# Stime non parametriche

# Perché?

Una strada alternativa all'uso di modelli lineari o di altre funzioni parametriche é quella di non usare alcuna formulazione parametrica di $f$, ma di stimarla in modo non parametrico, cioé senza assumere che appartenga ad una specifica classe di funzioni e assumendo solo alcune condizioni di regolaritá nel suo andamento.

Tale impostazione risulta efficace soprattutto quando si dispone di una considerevole massa di dati. Infatti in questi casi si ha quasi sempre abbastanza evidenza empirica per 'falsificare' il modello, tranne nel caso in cui non si tratti del modello 'vero', cosa che si verifica raramente.

Questo perché si cerca di riassumere tutta l'informazione dei dati in un numero ristretto di parametri. Tale difficoltá deve essere gestita con strumenti che garantiscono maggiore flessibilitá. Le possibilitá sono diverse e ancora una volta bisogna scegliere la migliore possibile.

# Perché?

Una strada alternativa all'uso di modelli lineari o di altre funzioni parametriche é quella di non usare alcuna formulazione parametrica di $f$, ma di stimarla in modo non parametrico, cioé senza assumere che appartenga ad una specifica classe di funzioni e assumendo solo alcune condizioni di regolaritá nel suo andamento.

Tale impostazione risulta efficace soprattutto quando si dispone di una considerevole massa di dati. Infatti in questi casi si ha quasi sempre abbastanza evidenza empirica per 'falsificare' il modello, tranne nel caso in cui non si tratti del modello 'vero', cosa che si verifica raramente.

Questo perché si cerca di riassumere tutta l'informazione dei dati in un numero ristretto di parametri. Tale difficoltá deve essere gestita con strumenti che garantiscono maggiore flessibilitá. Le possibilitá sono diverse e ancora una volta bisogna scegliere la migliore possibile.

# Perché?

Una strada alternativa all'uso di modelli lineari o di altre funzioni parametriche é quella di non usare alcuna formulazione parametrica di $f$, ma di stimarla in modo non parametrico, cioé senza assumere che appartenga ad una specifica classe di funzioni e assumendo solo alcune condizioni di regolaritá nel suo andamento.

Tale impostazione risulta efficace soprattutto quando si dispone di una considerevole massa di dati. Infatti in questi casi si ha quasi sempre abbastanza evidenza empirica per 'falsificare' il modello, tranne nel caso in cui non si tratti del modello 'vero', cosa che si verifica raramente.

Questo perché si cerca di riassumere tutta l'informazione dei dati in un numero ristretto di parametri. Tale difficoltá deve essere gestita con strumenti che garantiscono maggiore flessibilitá. Le possibilitá sono diverse e ancora una volta bisogna scegliere la migliore possibile.

# Regressione Locale

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

$\varepsilon$ termine di errore casuale non osservato e si puó sempre assumere che $E[\varepsilon] = 0$ perché un eventuale valore non nullo puó essere inserito in $f(x)$. In questo caso non si presume che $f$ appartenga a una specificata famiglia parametrica.

Si considera, ad esempio, un generico ma fissato punto $x_0$ dei numeri reali. Al primo step si cerca di stimare $f(x)$ solo in corrispondenza di questo valore $x_0$

# Regressione Locale

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

$\varepsilon$ termine di errore casuale non osservato e si puó sempre assumere che $E[\varepsilon] = 0$ perché un eventuale valore non nullo puó essere inserito in $f(x)$. In questo caso non si presume che $f$ appartenga a una specificata famiglia parametrica.

Si considera, ad esempio, un generico ma fissato punto $x_0$ dei numeri reali. Al primo step si cerca di stimare $f(x)$ solo in corrispondenza di questo valore $x_0$

Se $f(x)$ é una funzione derivabile con derivata continua in $x_0$ allora considerando lo sviluppo in serie di Taylor, $f(x)$ é localmente approssimabile con una retta passante per $(x_0, f(x_0))$ cioé:

$$f(x) = f(x_0) + (f'(x_0))(x - x_0) + \textit{resto}$$

$$f(x) = \alpha + \beta(x - x_0) + \textit{resto}$$

dove il resto é una quantitá di grandezza inferiore a $|x - x_0|$.

Tale risultato ci dice che una qualunque funzione $f(x)$ sufficientemente regolare puó essere approssimata localmente da una retta.
Trasferendo tale concetto nell'ambito della stima statistica, si cerca di stimare $f(x)$ in un intorno di $x_0$ attraverso un criterio, sulla base di $n$ coppie di osservazioni $(x_i, y_i)$.

Se $f(x)$ é una funzione derivabile con derivata continua in $x_0$ allora considerando lo sviluppo in serie di Taylor, $f(x)$ é localmente approssimabile con una retta passante per $(x_0, f(x_0))$ cioé:

$$f(x) = f(x_0) + (f'(x_0))(x - x_0) + \textit{resto}$$

$$f(x) = \alpha + \beta(x - x_0) + \textit{resto}$$

dove il resto é una quantitá di grandezza inferiore a $|x - x_0|$.

Tale risultato ci dice che una qualunque funzione $f(x)$ sufficientemente regolare puó essere approssimata localmente da una retta.
Trasferendo tale concetto nell'ambito della stima statistica, si cerca di stimare $f(x)$ in un intorno di $x_0$ attraverso un criterio, sulla base di $n$ coppie di osservazioni $(x_i, y_i)$.

Si introduce quindi un criterio analogo a quello dei minimi quadrati ma si pesano ora le osservazioni in base alla loro distanza da $x_0$:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} [y_i - \alpha - \beta(x_i - x_0)]^2 \, w_i$$

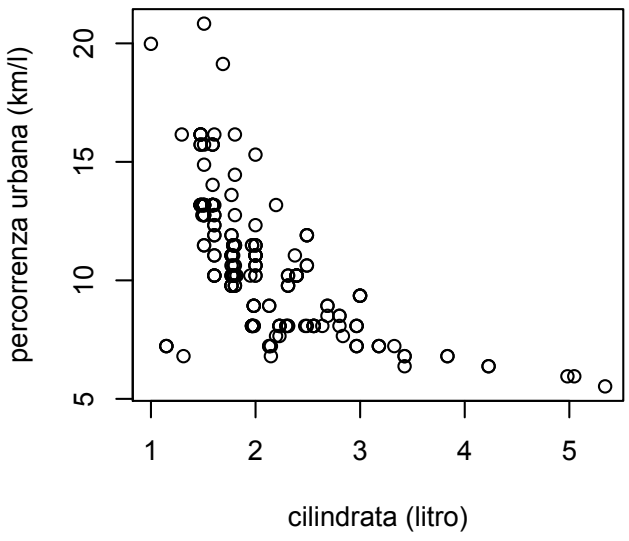
dove i pesi sono scelti in modo da essere piú alti quando $|x_i - x_0|$ é piú piccolo.
Si tratta di minimi quadrati pesati con pesi costruiti in un'ottica locale. Tale metodo é detto per questo motivo 'Regressione locale'.
Uno dei modi per scegliere i pesi é quello di porre:

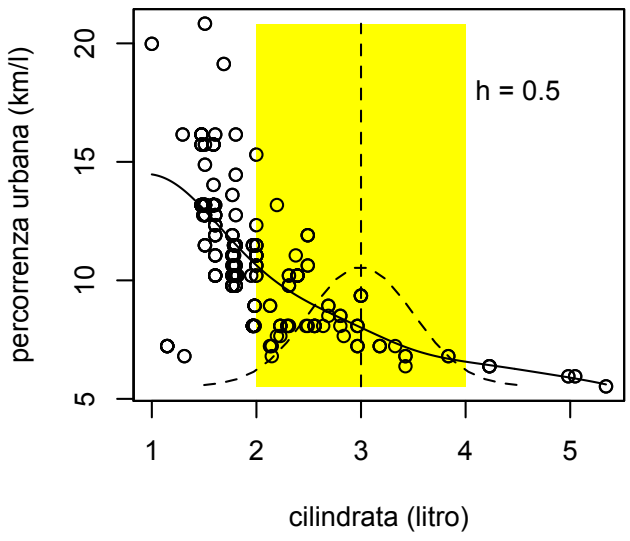
$$w_i = \frac{1}{h} w\left(\frac{x_i - x_0}{h}\right)$$

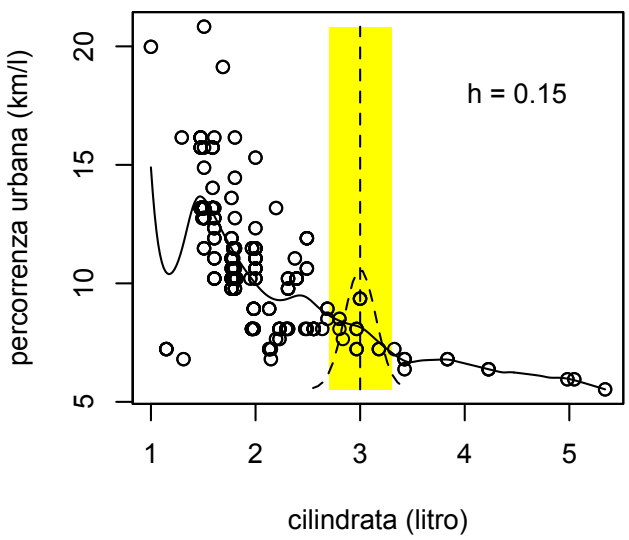
dove $w(\bullet)$ é una funzione di densitá asimmetrica attorno all'origine che in questo contesto é definita nucleo e $h > 0$ rappresenta un fattore di scala, definito *ampiezza di banda* o *parametro di lisciamento*.

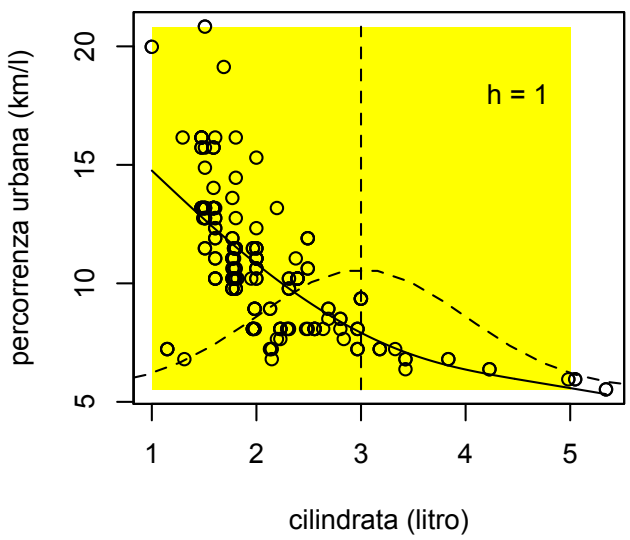
Per quanto riguarda $w$ si possono considerare le seguenti tipologie:

| nucleo | w(z) | supporto |
|---|---|---|
| normale | $\frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}z^2)$ | $\mathbb{R}$ |
| rettangolare | $1/2$ | (-1,1) |
| Epanechnikov | $\frac{3}{4}(1-z^2)$ | (-1,1) |
| biquadratico | $\frac{15}{16}(1-z^2)^2$ | (-1,1) |
| tricubico | $\frac{70}{81}(1-\lvert z\rvert^3)^3$ | (-1,1) |


percorrenza urbana (km/l)
cilindrata (litro)


h = 0.5
percorrenza urbana (km/l)
cilindrata (litro)


h = 0.15
percorrenza urbana (km/l)
cilindrata (litro)


h = 1
percorrenza urbana (km/l)
cilindrata (litro)

L'equazione per la stima dipende dai pesi che a loro volta dipendono da diversi elementi: $h, w(\bullet), x_0$. Anche fissando h e il nucleo $w(\bullet)$, il problema di minimo é comunque legato a $x_0$, mentre l'obiettivo é stimare $f(x)$ per diverse scelte di $x$. Bisogna quindi ripetere piú volte l'operazione di minimizzazione, ma non é un problema, perché si dimostra che si puóttenere:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{[a_2(x;h) - a_1(x;h)(x_i - x)]w_i y_i}{a_2(x;h)a_0(x;h) - a_1(x;h)^2}$$

dove $a_r(x;h) = [\sum (x_i - x)^r w_i]/n$ per $r = 0, 1, 2$. Si tratta di una stima non iterativa e lineare nelle $y_i$ che puó essere scritta come:

$$\hat{f}(x) = s_h^\top y$$

per un opportuno vettore $s_h$ dipendente da h, da $x$ e dalle $x_1, \ldots, x_n$ tramite le quantitá $a_r$.

In genere si stima quindi $f(x)$ non in un solo punto, ma su un intero insieme di $m$ (in genere equispaziati) che attraversano l'intervallo di interesse per la variabile $x$ e si calcolano tutte le stime secondo:

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

dove $S_h$ é ora una matrice $m \times n$ detta matrice di lisciamento, $x$ é il vettore in $\mathbb{R}^m$ delle ascisse in cui si stima la funzione e $\hat{f}(x)$ é il vettore delle stime.

# Scelta del parametro di lisciamento

Resta il problema della scelta di $h$ e di $w(\bullet)$. Per il nucleo in realtá é dimostrato che l'uso di un nucleo piuttosto che di un altro é praticamente indifferente.

La scelta importante riguarda invece $h$. Abbassando il valore di $h$ si produce una curva piú aderente al comportamento locale dei dati, poiché il sistema di pesi opera su una 'finestra' piú piccola, risentendo maggiormente della variabilitá locale. Se invece aumenta $h$ si ottiene una curva piú liscia.

Come procedere?

É necessario considerare le due seguenti quantitá:

$$E[\hat{f}(x)] \approx f(x) + \frac{h^2}{2}\sigma_w^2 f''(x)$$

$$var[\hat{f}(x)] \approx \frac{\sigma^2}{nh}\frac{\alpha(w)}{g(x)}$$

con $g(x)$) densitá da cui sono campionate le $x_i$, $\sigma_w^2 = \int z^2 w(z)dz$, $\alpha(w) = \int w(z)^2 dz$.

# Scelta del parametro di lisciamento

Resta il problema della scelta di $h$ e di $w(\bullet)$. Per il nucleo in realtá é dimostrato che l'uso di un nucleo piuttosto che di un altro é praticamente indifferente.

La scelta importante riguarda invece $h$. Abbassando il valore di $h$ si produce una curva piú aderente al comportamento locale dei dati, poiché il sistema di pesi opera su una 'finestra' piú piccola, risentendo maggiormente della variabilitá locale. Se invece aumenta $h$ si ottiene una curva piú liscia.

Come procedere?

É necessario considerare le due seguenti quantitá:

$$E[\hat{f}(x)] \approx f(x) + \frac{h^2}{2}\sigma_w^2 f''(x)$$

$$var[\hat{f}(x)] \approx \frac{\sigma^2}{nh}\frac{\alpha(w)}{g(x)}$$

con $g(x)$) densitá da cui sono campionate le $x_i$, $\sigma_w^2 = \int z^2 w(z)dz$, $\alpha(w) = \int w(z)^2 dz$.

# Scelta del parametro di lisciamento

Resta il problema della scelta di $h$ e di $w(\bullet)$. Per il nucleo in realtá é dimostrato che l'uso di un nucleo piuttosto che di un altro é praticamente indifferente.

La scelta importante riguarda invece $h$. Abbassando il valore di $h$ si produce una curva piú aderente al comportamento locale dei dati, poiché il sistema di pesi opera su una 'finestra' piú piccola, risentendo maggiormente della variabilitá locale. Se invece aumenta $h$ si ottiene una curva piú liscia.

Come procedere?

É necessario considerare le due seguenti quantitá:

$$E[\hat{f}(x)] \approx f(x) + \frac{h^2}{2}\sigma_w^2 f''(x)$$

$$var[\hat{f}(x)] \approx \frac{\sigma^2}{nh}\frac{\alpha(w)}{g(x)}$$

con $g(x)$) densitá da cui sono campionate le $x_i$, $\sigma_w^2 = \int z^2 w(z)dz$, $\alpha(w) = \int w(z)^2 dz$.

Dalle precedenti si evidenzia quindi che la distorsione é un multiplo di $h^2$ e la varianza é multipla di $\frac{1}{nh}$. Si dovrebbe quindi preferire h che tende a 0 per ridurre la distorsione, ma questo farebbe esplodere la varianza. Per $h$ tendente ad infinito succederebbe invece il contrario, quindi necessitá di un compromesso.
Cercando di minimizzare la somma di dispersione e varianza si ha:

$$h_{opt} = \left( \frac{\alpha(w)}{\sigma_w^4 f''(x)^2 g(x) n} \right)^{1/5}$$

Problema risolto?

No, perché ci sono termini non noti, come $f''$ e $g(x)$, ma informazioni importanti:

- $h$ deve tendere a 0 come $n^{-1/5}$, quindi molto lentamente
- l'errore quadratico medio invece tende a 0 con velocitá $n^{-4/5}$, mentre in un modello parametrico adeguato tale velocitá é pari a $n^{-1}$

Dalle precedenti si evidenzia quindi che la distorsione é un multiplo di $h^2$ e la varianza é multipla di $\frac{1}{nh}$. Si dovrebbe quindi preferire h che tende a 0 per ridurre la distorsione, ma questo farebbe esplodere la varianza. Per $h$ tendente ad infinito succederebbe invece il contrario, quindi necessitá di un compromesso.
Cercando di minimizzare la somma di dispersione e varianza si ha:

$$h_{opt} = \left( \frac{\alpha(w)}{\sigma_w^4 f''(x)^2 g(x) n} \right)^{1/5}$$

Problema risolto?

