Il **metodo dei minimi quadrati** fornisce le *stime dei minimi quadrati* ricercando il vettore $\hat{\boldsymbol{\beta}}$ tale che sia minima la funzione:

$$G(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big[y_i - h(x_i\,;\boldsymbol{\beta})\Big]^2 .$$

➤ Essendo $h(x_i\,;\boldsymbol{\beta})$ una *funzione non-lineare*, uguagliando a zero le derivata di $G(\boldsymbol{\beta})$ non si ottengono (generalmente) equazioni risolvibili in modo esplicito, per cui la ricerca delle soluzioni richiede approssimazioni numeriche disponibili in opportuni software statistici.

➤ Le proprietà degli stimatori (che non sono lineari e che risultano generalmente distorti) sono di tipo *asintotico*.

➤ Anche aggiungendo alle ipotesi classiche l'assunzione di Normalità e indipendenza delle v.c. errori, i test sui parametri (condotti nel modo usuale) hanno comunque solo *validità asintotica*.

Residui di un modello di regressione non lineare

➤ Dopo aver stimato un modello di regressione non lineare, si ottengono i ***residui*** $\hat{e}_i$ dalla definizione:

$$\hat{e}_i \;=\; y_i \;-\; \hat{y}_i \;=\; y_i \;-\; h(x_i\,;\hat{\boldsymbol{\beta}})\,, \qquad i = 1, 2, \ldots, n\,.$$

➤ Tali residui sono gli scarti tra le *osservazioni* y_i e quelle *calcolate* $\hat{y}_i = h(x_i\,;\hat{\boldsymbol{\beta}})$ utilizzando il miglior modello di regressione non lineare stimato sui dati campionari.

➤ Per le analisi sui ***residui*** (cioè gli scarti dal miglior modello di regressione non lineare stimato dai dati campionari) si seguono gli stessi approcci già discussi nel caso del modello di regressione lineare, con gli opportuni adattamenti per tenere conto della non linearità dl modello stimato.

Bontà di adattamento per un modello non lineare

➤ In particolare, per i modelli di regressione non lineare non ha significato calcolare l'indice R^2, non essendo valida una scomposizione della devianza totale.

➤ Per valutare la *bontà del modello stimato*, si propongono due possibilità.

1. Il confronto tra lo scarto quadratico medio campionario s_y della variabile dipendente Y e l'errore standard della regressione non lineare s_n (cioè lo scarto quadratico medio dei residui $\hat{e}_i$ stimati dal modello non lineare) permette di valutare di quanto la funzione non lineare ha ridotto la variabilità dei dati originali, contribuendo a "spiegare" la variabile Y.

2. La correlazione lineare al quadrato fra le osservazioni y_i e quelle teoriche (previste) $\hat{y}_i = h(x_i\,;\hat{\boldsymbol{\beta}})$ del modello stimato è un ulteriore criterio per misurare la bontà di adattamento del modello ai dati.

➤ ***Modello di Michaelis e Menten (1913)***. In origine, è stato introdotto per spiega la velocità di una reazione chimica ma, poi, è stato applicato in molti altri settori scientifici.

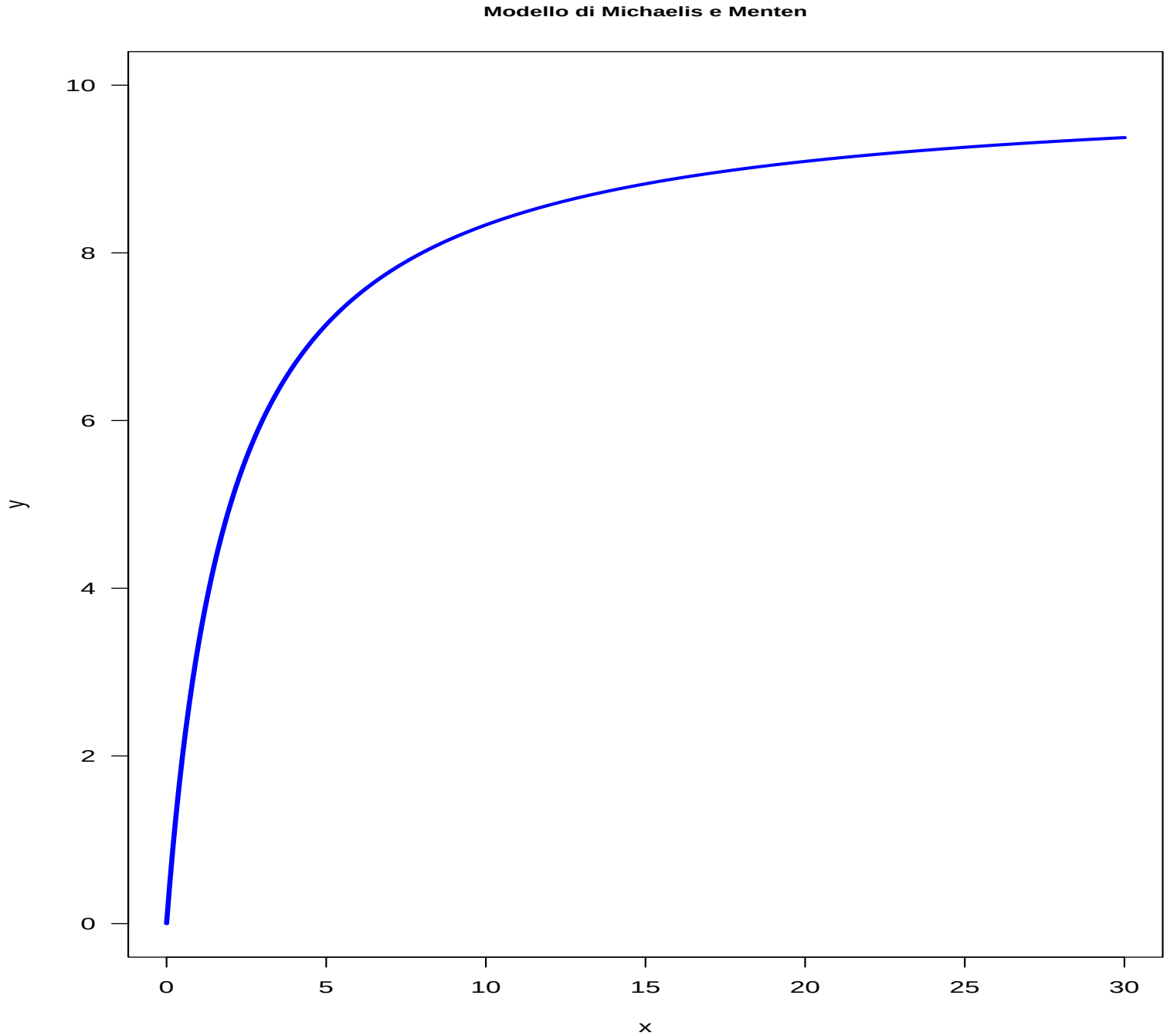
➤ Il ***modello di Michaelis e Menten*** rappresenta una *relazione iperbolica* tra la variabile dipendente e la variabile esplicativa che tende velocemente e asintoticamente ad un massimo β_0.

