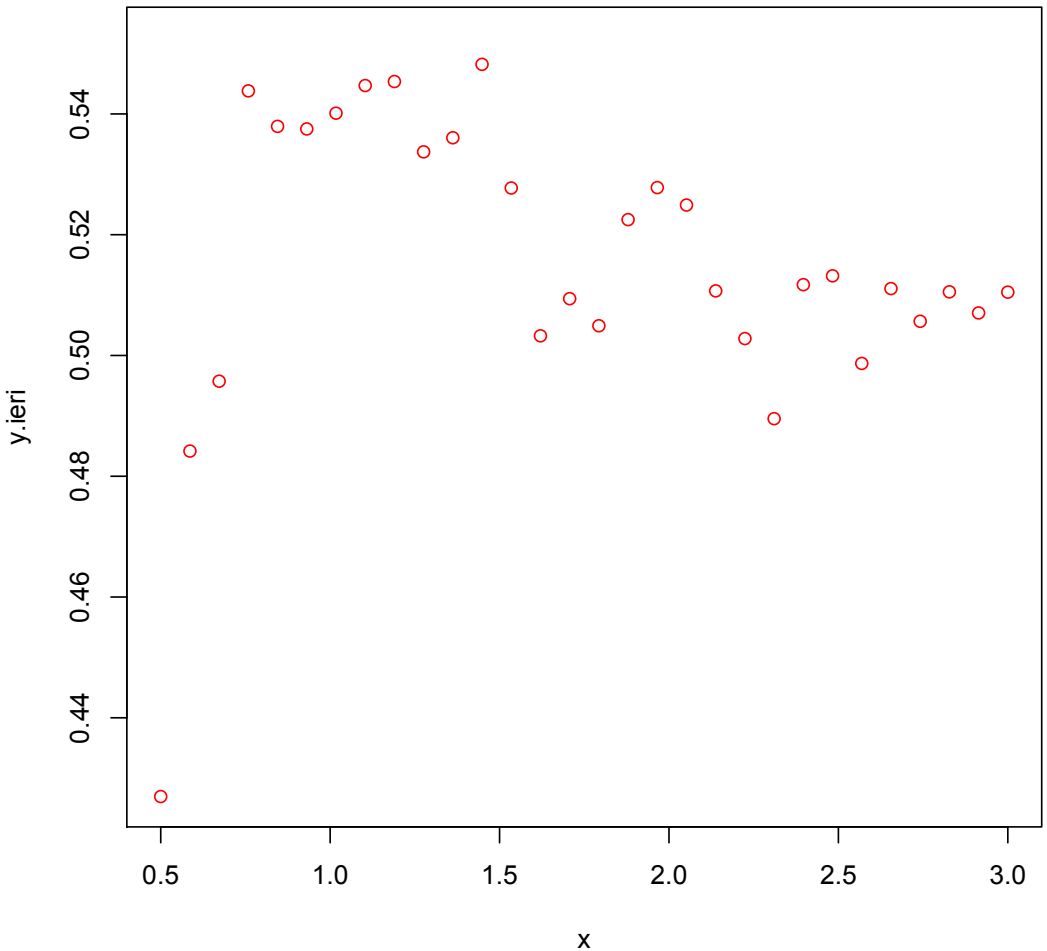
Ottimismo, conflitti e compromessi

Tecniche per la selezione del modello

Un problema tipo

Si supponga di avere generato $n = 30$ coppie di dati x_i, y_i, rappresentati nel seguente diagramma di dispersione.



I dati sono stati generati da una funzione

$$y = f(x) + \varepsilon$$

L'obiettivo é quello di trovare una stima di $f(x)$ che consenta di predire y quando saranno disponibili nuove osservazioni sulla x.
Una delle possibili soluzioni é quella di basarsi su una forma di tipo polinomiale:

$$f(x;\beta) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_{p-1} x^{p-1}$$

Come individuare il grado del polinomio??

I dati sono stati generati da una funzione

$$y = f(x) + \varepsilon$$

L'obiettivo é quello di trovare una stima di $f(x)$ che consenta di predire y quando saranno disponibili nuove osservazioni sulla x.
Una delle possibili soluzioni é quella di basarsi su una forma di tipo polinomiale:

$$f(x;\beta) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_{p-1} x^{p-1}$$

Come individuare il grado del polinomio??

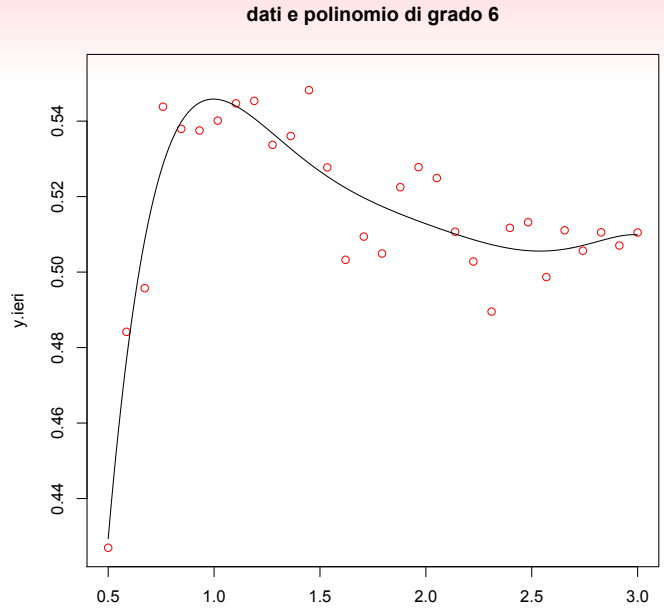
I dati sono stati generati da una funzione