No, perché ci sono termini non noti, come $f''$ e $g(x)$, ma informazioni importanti:

- $h$ deve tendere a 0 come $n^{-1/5}$, quindi molto lentamente
- l'errore quadratico medio invece tende a 0 con velocitá $n^{-4/5}$, mentre in un modello parametrico adeguato tale velocitá é pari a $n^{-1}$

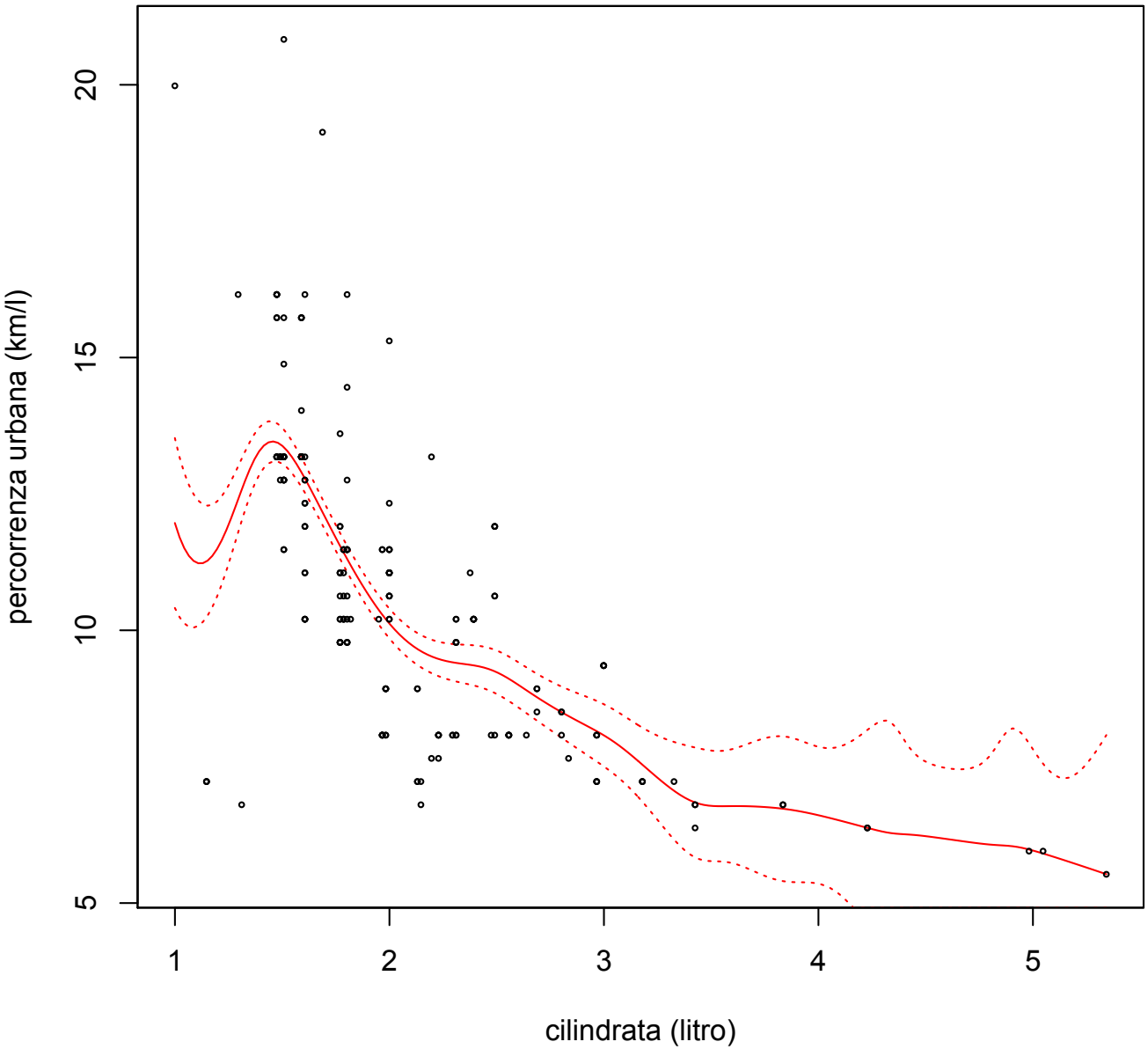
Praticamente per la stima di $h$ spesso si procede in maniera piú 'rudimentale': si usano la convalida incrociata e l'AIC.

$$AIC_c = log\hat{\sigma}^2 + 1 + \frac{2[tr(S_h) + 1]}{n - tr(S_h) - 2}$$

in cui

$$\hat{\sigma}^2 = \frac{1}{n}\sum(y_i - \hat{f}(x_i))^2 = \frac{1}{n}y^\top(I_n - S_h)^\top(I_n - S_h)y$$

é la stima della varianza residua e $tr(S_h)$ costituisce una misura sostitutiva del numero di parametri coinvolti.


percorrenza urbana (km/l)
5
10
15
20
1
2
3
4
5
cilindrata (litro)

# Bande di variabilitá

Strumenti simili agli intervalli di confidenza, necessitá di avere quindi una quantitá pivot del tipo:

$$\frac{\hat{f}(x) - f(x) - b(x)}{\sqrt{var[\hat{f}(x)]}} \sim N(0,1)$$

dove $b(x)$ indica la distorsione della stima
Poiché nella distorsione rientra anche la derivata seconda, si preferisce costruire le bande di variabilitá come segue:

$$\left(\hat{f}(x) - z_{\alpha/2} err.std(\hat{f}(x)), \hat{f}(x) + z_{\alpha/2} err.std(\hat{f}(x))\right)$$

Tali intervallo ha un livello di confidenza pari a $1-\alpha$ per ogni fissato valore di $x$ ma non per l'intera curva.

# Parametro di lisciamento variabile e *loess*

Variante alla regressione locale: ampiezza di banda non costante, ma funzione del grado di sparsitá dei punti osservati.
Si usano valori di $h$ piú elevati dove le $x_i$ sono piú disperse.
Una tecnica per ottenere tali risultati é quella del *loess*, che prevede di esprimere il parametro di lisciamento attraverso la frazione di osservazioni rilevanti per la stima ad una determinata ascissa.
Il *loess* allarga o restringe la 'finestra' in modo che la frazione di osservazioni coinvolte resti costante.

Un altro vantaggio di questa tecnica é quello di poter combinare l'idea di regressione locale con la stima robusta, sostituendo la funzione quadratica con un'altra che limiti l'effetto di osservazioni anomali.

# Parametro di lisciamento variabile e *loess*

Variante alla regressione locale: ampiezza di banda non costante, ma funzione del grado di sparsitá dei punti osservati.
Si usano valori di $h$ piú elevati dove le $x_i$ sono piú disperse.
Una tecnica per ottenere tali risultati é quella del *loess*, che prevede di esprimere il parametro di lisciamento attraverso la frazione di osservazioni rilevanti per la stima ad una determinata ascissa.
Il *loess* allarga o restringe la 'finestra' in modo che la frazione di osservazioni coinvolte resti costante.

Un altro vantaggio di questa tecnica é quello di poter combinare l'idea di regressione locale con la stima robusta, sostituendo la funzione quadratica con un'altra che limiti l'effetto di osservazioni anomali.

# Piú dimensioni

$$y = f(x_1, x_2) + \varepsilon$$

dove $f(x_1, x_2)$ é funzione da $\mathbb{R}^2$ in $\mathbb{R}$.
I dati saranno ovviamente costituiti dalle stesse $y_i$ e dai punti $(x_{i1}, x_{i2})$.
Per stimare la funzione sará necessario estendere il criterio precedente:

$$\min_{\alpha,\beta,\gamma} \sum_{i=1}^{n} [y_i - \alpha - \beta(x_{i1} - x_{01}) - \gamma(x_{i2} - x_{02})]^2 w_i$$

I pesi saranno da determinarsi in funzione di una opportuna distanza. Uno dei modi piú utilizzati é il seguente:

$$w_i = \frac{1}{h_1 h_2} w\left(\frac{x_{i1} - x_{01}}{h_1}\right) w\left(\frac{x_{i2} - x_{02}}{h_2}\right)$$

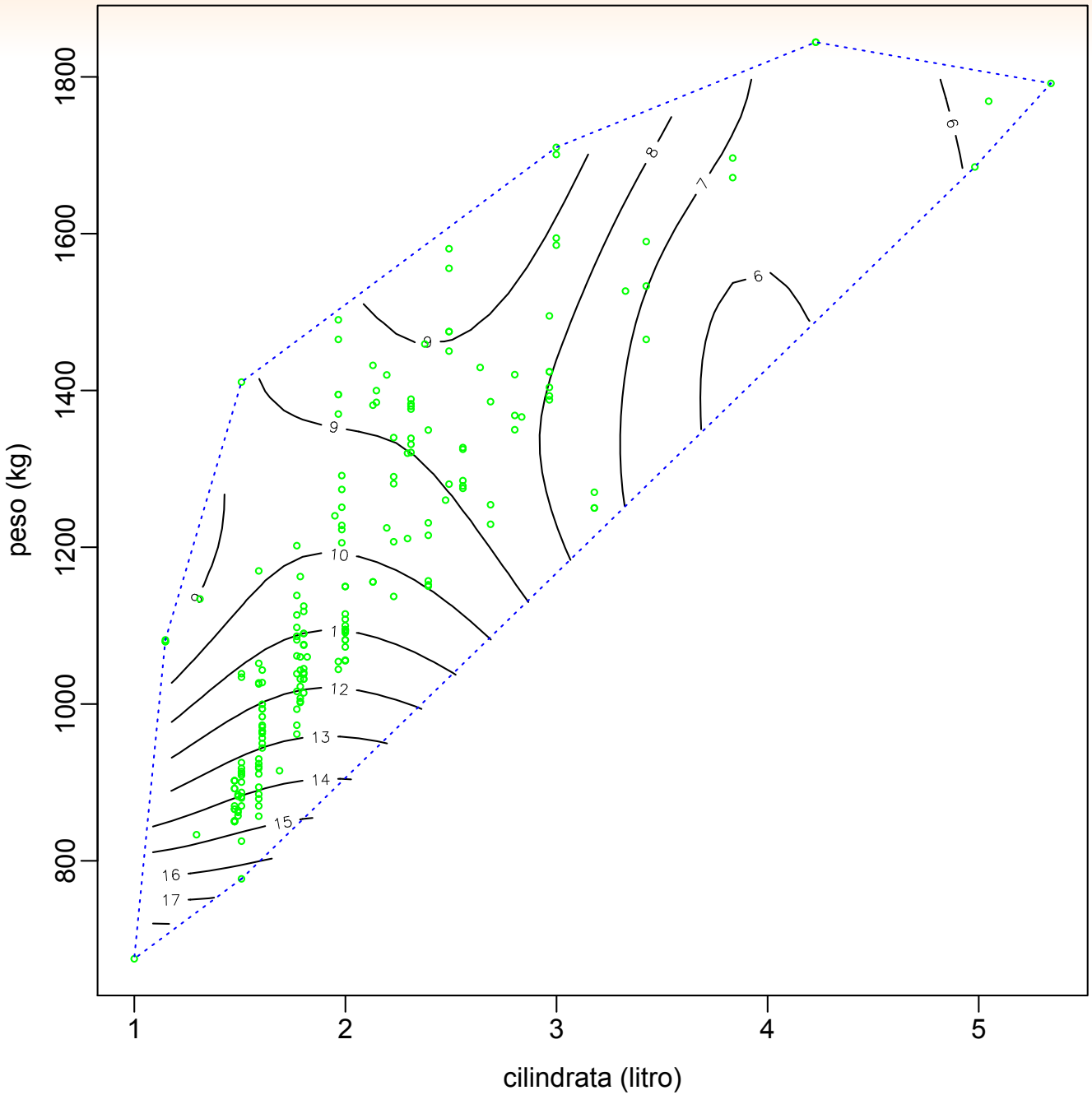
Naturalmente in questo caso abbiamo a che fare con due parametri di lisciamento, per tener conto della diversa variabilitá delle $x$ considerate.

Dal punto di vista computazionale, si tratta di una variante al problema dei minimi quadrati ponderati. Indicando con $X$ la generica matrice $(n \times 3)$ la cui i-ma riga é data da $(1, x_{i1} - x_{01}, x_{i2} - x_{02})$, $y = (y_1, \ldots, y_n)^\top$, $W = diag(w_1, \ldots, w_n)$
si ha che la soluzione del problema é il primo elemento (corrispondente ad $\alpha$) di

$$(X^\top W X)^{-1} X^\top W y$$

Il calcolo andrá ripetuto per ogni scelta di $x_0$

percorrenza urbana (km/l)
peso (kg)
cilindrata (litro)


peso (kg)
800
1000
1200
1400
1600
1800
1
2
3
4
5
cilindrata (litro)
9
10
11
12
13
14
15
16
17
8
7
6

# Caso multidimensionale

$$y = f(x) + \varepsilon = f(x_i, \ldots, x_p) + \varepsilon$$

# Maledizione della dimensionalitá

Raramente si va al di lá delle due dimensioni sia per problemi di visualizzazione grafica che per la difficoltá interpretativa del risultato.

Inoltre all'aumentare della dimensione i punti osservati si disperdono rapidamente.

Esempio:
n=500 nell'intervallo $(0, 1)$. Se utilizzati per stimare $f(x)$ in una dimensione si riescono ad ottenere stime abbastanza attendibili. Se gli stessi punti vengono distribuiti sul quadrato $(0, 1)^2$ del piano $(x_1, x_2)$ questi risultano molto meno fitti. Ovviamente la situazione si amplia se passiamo a dimensioni superiori e la dispersione aumenta molto rapidamente.

Per compensare tale aumento della dispersione sarebbe necessario avere dei campioni di dimensione $n^p$.
Ma anche nell'ambito del data-mining, considerando i 500 punti iniziali, sarebbe forse plausibile arrivare a $500^2$, ma giÃ  con 5 dimensioni ad esempio....

# Maledizione della dimensionalitá

Raramente si va al di lá delle due dimensioni sia per problemi di visualizzazione grafica che per la difficoltá interpretativa del risultato.

Inoltre all'aumentare della dimensione i punti osservati si disperdono rapidamente.

Esempio:
n=500 nell'intervallo $(0, 1)$. Se utilizzati per stimare $f(x)$ in una dimensione si riescono ad ottenere stime abbastanza attendibili. Se gli stessi punti vengono distribuiti sul quadrato $(0, 1)^2$ del piano $(x_1, x_2)$ questi risultano molto meno fitti. Ovviamente la situazione si amplia se passiamo a dimensioni superiori e la dispersione aumenta molto rapidamente.

Per compensare tale aumento della dispersione sarebbe necessario avere dei campioni di dimensione $n^p$.
Ma anche nell'ambito del data-mining, considerando i 500 punti iniziali, sarebbe forse plausibile arrivare a $500^2$, ma giÃ  con 5 dimensioni ad esempio....

Oltre ai problemi precedenti, se anche si riuscisse ad avere un campione molto numeroso, aumenterebbero i costi computazionali. Tutto ció vale non solo per la regressione locale, ma per ogni tecnica di stima non parametrica.

Per contrastare tale problema una delle strategie possibili é quella di effettuare un'operazione che riduca la dimensione delle variabili esplicative.

Oltre ai problemi precedenti, se anche si riuscisse ad avere un campione molto numeroso, aumenterebbero i costi computazionali. Tutto ció vale non solo per la regressione locale, ma per ogni tecnica di stima non parametrica.

Per contrastare tale problema una delle strategie possibili é quella di effettuare un'operazione che riduca la dimensione delle variabili esplicative.

# Spline

Il termine spline indicava delle asticciole di legno utilizzate per la progettazione degli scafi delle navi. Si fissavano alcuni punti sulla sezione trasversale dello scafo e il resto della curva veniva determinato forzando queste asticciole a passare in corrispondenza di questi punti e lasciandole invece libere di disporsi per il resto del profilo secondo la sua naturale tendenza. Si determinava quindi una curva regolare con comportamento pre-assegnato in certe posizioni.

Le funzioni matematiche definite 'spline' seguono una logica che replica il meccanismo delle navi. Si cerca di approssimare funzioni di cui si conosce il valore solo in alcuni punti o per interpolare coppie di punti.

# Spline

Il termine spline indicava delle asticciole di legno utilizzate per la progettazione degli scafi delle navi. Si fissavano alcuni punti sulla sezione trasversale dello scafo e il resto della curva veniva determinato forzando queste asticciole a passare in corrispondenza di questi punti e lasciandole invece libere di disporsi per il resto del profilo secondo la sua naturale tendenza. Si determinava quindi una curva regolare con comportamento pre-assegnato in certe posizioni.

Le funzioni matematiche definite 'spline' seguono una logica che replica il meccanismo delle navi. Si cerca di approssimare funzioni di cui si conosce il valore solo in alcuni punti o per interpolare coppie di punti.

# Funzioni spline

Sull'asse delle ascisse si scelgono $K$ punti $\xi_1 < \xi_2 < \ldots < \xi_K$ detti nodi. Si vuole costruire una funzione che passi esattamente per i nodi e che sia libera negli altri punti, purché presenti complessivamente un comportamento regolare.

Si procede nel seguente modo:

- tra due nodi successivi $(\xi_i, x_{i+1}$, la curva $f(x)$ coincide con un opportuno polinomio di grado prefissato $d$;
- si chiede che tali polinomi si congiungano 'bene' nei punti di giunzione $\xi_i \quad (i = 2, \ldots, K-1)$, nel senso che la funzione risultante $f(x)$ abbia derivate dal grado 0 al grado $d-1$ continue in ogni $\xi_i$.

In genere il grado scelto é 3 (perció si parla di spline cubiche). Il motivo di tale scelta é perché l'occhio umano non riesce a cogliere discontinuitá nella derivata terza.

