

Statistica Learning : ottenere conoscenza sui dati

• supervised learning : abbiamo la variabile risposta

- regressione => risposta continua

- classificazione => risposta discreta

• unsupervised learning : non abbiamo risposta

- clustering

- pattern discovery

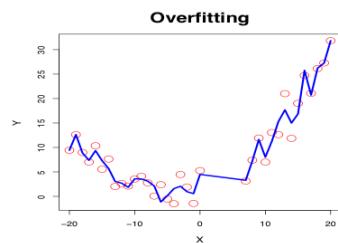
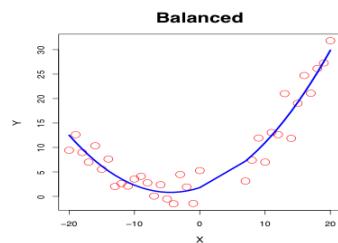
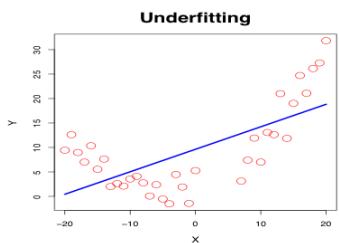
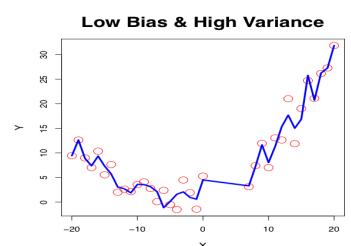
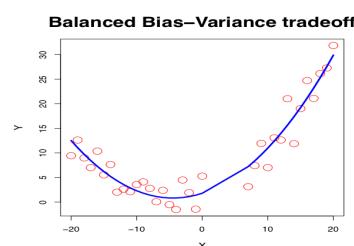
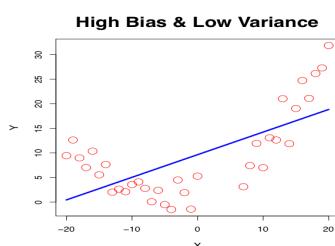
Obiettivi statistical learning:

Predizione: . predire le possibilità che qualcosa accadrà (o meno). => "modelli black-box"
. non siamo interessati a capire perché, ma solo il risultato.

Infanzia: capire perché qualcosa accadrà, o meno. => modelli più semplici e comprensibili

Bias - Variance tradeoff :

$$\text{valore atteso errore } (\hat{m}) = \frac{\text{Varianza } (\hat{m})}{\downarrow \text{varianza dei dati}} + \frac{\text{Bias}^2(\hat{m})}{\downarrow \text{differenza tra dato reale e dato predetto.}}$$



il nostro modello
 è troppo semplice
 e non riesce
 a modellare i
 dati in modo
 efficace

POCHE VARIABILI
NEL MODELLO

il nostro modello
 è troppo complesso,
 modella bene i dati
 ma è troppo sensibile
 alla varianza: le previsioni
 non saranno accurate.

TANTE VARIABILI NEL
MODELLO

Ripasso probabilità

$$E[x] = \begin{cases} \int x p(x) dx & \text{continue} \\ \sum_x x \cdot p(x) & \text{discrete} \end{cases}$$

$$E[h(x)] = \begin{cases} \int h(x) \cdot p(x) dx \\ \sum_x h(x) p(x) \end{cases}$$

$$E[h(x)] \neq h(E[x]) \quad \text{occhio!}$$

$$\text{Var}[x] = E[(X - E[x])^2]$$

$$\text{Cov}[x, y] = E[(x - E[x]) \cdot (y - E[y])]$$

$$\text{Corr}[x, y] = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}[x] \text{Var}[y]}}$$

Modello di regressione lineare

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Diagramma delle componenti del modello:

- Y : variabile risposta
- β_0 : intercetta
- β_1 : coefficiente angolare
- X : variabile predittiva
- ϵ : errore