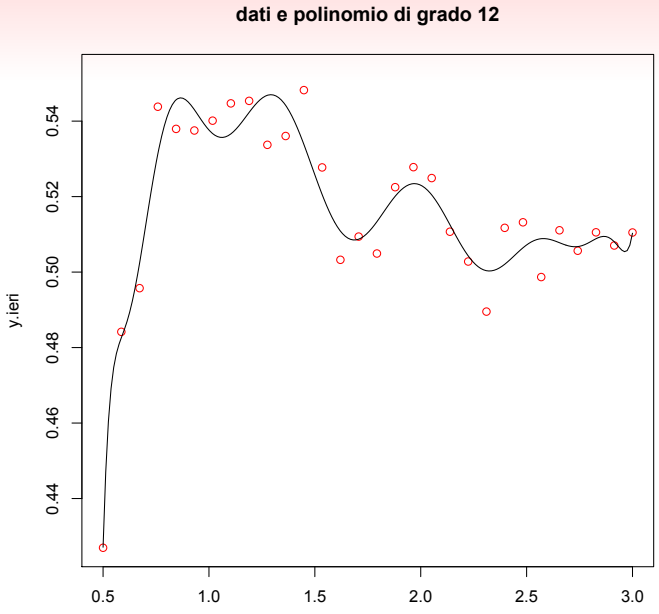
$$y = f(x) + \varepsilon$$

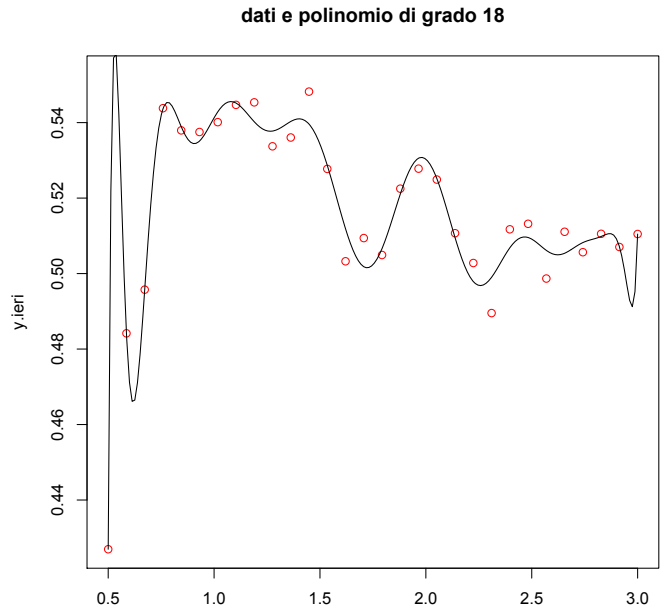
L'obiettivo é quello di trovare una stima di $f(x)$ che consenta di predire y quando saranno disponibili nuove osservazioni sulla x.
Una delle possibili soluzioni é quella di basarsi su una forma di tipo polinomiale:

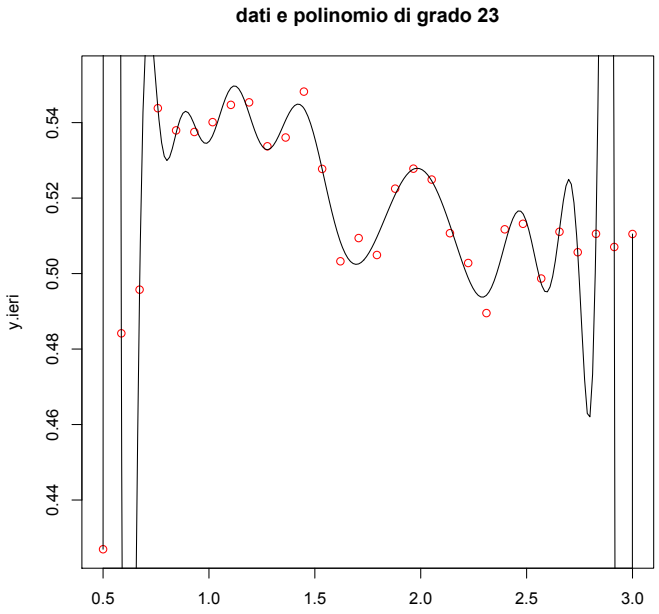
$$f(x;\beta) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_{p-1} x^{p-1}$$

Come individuare il grado del polinomio??

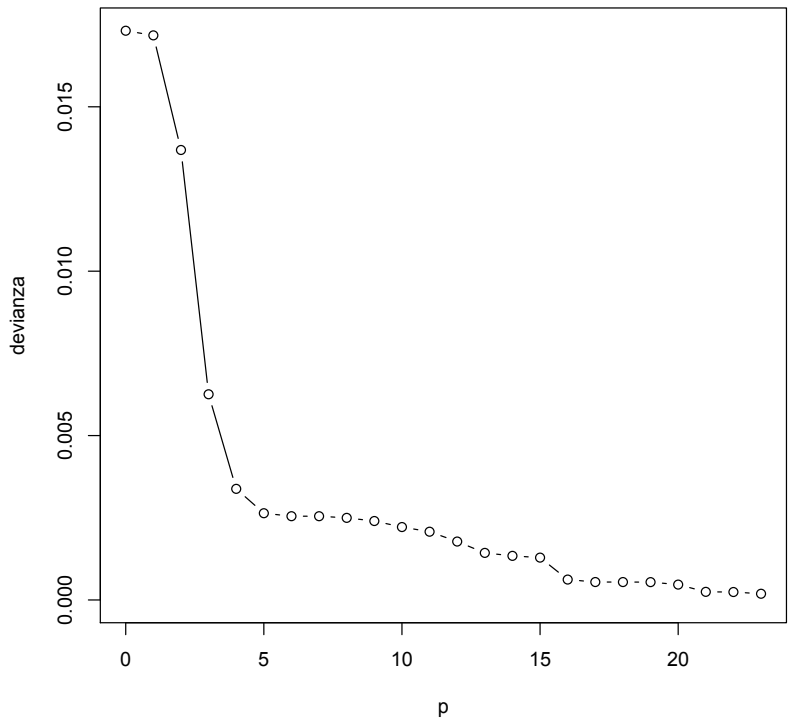

dati e polinomio di grado 6
y.ieri
0.44
0.46
0.48
0.50
0.52
0.54
0.5
1.0
1.5
2.0
2.5
3.0

