

Probabilità e Processi Stocastici (455AA)

Lezione 12

Dario Trevisan

2/11/2023

Section 1

Stime dei parametri: caso vettoriale

Richiami

- Variabile gaussiana vettoriale $X \in \mathbb{R}^d$

$$p(X = x) \propto \exp\left(-\frac{1}{2}(x - m) \cdot \Sigma^{-1}(x - m)\right)$$

Richiami

- Variabile gaussiana vettoriale $X \in \mathbb{R}^d$

$$p(X = x) \propto \exp\left(-\frac{1}{2}(x - m) \cdot \Sigma^{-1}(x - m)\right)$$

- $\mathbb{E}[X] = m \in \mathbb{R}^d$, $\Sigma_X = \Sigma \in \mathbb{R}^{d \times d}$

Richiami

- Variabile gaussiana vettoriale $X \in \mathbb{R}^d$

$$p(X = x) \propto \exp\left(-\frac{1}{2}(x - m) \cdot \Sigma^{-1}(x - m)\right)$$

- $\mathbb{E}[X] = m \in \mathbb{R}^d$, $\Sigma_X = \Sigma \in \mathbb{R}^{d \times d}$
- Stima MLE dei parametri (caso scalare) da n osservazioni i.i.d. $X_1 = x_1, \dots, X_n = x_n$:

$$m_{\text{MLE}} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\sigma_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Stime nel caso vettoriale

I risultati della sezione precedente si possono estendere al caso vettoriale, ossia di n osservazioni di variabili aleatorie $X_1, \dots, X_n \in \mathbb{R}^d$, tutte indipendenti tra loro e ciascuna con densità gaussiana vettoriale di parametri comuni $\mathcal{N}(m, \Sigma)$.

Verosimiglianza

$$L(m, \Sigma; x) \propto \exp\left(-\frac{1}{2} \sum_{i=1}^n (x_i - m) \cdot \Sigma^{-1} (x_i - m)\right) \frac{1}{(\det \Sigma)^{n/2}}.$$

- Se la varianza $\Sigma = \Sigma_0$ è nota (rispetto all'informazione prima di osservare le X_i), la stima di massima verosimiglianza per il parametro di media è la media campionaria

$$m_{\text{MLE}} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

(qualsiasi sia Σ_0).

- Se il parametro di media $m = m_0 \in \mathbb{R}^d$ è noto (rispetto all'informazione a priori), allora la stima di massima verosimiglianza per la covarianza è

$$\Sigma_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n (x_i - m_0)(x_i - m_0)^T,$$

dove T indica l'operazione di trasposizione (quindi il prodotto righe per colonne risulta in una matrice $d \times d$); più esplicitamente, la stima della covarianza tra la componente j e k è

$$(\Sigma_{\text{MLE}})_{jk} = \frac{1}{n} \sum_{i=1}^n (x_i - m_0)_j (x_i - m_0)_k.$$

La stima (congiunta) di massima verosimiglianza per (m, Σ) è data dalla media e dalla covarianza campionarie:

$$m_{\text{MLE}} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \Sigma_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T.$$

- Osserviamo che Σ_{MLE} è una matrice simmetrica e semi-definita positiva. La si può anche interpretare come la matrice di covarianza della variabile aleatoria vettoriale che sceglie uno degli n valori osservati con probabilità uniforme discreta.

Coefficiente di correlazione campionario

È utile anche considerare la matrice delle correlazioni, in cui al posto delle covarianze è calcolato il coefficiente di correlazione campionario,

$$\bar{\rho}_{jk} = \frac{\Sigma_{jk}}{\sqrt{\Sigma_{jj}\Sigma_{kk}}},$$

che è sempre compreso tra -1 e 1 (segue dal fatto che la matrice $\Sigma = \Sigma_{MLE}$ è semi-definita positiva). Il comando in questo caso è `cor()`.

```
##           Sepal.Length Sepal.Width Petal.Length Petal.Wi
## Sepal.Length           1.00          -0.12           0.87
## Sepal.Width           -0.12           1.00          -0.43
## Petal.Length           0.87          -0.43           1.00
## Petal.Width            0.82          -0.37           0.96
```

Correlogramma

Per visualizzare la correlazione si può usare un *correlogramma*.