- La distribuzione di X è arbitraria
- ϵ è randomico
- Y è lineare (vedi sopra)
- $E[\epsilon | X=x] = 0$ (non importa a che x corrisponde)
- $\text{Var}[\epsilon | X=x] = \sigma^2$ (costante! non importa a che x corrisponde)
- $\forall i, j \quad \text{cov}[\epsilon_i, \epsilon_j] = 0 \quad , \text{ con } i \neq j$

Quali sono i parametri β_0 e β_1 ottimali?

$$\begin{aligned} \text{MSE}(m) &= E[(Y - m)^2] \\ &= (E[Y] - m)^2 + \text{Var}[Y] \\ &= \text{Bias}^2 + \text{Variance} \quad \rightarrow \text{Bias Variance tradeoff} \end{aligned}$$

m è la nostra
predizione (scalare)

Puondiamo m che lo minimizzi:

$$m^* = \underset{m}{\operatorname{argmin}} \text{MSE}(m)$$

(derivata prima e seconda in funzione di m)

$$m^* = E[Y] \Rightarrow \text{La migliore stima di } Y \text{ è } E[Y]$$

sapendo che m è una funzione lineare:

$$m(x) = b_0 + b_1 x \quad \text{minimizziamo MSE in funzione di } b_0 \text{ e } b_1$$

$$\text{MSE}(b_0, b_1) = E \left[(Y - (b_0 + b_1 x))^2 \right]$$

(derivate prima e seconda in due variabili)

troviamo:

$$\beta_0 = E[Y] - \beta_1 \cdot E[X]$$

$$\beta_1 = \frac{\text{Cov}[X, Y]}{\text{Var}[X]}$$

Stima dei parametri β_0 e β_1

- non possiamo utilizzare i parametri ottimali:

$E[Y], E[X], \text{Cov}[X, Y]$ e $\text{Var}[X]$ sono funzioni della vera distribuzione!

abbiamo un campione di coppie (x_i, y_i)

NON SAPPIAMO LA VERA
DISTRIBUZIONE!

Plug-in estimate:

- utilizziamo i valori campionari
- assumiamo che gli stimatori siano consistenti

con $n \rightarrow \infty$ abbiamo: $\bar{Y} \rightarrow E[Y]$, $\bar{X} \rightarrow E[X]$, $C_{xy} \rightarrow \text{Cov}[X, Y]$

$$S_x^2 \rightarrow \text{Var}[X]$$

quindi:

$$\hat{\beta}_1 \rightarrow \beta_1 \quad \text{e} \quad \hat{\beta}_0 \rightarrow \beta_0$$

Troviamo:

$$\hat{\beta}_1 = \frac{c_{xy}}{s_x^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Least squares estimates:

• con $n \rightarrow \infty$ $\hat{MSE}(b_0, b_1) \rightarrow MSE(b_0, b_1)$

$$MSE(b_0, b_1) = \frac{1}{n} \sum_i (y_i - (b_0 + b_1 x_i))^2$$

(derivate prima e seconda in due variabili)

Troviamo:

$$\hat{\beta}_1 = \frac{c_{xy}}{s_x^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

L'equivalenza fra plug-in estimates e least-squares estimates è tale solo per il modello lineare semplice!

L'uso di uno o l'altro dipende dal modello!

Residui vs Errori

Errori : $e_i = y_i - (\beta_0 + \beta_1 x_i)$ parametri vari

Residui : $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ parametri stimati

- i residui sono gli errori "empirici"
- sono abbastanza simili da utilizzarli per model-checking e diagnostica.

• dobbiamo controllare che i residui:

- che abbiano media 0
- che abbiano varianza costante
- che siano incorrelati o molto poco

} \rightarrow come da assunzioni degli errori

=> per farlo si plotta
residui $\sim x$

Il coefficiente R^2

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

• valori fra 0 e 1

• è lo per centuale di variabilità spiegata dal modello

$$SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2$$

differenza tra osservazioni
e media
(variabilità totale)

$$SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

differenza tra il valore
stimato dal modello e
la media
(variabilità spiegata dal modello)

$$SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

differenza tra valore
osservato e stima
del modello
(variabilità spiegata
dai residui)