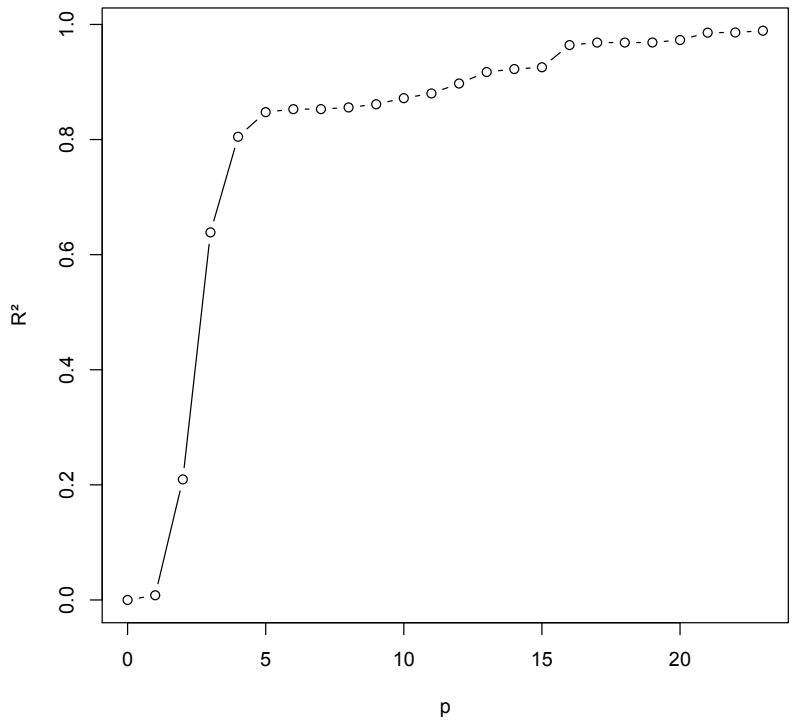

dati e polinomio di grado 12
y.ieri
0.44
0.46
0.48
0.50
0.52
0.54
0.5
1.0
1.5
2.0
2.5
3.0


dati e polinomio di grado 18
y.ieri
0.44
0.46
0.48
0.50
0.52
0.54
0.5
1.0
1.5
2.0
2.5
3.0

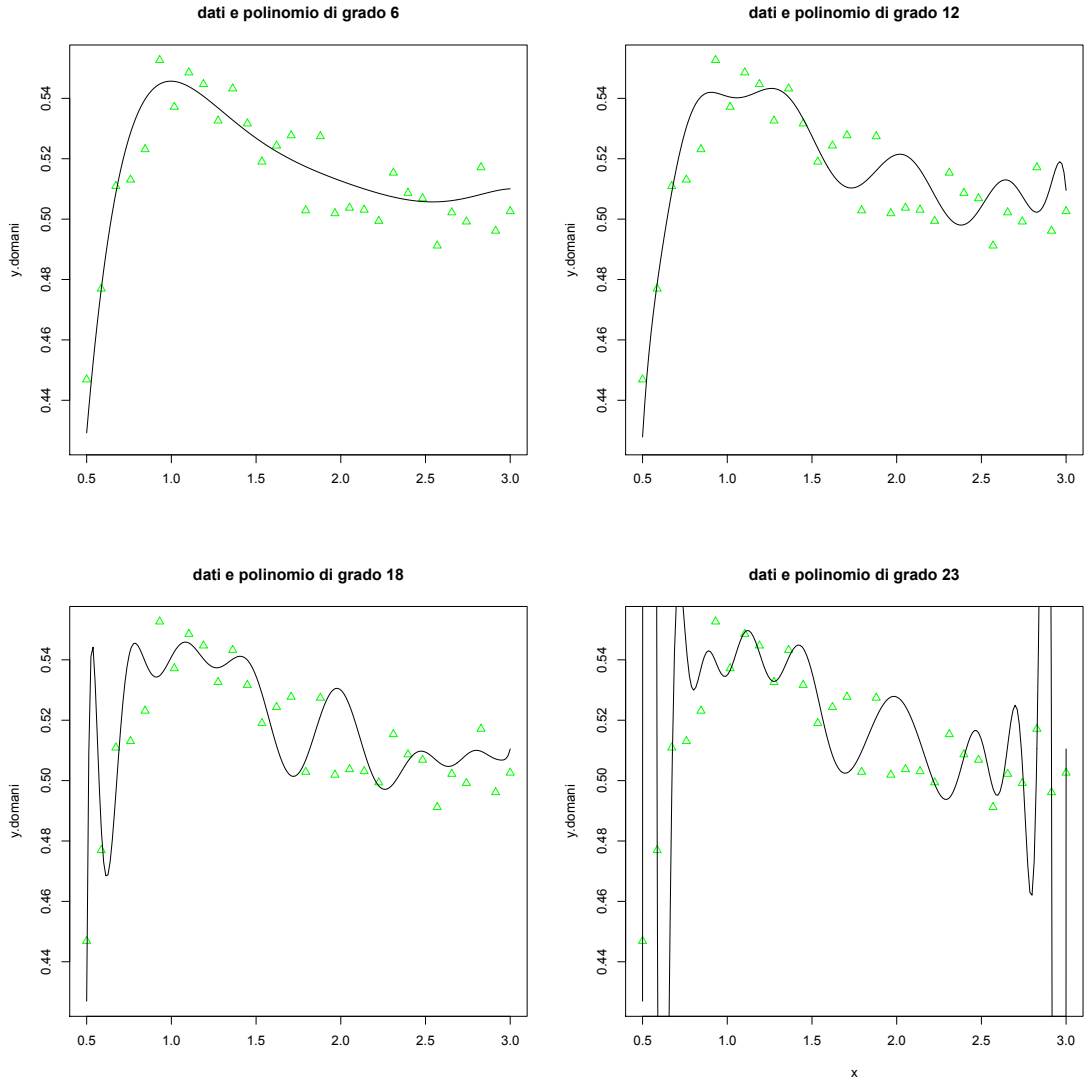

dati e polinomio di grado 23
y.ieri
0.44
0.46
0.48
0.50
0.52
0.54
0.5
1.0
1.5
2.0
2.5
3.0
x