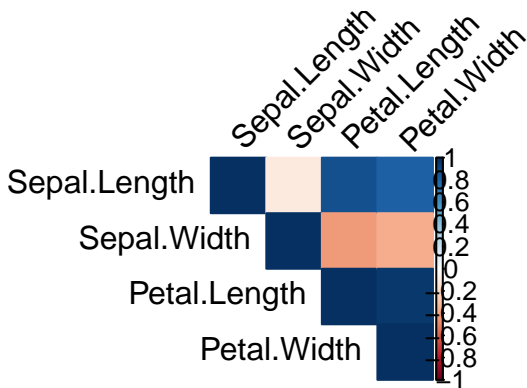
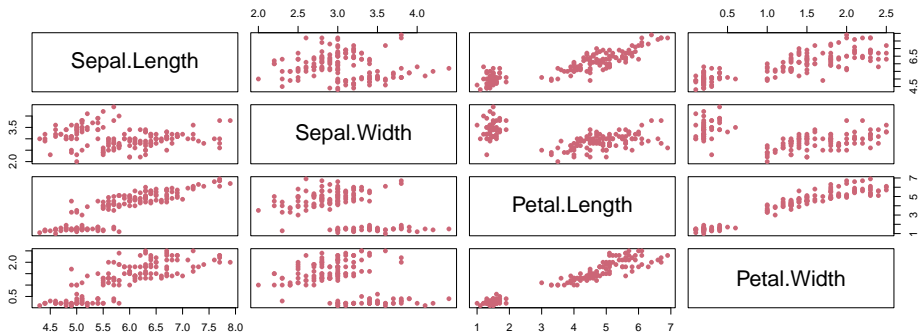


Figure 1: Correlogramma del dataset Iris

Diagramma a dispersione

Il comando `plot()` applicato direttamente ai dati fornisce invece un diagramma a dispersione di tutte le possibili coppie.



Section 2

PCA

Analisi delle componenti principali (PCA)

Problema: *ridurre la dimensionalità* di una variabile $Y \in \mathbb{R}^d$ (o di un campione di taglia n), con $d \gg 1$, definendo una variabile $X \in \mathbb{R}^k$, con $k \ll d$.

- Questo può essere utile per rappresentare graficamente Y (ad esempio se $k = 2$) ma soprattutto anche per velocizzare l'esecuzione di algoritmi che in dimensione alta possono risultare lenti.

Analisi delle componenti principali (PCA)

Problema: *ridurre la dimensionalità* di una variabile $Y \in \mathbb{R}^d$ (o di un campione di taglia n), con $d \gg 1$, definendo una variabile $X \in \mathbb{R}^k$, con $k \ll d$.

- Questo può essere utile per rappresentare graficamente Y (ad esempio se $k = 2$) ma soprattutto anche per velocizzare l'esecuzione di algoritmi che in dimensione alta possono risultare lenti.
- Problema estremamente comune e molteplici tecniche per affrontarlo. L'**analisi delle componenti principali** (in inglese *principal component analysis*, abbreviato PCA) è forse la più semplice, ma spesso efficace.

Punto di vista teorico

La PCA si spiega a partire dalla *standardizzazione* di un vettore aleatorio.

- Data $Y \in \mathbb{R}^d$, la matrice delle covarianze Σ_Y può essere diagonalizzata ossia esiste $U_Y \in \mathbb{R}^{d \times d}$ ortogonale ($U_Y^T = U_Y^{-1}$) tale che

$$U_Y \Sigma_Y U_Y^T = D_Y$$

è diagonale (e contiene gli autovalori di Σ_Y).

Punto di vista teorico

La PCA si spiega a partire dalla *standardizzazione* di un vettore aleatorio.

- Data $Y \in \mathbb{R}^d$, la matrice delle covarianze Σ_Y può essere diagonalizzata ossia esiste $U_Y \in \mathbb{R}^{d \times d}$ ortogonale ($U_Y^T = U_Y^{-1}$) tale che

$$U_Y \Sigma_Y U_Y^T = D_Y$$

è diagonale (e contiene gli autovalori di Σ_Y).

- Posta $Y' = U_Y Y$, la matrice delle covarianze di Y' è diagonale.

Punto di vista teorico

La PCA si spiega a partire dalla *standardizzazione* di un vettore aleatorio.

- Data $Y \in \mathbb{R}^d$, la matrice delle covarianze Σ_Y può essere diagonalizzata ossia esiste $U_Y \in \mathbb{R}^{d \times d}$ ortogonale ($U_Y^T = U_Y^{-1}$) tale che

$$U_Y \Sigma_Y U_Y^T = D_Y$$

è diagonale (e contiene gli autovalori di Σ_Y).

- Posta $Y' = U_Y Y$, la matrice delle covarianze di Y' è diagonale.
- Si definisce $X \in \mathbb{R}^k$ come la variabile congiunta delle k coordinate di Y' che hanno varianza maggiore. Vale quindi

$$X = \Pi_Y Y.$$

dove Π_Y è la proiezione ortogonale sul sottospazio k -dimensionale di “maggior variabilità”.

Due problemi

- 1 spesso si dispone solamente di n di osservazioni (y_1, \dots, y_n) associate a variabili aleatorie (Y_1, \dots, Y_n) , tutte indipendenti tra loro e con la stessa legge di Y (un campione). Come stimare Π_Y ?

Due problemi

- 1 spesso si dispone solamente di n di osservazioni (y_1, \dots, y_n) associate a variabili aleatorie (Y_1, \dots, Y_n) , tutte indipendenti tra loro e con la stessa legge di Y (un campione). Come stimare Π_Y ?
- 2 la PCA è una procedura *ad-hoc* per questo problema oppure si può giustificare mediante le regole del calcolo della probabilità?