Modello di regressione lineare a rumore gaussiano

- Mantenendo le assunzioni fatte per il modello lineare semplice assumiamo:

- $\varepsilon \sim N(0, \sigma^2)$, indipendente da X

$\overbrace{\sigma^2}$ varianza costante

Sotto la condizione di rumore gaussiano abbiamo che:

$$Y \sim N(B_0 + B_1 X, \sigma^2)$$

Stima di B_0 e B_1 (stima della massima verosimiglianza)

$$\mathcal{L}(B_0, B_1, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(y_i - (B_0 + B_1 x_i))^2}{2\sigma^2}}$$

(tramite derivate prima e seconde)

Troviamo i medesimi valori del metodo dei minimi quadrati



metodo minimi quadrati	$=$ metodo massima verosimiglianza (sotto l'ipotesi di rumore gaussiano)
------------------------------	--

\hat{B}_0 e \hat{B}_1 sotto l'ipotesi di rumore gaussiano:

$$\hat{B}_0 \sim N(B_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right))$$

$\hat{B}_1 \sim N(B_1, \left(\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right))$

\Leftarrow in quanto sia \hat{B}_0 che \hat{B}_1 sono
combinazioni lineari di y_i e
ogni y_i è distribuita normalmente

possiamo quindi standardizzarli:

$$\frac{\hat{\beta} - \beta}{SD[\hat{\beta}]} \sim N(0, 1)$$

$$SD[\hat{\beta}_0] = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$SD[\hat{\beta}_1] = \sigma \cdot \frac{1}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

non sappiamo σ ma possiamo campionarlo:

$$s_e^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}$$

e quindi ottenerlo:

$$SE[\hat{\beta}_0]/SD[\hat{\beta}_0] = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$SE[\hat{\beta}_1]/SD[\hat{\beta}_1] = s_e \cdot \frac{1}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

quindi se dividiamo per SD o SE ottieniamo:

$$\frac{\hat{\beta} - \beta}{SE[\hat{\beta}]} \sim t_{n-2} \quad \text{con } t :$$

- distribuzione simile a N

- con gradi di libertà che aumentano:

$$t \rightarrow N(0, 1)$$

Intervalli di confidenza per $\hat{\beta}_0$ e $\hat{\beta}_1$

- utilizzando il risultato di prima:

$$\begin{aligned}\hat{\beta}_0 &\pm t_{\alpha/2, n-2} \cdot SE[\hat{\beta}_0] \\ \hat{\beta}_1 &\pm t_{\alpha/2, n-2} \cdot SE[\hat{\beta}_1]\end{aligned}$$

↓
 gradi libertà

$1 - \alpha =$ probabilità che l'intervallo contenga
 β

es: se $1 - \alpha = 0.90$ (confidenza al 0.90)

allora $\alpha = 1 - 0.9$

$\alpha = 0.1 \rightarrow$ probabilità che il valore cada fuori dall'intervallo

$\frac{\alpha}{2} = 0.05 \rightarrow$ probabilità che il valore cada solo da una parte dell'intervallo.

Intervalli di confidenza per $\hat{m}(x)$

Sapendo che $\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

- β_0 e β_1 sono normali

⇓

$$\hat{m}(x) \sim N\left(\beta_0 + \beta_1 x, \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)\right)$$

utilizzando $SE(\hat{m}(x))$:

$$\hat{m}(x) \pm t_{\alpha/2, n-2} \cdot SE \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

intervallo
di confidenza

Intervalli di predizione per $\hat{m}(x)$

quando vogliamo prevedere una nuova osservazione abbiamo che:

$$Y - \hat{m}(x) \mid x, x_1, \dots, x_n \sim N\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right)$$

utilizzando SE:

$$\hat{m}(x) \pm t_{\alpha/2, n-2} \cdot \text{SE} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

intervallo di
predizione ↴
variabilità data dall'
errore della predizione

L'intervallo di predizione è SEMPRE più grande rispetto a un intervallo di confidenza