Al crescere di p l'adattamento ai punti migliora. Il tutto é confermato anche dai due grafici seguenti:

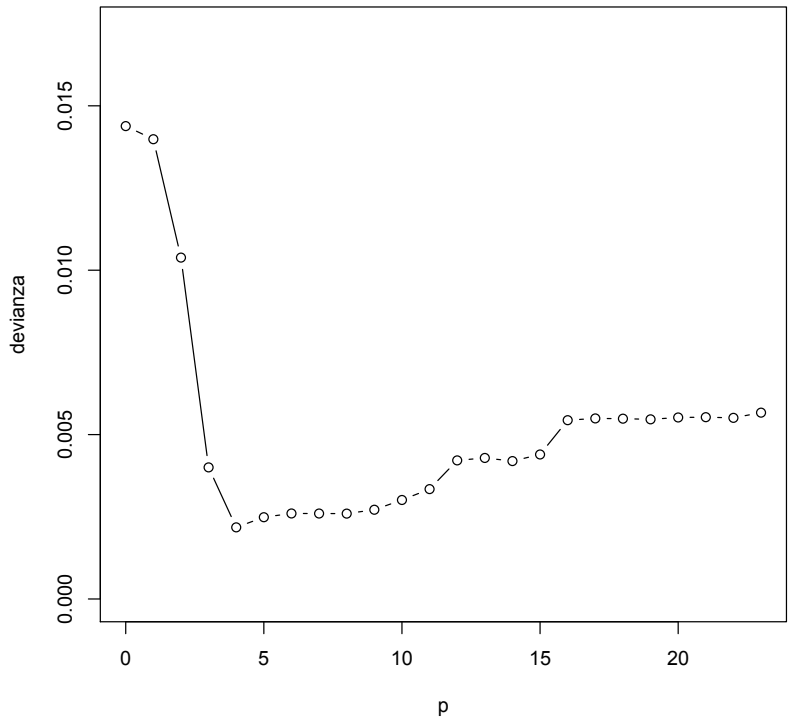


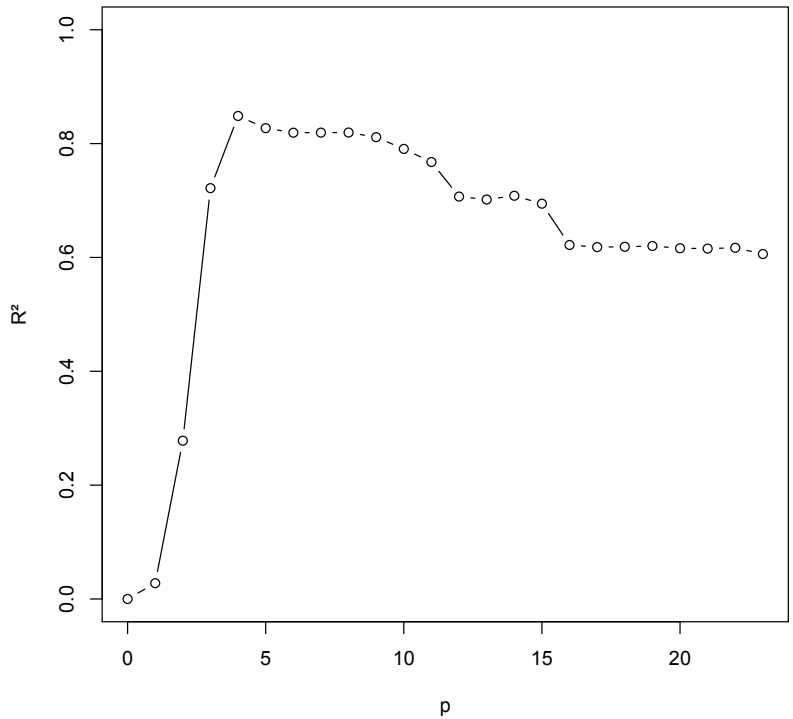


L'obiettivo é peró quello di predire dati futuri. Cosa succede usando i polinomi calcolati in precedenza?



Cosa succede ora alla devianza residua e al R^2?





Se si conoscesse $f(x)$...

L'obiettivo é quello di stimare $f(x)$ usando un generico stimatore

$$\hat{y} = \hat{f}(x)$$

che nell'esempio precedente puó essere uno dei polinomi utilizzati.
Come procedere?
Si puó iniziare considerando uno specifico valore x' per x.
Se si conoscesse completamente il meccanismo generatore dei dati, cioé anche $f(x')$ si potrebbero calcolare delle quantitá di interesse relative alla qualitá dello stimatore $\hat{y}$. Un importante indice della bontá di una stima é dato dall'errore quadratico medio:

$$E\{[\hat{y} - f(x')]^2\} = [E\{\hat{y}\} - f(x')]^2 + var\{\hat{y}\}$$

Se si conoscesse $f(x)$...

L'obiettivo é quello di stimare $f(x)$ usando un generico stimatore

$$\hat{y} = \hat{f}(x)$$

che nell'esempio precedente puó essere uno dei polinomi utilizzati.
Come procedere?
Si puó iniziare considerando uno specifico valore x' per x.
Se si conoscesse completamente il meccanismo generatore dei dati, cioé anche $f(x')$ si potrebbero calcolare delle quantitá di interesse relative alla qualitá dello stimatore $\hat{y}$. Un importante indice della bontá di una stima é dato dall'errore quadratico medio:

$$E\{[\hat{y} - f(x')]^2\} = [E\{\hat{y}\} - f(x')]^2 + var\{\hat{y}\}$$

Se si conoscesse $f(x)$...