➤ La specificazione del modello è:

$$Y_i = \frac{\beta_0 \, x_i}{x_i \, + \, \beta_1} + \epsilon_i \, , \qquad i = 1, 2, \ldots, n \, ,$$

ed include due parametri (oltre alla varianza delle v.c. errori).

➤ Generalmente, si assume che le v.c. errori rispettano le ipotesi classiche del modello di regressione lineare.


Modello di Michaelis e Menten
y
x
0
2
4
6
8
10
0
5
10
15
20
25
30

Un'applicazione del modello di Michaelis e Menten .....(1)

- Il dataset `POVERTY` include le variabili principali di uno studio del 1994 riguardante il *livello di povertà* misurato in ciascuna delle 58 contee dello Stato della California delle famiglie residenti in California (USA).

- I dati sono proposti nel testo di Ramanathan, R. (2002). *Introductory Econometrics with Applications*, Fort Worth: Harcourt, fifth edition.

- L'unità statistica di riferimento è la *contea*.

- Nello specifico, si intende valutare come il livello di povertà `Povrate` sia funzione del numero di laureati presenti nella contea `Laurea`.

- La Tabella riassume le informazioni essenziali di queste due variabili, la cui correlazione lineare è negativa ed è importante in valore assoluto: $Corr(\texttt{Povrate}, \texttt{Laurea}) = -0.619$.

Variabile	Descrizione della variabile	min	max
Povrate	Famiglie con reddito sotto il livello di povertà	3.0	20.8
Laurea	Popolazione (> 25 anni) con titolo universitario	9.0	44.0

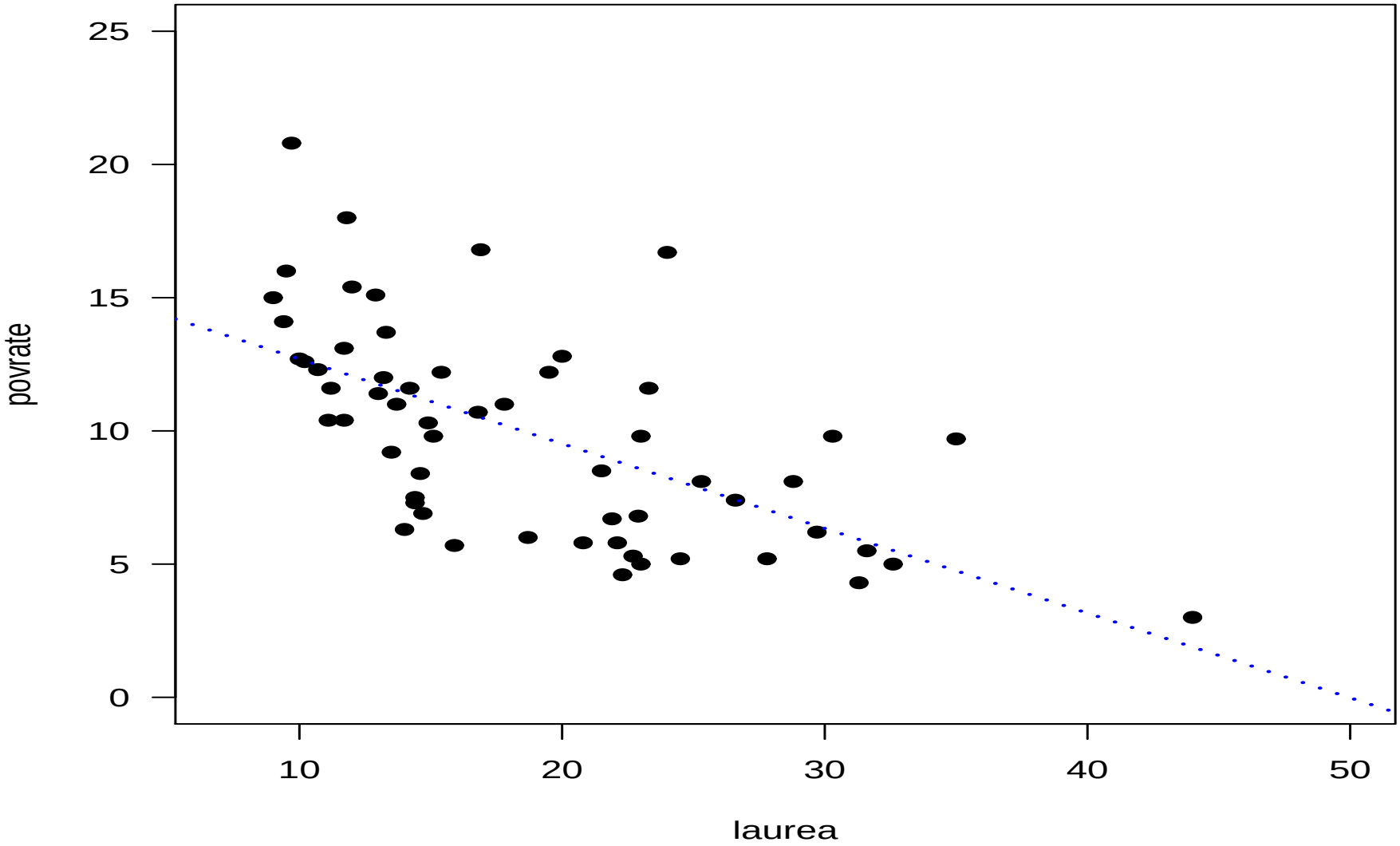
➤ La Figura successiva conferma che assumere una relazione lineare, da stimare mediante la retta di regressione, è solo una prima approssimazione, anche se indubbiamente vi è una relazione decrescente fra queste due variabili.