Test di ipotesi per $\hat{\beta}_0$ e $\hat{\beta}_1$

possiamo testare se β ha un certo valore β^*

$$H_0: \beta = \beta^* \quad \text{vs} \quad H_A: \beta \neq \beta^*$$

$$t = \frac{\hat{\beta} - \beta^*}{SE[\hat{\beta}]}$$

• utile quando $\beta^* = 0$ per testare la significatività della relazione lineare

H_0 : non c'è una relazione lineare significativa fra x e y

H_A : c'è una relazione lineare significativa fra x e y

Regressione lineare multipla

con p variabili predittive:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i(p-1)} + \epsilon_i \quad \text{con } i=1, \dots, n$$

\downarrow
n° osservazioni

• assumiamo che $\epsilon_i \sim N(0, \sigma^2)$

σ^2
costante

• usando una notazione matriciale:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1(p-1)} \\ 1 & x_{21} & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & & x_{n(p-1)} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

\downarrow
vettore
di size n \downarrow
matrice
n × p \downarrow
vettore
size p \downarrow
vettore
size n

$$Y = X \cdot \beta + \epsilon$$

Stima di β

possiamo stimare β minimizzando SSE:

$$\sum_{i=1}^n \left(\underbrace{Y_i}_{\text{dato osservato}} - \underbrace{\beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i(p-1)}}_{\text{dato stimato}} \right)^2$$

prendiamo p derivate: (p equazioni di stima)

$$X^\top X \beta = X^\top Y$$

Troviamo:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

Stima σ : $\hat{\sigma}/Se^2$

sotto l'ipotesi di rumore gaussiano:

$$Se^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-p}$$

↓
 parametri stimati

$$\text{dove } e = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

Intervallo di confidenza e test ipotesi per $\hat{\beta}$

possiamo dire che:

$$\hat{\beta} \sim N_p(\beta, \sigma^2 (x^\top x)^{-1})$$

dove N_p è una normale multivariata se ognuna delle p componenti di $\hat{\beta}$ si distribuisce come una normale:

$$\forall \hat{\beta}_j \in \hat{\beta} : \quad \hat{\beta}_j \sim N(\beta_j, \sigma^2 \cdot (x^\top x)_{jj}^{-1})$$

reali

empirici

$$\text{Var}[\hat{\beta}] = \sigma^2 (x^\top x)^{-1} \quad SE[\hat{\beta}] = Se \cdot \sqrt{\text{diag}(x^\top x)^{-1}}$$

$$\text{Var}[\hat{\beta}_j] = \sigma^2 (x^\top x)_{jj}^{-1} \quad SE[\hat{\beta}_j] = Se \cdot \sqrt{(x^\top x)_{jj}^{-1}}$$

$\forall \beta_j$ possiamo definire

$$t = \frac{\hat{\beta}_j - \beta_j}{SE[\hat{\beta}_j]}$$

$$\hat{\beta}_j \pm t_{\alpha/2, n-p} \cdot se \sqrt{(x^T x)^{-1}_{jj}}$$

Intervallo di confidenza e predizione per $\hat{y}(x_0)$:

PER UNA SOLA VARIABILE PREDITTIVA

$$\hat{y}(x_0) \pm t_{\alpha/2, n-p} \cdot se \sqrt{x_0^T (x^T x)^{-1} x_0} \quad \text{intervallo confidenza}$$

$$\hat{y}(x_0) \pm t_{\alpha/2, n-p} \cdot se \sqrt{1 + x_0^T (x^T x)^{-1} x_0} \quad \text{intervallo predizione}$$

Selezione del modello

come possiamo capire quali variabili sono utili per prevedere la risposta?