# Funzioni spline

Sull'asse delle ascisse si scelgono $K$ punti $\xi_1 < \xi_2 < \ldots < \xi_K$ detti nodi. Si vuole costruire una funzione che passi esattamente per i nodi e che sia libera negli altri punti, purché presenti complessivamente un comportamento regolare.

Si procede nel seguente modo:

- tra due nodi successivi $(\xi_i, x_{i+1}$, la curva $f(x)$ coincide con un opportuno polinomio di grado prefissato $d$;
- si chiede che tali polinomi si congiungano 'bene' nei punti di giunzione $\xi_i \quad (i = 2, \ldots, K-1)$, nel senso che la funzione risultante $f(x)$ abbia derivate dal grado 0 al grado $d-1$ continue in ogni $\xi_i$.

In genere il grado scelto é 3 (perció si parla di spline cubiche). Il motivo di tale scelta é perché l'occhio umano non riesce a cogliere discontinuitá nella derivata terza.

Le precedenti condizioni si possono esplicitare come:

- $f(\xi_i) = y_i$ per $i = 1, \ldots, K$
- $f(\xi_i^-) = f(\xi_i^+)$ per $i = 2, \ldots, K-1$
- $f'(\xi_i^-) = f'(\xi_i^+)$ per $i = 2, \ldots, K-1$
- $f''(\xi_i^-) = f''(\xi_i^+)$ per $i = 2, \ldots, K-1$

dove in generale $g(x^-)$ e $g(x^+)$ indicano il limite da sinistra e da destra di una funzione $g(\bullet)$ nel punto $x$.

Tale impostazione comporta le seguenti condizioni:

- ognuna delle $K-1$ cubiche richiede 4 parametri;
- ci sono $K$ vincoli del tipo $f(\xi_i) = y_i$ e $3(K-2)$ vincoli di continuitá della funzione e delle prime due derivate.

La differenza tra coefficienti e vincoli é di 2 unitá. quindi il sistema di condizioni non identifica univocamente una funzione. C'é bisogno quindi di introdurre due vincoli aggiuntivi.
Sono state fatte diverse proposte per la definizione di tali vincoli, la maggior parte dei quali riguarda gli intervalli o i punti estremi della funzione.

Una scelta semplice consiste nel vincolare le derivate seconde dei polinomi nei due nodi estremi ad essere nulle.

Tale impostazione comporta le seguenti condizioni:

- ognuna delle $K-1$ cubiche richiede 4 parametri;
- ci sono $K$ vincoli del tipo $f(\xi_i) = y_i$ e $3(K-2)$ vincoli di continuitá della funzione e delle prime due derivate.

La differenza tra coefficienti e vincoli é di 2 unitá. quindi il sistema di condizioni non identifica univocamente una funzione. C'é bisogno quindi di introdurre due vincoli aggiuntivi.
Sono state fatte diverse proposte per la definizione di tali vincoli, la maggior parte dei quali riguarda gli intervalli o i punti estremi della funzione.

Una scelta semplice consiste nel vincolare le derivate seconde dei polinomi nei due nodi estremi ad essere nulle.

# Spline di regressione

Riprendendo

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

dove la funzione viene ipotizzata di tipo spline. Si divide l'asse delle ascisse in $K + 1$ intervalli separati da $K$ ascisse $\xi_1, \ldots, \xi_K$ detti nodi.

Si interpolano gli $n$ punti con il criterio dei minimi quadrati dove i $\beta$ sono i parametri non vincolati dei $K + 1$ polinomi costituenti.

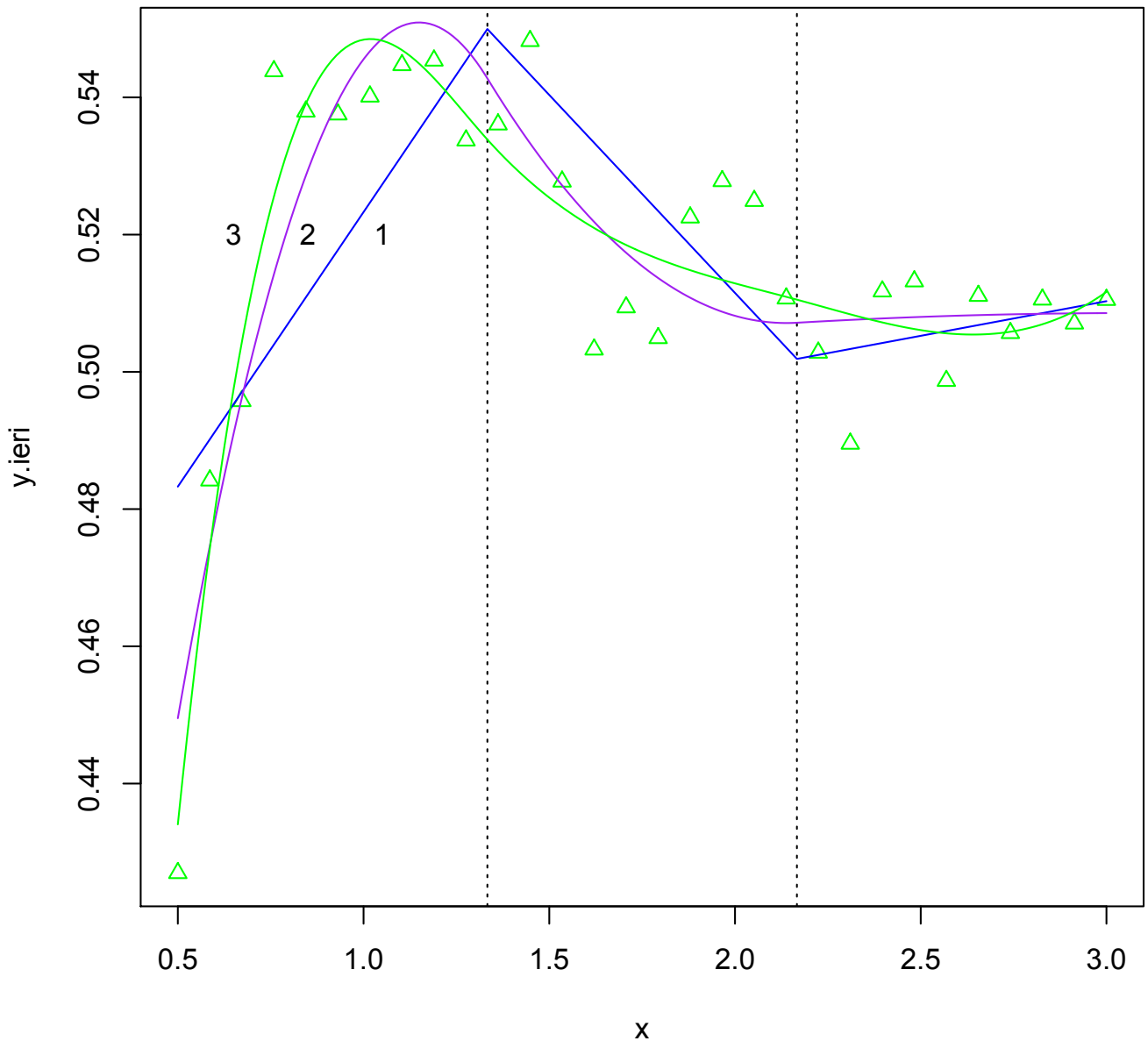
La soluzione del problema di minimo puó essere scritta come:

$$f(x;\beta) = \sum_{j=1}^{K+4} \hat{\beta}_j h_j(x)$$

dove

- $h_j(x) = x^{j-1}$ per $j = 1, \ldots, 4$
- $h_{j+4}(x) = (x - \xi_j)_+^3$ per $j = 1, \ldots, K$

La soluzione sará quindi costituita da una opportuna combinazione lineare di una base di funzioni costituita in parte da polinomi elementari e in parte da funzioni del tipo $max(0, (x - \xi)^3)$


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
3
2
1

# Spline di lisciamento

Un altro approccio é quello dei minimi quadrati penalizzati.

$$D(f, \lambda) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int_{-\infty}^{\infty} [f''(t)]^2 dt$$

dove $\lambda$ é un parametro positivo di penalizzazione del grado di irregolaritá della curva $f$ che agisce come parametro di lisciamento.

Se $\lambda = 0$ non vi é alcuna penalitá per l'irregolaritá di $f(x)$ per cui il criterio precedente non risente di $f(x)$ fuori dalle ascisse $x_1, \ldots, x_n$ e la soluzione é la media aritmetica delle $y_i$ corrispondenti a quella data ascissa, per ciascuna delle $x_i$ osservate, mentre non é determinata per gli altri valori di $x$. Se invece $\lambda \to \infty$ la penalitá é massima e comporta di adottare una retta perché si impone $f''(x) = 0$ e il risultato complessivo é la retta dei minimi quadrati.

# Spline di lisciamento

Un altro approccio é quello dei minimi quadrati penalizzati.

$$D(f, \lambda) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int_{-\infty}^{\infty} [f''(t)]^2 dt$$

dove $\lambda$ é un parametro positivo di penalizzazione del grado di irregolaritá della curva $f$ che agisce come parametro di lisciamento.

Se $\lambda = 0$ non vi é alcuna penalitá per l'irregolaritá di $f(x)$ per cui il criterio precedente non risente di $f(x)$ fuori dalle ascisse $x_1, \ldots, x_n$ e la soluzione é la media aritmetica delle $y_i$ corrispondenti a quella data ascissa, per ciascuna delle $x_i$ osservate, mentre non é determinata per gli altri valori di $x$. Se invece $\lambda \to \infty$ la penalitá é massima e comporta di adottare una retta perché si impone $f''(x) = 0$ e il risultato complessivo é la retta dei minimi quadrati.

Si dimostra che la soluzione del problema di minimizzazione é costituita da una funzione di tipo spline cubica naturale, i cui nodi sono i punti $x_i$ distinti. Si puó quindi scrivere:

$$\hat{f}(x) = \sum_{j=1}^{n_0} \theta_j N_j(x)$$

dove $n_0$ é il numero di $x_i$ distinti e gli $N_j(x)$ sono basi delle spline cubiche naturali.

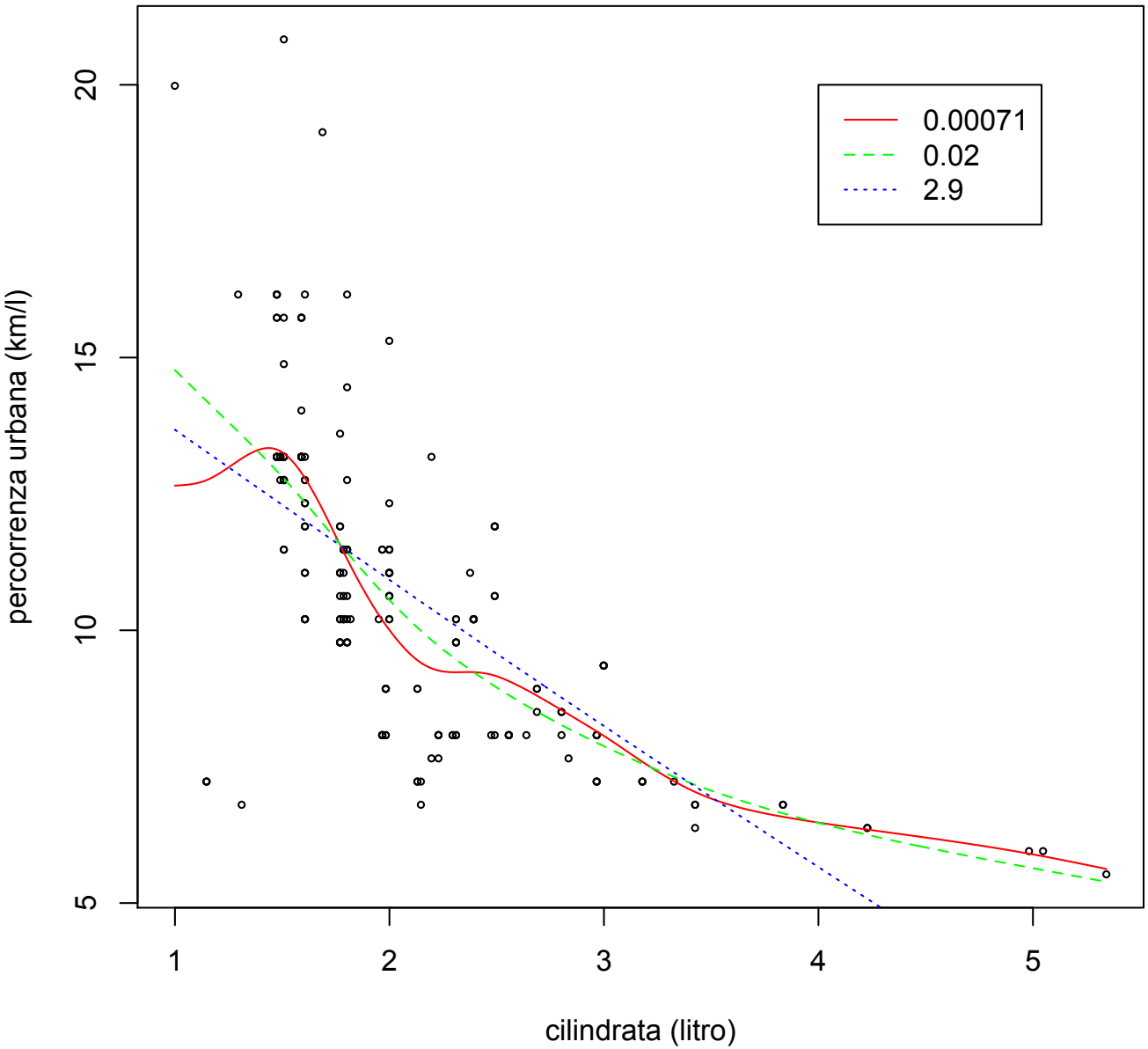
Si puó quindi riscrivere

$$D(f, \lambda) = (y - N\theta)^T (y - N\theta) + \lambda \theta^T \Omega \theta$$

con $N$ che indica la matrice la cui j-ma colonna contiene i valori di $N_j$ in corrispondenza dei valori di $x_i$ e la matrice $\Omega$ ha come generico elemento $\int N_j^{''}(t) N_k^{''}(t) dt$ La soluzione del problema di minimo sará data da:

$$\hat{\theta} = (N^T N + \lambda \Omega)^{-1} N^T y$$

che dipende dalla scelta del parametro di lisciamento $\lambda$.

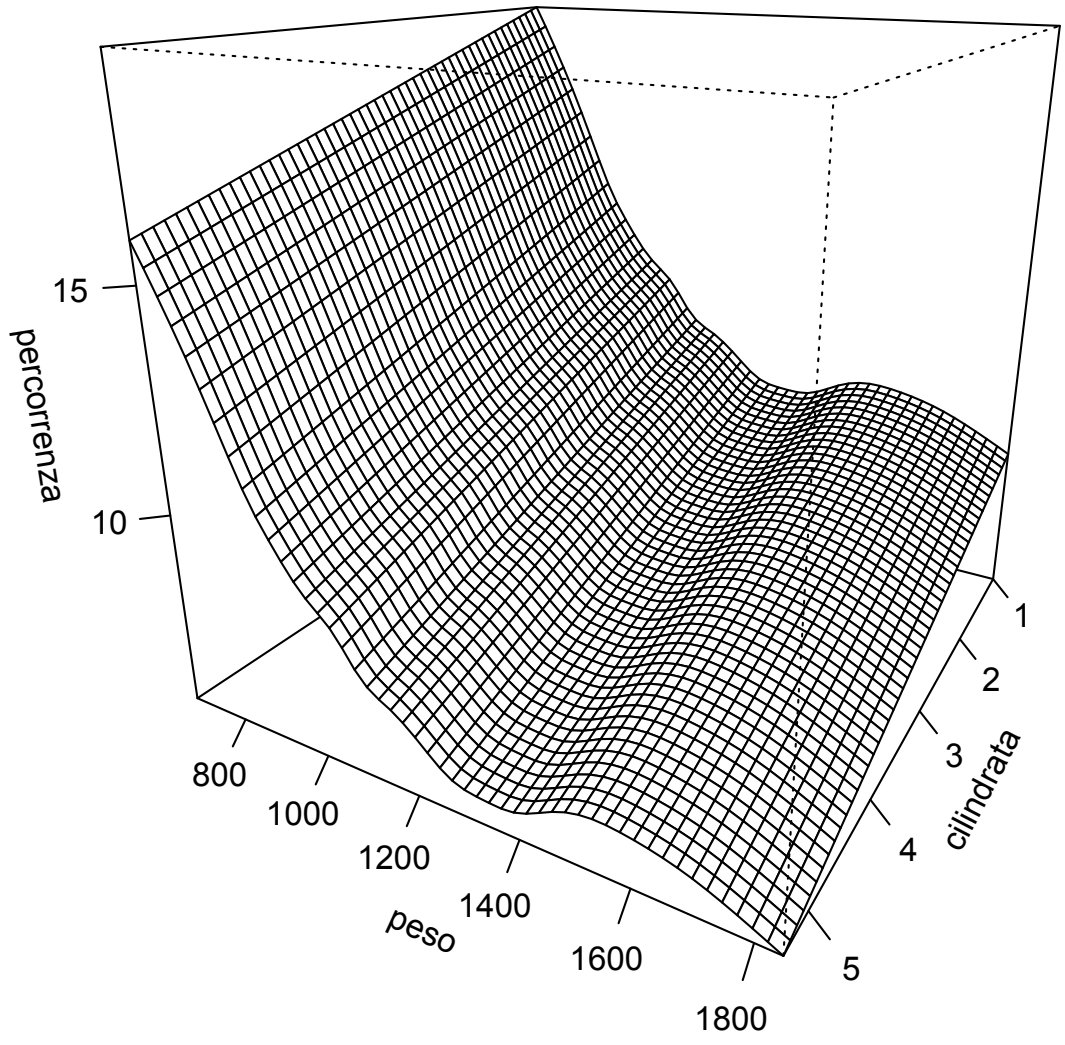

0.00071
0.02
2.9
percorrenza urbana (km/l)
cilindrata (litro)
5
10
15
20
1
2
3
4
5