- *Dal punto di vista statistico*, sembra che la linea retta –pur adeguandosi tendenzialmente e coerentemente alla decrescita di `Povrate` quando aumenta `Laurea`– non coglie appieno la tipologia dell'andamento osservato nei dati.

- *Dal punto di vista interpretativo*, il modello lineare stimato presuppone che, quando la quota di laureati supera un certo limite, il tasso di povertà diventerebbe addirittura negativo: il che mostra che tale modello NON sarebbe coerente.

➤ Tali obiezioni suggeriscono *una relazione media non lineare* fra tali variabili essendo anche opportuno considerare un asintoto per meglio adattarsi alle osservazioni.

Un'applicazione del modello di Michaelis e Menten .....(3)



➤ Pertanto, si può proporre un *modello di Michaelis e Menten* che, nel caso specifico, diventa:

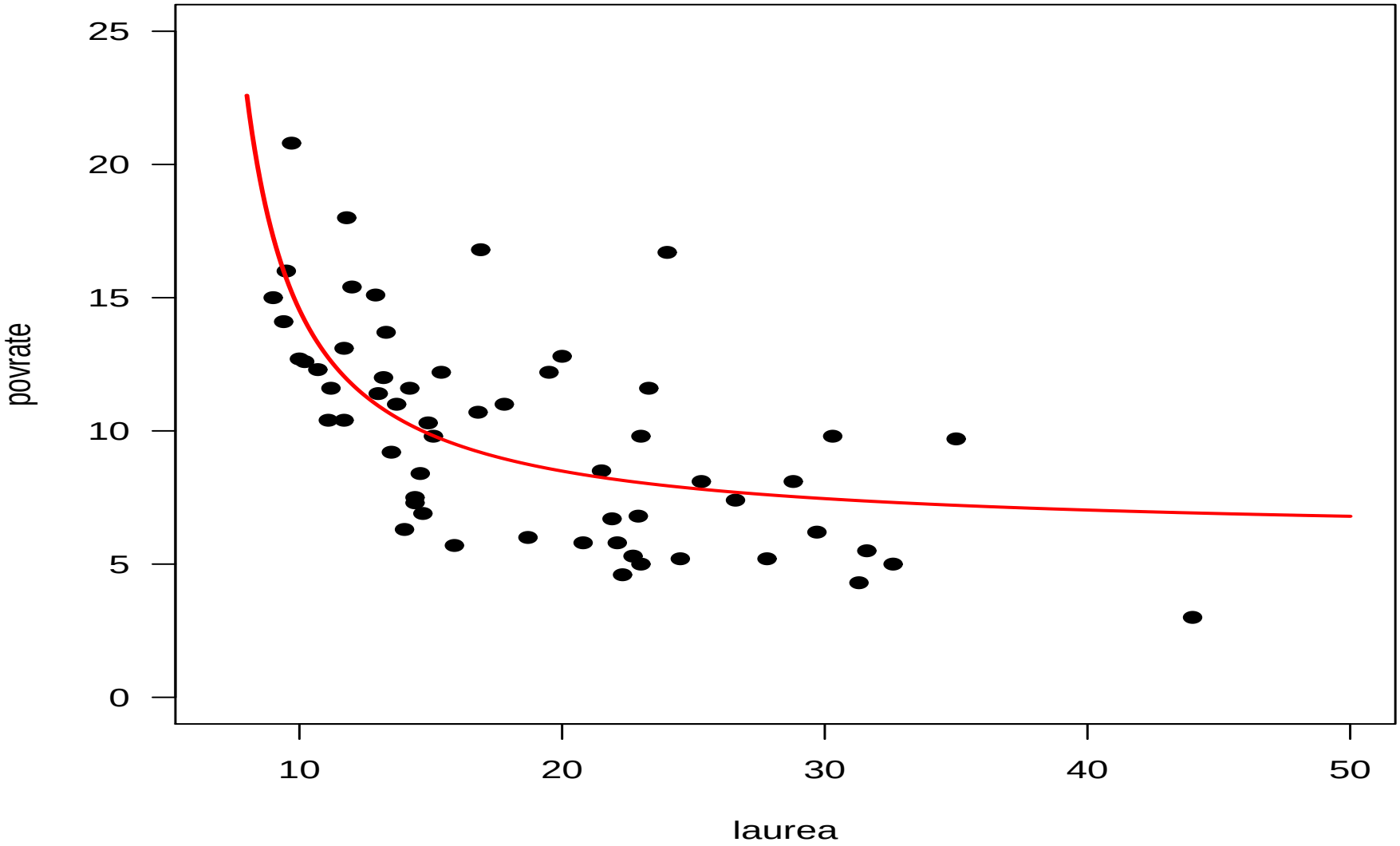
$$Povrate_i = \frac{\beta_0 \, Laurea_i}{Laurea_i + \beta_1} + \epsilon_i \,, \qquad i = 1, 2, \ldots, n \,.$$