- un R^2 potrebbe darci un'indicazione sull'ammontare di variabilità spiegata ma non possiamo dire con chiarezza quando è "abbastanza alto"

||

abbiamo bisogno di un test che misuri la significatività di un modello rispetto a un modello più semplice, senza la nostra variabile di interesse.

necessitiamo che il modello più semplice sia "contenuto" in quello complesso:

modello completo: $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{(p-1)} x_{i(p-1)} + \epsilon_i$

modello semplice: $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{(q-1)} x_{i(q-1)} + \epsilon_i$, dove $q < p$

costruiamo il test:

$$H_0: \beta_1 = \dots = \beta_{p-1} = 0 \quad H_A: \exists \beta_j \neq 0 \text{ con } j = q, \dots, (p-1)$$

Utilizzando il modello nullo:

$$H_0: \beta_1 = \dots = \beta_{p-1} = 0 \quad H_A: \exists \beta_j \neq 0 \text{ con } j = 1, \dots, (p-1)$$

$$y_i = \beta_0 + \epsilon_i$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i(p-1)} + \epsilon_i$$

→ modello nullo (contenuto in ogni altro)

quale test statist usiamo per il test?

$SS_{\text{res}}(H_0)$: ammontare di variabilità spiegata dai residui del modello semplice

$SS_{\text{res}}(H_A)$: ammontare di variabilità spiegata dai residui del modello complesso

se $SS_{\text{res}}(H_0) - SS_{\text{res}}(H_A)$ è piccolo allora i residui spiegano lo stesso ammontare di variabilità del modello semplice.

↑
preferiremo quindi il
modello semplice

se $SS_{\text{res}}(H_0) - SS_{\text{res}}(H_A)$ allora preferiremo l'altro.

$$\frac{SS_{\text{res}}(H_0) - SS_{\text{res}}(H_A)}{SS_{\text{res}}(H_A)}$$

F-distribution

$$x_1 \sim \chi^2_{d_1} \quad \frac{x_1/d_1}{x_2/d_2} \sim F_{d_1, d_2}$$

$$x_2 \sim \chi^2_{d_2}$$

diventa:

$$F = \frac{\left(SS_{\text{res}}(H_0) - SS_{\text{res}}(H_A) \right) / (p-q)}{SS_{\text{res}}(H_A) / (n-p)}$$

calcoliamo il p-value:

p-value basso: rifiuto H_0

p-value alto: non rifiuto H_0

anova (mod-semplice, mod-complex)

Come posso trovare il modello migliore fra tutti i possibili?

Criteri di qualità del modello:

. R^2 aumenta sempre anche aggiungendo variabili poco significative

↓
abbiamo bisogno di dei criteri che prendono in considerazione anche la complessità del modello.

. IC: bontà modello + penalità complessità
(MINIMIZZA MEGLIO)

$$AIC(M) = n \cdot \log SS_{\text{res}}(M) + 2p(M)$$

$$BIC(M) = n \cdot \log SS_{\text{res}}(M) + (\log(n)p(M))$$

in quanto
 $\log(n) > 2$ if $n \geq 8$
il BIC penalizza di più la complessità del modello.

in R: AIC(model) BIC(model)

• Adjusted R^2 (MASSIMIZZARE MEGLIO)

$$1 - \frac{SS_{\text{res}} / (n - p - 1)}{SS_{\text{tot}} / (n - 1)}$$

prnde i c
onsiderazione
la complessità

- idealmente vorremo il più semplice modello che spieghi i nostri dati

- metodi automatici di ricerca:

• forward stepwise selection: modello nullo \rightarrow modello completo

• backward stepwise selection: modello completo \rightarrow modello nullo

• stepwise search: sia backward che forward

in ogni momento valuta l'aggiunta / rimozione di tutte le variabili non presenti / presenti nel modello, utilizzando il criterio di qualità scelto. In seguito inserisce / rimuove la variabile con valore migliore non per stepwise search

in R: `step(modello-di-partenza, scope = list(lower = modello-di-partenza),`

`upper = modello-di-arrivo), direction = " ",`

`K = " ", trace = 1)`

\downarrow
 $2 \rightarrow \text{AIC}$

$\log(n) \rightarrow \text{BIC}$

forward
backward
both

Exhaustive search

- considera tutti i possibili sottoinsiomi di variabili esplicative e trova il modello con valore del criterio scelto migliore

In R: `all_mod <- summary(regsubsets(formula, data = autompg))`

y_2 .