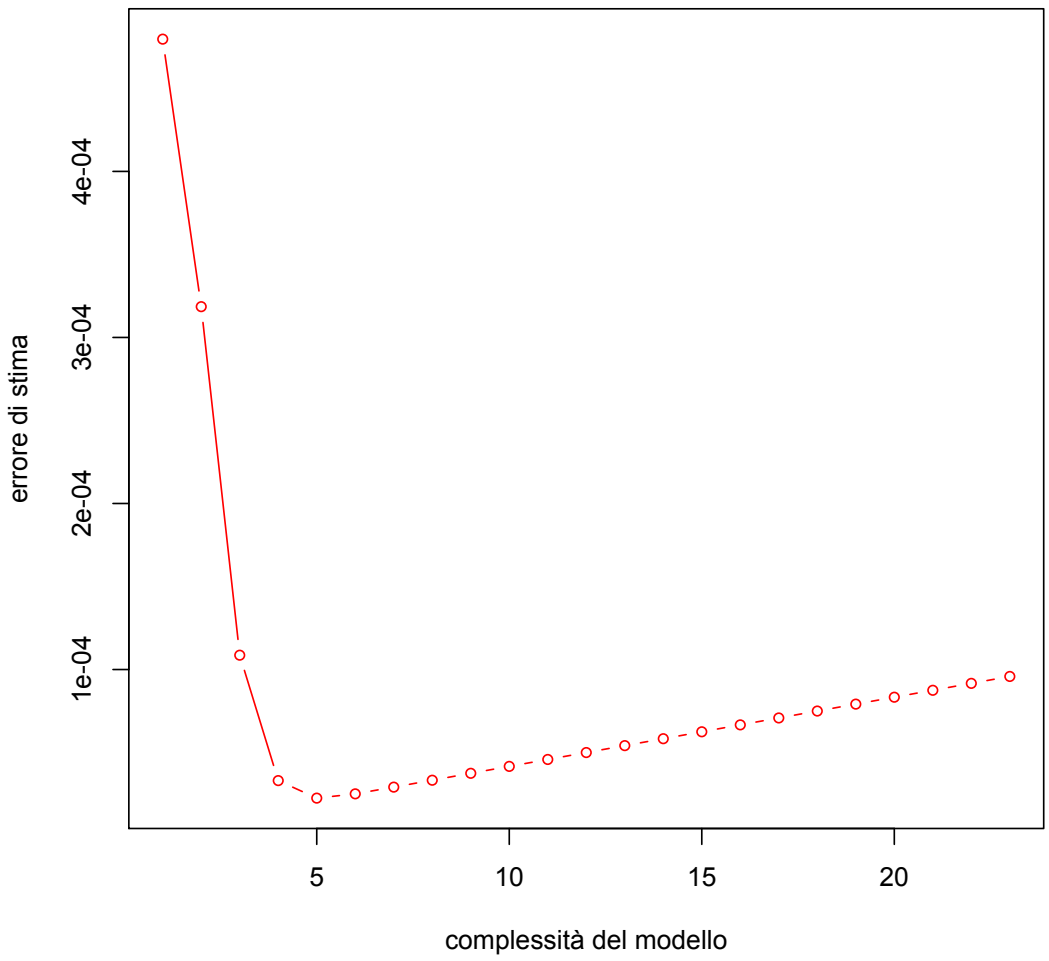
L'obiettivo é quello di stimare $f(x)$ usando un generico stimatore

$$\hat{y} = \hat{f}(x)$$

che nell'esempio precedente puó essere uno dei polinomi utilizzati.
Come procedere?
Si puó iniziare considerando uno specifico valore x' per x.
Se si conoscesse completamente il meccanismo generatore dei dati, cioé anche $f(x')$ si potrebbero calcolare delle quantitá di interesse relative alla qualitá dello stimatore $\hat{y}$. Un importante indice della bontá di una stima é dato dall'errore quadratico medio:

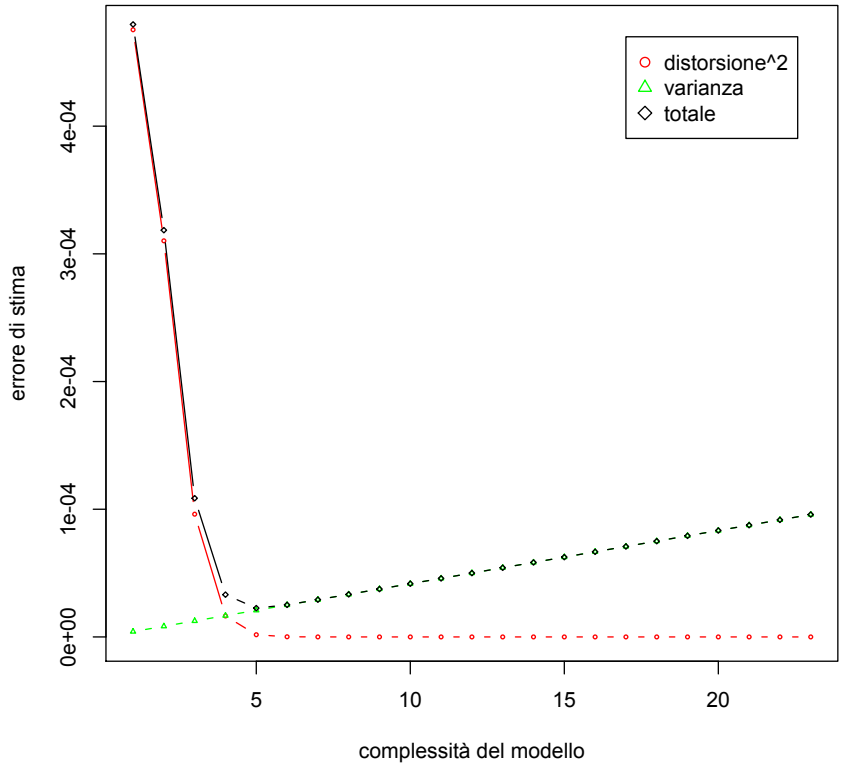
$$E\{[\hat{y} - f(x')]^2\} = [E\{\hat{y}\} - f(x')]^2 + var\{\hat{y}\}$$

Non si é interessati ovviamente solo ad un punto x' e quindi si calcola la somma degli errori quadratici relativi a tutti gli n valori di x. Rappresentando il valore risultante in funzione di p si ottiene:



In forma generale, non solo per i polinomi, si puó scrivere:

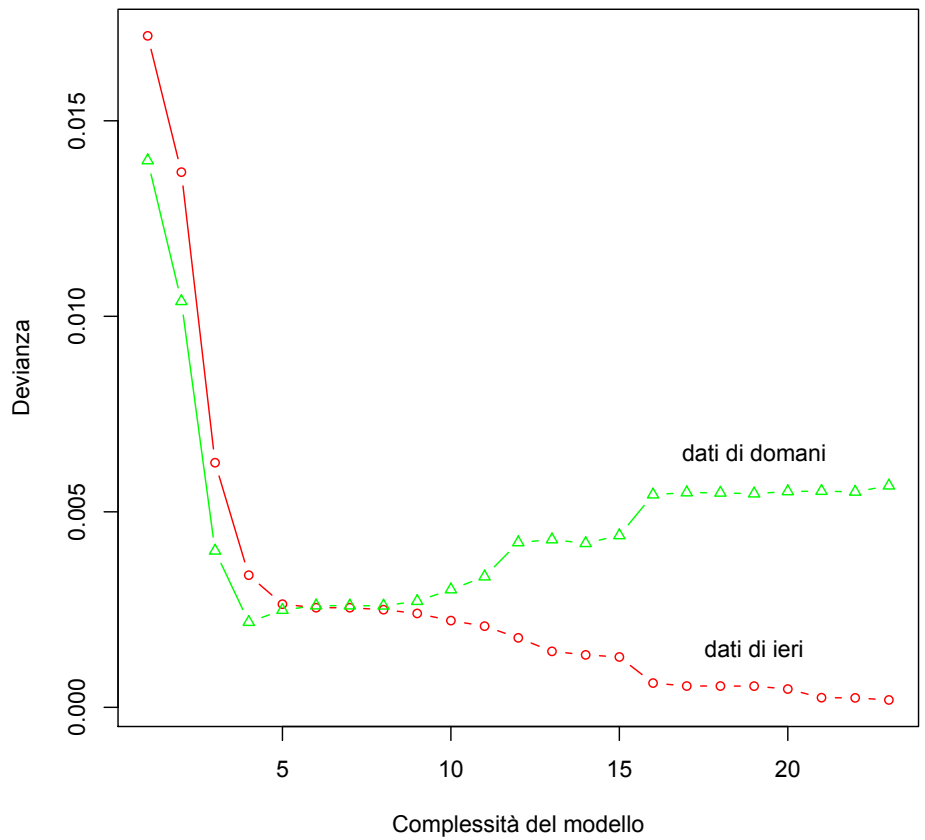
$$E\{[\hat{y} - f(x')]^2\} = distorsione^2 + varianza$$



Distorsione e varianza sono entitá in conflitto, non possono essere minimizzate congiuntamente: situazione ancora piú evidente se la complessitá del modello cresce.

In realtá non conosciamo $f(x)\ldots$

Cosa fare? La prima cosa da evitare é il sovradattamento ai dati. Problema risolvibile con i "dati di ieri" e "di oggi" nell'esempio precedente.



Insieme di stima e di prova

Una parte dei dati viene usata per la stima dei vari modelli candidati, mentre la restante parte (insieme di prova o di verifica) é usata per valutare le loro prestazioni e scegliere quindi quello preferibile.

Un possibile problema é che questa procedura abbassa ovviamente la numerositá del campione sul quale si effettua la stima. Questa problematica peró non é importante nell'ambito del data mining, dove si ha a disposizione una grossa mole di dati.

Quando i dati abbondano si potrebbe usare anche un terzo insieme di dati, quello che viene definito di controprova.

Stima	Prova	Controprova
50%	25%	25%
75%	25%	

Insieme di stima e di prova

Una parte dei dati viene usata per la stima dei vari modelli candidati, mentre la restante parte (insieme di prova o di verifica) é usata per valutare le loro prestazioni e scegliere quindi quello preferibile.

Un possibile problema é che questa procedura abbassa ovviamente la numerositá del campione sul quale si effettua la stima. Questa problematica peró non é importante nell'ambito del data mining, dove si ha a disposizione una grossa mole di dati.

Quando i dati abbondano si potrebbe usare anche un terzo insieme di dati, quello che viene definito di controprova.

Stima	Prova	Controprova
50%	25%	25%
75%	25%	

Insieme di stima e di prova

Una parte dei dati viene usata per la stima dei vari modelli candidati, mentre la restante parte (insieme di prova o di verifica) é usata per valutare le loro prestazioni e scegliere quindi quello preferibile.

Un possibile problema é che questa procedura abbassa ovviamente la numerositá del campione sul quale si effettua la stima. Questa problematica peró non é importante nell'ambito del data mining, dove si ha a disposizione una grossa mole di dati.

Quando i dati abbondano si potrebbe usare anche un terzo insieme di dati, quello che viene definito di controprova.

Stima	Prova	Controprova
50%	25%	25%
75%	25%	

Insieme di stima e di prova

Una parte dei dati viene usata per la stima dei vari modelli candidati, mentre la restante parte (insieme di prova o di verifica) é usata per valutare le loro prestazioni e scegliere quindi quello preferibile.

Un possibile problema é che questa procedura abbassa ovviamente la numerositá del campione sul quale si effettua la stima. Questa problematica peró non é importante nell'ambito del data mining, dove si ha a disposizione una grossa mole di dati.

Quando i dati abbondano si potrebbe usare anche un terzo insieme di dati, quello che viene definito di controprova.

Stima	Prova	Controprova
50%	25%	25%
75%	25%	

La convalida incrociata

Come migliorare l'accuratezza?

Bisogna svincolarsi dalla scelta di quel 75% e 25%.
Per superare questa problematica, una possibilitá é quellla di dividere l'insieme in 4 parti e usare a 'rotazione' una parte per la verifica e le altre tre per la stima.

In tal modo si avrebbero 4 stime diverse da combinare insieme. Lo stesso per le figure ottenute in precedenza, da cui si potrebbe riuscire ad ottenere una 'curva media'.