# Spline a piú dimensioni

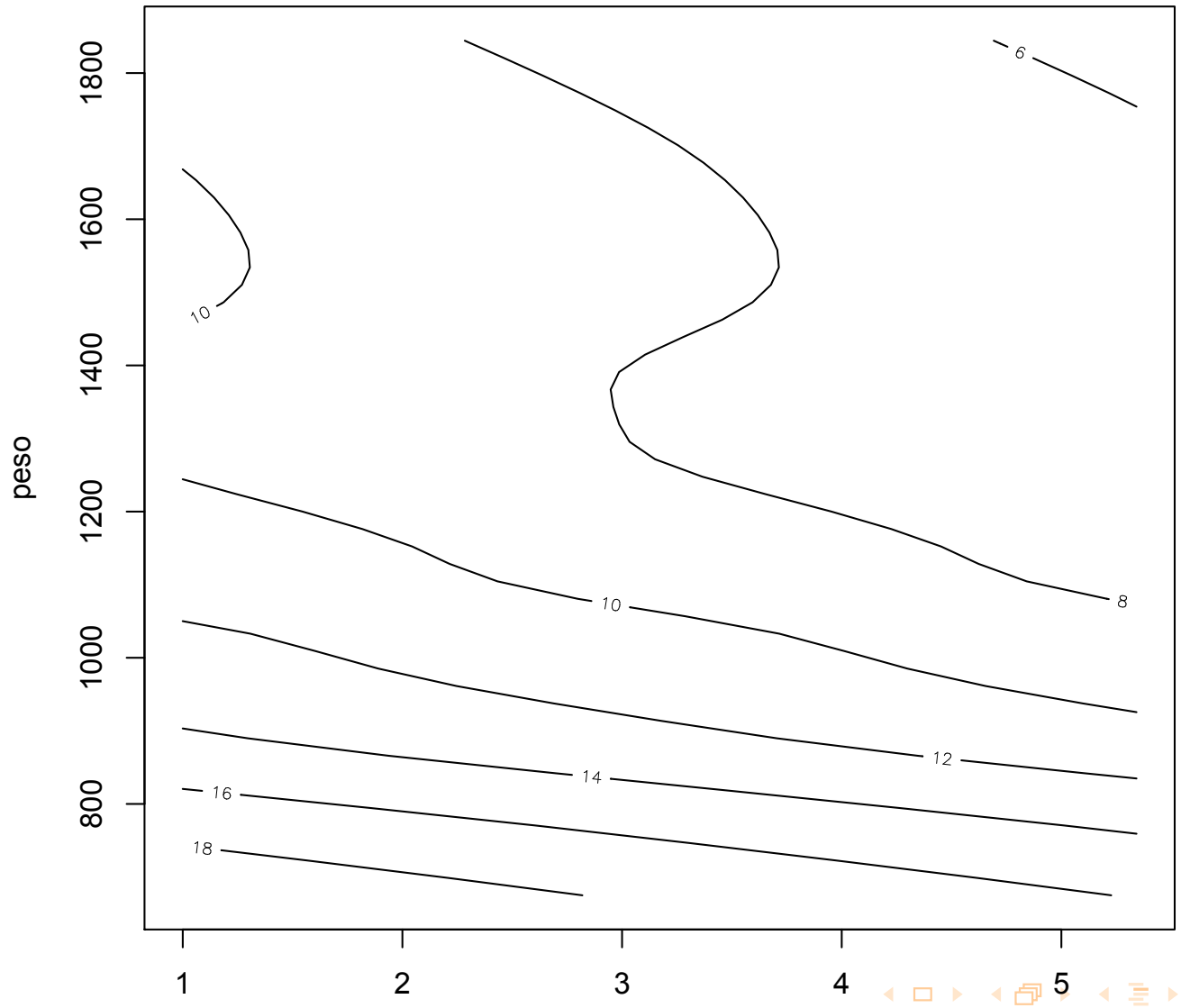
La generalizzazione delle spline, quando si ha a che fare con due o piú variabili esplicative non é cosi automatica come negli altri casi.
Le due estensioni piú note, ma comunque non molto utilizzate sono:

- Thin-plate spline;
- spline prodotto tensoriale;

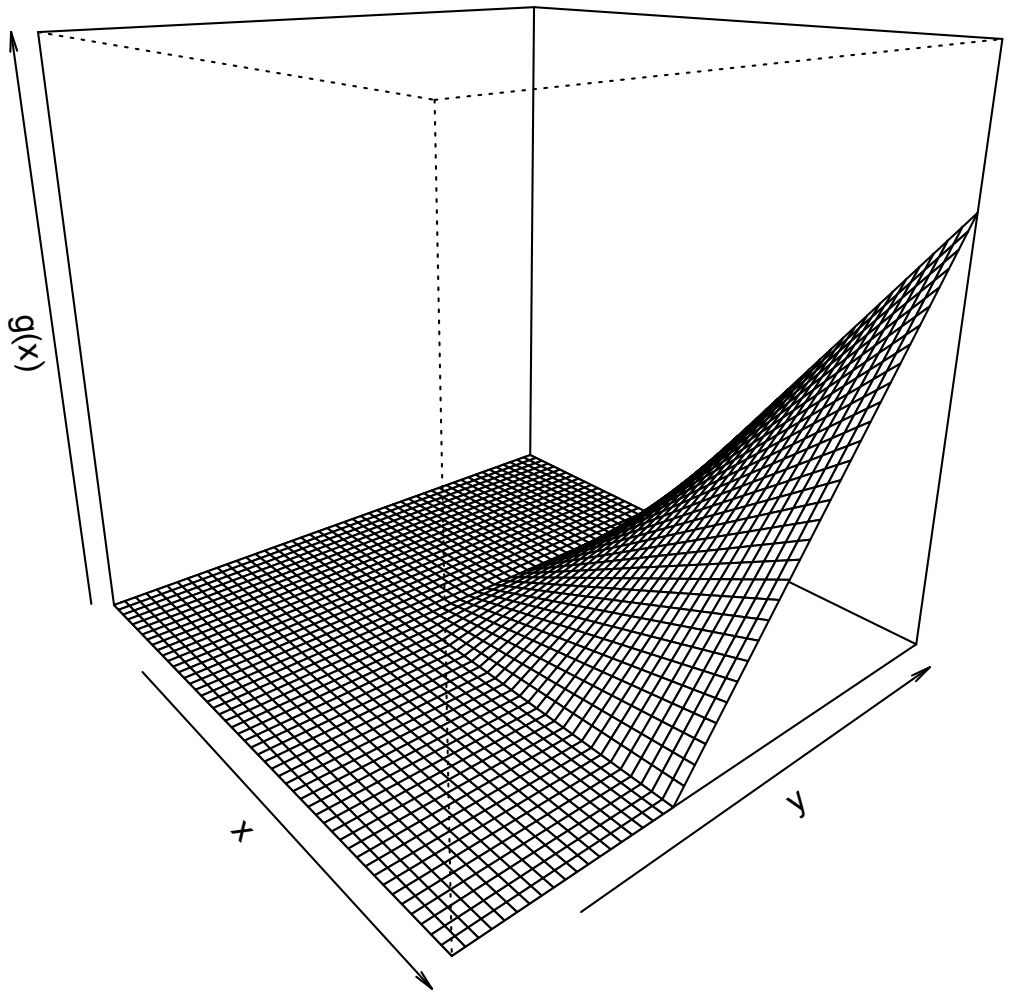
# Thin-plate spline



# Thin-plate spline



# Spline prodotto tensoriale



# MARS

Multivariate Adaptive Regression Splines
Sono una particolare specificazione iterativa delle spline di regressione, finalizzata alla modellazione di problemi con molte variabili esplicative.
Le basi utilizzate sono coppie di funzioni lineari a tratti del tipo $(x - \xi)_+$ e $(\xi - x)_+$ con un solo nodo nel punto $\xi$. Si cerca una relazione tra la y e le p variabili esplicative. Per ogni variabile esplicativa si determina una coppia di basi con il nodo in ciascun valore osservato su quella variabile $x_{ij}$. L'insieme di basi di funzioni sará:

$$C = \{(x_j - \xi)_+, (\xi - x_j)_+\}$$

con $\xi \in \{x_{1j}, x_{2j}, \ldots, x_{pj}\}$

# MARS

Il modello quindi si puó scrivere come:

$$f(x) = \beta_0 + \sum_{k=1}^{K} \beta_k h_k(x)$$

dove $h_k(x)$ sono le funzioni appartenenti a $C$ o prodotti di due o piú funzioni.
Una volta scelte le $h_k$, i parametri $\beta_k$ sono stimati minimizzando la somma dei quadrati dei residui. Il problema si riduce quindi alla scelta delle basi da utilizzare e del loro numero. Lo si fa attraverso un procedimento ricorsivo.

- Si parte con K=0. Si introduce per prima la funzione costante $h_0(x) = 1$ e tutte le funzioni sono candidate ad entrare nel modello.
- Al passo generico K+1 si considera come nuova coppia di basi ciascuna delle possibili coppie di prodotti di una funzione $h_k$ presente nel modello con un'altra coppia di funzioni in $C$ e si sceglie la coppia di basi che aggiungerá alla formulazione MARS il termine $\hat{\beta}_{K+1} h_m(x)(x_j - \xi)_+ + \hat{\beta}_{K+2} h_m(x)(\xi - x_j)_+$ che minimizza il criterio dei minimi quadrati. $h_m$ indica una funzione giá presente nel modello e i beta sono parametri stimati ai minimi quadrati insieme agli altri parametri del modello.
- Il processo di selezione e aggiunta di nuove basi continua fino al raggiungimento di un prefissato numero massimo di termini.

Il modello che si ottiene é in genere molto grande e di solito si sovra-adatta ai dati. É opportuno impostare una procedura all'indietro in cui si selezionano e si eliminano dal modello i termini col minor apporto alla somma dei quadrati dei residui. La scelta del numero di termini da contemplare nel modello ($\lambda$) puó avvenire attraverso un insieme di prova (come per le tecniche parametriche) oppure con la cross validation. In tal caso i costi computazionali aumentano notevolmente e si preferisce un'ulteriore strada, la convalida incrociata generalizzata (GCV).

- Si parte con K=0. Si introduce per prima la funzione costante $h_0(x) = 1$ e tutte le funzioni sono candidate ad entrare nel modello.
- Al passo generico K+1 si considera come nuova coppia di basi ciascuna delle possibili coppie di prodotti di una funzione $h_k$ presente nel modello con un'altra coppia di funzioni in $C$ e si sceglie la coppia di basi che aggiungerá alla formulazione MARS il termine $\hat{\beta}_{K+1} h_m(x)(x_j - \xi)_+ + \hat{\beta}_{K+2} h_m(x)(\xi - x_j)_+$ che minimizza il criterio dei minimi quadrati. $h_m$ indica una funzione giá presente nel modello e i beta sono parametri stimati ai minimi quadrati insieme agli altri parametri del modello.
- Il processo di selezione e aggiunta di nuove basi continua fino al raggiungimento di un prefissato numero massimo di termini.

Il modello che si ottiene é in genere molto grande e di solito si sovra-adatta ai dati. É opportuno impostare una procedura all'indietro in cui si selezionano e si eliminano dal modello i termini col minor apporto alla somma dei quadrati dei residui. La scelta del numero di termini da contemplare nel modello ($\lambda$) puó avvenire attraverso un insieme di prova (come per le tecniche parametriche) oppure con la cross validation. In tal caso i costi computazionali aumentano notevolmente e si preferisce un'ulteriore strada, la convalida incrociata generalizzata (GCV).

# GCV

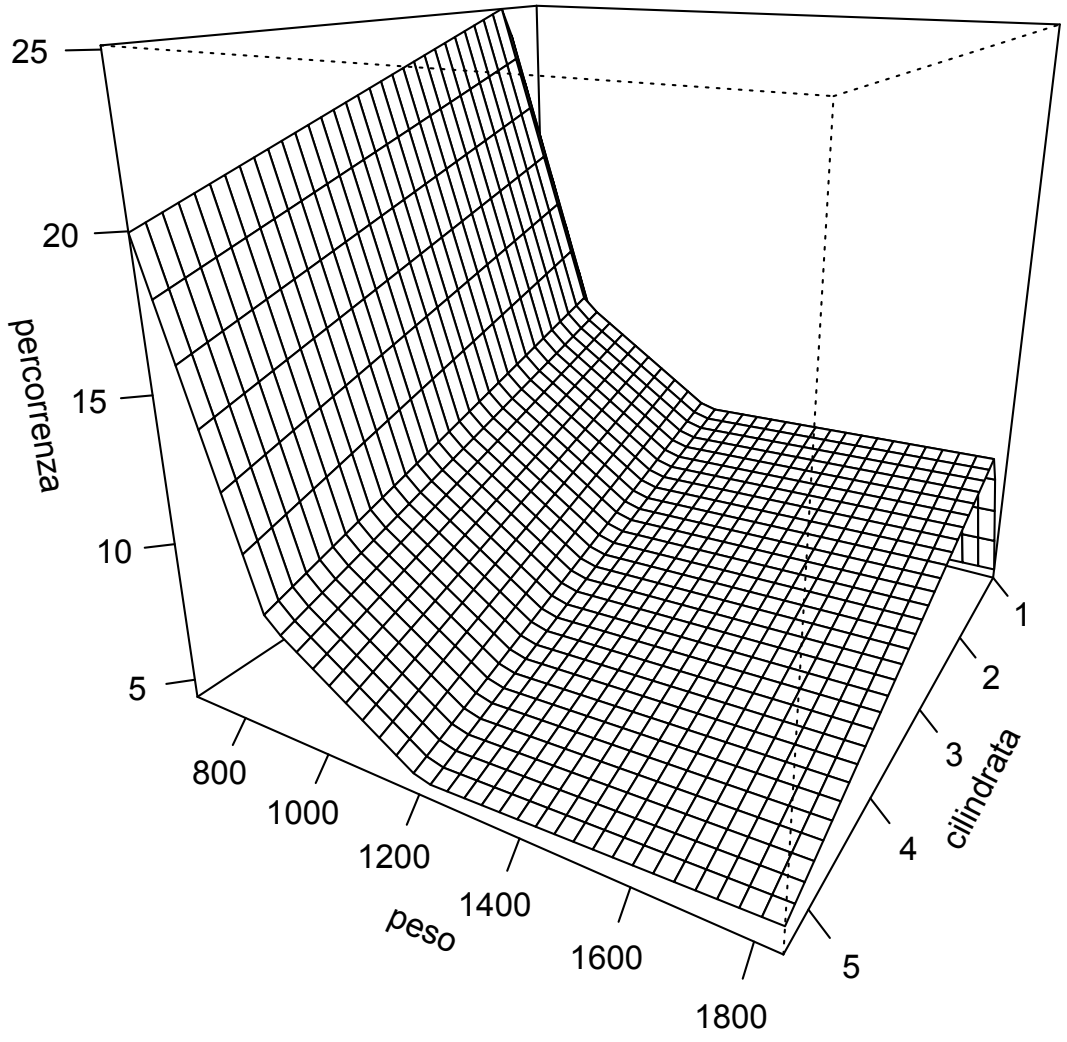
$$GCV(\lambda) = \frac{\sum_{i=1}^{n}[y_i - \hat{f}_\lambda(x_i)]^2}{[1 - d(\lambda)/n]^2}$$

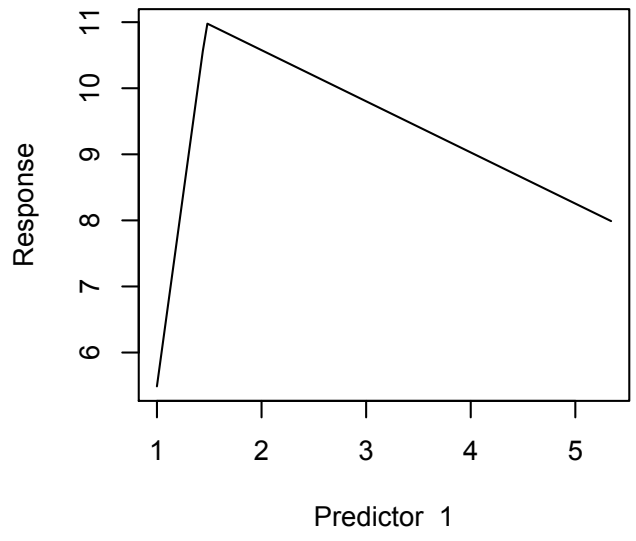
con $d(\lambda)$ indicatore del numero effettivo di parametri nel modello, funzione del numero di termini del modello e del numero di parametri utilizzati per selezionare la posizione ottima dei nodi.
Studi di simulazione mostrano che la scelta di un nodo nella regressione lineare a tratti corrisponde a circa tre parametri nella formulazione del modello. Quindi, indicando con $r$ il numero di basi di funzioni linearmente indipendenti presenti nel modello e con $K$ il numero di nodi, si ottiene