`all_mod$which[wchick.max(all_mod$bic),]`

Inferenza dopo la selezione

utilizziamo i dati 2 volte : prima per fare selezione e poi per fare inferenza 



selezioneremo un modello che modella bene i dati, potrebbe non scovare la vera relazione tra x e y non osservati.



soltuzione: cross-validation

Validation based selection

$$RMSE_{200cv} = \sqrt{\frac{1}{n} \sum_{i=1}^n e_{[i]}^2}$$

[MINIMIZZARE È MEGLIO]

Root mean squared error, leave one out cross-validation

1) filtriamo il modello con una osservazione (x_i, y_i) in meno,

2) utilizziamo il modello per predirne \hat{y}_i , chiamandola $\hat{y}_{[i]}$

3) calcoliamo $e_{[i]} = y_i - \hat{y}_{[i]}$

4) ripeti n volte \neq osservazione

5) ottieniamo il RMSE



Variabili categoriali:

R tratta le variabili categoriali come fattori

`is.factor(variable)` → c'è una variabile fattore?
TRUE / FALSE

`var ← as.factor(var)` → converte in variabile fattore

Il valore di una variabile categoriale viene creato una variabile binaria (one hot encoding)



si crea quindi un modello diverso

al variare della categoria (cambia il valore di β_0)

es: • sia x_2 una variabile binaria

• sia $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

$$\begin{cases} \text{se } x_2=0 \Rightarrow y = \beta_0 + \beta_1 x_1 + \epsilon \\ \text{se } x_2=1 \Rightarrow y = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon \end{cases}$$

Interazioni:

es: • sia x_2 una variabile binaria

• sia $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$

$$\begin{cases} \text{se } x_2=0 \Rightarrow y = \beta_0 + \beta_1 x_1 + \epsilon \\ \text{se } x_2=1 \Rightarrow y = (\underbrace{\beta_0 + \beta_2}_{\text{intervetta}}) + \underbrace{(\beta_1 + \beta_3)}_{\text{coefficiente angolare}} x_1 + \epsilon \end{cases}$$

In R:

1) `lm(y ~ x1 + x2 + x1:x2, data=...)`
stessa cosa

2) `lm(y ~ x1 * x2, data=...)` 

Controllo del modello:

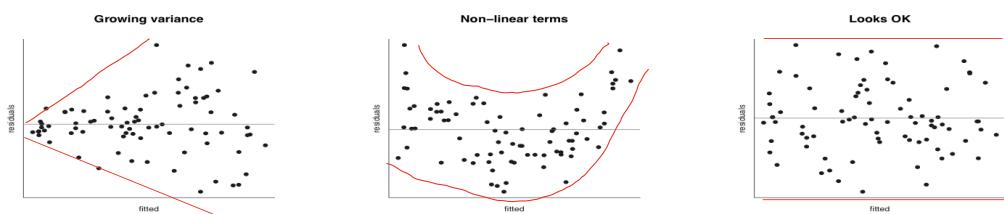
- se le assunzioni teoriche non sono valide l'inferenza potrebbe essere ambigua.
(meglio tenere una variabile che migliora le assunzioni anche se poco significativa)
- costruire un modello è un processo iterativo :
 - controllo obiettivo
 - costruzione modello
 - controllo assunzioni

Assunzioni:

- Linearity : $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i$
- Independence : errori sono indipendenti
- Normality : $\epsilon_i \sim N(0, \sigma^2) \Rightarrow y_i \sim N(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, \sigma^2)$ se
- Equal variance : $\text{Var}[\epsilon_i] = \sigma^2$, se

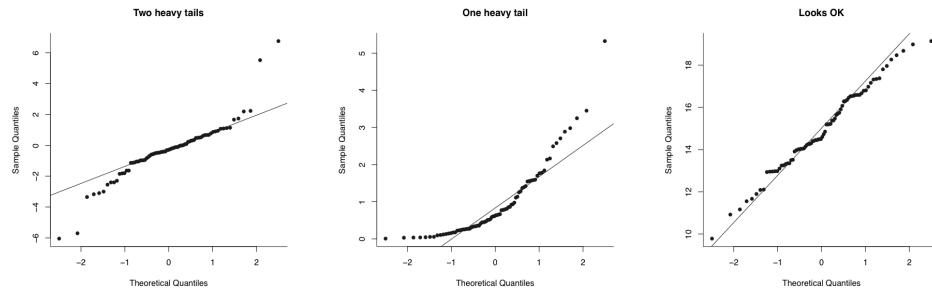
Usiamo i residui (errori empirici) per verificare le assunzioni