La convalida incrociata

Come migliorare l'accuratezza?

Bisogna svincolarsi dalla scelta di quel 75% e 25%.
Per superare questa problematica, una possibilitá é quellla di dividere l'insieme in 4 parti e usare a 'rotazione' una parte per la verifica e le altre tre per la stima.

In tal modo si avrebbero 4 stime diverse da combinare insieme. Lo stesso per le figure ottenute in precedenza, da cui si potrebbe riuscire ad ottenere una 'curva media'.

La convalida incrociata

Come migliorare l'accuratezza?

Bisogna svincolarsi dalla scelta di quel 75% e 25%.
Per superare questa problematica, una possibilitá é quellla di dividere l'insieme in 4 parti e usare a 'rotazione' una parte per la verifica e le altre tre per la stima.

In tal modo si avrebbero 4 stime diverse da combinare insieme. Lo stesso per le figure ottenute in precedenza, da cui si potrebbe riuscire ad ottenere una 'curva media'.

La convalida incrociata

Il procedimento diventa ancor piú accurato se invece di ottenere 4 porzioni di dimensione $n/4$, si decide di ottenere k porzioni di dimensione n/k, con k elevato, al massimo pari a n.
Nel caso estremo si utilizzano $n-1$ dati per la stima e la restante osservazione per la verifica.

PROCEDURA

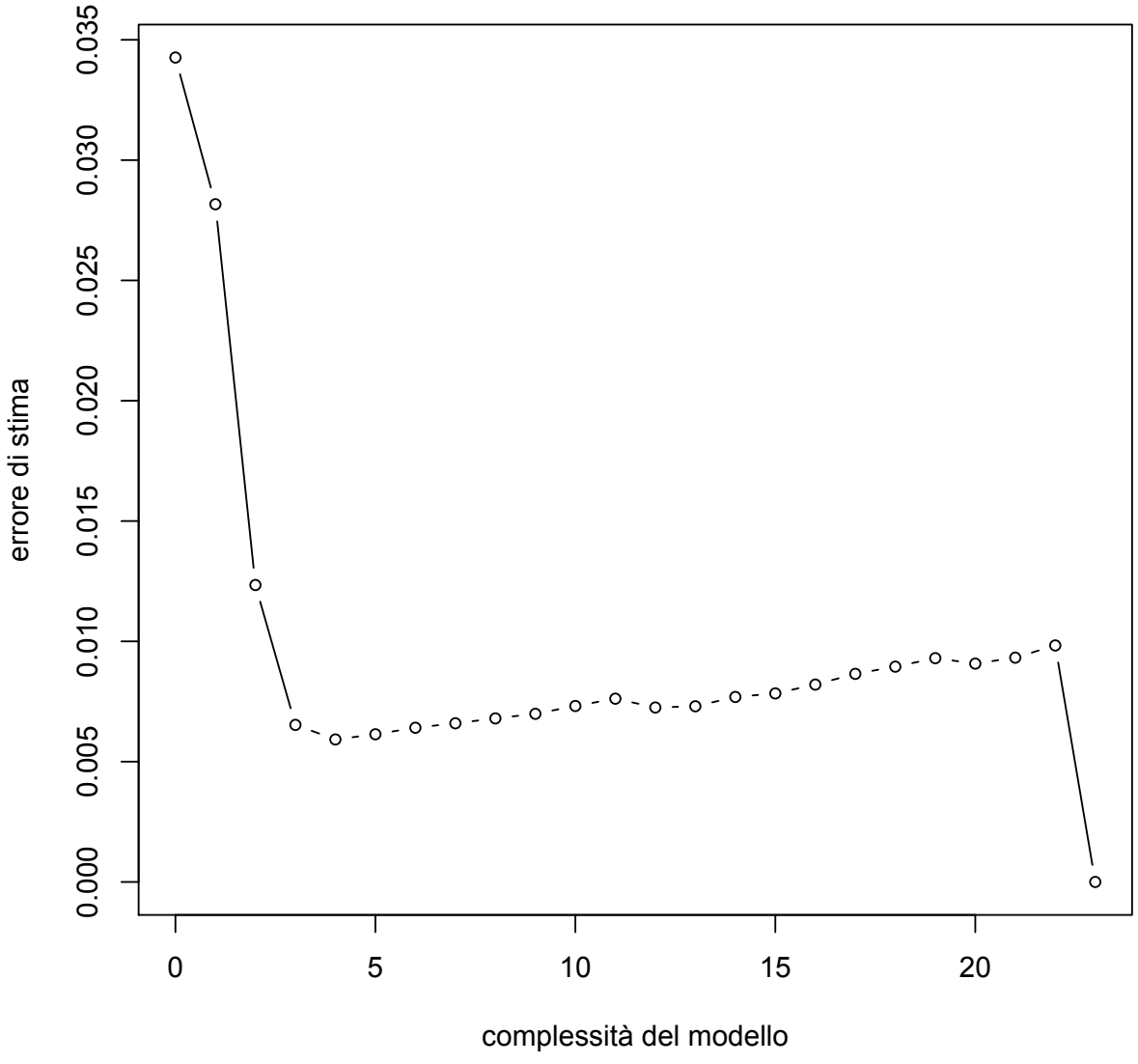
1. Far variare p da 1 al massimo;
2. Far variare i da 1 a n;
3. stimare il modello di grado p eliminando il dato i-mo;
4. ottenere la previsione $\hat{y}_{-1}$ per y_i, in corrispondenza di x_i;
5. calcolare l'errore $e_i = (y_i - \hat{y}_{-1}$
6. calcolare $D^*(p) = \sum_{i=1}^{n} e_i^2$
7. scegliere il valore di p per cui $D^*(p)$ é minimo

La convalida incrociata

Il procedimento diventa ancor piú accurato se invece di ottenere 4 porzioni di dimensione $n/4$, si decide di ottenere k porzioni di dimensione n/k, con k elevato, al massimo pari a n.
Nel caso estremo si utilizzano $n-1$ dati per la stima e la restante osservazione per la verifica.