➤ Utilizzando il comando `nls(...)` nell'ambiente R si ottengono le seguenti stime:

$$\widehat{Povrate}_i = \frac{5.9963 \, Laurea_i}{Laurea_i - 5.8769} \,, \qquad i = 1, 2, \ldots, n \,.$$

➤ *Nota bene: l'algoritmo richiede valori iniziali accurati per assicurare la convergenza verso il minimo.*

Un'applicazione del modello di Michaelis e Menten .....(5)



➤ Le stime dei parametri sono tutte molto significative e l'errore standard di questo modello non lineare è 3.006, più basso dell'errore standard del modello lineare, che è 3.135.

➤ Il coefficiente di correlazione al quadrato fra valori osservati e valori previsti dal modello non lineare è 0.434 mentre l'analoga quantità, calcolata per il modello lineare (che equivale a R^2) è 0.383.

➤ Il modello di Michaelis e Menten fornisce un migliore adattamento ai dati ed è maggiormente coerente con l'andamento previsto per tale fenomeno: *il valore asintotico stimato per il tasso di povertà, quando la quota dei laureati continua a crescere, è circa il* 6%*, il che appare ragionevole.*

➤ L'osservazione che si presenta isolata (in bassa a destra nella rappresentazione grafica) potrebbe essere considerata un valore atipico rispetto al complesso dei dati e meritare un ulteriore approfondimento del modello stimato.

```
> dati=read.table("POVERTY.R",header=TRUE)
> names(dati)
# "povrate"  "urban"   "compon"  "disocc"
# "diploma" "laurea" "reddfam"
#------------------
> povrate=dati$povrate
> laurea=dati$laurea

#----LINEARE ---------
> plot(laurea,povrate,pch=19,las=1,xlim=c(7,50),ylim=c(0,25))
> mod0=lm(povrate~laurea)
> summary(mod0) ## R^2=0.383
> abline(mod0,lwd=2,lty=3,col="blue")
# Residual standard error: 3.135 on 56 degrees of freedom
```

```
###----MICHAELIS E MENTEN -----
> michmer=povrate~I(bet0*laurea/(bet1+laurea))
> outmod=nls(michmer,start=list(bet0=0.1,bet1=0.1))
> out1=summary(outmod)
> out1$residuals
> outp=predict(outmod)


# > cbind(povrate,outp,out1$residuals,povrate-outp)
# Residual standard error: 3.003 on 56 degrees of freedom

> laureax=seq(8,50,by=0.1)
> povratey=predict(outmod,list(laurea=laureax))
> lines(laureax, povratey, type='l', lwd=2,col='red')

#-----------------------------------
> (cor(povrate,outp))^2     # 0.4343311
#-----------------------------------
```

Modello della funzione logistica

➤ ***Modello della funzione logistica (1838)***. Vari studiosi (a cominciare da Verhulst) lo proposero per studiare l'accrescimento di popolazioni che trovano un ostacolo ad una crescita illimitata. Poi, è stato utilizzato in molti altri contesti: Biologia, Demografia, Economia, Tecnologia, etc.

➤ Nella forma più completa, il ***modello della funzione logistica*** è specificato, da:

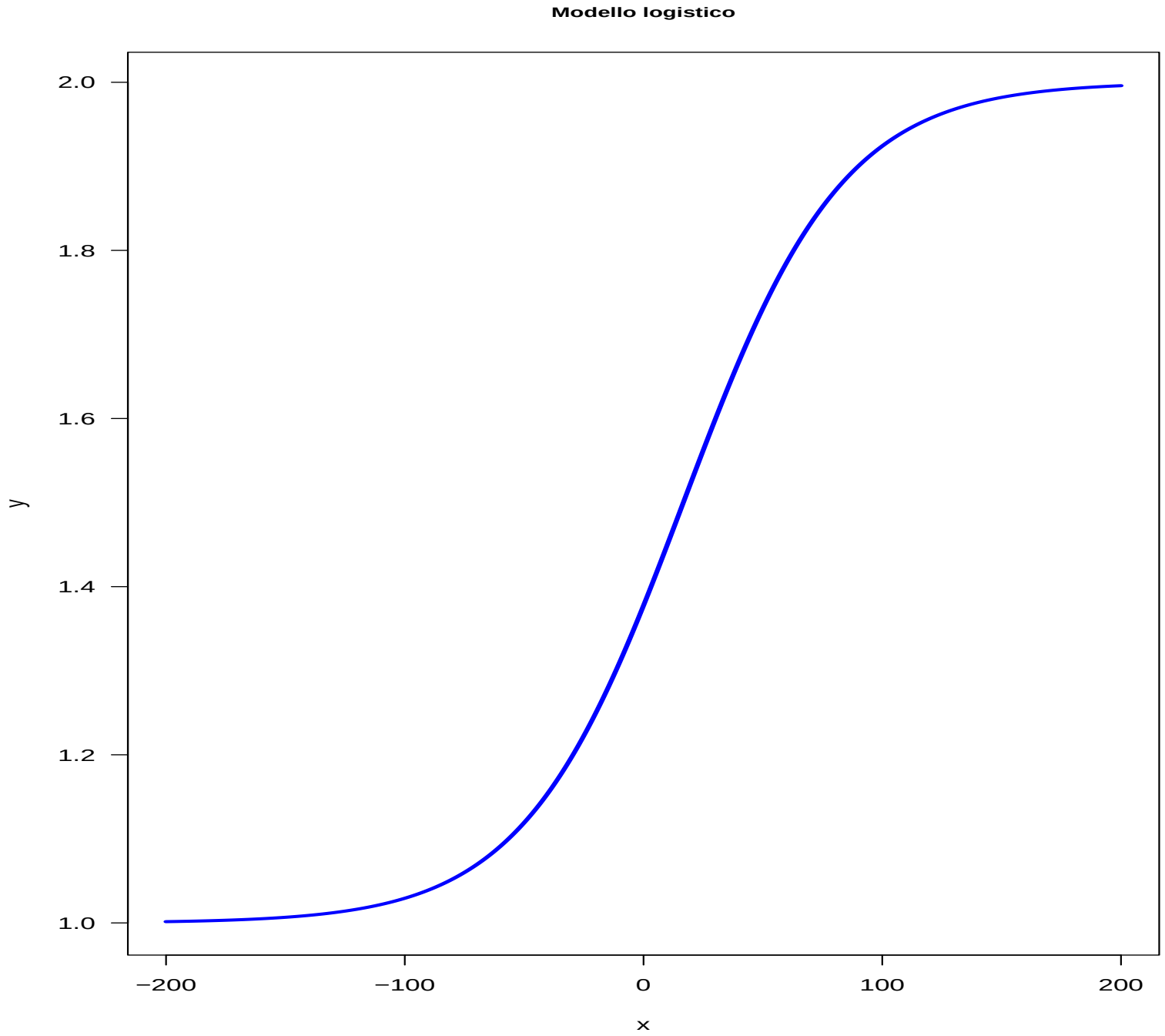
$$Y_i = \beta_0 + \frac{\beta_1}{1 + e^{\beta_2 + \beta_3 x_i}} + \epsilon_i \,, \qquad i = 1, 2, \ldots, n \,,$$

e richiede la stima di quattro parametri espliciti.

➤ **Se** $\beta_3 < 0$, la relazione media tra le variabili rappresenta una *funzione sigmoide* **crescente** (cioè a forma di S allungata), simmetrica rispetto al suo punto di flesso e compresa fra un minimo β_0 ed un massimo $\beta_0 + \beta_1$.

➤ Nel punto di flesso $t_{flex} = \dfrac{\beta_2}{-\beta_3}$ la funzione logistica assume un valore esattamente a metà tra l'asintoto inferiore e quello superiore, cioè: $Y_{flex} = \beta_0 + \beta_1/2$.

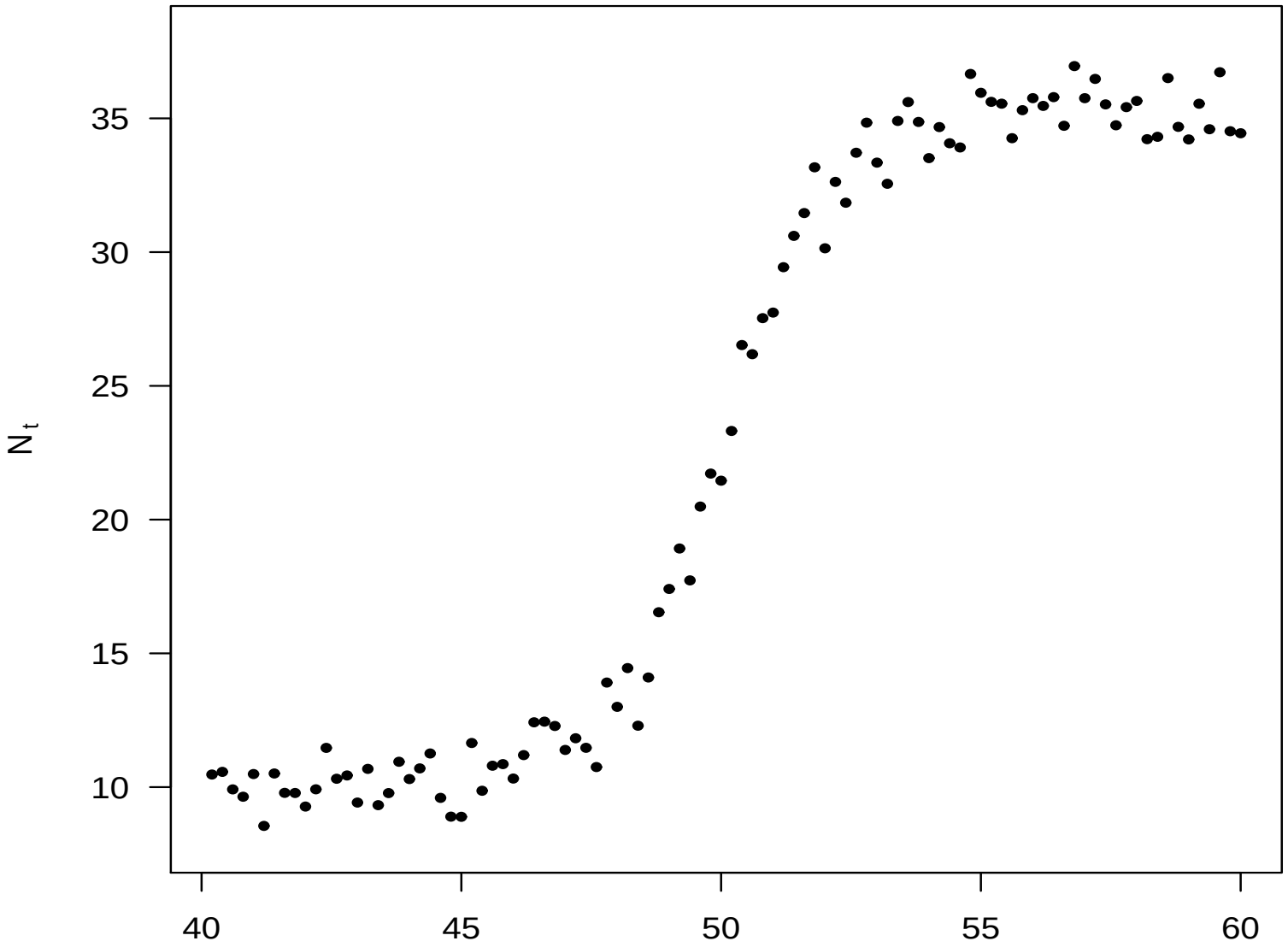
➤ **Se** $\beta_3 > 0$, la relazione media è **decrescente** in modo speculare. resta invariate le altre caratteristiche.