Plot v_i vs \hat{y}_i



- se la varianza è costante i punti devono essere distribuiti intorno alla media (0), senza assumere forme strane
omoschedasticità dei residui

QQ plot



- controllo assunzione di normalità

In R:

`plot(lm.object)` → tutti grafici utili

Trasformazioni:

- per risolvere assunzione di Linearietà → trasformo x .
- per risolvere assunzione di varianza costante → trasformo y .

Trasformazioni della risposta:

- una funzione stabilizzatrice della varianza comune è il logaritmo.
- utile se y :
 - range di valori molto ampio
 - positiva

$$\log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$$

||

$$y_i = e^{\beta_0 + \beta_1 x_i} + \epsilon_i$$

su β_0 scala originale

• un'altra funzione : Box-Cox transform

$$y_\lambda = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0 \\ \log(y) & \text{se } \lambda = 0 \end{cases}$$

- utile se $y > 0$ sempre (possiamo trasformare la y sui positivi con una costante).

λ scelgo massimizzando la log verosimiglianza

- $\lambda < 1$ per dati "positive skew"
- $\lambda > 1$ per dati "negative skew"

In R :

$$1) \text{ lm}(\log(y) \sim x, \text{ data} = \dots)$$

$$2) \text{ boxcox}(\text{lm-object}, \text{ plotit=TRUE})$$

Trasformazioni della variabile predittiva

IMPORTANTE RICORDARE : il modello è chiamato lineare non perché la curva di regressione è un piano, ma perché gli effetti dei parametri sono lineari



prendiamo il risultato della trasformazione di x :

$$Y = \beta_0 + \beta_1 \cdot f(x) + \epsilon$$

dove $f(x)$ può essere:

- $\log(\cdot)$
- trasformazione polinomiale



||

✓ utilizziamo la serie di Taylor per approssimare con un polinomio, una funzione sconosciuta.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_K x_i^K + \epsilon_i$$

$\beta f(x)$

plot polinomio grado K

In R:

$$\text{lm}(Y \sim x + I(x^2) + I(x^3) + I(x^4))$$

oppure

$$\text{lm}(Y \sim \text{poly}(x, 4), \text{raw} = \text{TRUE})$$

Multicollinearità:

quando esiste una relazione lineare tra 2 o più predittori.
↓
multicollinearità

- le stime dei parametri diventano non affidabili e molto variabili

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad \text{assume che } (X^T X)^{-1} \text{ sia invertibile (non singolare)}$$

↓

se $(X^T X)$ è singolare la varianza diventa molto grande

- possiamo eliminare il problema cancellando una o più variabili predittive

quale variabile eliminare?

- collinearità
- le plotto e cerco relazioni lineari
 - quando alla matrice di correlazione

per quanto riguarda la multicollinearità potrebbe essere invisibile:

- guardiamo a:
 - grossi cambiamenti in β_i quando una variabile è aggiunta/tolta
 - β_i poco significativo preso insieme a altre variabili, ma significativo singolarmente.
- Variance Inflation Factor (VIF)

$$\begin{aligned} VIF_i &= \frac{(x^T x)^{-1}_{i+1, i+1}}{n \cdot s^2 x_i} && \forall i \in 0, \dots, (p-1) \\ &= \frac{\text{Var}[\hat{\beta}_i] \text{ covoletata}}{\text{Var}[\hat{\beta}_i] \text{ non covoletata}} = \frac{1}{1 - R_i^2} && \begin{array}{l} \text{ottenuto dal modello} \\ X_i \sim x_1, \dots, x_j \end{array} \end{aligned}$$