PROCEDURA

1. Far variare p da 1 al massimo;
2. Far variare i da 1 a n;
3. stimare il modello di grado p eliminando il dato i-mo;
4. ottenere la previsione $\hat{y}_{-1}$ per y_i, in corrispondenza di x_i;
5. calcolare l'errore $e_i = (y_i - \hat{y}_{-1}$
6. calcolare $D^*(p) = \sum_{i=1}^{n} e_i^2$
7. scegliere il valore di p per cui $D^*(p)$ é minimo


errore di stima
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0
5
10
15
20
complessità del modello

Criteri basati sull'informazione

Il procedimento statistico principe per stimare un modello é quello di massimizzare la log- verosimiglianza A volte peró non é sufficiente, soprattutto quando si sceglie fra molti modelli alternativi. Bisogna tener conto del diverso numero di parametri introducendo una penalizzazione.

Una famiglia di criteri che rispetta questa logica é riconducibile a funzioni obiettivo del seguente tipo.

$$IC = -2 log L(\hat{\Theta}) + penalitá(p)$$

La scelta del tipo di penalitá identifica un particolare criterio

Criteri basati sull'informazione

Il procedimento statistico principe per stimare un modello é quello di massimizzare la log- verosimiglianza A volte peró non é sufficiente, soprattutto quando si sceglie fra molti modelli alternativi. Bisogna tener conto del diverso numero di parametri introducendo una penalizzazione.

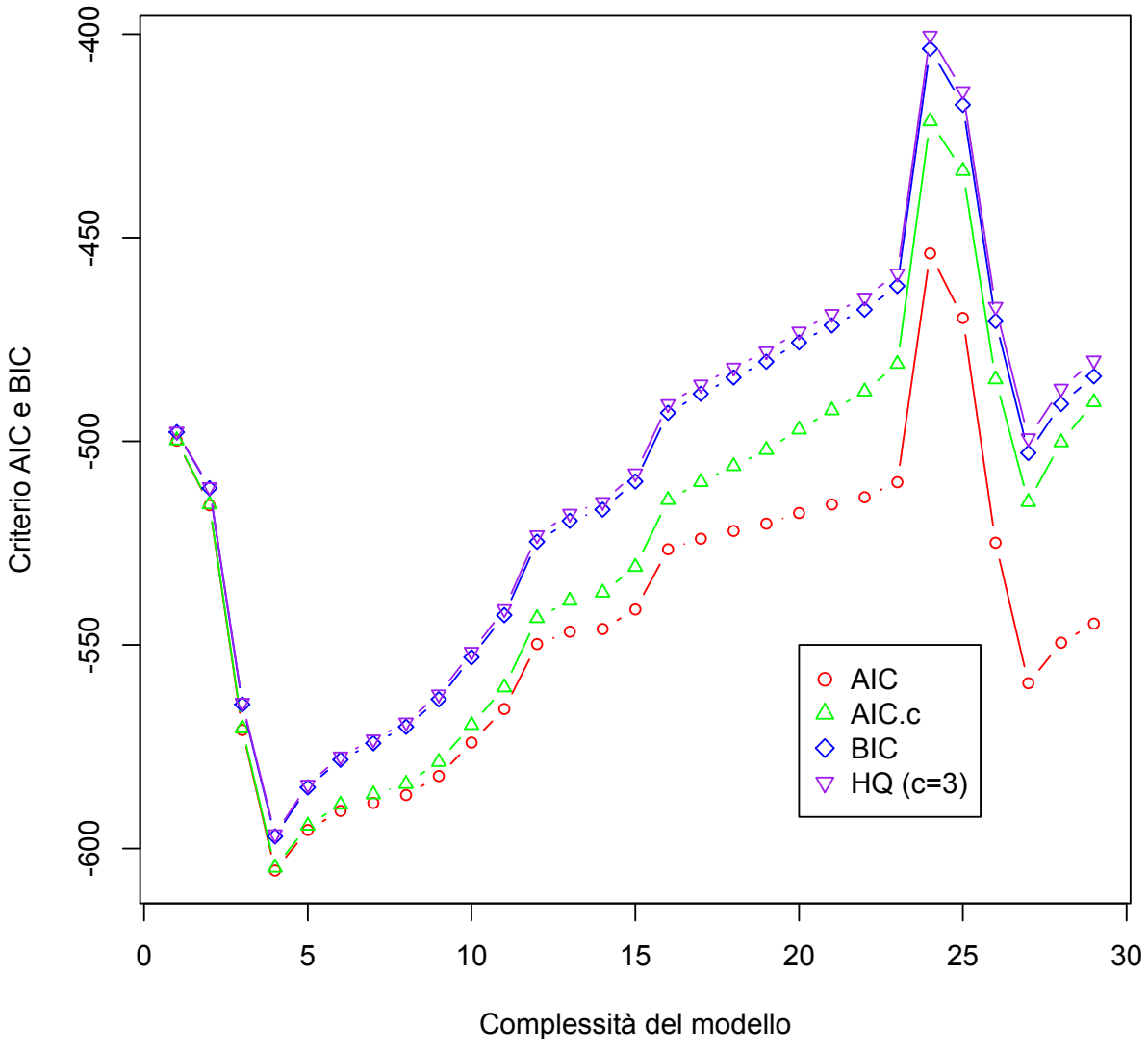
Una famiglia di criteri che rispetta questa logica é riconducibile a funzioni obiettivo del seguente tipo.

$$IC = -2logL(\hat{\Theta}) + penalitá(p)$$

La scelta del tipo di penalitá identifica un particolare criterio

Alcune possibili penalizzazioni

Criterio	Autore	Penalitá(p)
AIC	Akaike	2p
AIC_c	Sugiura, Hurvich-Tsay	$2p + \frac{2p(p+1)}{n-(p+1)}$
BIC/SIC	Akaike, Schwarz	p log n
HQ	Hannan-Quinn	c p log log n


Criterio AIC e BIC
-400
-450
-500
-550
-600
0
5
10
15
20
25
30
Complessità del modello
AIC
AIC.c
BIC
HQ (c=3)