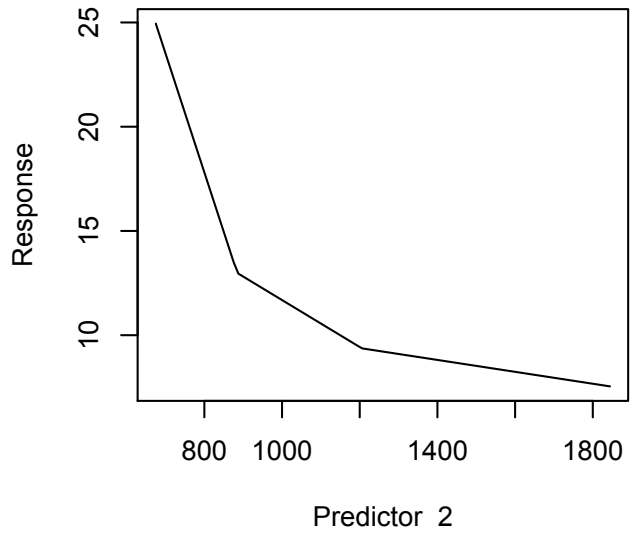
$$d(\lambda) = r + 3K$$

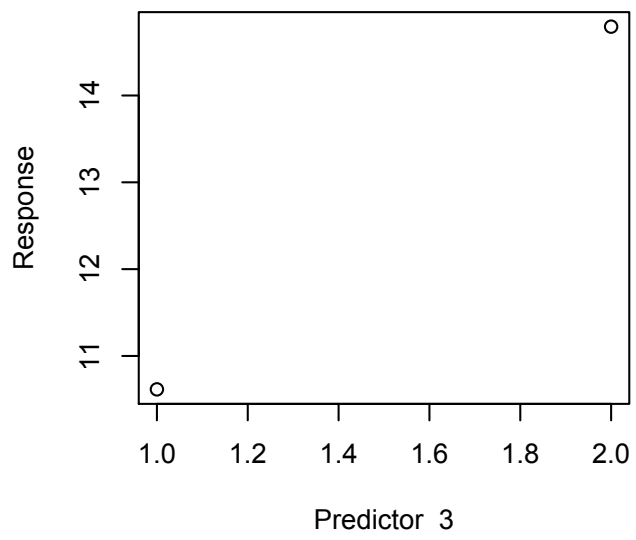
Vantaggi

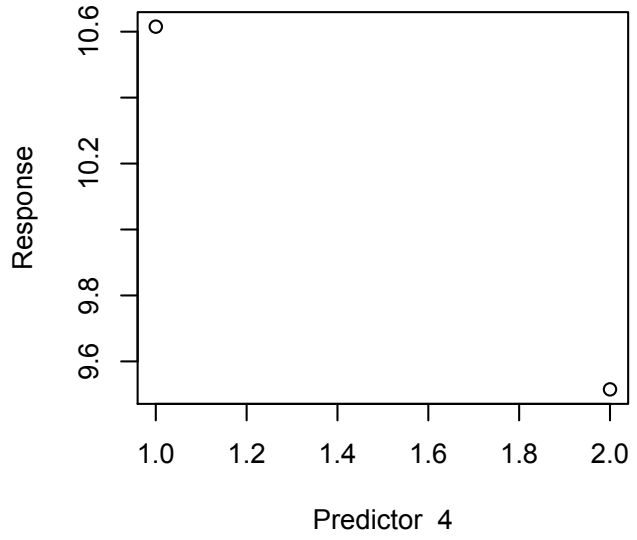
- Le coppie di funzioni lineari scelte come basi sono caratterizzate dalla semplicitá di operare localmente: quando sono moltiplicate tra loro sono diverse da zero solo nella parte di spazio in cui tutte le funzioni sono positive.
- Le funzioni sono inoltre facilmente moltiplicabili fra loro e quindi complessitá computazionale ridotta.
- Per modelli MARS si possono introdurre anche variabili qualitative.

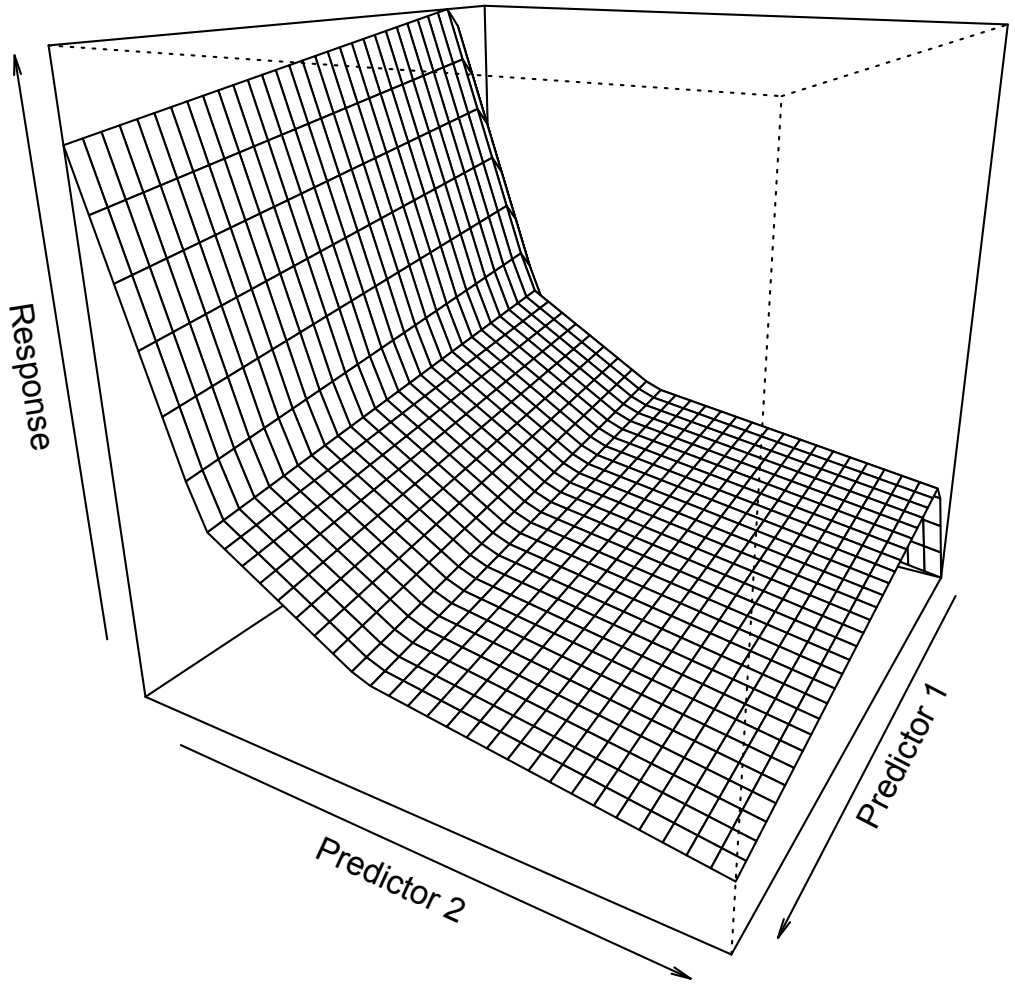

25
20
15
10
5
percorrenza
800
1000
1200
1400
1600
1800
peso
1
2
3
4
5
cilindrata


Response
6
7
8
9
10
11
1
2
3
4
5
Predictor 1


Response
10
15
20
25
800
1000
1400
1800
Predictor 2


Response
11
12
13
14
1.0
1.2
1.4
1.6
1.8
2.0
Predictor 3


Response
9.6
9.8
10.2
10.6
1.0
1.2
1.4
1.6
1.8
2.0
Predictor 4


Response
Predictor 2
Predictor 1

| Variabile | Nodo | Livelli | Parametri | Err.Std. |
|---|---|---|---|---|
| Costante | | | 57.0798 | 4.4884 |
| Peso | | | -0.0639 | 0.0063 |
| Alimentazione | | 1 | -4.0680 | 0.2768 |
| Aspirazione | | 1 | 1.3412 | 0.2287 |
| Peso | 861.84 | | 0.0510 | 0.0067 |
| Peso | 1149.88 | | 0.0069 | 0.0013 |
| Cilindrata | | | 11.6215 | 1.7015 |
| Cilindrata | 1.47 | | -12.1585 | 1.7581 |

# Modelli additivi

Tutte le tecniche finora introdotte soffrono del problema della dimensionalitá. Per ovviare a ció é necessario introdurre una forma di 'struttura', sulla forma della funzione di regressione, senza peró perdere la flessibilitá. Una soluzione molto utilizzata é la seguente:

$$f(x_1, \ldots, x_p) = \alpha + \sum_{j=1}^{p} f_J(x_j)$$

dove le funzioni $f_1, \ldots, f_p$ sono funzioni in una variabile dall'andamento sufficientemente regolare e $\alpha$ é una costante. La formulazione $y = f(x) + \varepsilon$ con $f(x)$ rappresentata in tal modo, é definito modello additivo.

Per evitare problemi di identificabilitá del modello é necessario che

$$\sum_{i=1}^{n} f_j(x_{ij}) = 0$$

Per ottenere la stima delle funzioni relative al modello additivo é necessaria una procedura iterativa, che si appoggia ad un metodo di stima non parametrica di funzioni in una variabile. Tale procedura é definita *backfitting*.

Una generalizzazione é:

$$f(x_1, \ldots, x_p) = \alpha + \sum_{j=1}^{p} f_j(x_j) + \sum_{j=1}^{p} \sum_{k<j} f_{kj}(x_k, x_j) +$$

$$+ \sum_{j=1}^{p} \sum_{k<j} \sum_{h<k<j} f_{hkj}(x_h, x_k, x_j) + \ldots$$

che permette di tenr conto anche dell'effetto di interazione tra coppie di variabili, o anche terne o interazioni di ordine superiore.