$\overline{VIF} \rightarrow$ media di tutti i coefficienti

se $VIF_i \geq 1$ predittori sono covaletati

se $\overline{VIF} \gg 1$ indica una forte multicollinearità

Punti di influenza:

GLM:

due proprietà:

- una distribuzione di Y appartenente alla famiglia esponenziale $Y \sim EF(\theta_i, \phi, a, b, c)$
 - una link function $g(\cdot)$ che definisca come la combinazione lineare dei predittori $\beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}$ è legata alla media della risposta condizionata ai predittori $\beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}$
- $g(E[Y|X_1=x_1, \dots, X_{p-1}=x_{p-1}]) = \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}$
- preserva la linearità

- il modello lineare semplice è un caso specifico del GLM dove:

$$- Y \sim N(\beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}, \sigma^2)$$

$$- \text{link fun: } I(\cdot) \rightarrow \text{identità} \quad \text{ovvero} \quad I(E[Y|X_1=x_1, \dots, X_{p-1}=x_{p-1}]) = \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}$$

Famiglia esponenziale:

- tutte le distribuzioni della forma

$$f(y; \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right\}$$

θ : parametro canonico $\mu = E[Y] = b'(\theta)$

ϕ : dispersione $\text{Var}[Y] = b''(\theta) \phi$

La funzione di link canonico è una funzione $g(\cdot)$ che trasforma $E[Y] = b'(\theta)$ in $g(E[Y]) = \theta$:

$$\theta = b'^{-1}(E[Y])$$

Distribuzioni esponenziali:

• Normale

$$g(\xi) = \xi$$

• Gamma

$$g(\xi) = \sqrt{\xi}$$

• Bernoulli

$$g(\xi) = \log\left(\frac{\xi}{1-\xi}\right)$$

• Poisson

$$g(\xi) = \log(\xi)$$

Stima dei parametri β :

• metodo della massima verosimiglianza



nessuna soluzione analitica



IRLS \rightarrow Algoritmo che stima
soluzione

$$\hat{\beta} \stackrel{\text{approx}}{\sim} N(\beta, I(\beta)^{-1}) \quad \text{con } n \rightarrow \infty \quad (\text{solo asintoticamente vero})$$

$I(\beta)^{-1}$: matrice di informazione di Fisher.

quindi:

$$z_j = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \stackrel{\text{approx}}{\sim} N(0, 1)$$

$$\hat{\beta}_j \pm z_{1-\alpha/2} \cdot SE(\hat{\beta}_j)$$

• In R:

`glm(formula, family = "data = "...")`

↓
 poisson
 binomial
 gamma
 gaussian

Likelihood-Ratio Test (LRT)

- come possiamo fare model selection sui GLM?

LRT → analogo F-test ANOVA per la regressione

$$LR = 2 \left(\ell(\hat{\beta}_{\text{full}}) - \ell(\hat{\beta}_{\text{NLL}}) \right)$$

$$LR \underset{\text{approx}}{\sim} \chi^2_{p-q}$$

$\hat{\beta}_{\text{NLL}}$ modello nested di $\hat{\beta}_{\text{full}}$

$$\text{Devianza} = SS_{\text{TOT}}$$

in R:

`anova(null, full, test = "LRT")`

Residui GLM

$$r_i^P = \frac{y_i - \hat{y}_i}{\sqrt{s_i}}$$

(Classification:

Fare predizioni nel contesto della regressione logistica.

Bayes classifier:

$$\hat{C}(x_1, \dots, x_{p-1}) = \begin{cases} 1 & \hat{p}(x_1, \dots, x_{p-1}) > 0.5 \\ 0 & \hat{p}(x_1, \dots, x_{p-1}) \leq 0.5 \end{cases}$$

Usiamo la regressione logistica per stimare \hat{p}

Misclassification rate

$$\text{Misclass}(\hat{C}, \text{Dato}) = \frac{1}{n} \sum_{i=1}^n I(x_i \neq \hat{C}(x_i))$$

1 se vero
0 se falso

Devianza: Stesso ruolo delle SST nel modello Gaussiano.