Modello logistico
2.0
1.8
1.6
1.4
1.2
1.0
y
−200
−100
0
100
200
x

➤ É stato condotto un esperimento per $n = 100$ giorni consecutivi durante i quali si è controllata la numerosità di una certa popolazione batterica. Le osservazioni hanno fornito i conteggi (cioè le realizzazioni di N_t) presentati nella Tabella (da leggere per riga).

10.46981	10.56927	9.91463	9.63973	10.48779
8.54940	10.50868	9.78467	9.77954	9.27293
9.91724	11.46635	10.31177	10.43482	9.42235
10.68133	9.32488	9.77621	10.94720	10.30079
10.69916	11.25673	9.59694	8.89465	8.88915
11.64814	9.86348	10.79951	10.85905	10.31981
11.19653	12.42239	12.44674	12.28510	11.38906
11.82606	11.46919	10.74968	13.90795	13.00071
14.44909	12.29525	14.09719	16.53659	17.40710
18.91922	17.72811	20.48673	21.72070	21.45456
23.31354	26.52535	26.18428	27.53069	27.73968
29.43548	30.60753	31.45950	33.16978	30.14237
32.62591	31.84666	33.71418	34.83887	33.34485
32.55496	34.90312	35.60987	34.86645	33.51381
34.67434	34.06907	33.91172	36.66035	35.95484
35.61692	35.54860	34.25866	35.30554	35.75519
35.46567	35.79304	34.72254	36.95388	35.75420
36.47503	35.52236	34.74113	35.41898	35.65151
34.22201	34.31031	36.50496	34.68278	34.21241
35.54734	34.59390	36.72448	34.51878	34.44323

Un'applicazione del modello della funzione logistica ..... (2)



➤ Si vuole studiare l'accrescimento di una popolazione batterica di N_t batteri al tempo t, in funzione della variabile *tempo* t, misurata in giorni.

➤ Il modello di regressione da specificare è del tipo:

$$N_t \;=\; f(t;\,\boldsymbol{\beta}) + \epsilon_t\,, \qquad\qquad t = 1, 2, \ldots, n\,.$$

➤ I dati rappresentano una serie di conteggi misurati rispetto al tempo, cioè *una serie storica di conteggi*, per cui la notazione utilizzata è più immediata.

➤ Motivazioni biologiche ed una lunga serie di esperimenti nel settore mostrano che una funzione adeguata per tale rappresentazione è la ***funzione logistica***, per cui la specificazione del modello è:

$$N_t \;=\; \beta_0 \;+\; \frac{\beta_1}{1 \;+\; e^{\beta_2 + \beta_3\, t}} + \epsilon_t\,, \qquad\qquad t = 1, 2, \ldots, n\,.$$