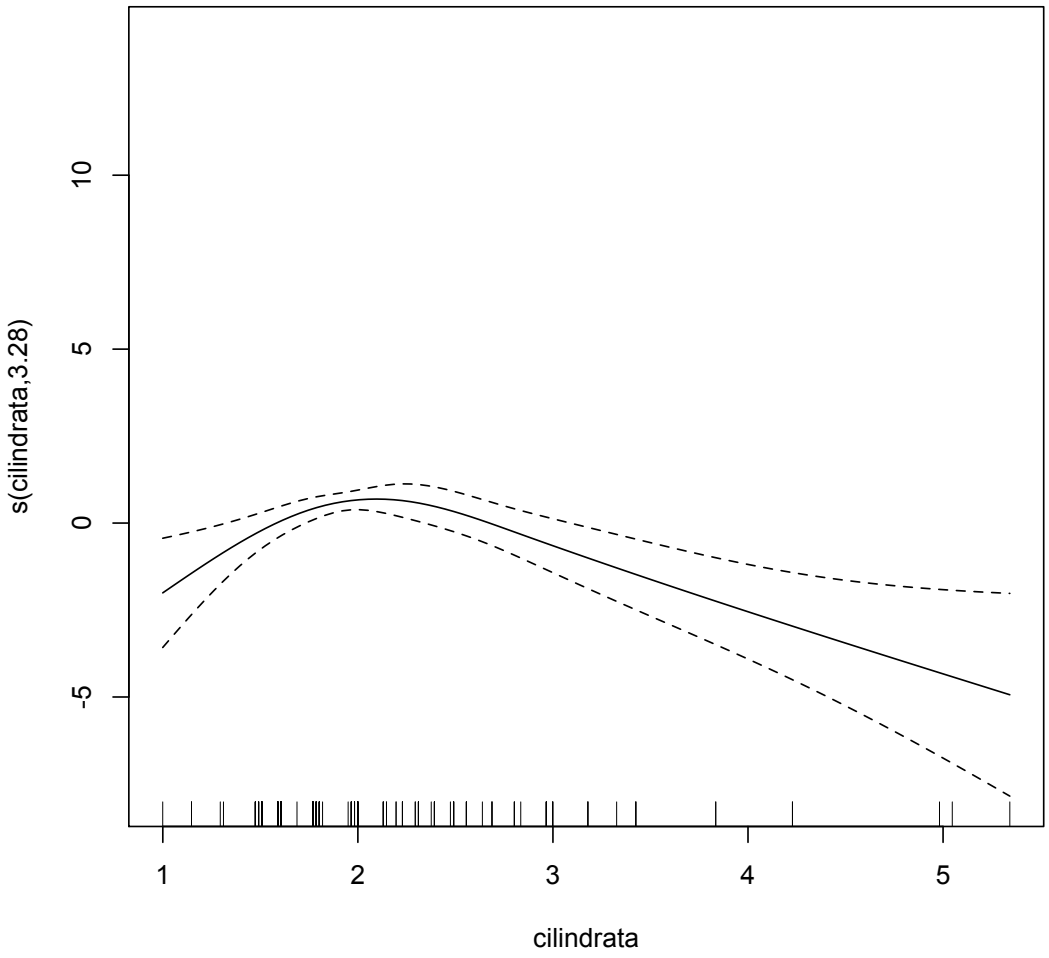



Figura: Percorrenza in funzione della cilindrata

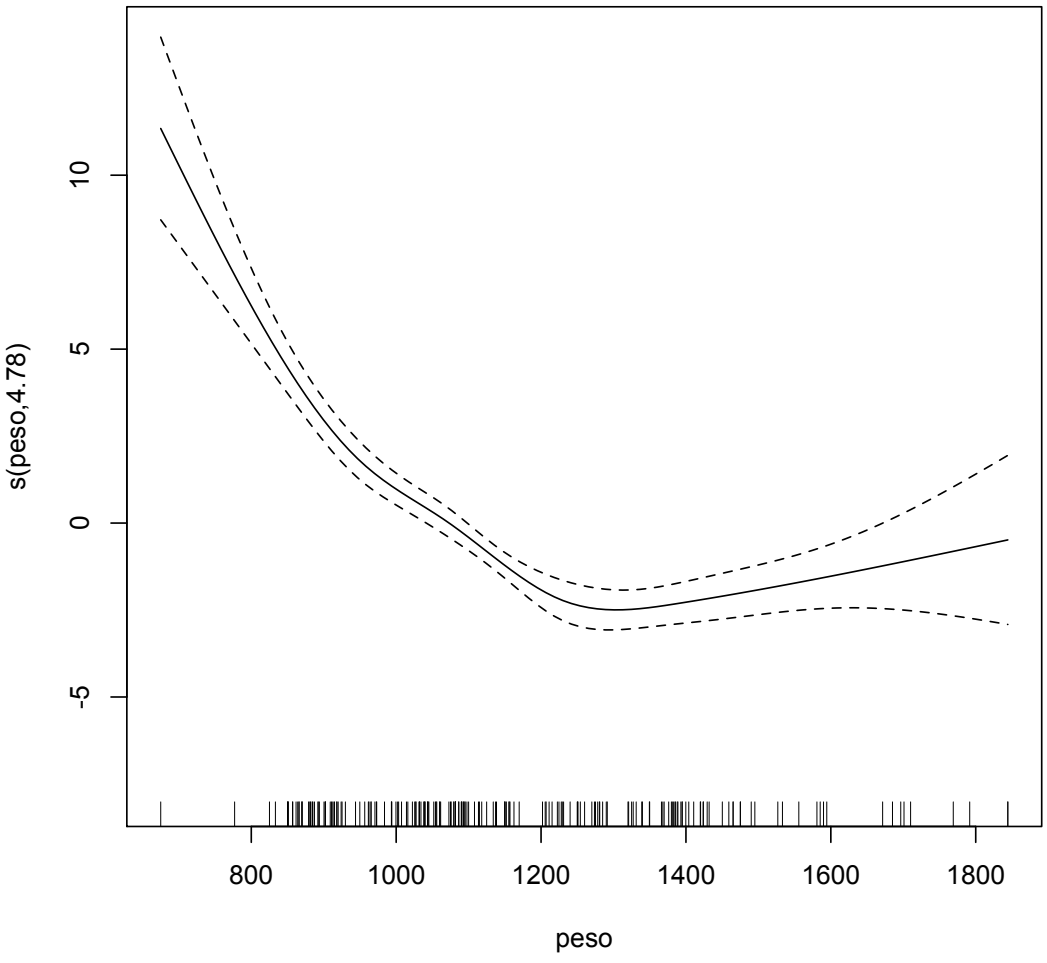



Figura: Percorrenza in funzione del peso

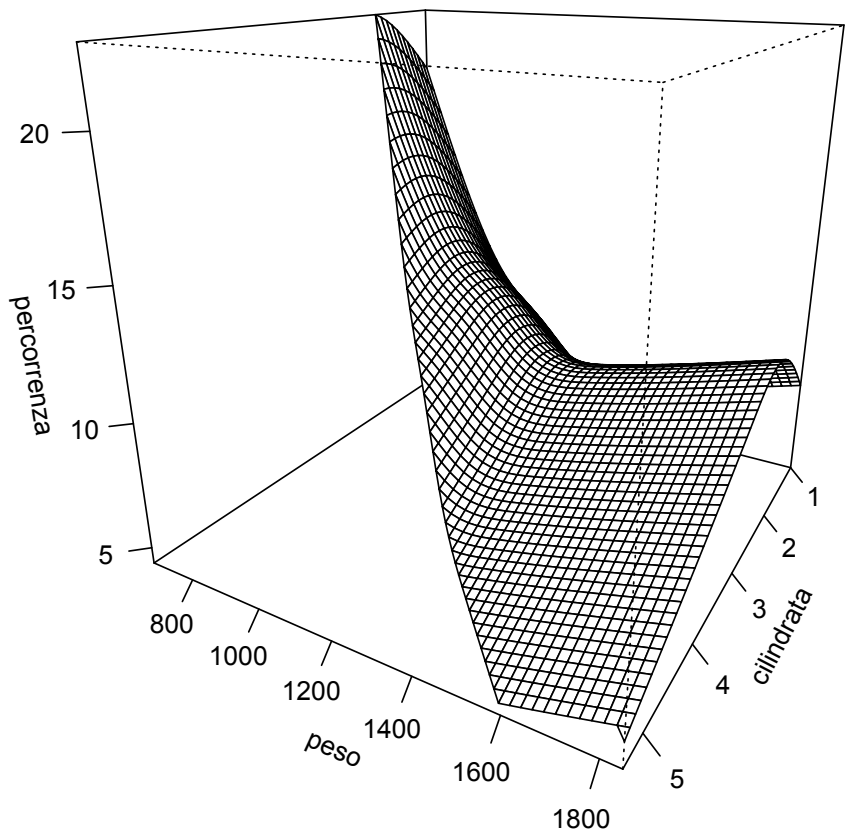



Figura: Modello additivo con lisciatore spline

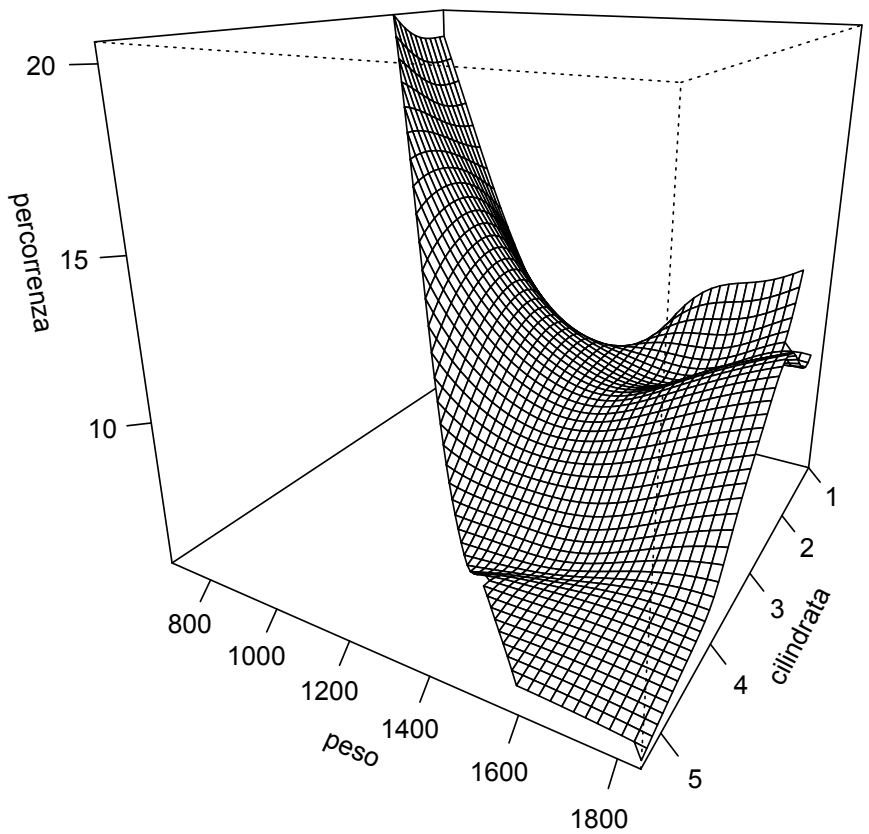



Figura: Modello senza ipotesi di additivitá

# GAM

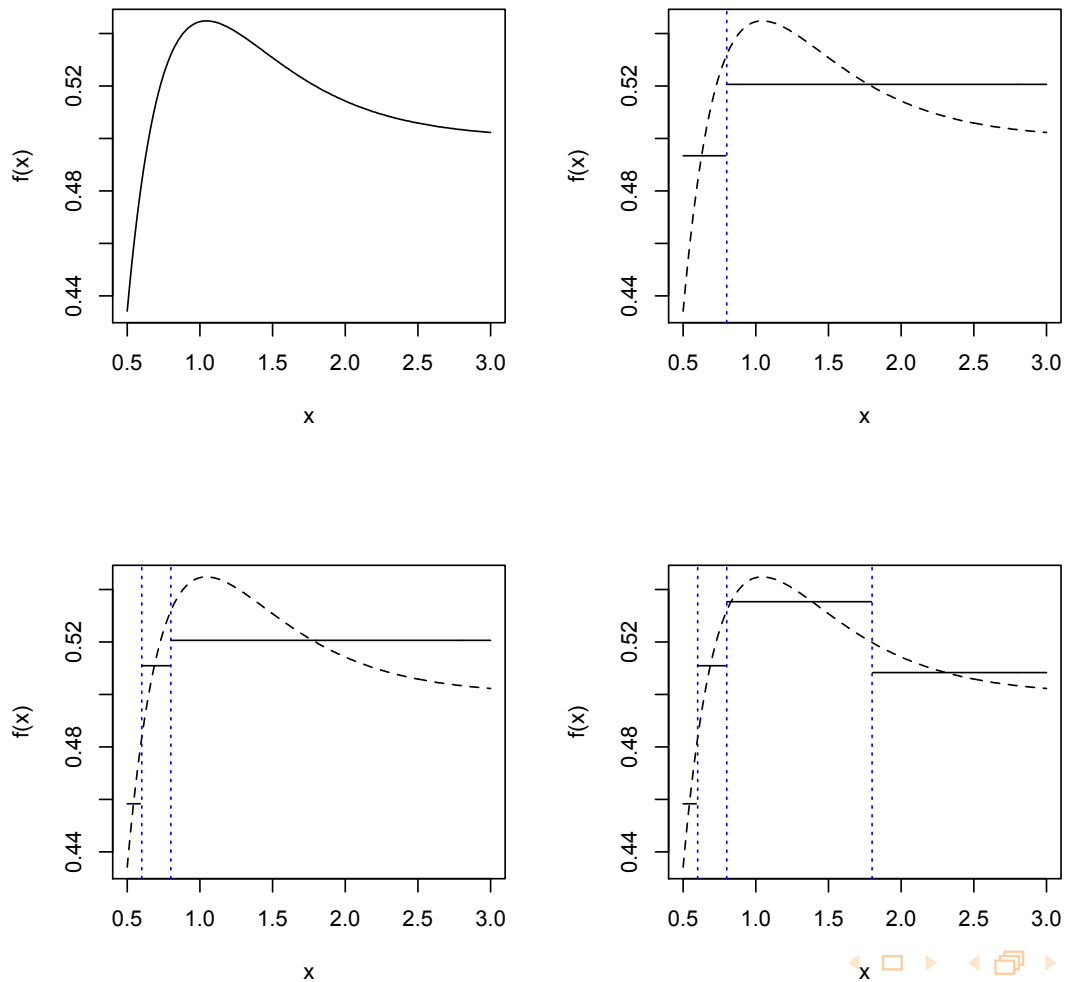
Un'ulteriore generalizzazione é la seguente:

$$g(E\{Y|x_1, \ldots, x_p\}) = \alpha + \sum_{j=1}^{p} f_j(x_j)$$

definita come Generalized Additive Model (GAM).
Per la stima di tali modelli si usa una opportuna combinazione dell'algoritmo backfitting e dei minimi quadrati pesati.

# Alberi di regressione

Il modo piú semplice per approssimare una qualsiasi funzione $y = f(x)$ é quello di usare una funzione approssimante a gradini, cioé una funzione costante a tratti su intervalli.



Aspetti da stabilire:

- quante suddivisioni dell'asse considerare;
- dove scegliere i punti di suddivisione;
- quale valore di ordinata assegnare ad ogni intervallo.

L'ultimo aspetto é il piú semplice, si sceglie il valore $\int f(x)dx/|R_j|$ per il generico intervallo $R_j$ e indicando con $|R_j|$ la lunghezza dell'intervallo.
Per la scelta del posizionamente dei punti di suddivisione conviene scegliere intervalli piú piccoli quando $f(x)$ é piú ripida.
Il numero di suddivisioni é la scelta piú soggettiva ricordando che all'aumentare dei gradini aumenta la qualitá dell'approssimazione, ma diminuisce la parsimonia.

Aspetti da stabilire:

- quante suddivisioni dell'asse considerare;
- dove scegliere i punti di suddivisione;
- quale valore di ordinata assegnare ad ogni intervallo.

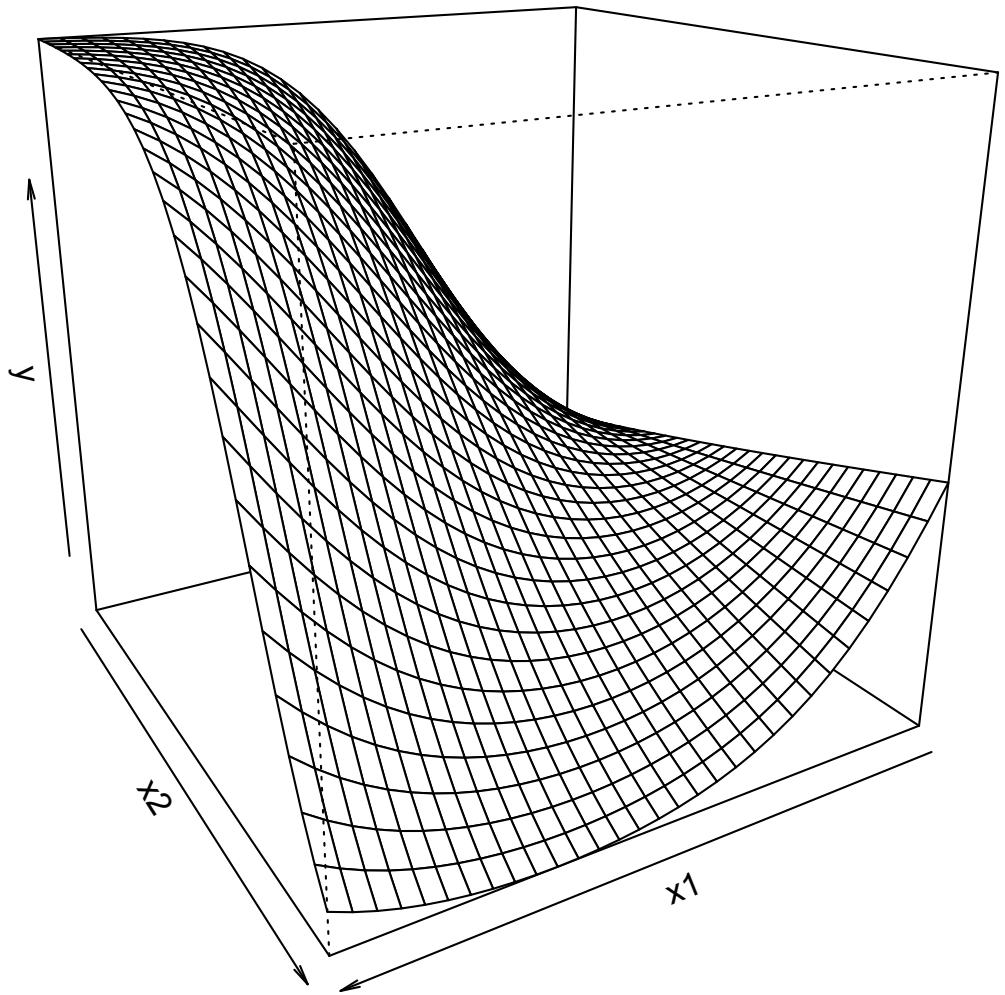
L'ultimo aspetto é il piú semplice, si sceglie il valore $\int f(x)dx/|R_j|$ per il generico intervallo $R_j$ e indicando con $|R_j|$ la lunghezza dell'intervallo.
Per la scelta del posizionamente dei punti di suddivisione conviene scegliere intervalli piú piccoli quando $f(x)$ é piú ripida.
Il numero di suddivisioni é la scelta piú soggettiva ricordando che all'aumentare dei gradini aumenta la qualitá dell'approssimazione, ma diminuisce la parsimonia.

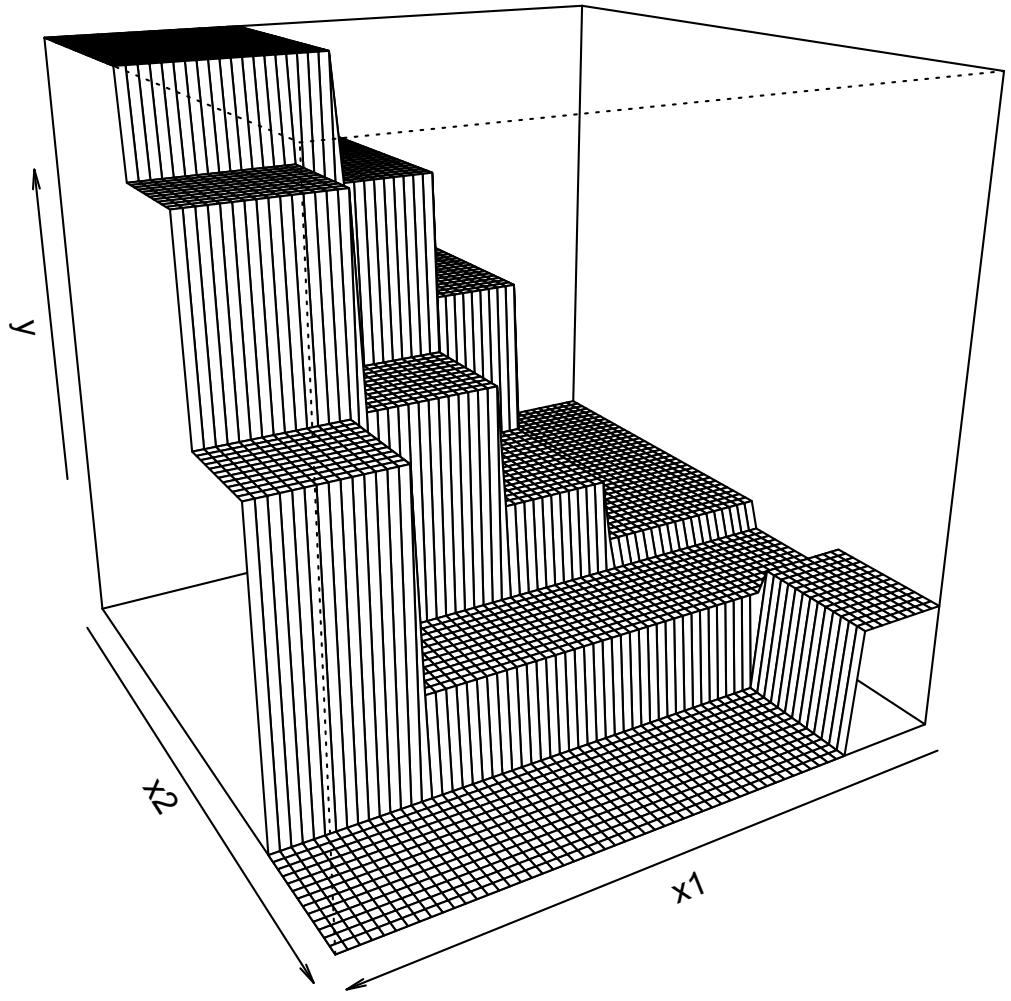
Aspetti da stabilire:

- quante suddivisioni dell'asse considerare;
- dove scegliere i punti di suddivisione;
- quale valore di ordinata assegnare ad ogni intervallo.

L'ultimo aspetto é il piú semplice, si sceglie il valore $\int f(x)dx/|R_j|$ per il generico intervallo $R_j$ e indicando con $|R_j|$ la lunghezza dell'intervallo.
Per la scelta del posizionamente dei punti di suddivisione conviene scegliere intervalli piú piccoli quando $f(x)$ é piú ripida.
Il numero di suddivisioni é la scelta piú soggettiva ricordando che all'aumentare dei gradini aumenta la qualitá dell'approssimazione, ma diminuisce la parsimonia.

# Piú dimensioni




y
x2
x1

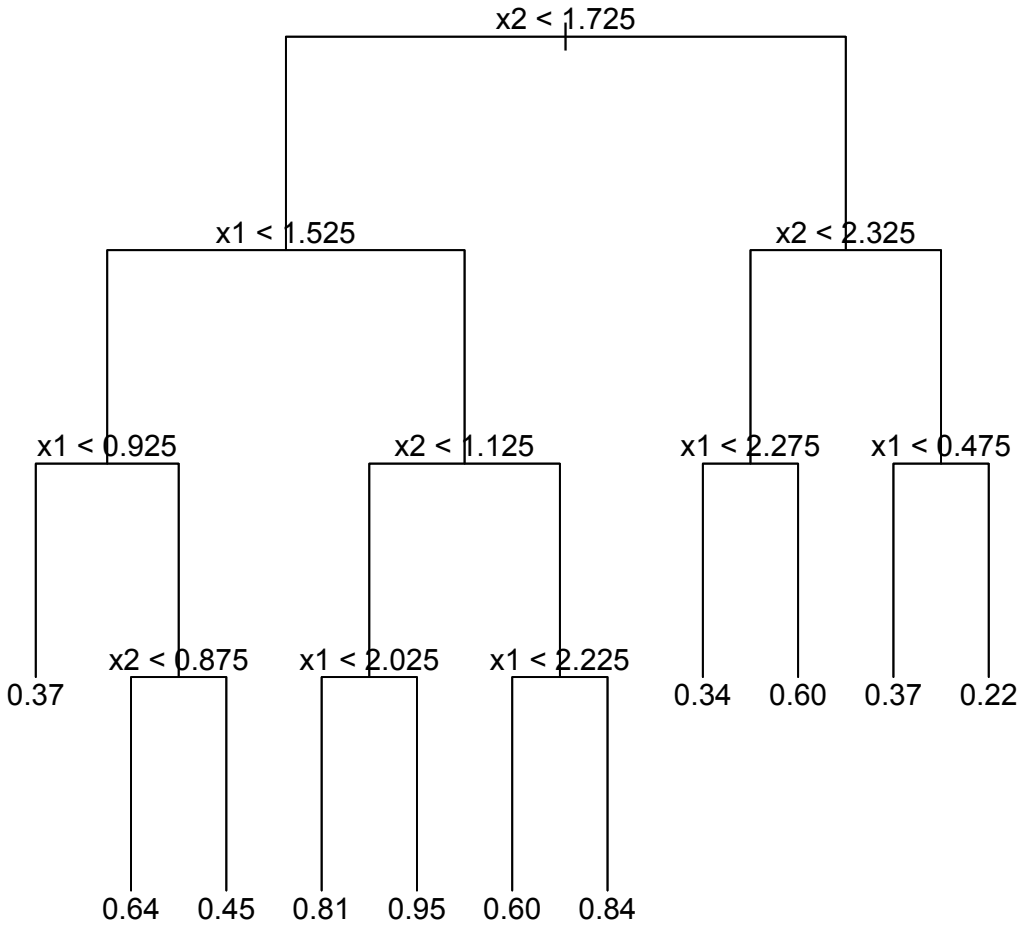



Figura: Albero binario

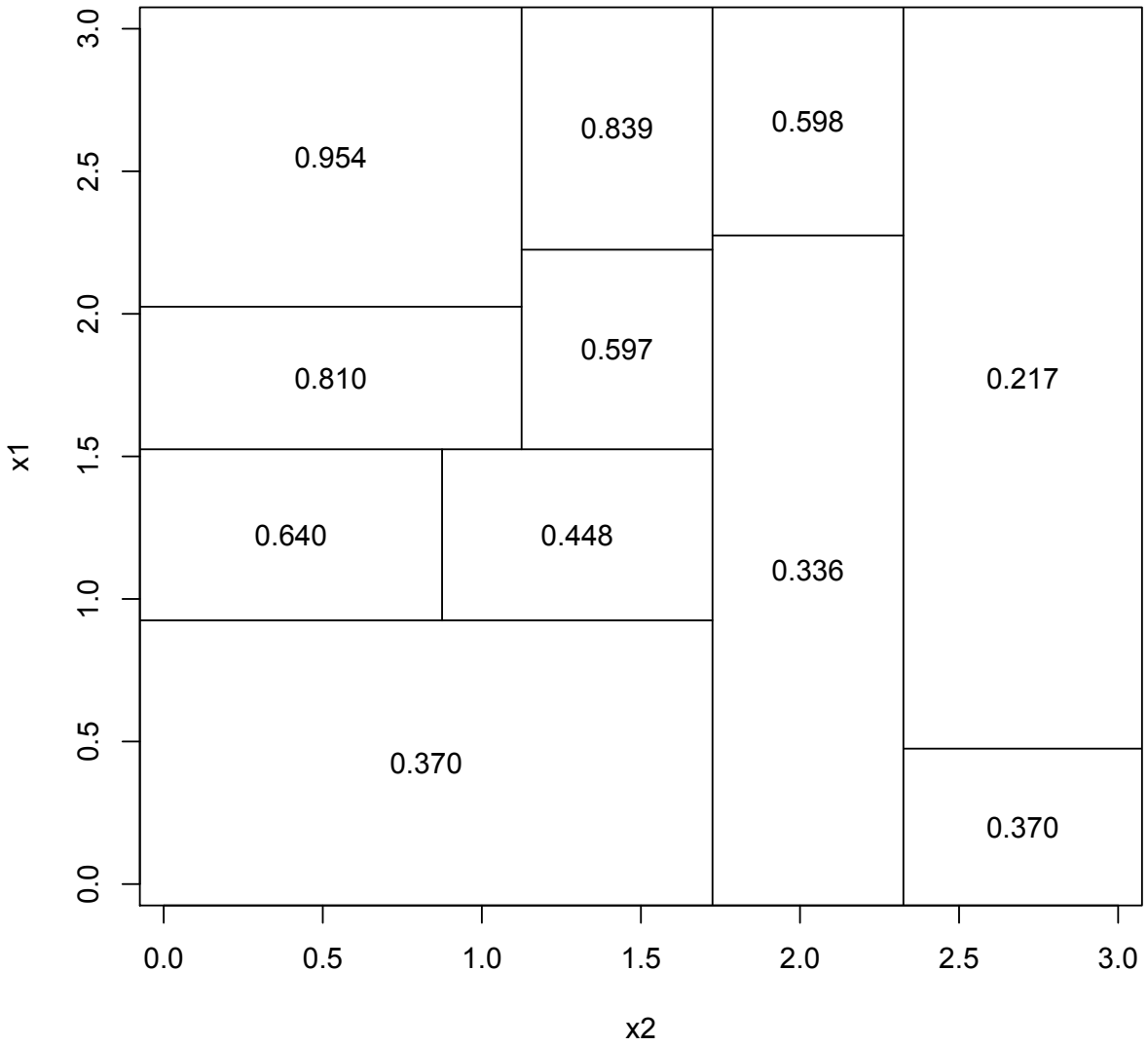



Figura: Partizione albero

# Alberi di regressione: crescita

L'obiettivo é quello di stimare la curva di regressione tramite una funzione a gradini del tipo:

$$\hat{f}(x) = \sum_{j=1}^{J} c_j I(x \in R_j)$$

con $I(x \in A)$ funzione indicatrice dell'insieme $A$ e $c_1, \ldots, c_J$ sono costanti. Per la scelta di $R_j$ e di $c_j$ si ha bisogno di una funzione obiettivo.

In genere il criterio potrebbe essere quello della minimizzazione della devianza, ma, anche con numero di gradini fissati, diventa computazionalmente molto complesso.
Si preferisce quindi una ottimizzazione passo a passo.

# Alberi di regressione: crescita

L'obiettivo é quello di stimare la curva di regressione tramite una funzione a gradini del tipo:

$$\hat{f}(x) = \sum_{j=1}^{J} c_j I(x \in R_j)$$

con $I(x \in A)$ funzione indicatrice dell'insieme $A$ e $c_1, \ldots, c_J$ sono costanti. Per la scelta di $R_j$ e di $c_j$ si ha bisogno di una funzione obiettivo.

In genere il criterio potrebbe essere quello della minimizzazione della devianza, ma, anche con numero di gradini fissati, diventa computazionalmente molto complesso.
Si preferisce quindi una ottimizzazione passo a passo.

Si costruisce una sequenza di approssimazioni via via piú raffinate e ad ogni passo si minimizza la devianza del passo successivo. Praticamente si divide in due uno dei rettangoli costruiti e si ottimizza la devianza rispetto a questa operazione.

In questo modo non si minimizza la devianza globale, ma almeno si riduce la complessitá computazionale.
La procedura si puó ripetere fin quando si ottengono rettangoli con una sola osservazione, ma in tal caso non sarebbe di grossa utilitá. Per tale motivo dopo la crescita dell'albero si parla di potatura dello stesso.

Si costruisce una sequenza di approssimazioni via via piú raffinate e ad ogni passo si minimizza la devianza del passo successivo. Praticamente si divide in due uno dei rettangoli costruiti e si ottimizza la devianza rispetto a questa operazione.

In questo modo non si minimizza la devianza globale, ma almeno si riduce la complessitá computazionale.
La procedura si puó ripetere fin quando si ottengono rettangoli con una sola osservazione, ma in tal caso non sarebbe di grossa utilitá. Per tale motivo dopo la crescita dell'albero si parla di potatura dello stesso.

Per lo sviluppo dell'algoritmo di crescita si puó scomporre la devianza come segue:

$$D = \sum_{i=1}^{n} [y_i - \hat{f}(x_i)]^2 = \sum_{j=1}^{J} \left[ \sum_{i \in R_j} (y_i - \hat{c}_j)^2 \right] = \sum_j D_j$$

Il procedimento di crescita parte con $J = 1$, $R_j = \mathbb{R}^p$, $D = \sum_i (y_i - M(y))^2$. Si procede poi iterativamente per piú cicli, nel modo seguente:

- Individuato un rettangolo $R_j$, il valore di $c_j$ sará la media dei valori corrispondenti: $\hat{c}_j = M(y_i : x_i \in R_j)$
- una volta suddivisa la regione in due parti l'addendo della devianza viene sostituito da $D_j^* = \sum_{i \in R_j'} (y_i - \hat{c}_j')^2 + \sum_{i \in R_j''} (y_i - \hat{c}_j'')^2$ con un guadagno in termini di abbassamento della devianza.
- si possono ispezionare tutte le variabili esplicative e i diversi punti di suddivisione, selezionando quella combinazione che massimizza la differenza $D_j - D_j^*$.

Ci si ferma teoricamente quando $J = n$, ma in realtá, soprattutto per n grande, ci si ferma quando tutte le foglie hanno un numero di elementi inferiore ad un valore prefissato o quando la diminuzione di devianza é inferiore ad un valore prefissato.

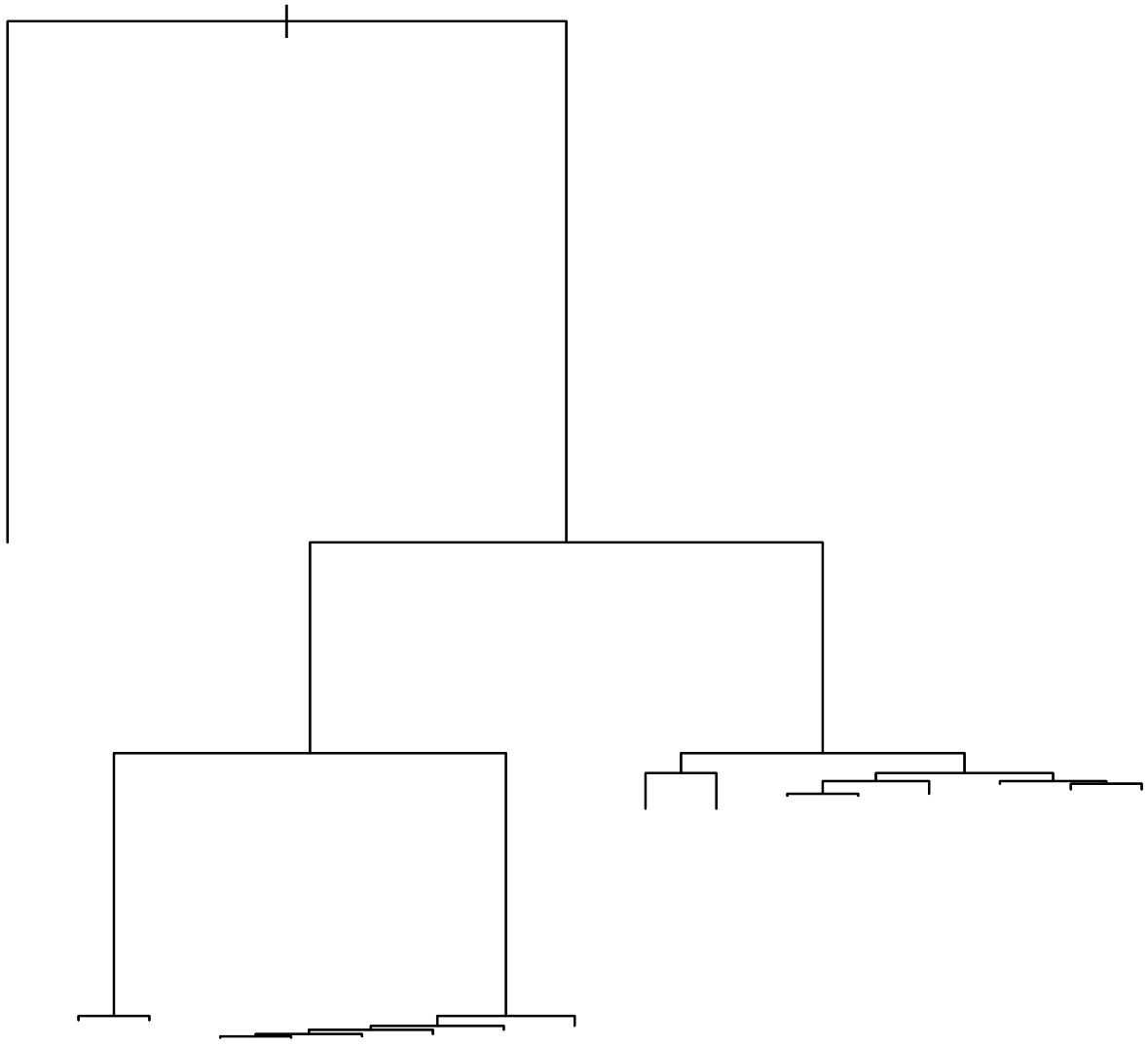
# Potatura dell'albero

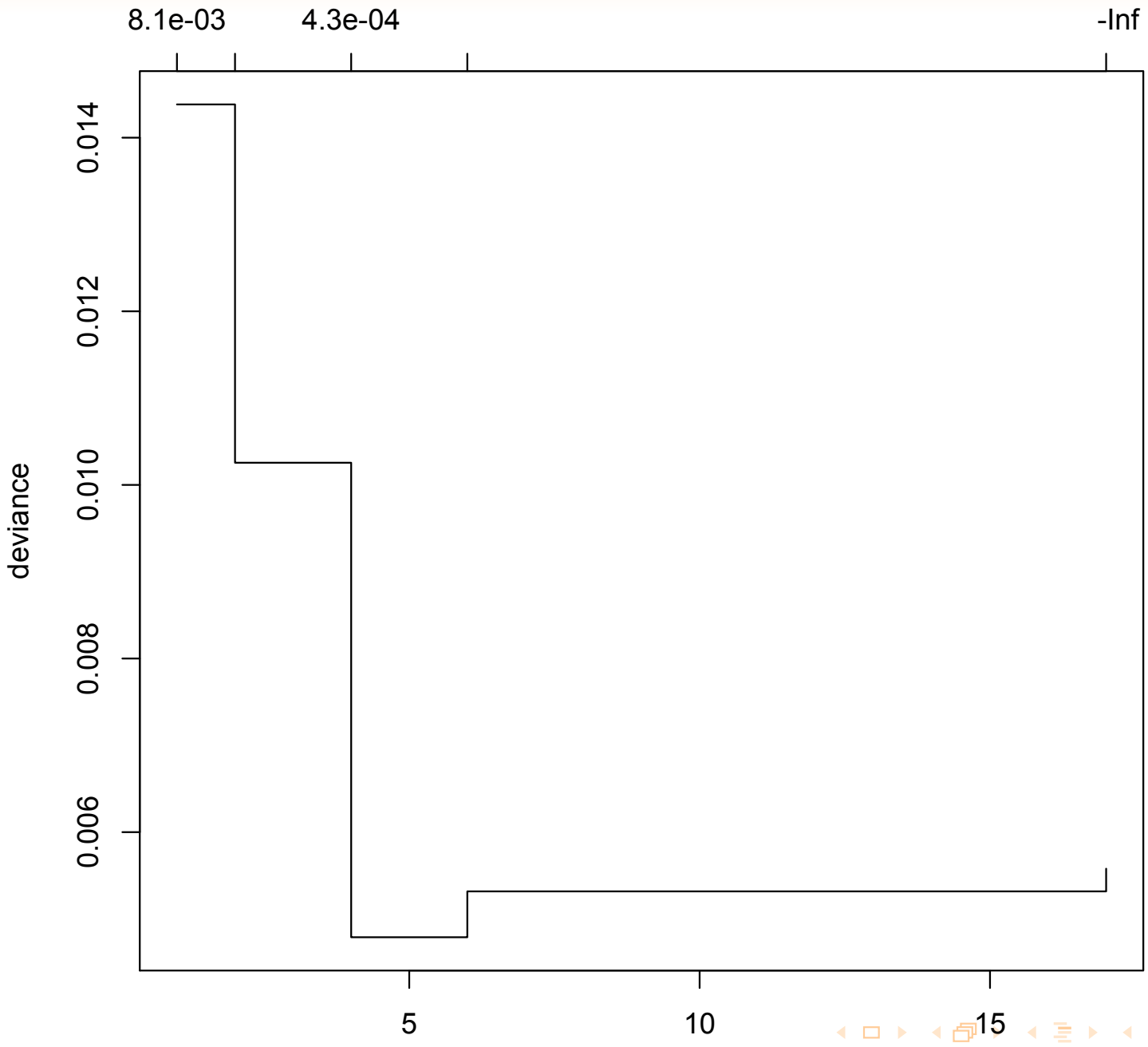
Bisogna potare l'albero dei rami inutili o poco utili. Si introduce a tal fine una funzione obiettivo che include una penalizzazione per il costo-complessitá dell'albero, cioé per la dimensione dell'albero.