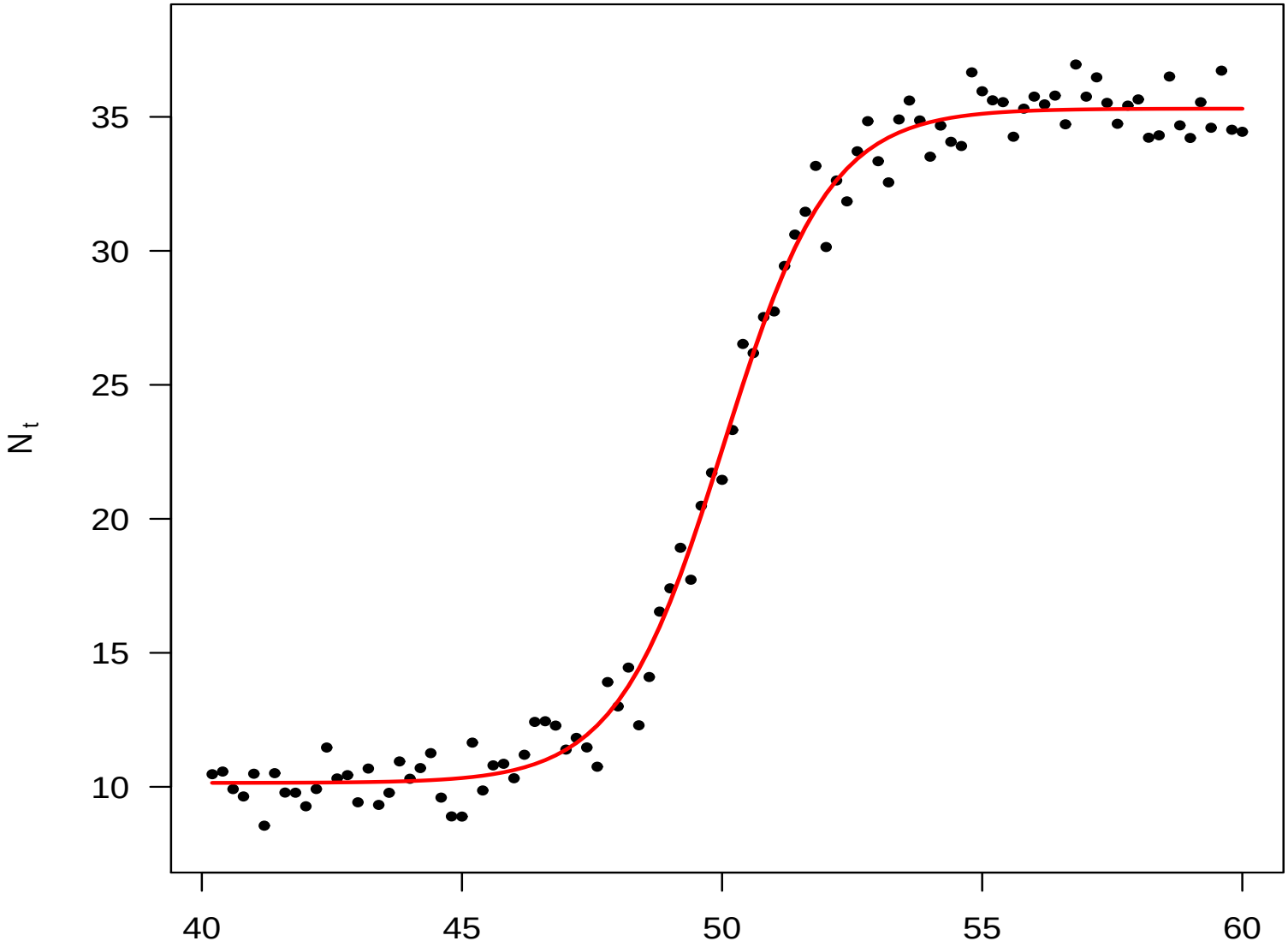
➤ Essendo in presenza di una relazione media che prevede un accrescimento tendenziale deve essere $\beta_3 < 0$. L'*asintoto superiore* è $\beta_0 + \beta_1$ e l'*asintoto inferiore* è β_0.

➤ Il modello stimato presenta *parametri tutti molto significativi* (con $p\text{-}value$ quasi nulli) e con segni che sono coerenti con l'andamento dei conteggi.

Parametri	stime	err. stand.	test t	$p\text{-}value$
$\hat{\beta}_0$	10.1456	0.1631	62.19	$< 2 \times 10^{-16}$
$\hat{\beta}_1$	25.1608	0.2414	104.23	$< 2 \times 10^{-16}$
$\hat{\beta}_2$	48.9816	2.0011	24.48	$< 2 \times 10^{-16}$
$\hat{\beta}_3$	-0.9791	0.0400	-24.49	$< 2 \times 10^{-16}$

➤ Il grafico della funzione stimata si adatta con notevole approssimazione allo sviluppo di questa popolazione batterica.

Un'applicazione del modello della funzione logistica ..... (5)