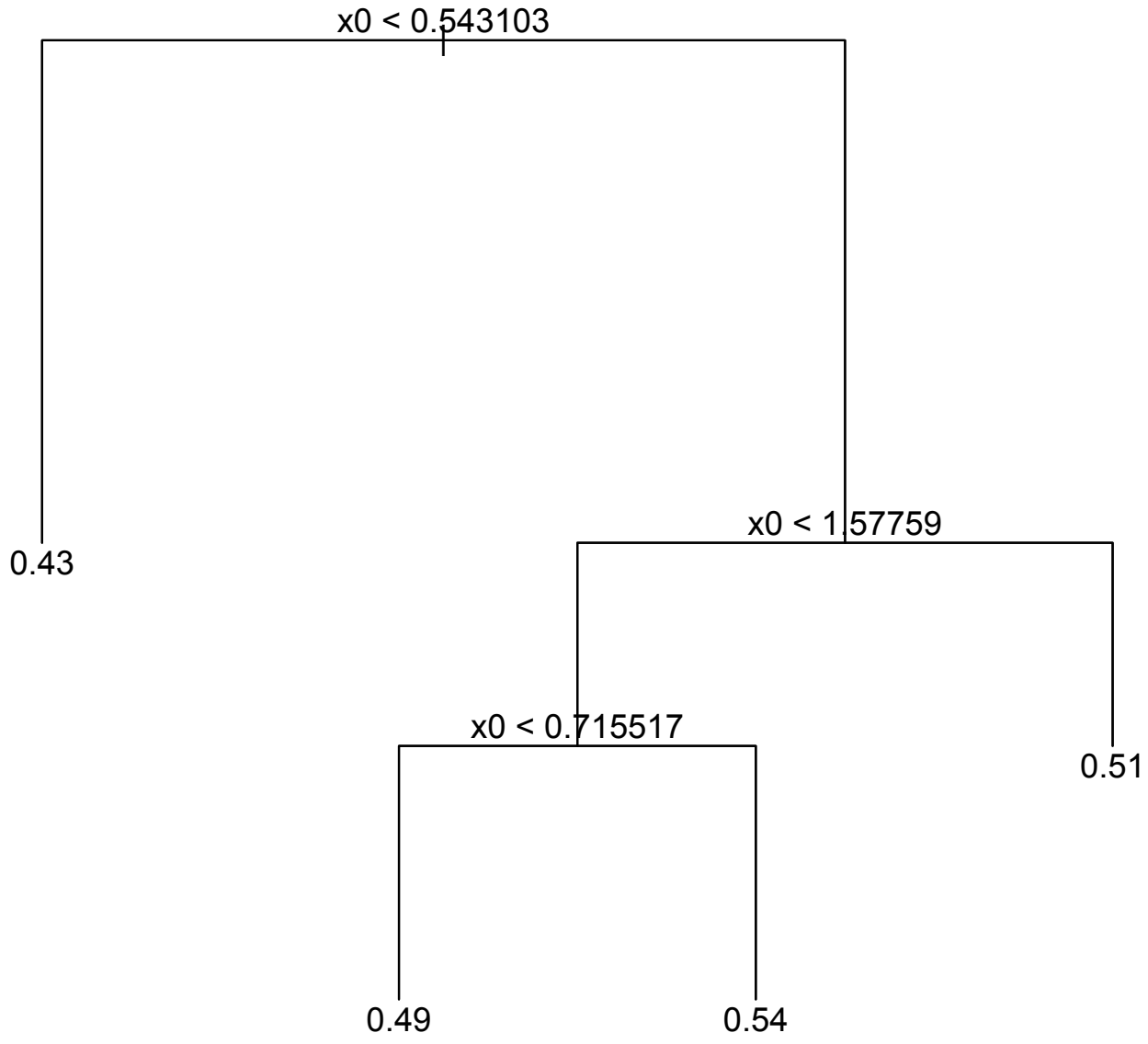
$$C_\alpha(J) = \sum_{j=1}^{J} D_j + \alpha J$$

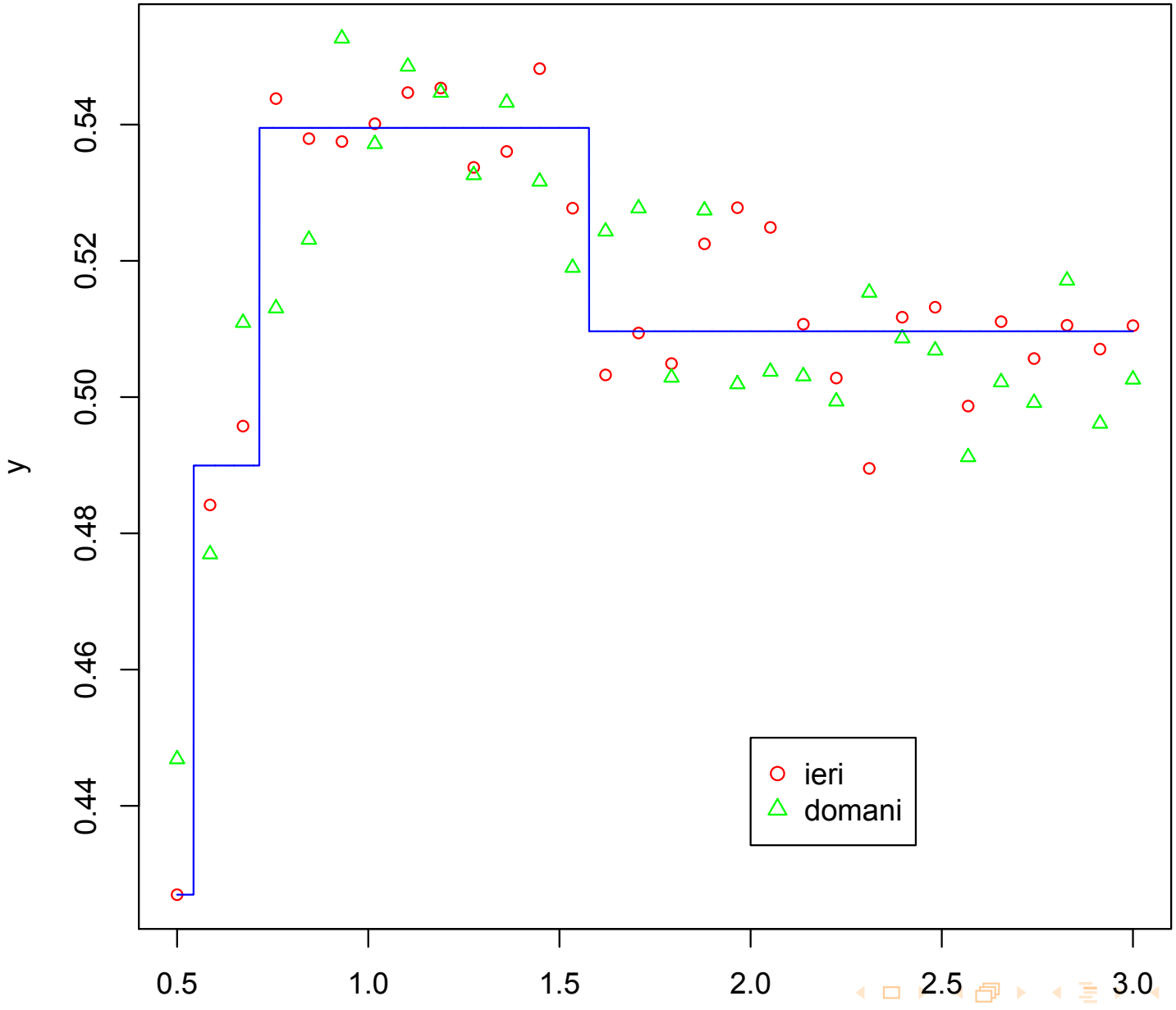
dove $\alpha$ é il parametro di penalizzazione. Per un fissato $\alpha$ si selezione l'albero che minimizza $C_\alpha(J)$.
Per ottenere la minimizzazione si procede eliminando una foglia alla volta. Si prende quella foglia la cui eliminazione comporta il minor incremento di $\sum_j D_j$.
Il problema sostanziale é quindi la scelta di $\alpha$. Per tale scelta si puó mostrare che il criterio dell'AIC porterebbe a scegliere $\alpha = 2\hat{\sigma}^2$, con $\sigma^2$ stima dell'errore residuo. Non essendo peró ben chiaro come questo possa essere determinato si preferisce utilizzare la cross validation o la divisione dei dati in campione base e campione test.

Per effettuare le previsioni é sufficiente far "cadere" la nuova osservazione dalla radice dell'albero disponibile e vedendo come si "incanala".




8.1e-03
4.3e-04
-Inf
deviance
0.014
0.012
0.010
0.008
0.006
5
10
15


x0 < 0.543103
0.43
x0 < 1.57759
x0 < 0.715517
0.51
0.49
0.54


y
0.44
0.46
0.48
0.50
0.52
0.54
0.5
1.0
1.5
2.0
2.5
3.0
ieri
domani

# Alberi di regressione: vantaggi

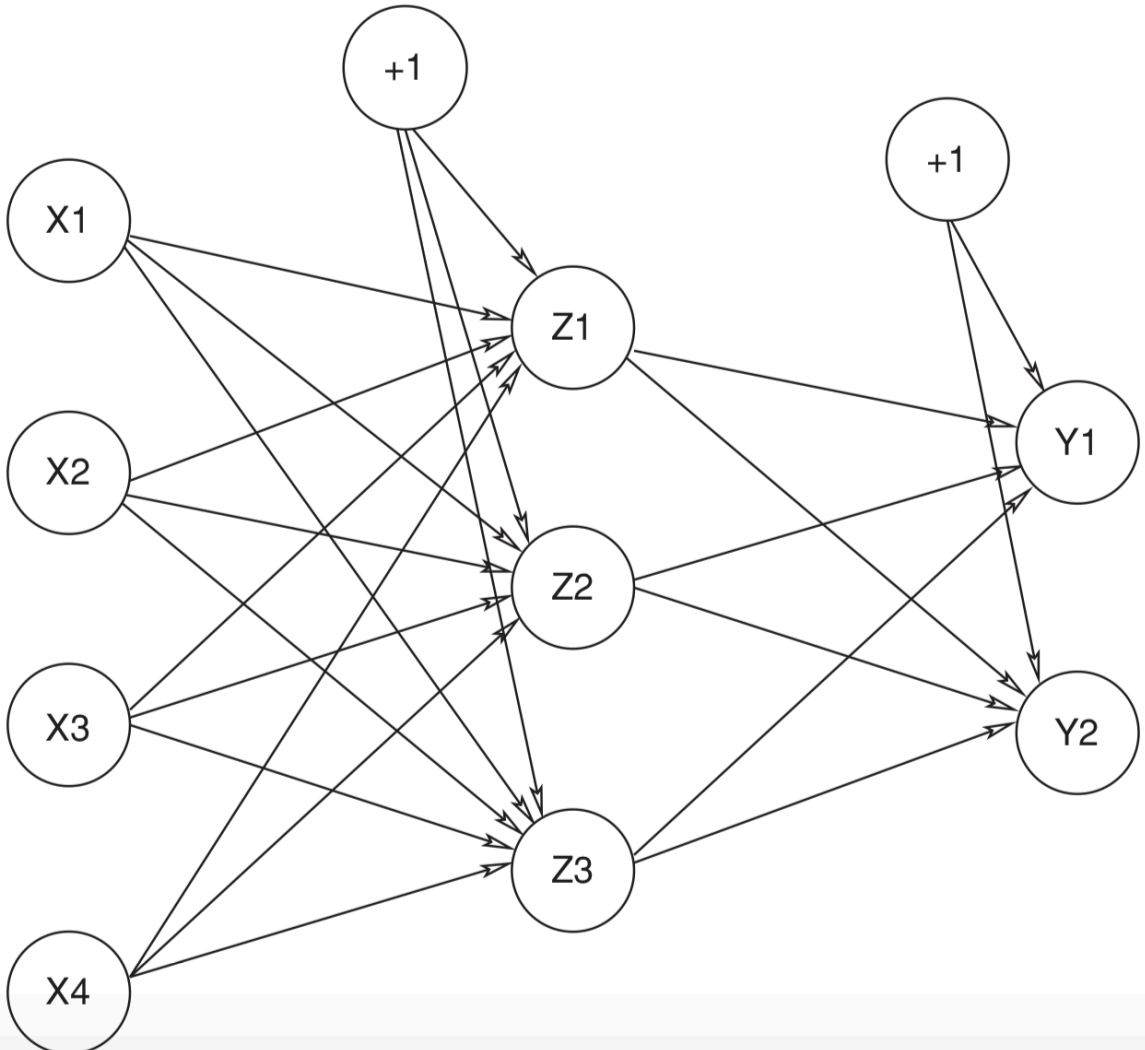
- Semplicitá e facilitá di comunicazione.
- La funzione a gradini ha una rappresentazione matematica semplice e compatta, in termini di informazione da memorizzare.
- Rapiditá di calcolo. La procedura non é onerosa dal punto di vista computazionale.
- Uso di variabili discrete e categoriali.
- Possibilitá di introdurre forme robuste di devianza.
- Possibilitá di introdurre varianti per trattare i dati mancanti.
- Selezione delle variabili: il modello produce un'automatica selezione delle variabili rilevanti.

# Alberi di regressione: svantaggi

- Instabilitá del risultato: albero molto sensibile all'inserimento di nuovi dati o alla perturbazione di quelli esistenti.
- Non si puó aggiornare l'albero se arrivano nuovi dati, bisogna ripetere la procedura.
- Difficoltá ad approssimare funzioni matematicamente semplici (soprattutto ripide), ma che si approssimerebbero molto bene attraverso una retta o altra funzione semplice.
- Non ci sono procedimenti formali di inferenza statistica
- Non é semplice valutare realmente l'ordine di importanza delle variabili mantenute nell'albero potato.

# Reti neurali

Il termine reti neurali abbraccia un insieme di tecniche sviluppate nell'ambito del machine learning.



In tali sistemi esistono variabili di input, variabili latenti e variabili di output, oltre a delle costanti. Il termine rete neurale ha origine come modello matematico di quello che in passato si riteneva il meccanismo di funzionamento di cervello animale. Ogni unitá rappresenta un neurone e le connessioni le sinapsi.

La rete neurale puó essere visto come un sistema di regressione a due stadi. Si puó scrivere:

$$z_j = f_0\left(\sum_{h \to j} \alpha_{hj} x_h\right)$$

$$y_k = f_1\left(\sum_{j \to k} \beta_{jk} z_j\right)$$

con $\alpha_{hj}$ e $\beta_{jk}$ parametri da stimare.

In tali sistemi esistono variabili di input, variabili latenti e variabili di output, oltre a delle costanti. Il termine rete neurale ha origine come modello matematico di quello che in passato si riteneva il meccanismo di funzionamento di cervello animale. Ogni unitá rappresenta un neurone e le connessioni le sinapsi.

La rete neurale puó essere visto come un sistema di regressione a due stadi. Si puó scrivere:

$$z_j = f_0\left(\sum_{h \to j} \alpha_{hj} x_h\right)$$

$$y_k = f_1\left(\sum_{j \to k} \beta_{jk} z_j\right)$$

con $\alpha_{hj}$ e $\beta_{jk}$ parametri da stimare.

In genere, in problemi di regressione si pone inoltre

$$f_0(u) = \frac{e^u}{1 + e^u}$$

$$f_1(u) = u$$

Sono possibili anche estensioni in piú direzioni. Per esempio considerare piú strati di variabili latenti oppure costruire archi che saltano unostrato.

In genere, in problemi di regressione si pone inoltre

$$f_0(u) = \frac{e^u}{1 + e^u}$$

$$f_1(u) = u$$

Sono possibili anche estensioni in piú direzioni. Per esempio considerare piú strati di variabili latenti oppure costruire archi che saltano unostrato.

Prima di stimare i coefficienti é anche necessario determinare il numero di variabili latenti da considerare. In genere lo si fa attraverso piú prove.
Una volta determinato il numero di variabili latenti bisogna minimizzare:

$$D = \sum_i ||y^{(i)} - f(x^{(i)})||^2$$

dove $y^{(i)}$ é il vettore q-dimensionale di variabili risposta relative alla i-esima osservazione.
Ci sono anche varianti piú elaborate della funzione obiettivo che considerano inserendo un termine di penalizzazione:

$$D_0 = D + \lambda J(\alpha, \beta)$$

con $\lambda$ parametro di regolazione e $J(\alpha, \beta)$ funzione di penalizzazione.

La minimizzazione di $D_0$ richiede un procedimento di ottimizzazione numerica. Il metodo in genere piú utilizzato é quello della back-propagation, che gode di alcune proprietá. Una variante dell'algoritmo consente inoltre di aggiornare successivamente le stime dei parametri, man mano che nuovi dati vengono immessi.

La funzione obiettivo ha spesso punti di minimo locale e quindi é opportuno effettuare varie prove avviando l'algoritmo di ottimizzazione da diversi punti iniziali.

La minimizzazione di $D_0$ richiede un procedimento di ottimizzazione numerica. Il metodo in genere piú utilizzato é quello della back-propagation, che gode di alcune proprietá. Una variante dell'algoritmo consente inoltre di aggiornare successivamente le stime dei parametri, man mano che nuovi dati vengono immessi.

La funzione obiettivo ha spesso punti di minimo locale e quindi é opportuno effettuare varie prove avviando l'algoritmo di ottimizzazione da diversi punti iniziali.

# Reti neurali: vantaggi

- Flessibilitá: il metodo consente di approssimare ogni funzione di regressione.
- Compattezza della rappresentazione: la funzione di regressione stimata é identificata da un numero limitato di componenti.
- Aggiornabilitá sequenziale: i coefficienti possono essere aggiornati sequenzialmente via via che arrivano nuovi dati.

# Reti neurali: svantaggi

- Arbitrarietá: non ci sono criteri forti per scegliere il numero di nodi latenti. Anche per la scelta di $\lambda$ esistono solo indicazioni di massima.
- Difficoltá di stima: non esiste un solo punto di minimo, presenza di molti minimi locali e avviando l'algoritmo da punti diversi si ottengono risultati diversi.
- Inferenza: non ci sono errori standard associati ai coefficienti o altre procedure per ridurre il numero di coefficienti.
- Interpretazione: difficoltá di interpretazione, soprattutto quando il numero di variabili latenti cresce.