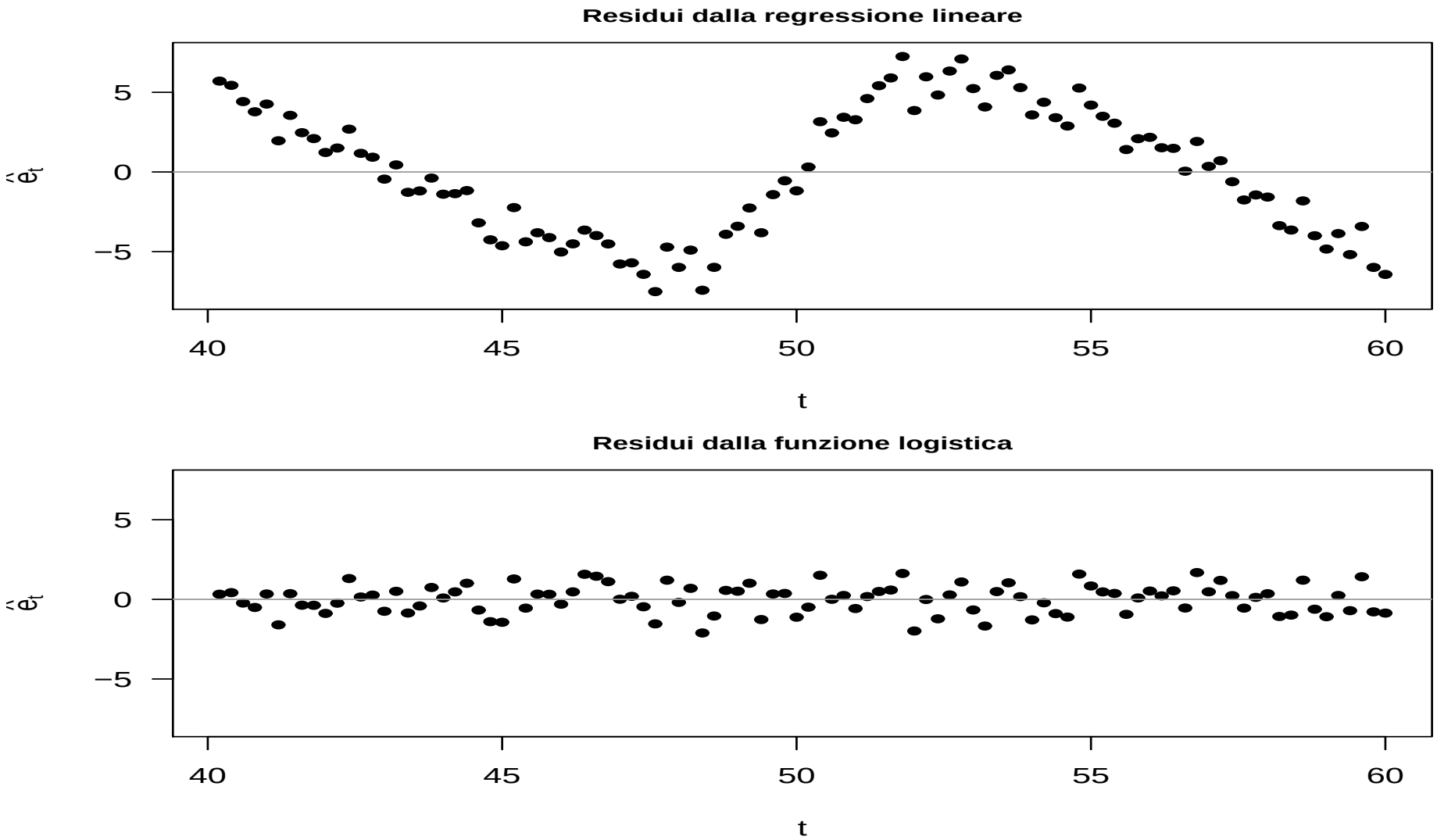
➤ L'asintoto inferiore è stimato mediante $\hat{\beta}_0 = 10.15$ e l'asintoto superiore mediante $\hat{\beta}_0 + \hat{\beta}_1 = 35.31$.

➤ Il punto di flesso è stimato mediante: $t_{flex} = \dfrac{48.9816}{-(-0.9791)} = 50.02717$ cioè circa al tempo $t = 50$ giorni, in aderenza con quanto appare nella Figura.

➤ Confrontando la variabilità dei dati di conteggi N_t (misurata mediante lo scarto quadratico medio campionario $s_N = 11.3135$) con la variabilità dei residui $\hat{e}_t$ stimati dalla regressione non lineare (misurata mediante lo scarto quadratico medio di tali residui campionario $s_n = 0.9022$), *il modello spiega una grandissima parte della variabilità dei dati originali*.

➤ La bontà di adattamento è confermata calcolando il quadrato del coefficiente di correlazione tra valori osservati dei conteggi e valori stimati mediante la funzione logistica, che risulta essere: 0.9938.

Un'applicazione del modello della funzione logistica ..... (7)



➤ *La preferenza a favore del modello non lineare rispetto a quello lineare è molto netta*.

➤ Per questo, è interessante confrontare i residui che si sarebbero ottenuti adattando a questi dati una retta di regressione (per la quale l'indice R^2 è 0.8748) con i corrispondenti residui che si ottengono dalla stima di una funzione logistica.

- I residui dal modello lineare primi sono molto più variabili e manifestano una forte correlazione (violando così la quarta ipotesi classica).
- I residui dal modello non lineare non presentano particolari problemi rispetto alle ipotesi classiche.

➤ Questo esempio conferma come l'analisi dei residui sia un momento fondamentale nella costruzione di un modello di regressione perché, anche **il rifiuto del modello stimato consente di derivare dai residui stimati delle indicazioni a favore di una specificazione alternativa più convincente**.

➤ Le esemplificazioni hanno mostrato come un modello non lineare che derivi da una teoria consolidata deve essere preferito rispetto ad un modello di tipo lineare.

➤ Quello che diventa problematico nella prassi della costruzione dei modelli non lineari è la individuazione di *una* forma funzionale che sia preferibile per i dati che si possiedono rispetto alle infinite altre possibili.

➤ La difficoltà deriva dal fatto che anche una variazione di poche osservazioni potrebbero minare la qualità del modello, un ottimo adattamento e la stabilità previsiva. Sono quindi opportune analisi di sensitività.

➤ A questi rischi, di natura statistica, bisogna aggiungere i costi computazionali perché modelli complessi (con molti parametri e una pluralità di variabili esplicative) richiedono dei valori iniziali delle stime che siano abbastanza vicini alle stime finali dei minimi quadrati: il che è possibile se esiste una *pre-conoscenza* sui fenomeni che si intende studiare.

Considerazioni finali

➤ *La modellistica lineare non andrebbe mai esclusa dalle indagini statistiche* anche nelle situazioni per le quali è evidente che la relazione fra le variabili è di tipo non lineare.

➤ Infatti, il modello lineare è una *prima approssimazione* rispetto ad una relazione più accurata.

➤ Soprattutto, il modello lineare è un ***punto di riferimento*** obbligato e molto utile rispetto al quale modelli più elaborati debbono essere confrontati.