PCA empirica

Soluzione al primo problema:

- invece di considerare la matrice delle covarianze teorica Σ_Y (non nota), partendo dalle osservazioni $y = (y_1, \dots, y_n)$, si procede partendo però dalla matrice delle covarianze campionarie,

$$\Sigma_y = \frac{1}{n} \sum_{i=1} (y_i - \bar{y})(y_i - \bar{y})^T.$$

PCA empirica

Soluzione al primo problema:

- invece di considerare la matrice delle covarianze teorica Σ_Y (non nota), partendo dalle osservazioni $y = (y_1, \dots, y_n)$, si procede partendo però dalla matrice delle covarianze campionarie,

$$\Sigma_y = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T.$$

- Il teorema spettrale applicato Σ_y determina allora una matrice ortogonale $U_y \in \mathbb{R}^{d \times d}$ e una matrice diagonale $D_y \in \mathbb{R}^{d \times d}$ (contenente gli autovalori di Σ_y) tali che

$$U_y \Sigma_y U_y^T = D_y.$$

PCA empirica

Soluzione al primo problema:

- invece di considerare la matrice delle covarianze teorica Σ_Y (non nota), partendo dalle osservazioni $y = (y_1, \dots, y_n)$, si procede partendo però dalla matrice delle covarianze campionarie,

$$\Sigma_y = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T.$$

- Il teorema spettrale applicato Σ_y determina allora una matrice ortogonale $U_y \in \mathbb{R}^{d \times d}$ e una matrice diagonale $D_y \in \mathbb{R}^{d \times d}$ (contenente gli autovalori di Σ_y) tali che

$$U_y \Sigma_y U_y^T = D_y.$$

- Si definisce $\Pi_y \in \mathbb{R}^{k \times d}$ come la matrice di proiezione nel sottospazio associato alle direzioni dei k vettori di U_y per cui le componenti nella diagonale (le varianze) sono il più grande possibile.

PCA empirica

Soluzione al primo problema:

- invece di considerare la matrice delle covarianze teorica Σ_Y (non nota), partendo dalle osservazioni $y = (y_1, \dots, y_n)$, si procede partendo però dalla matrice delle covarianze campionarie,

$$\Sigma_y = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T.$$

- Il teorema spettrale applicato Σ_y determina allora una matrice ortogonale $U_y \in \mathbb{R}^{d \times d}$ e una matrice diagonale $D_y \in \mathbb{R}^{d \times d}$ (contenente gli autovalori di Σ_y) tali che

$$U_y \Sigma_y U_y^T = D_y.$$

- Si definisce $\Pi_y \in \mathbb{R}^{k \times d}$ come la matrice di proiezione nel sottospazio associato alle direzioni dei k vettori di U_y per cui le componenti nella diagonale (le varianze) sono il più grande possibile.

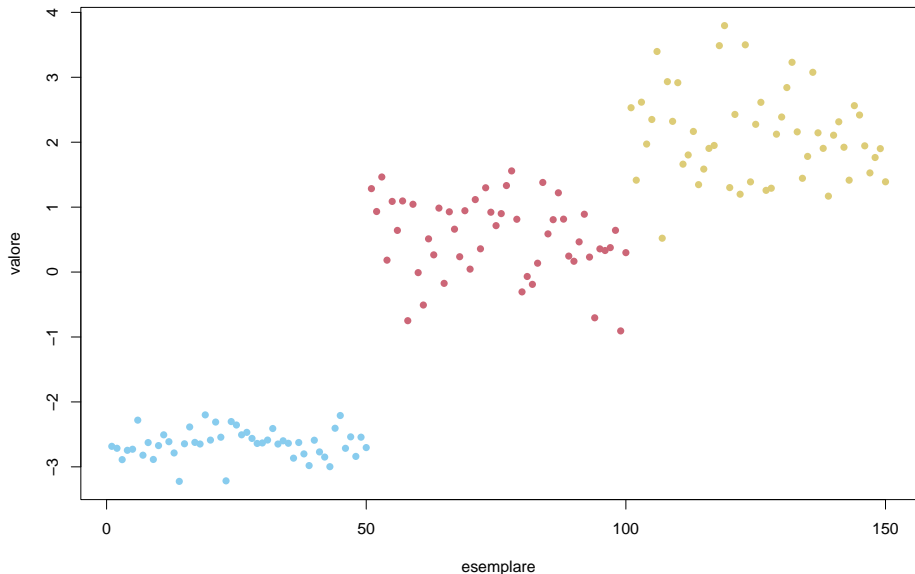
Applichiamo la PCA in R usando la funzione specifica `prcomp()`. Vediamo ad esempio sul dataset Iris (applicandolo solo ai dati numerici, escludendo la colonna della specie) :

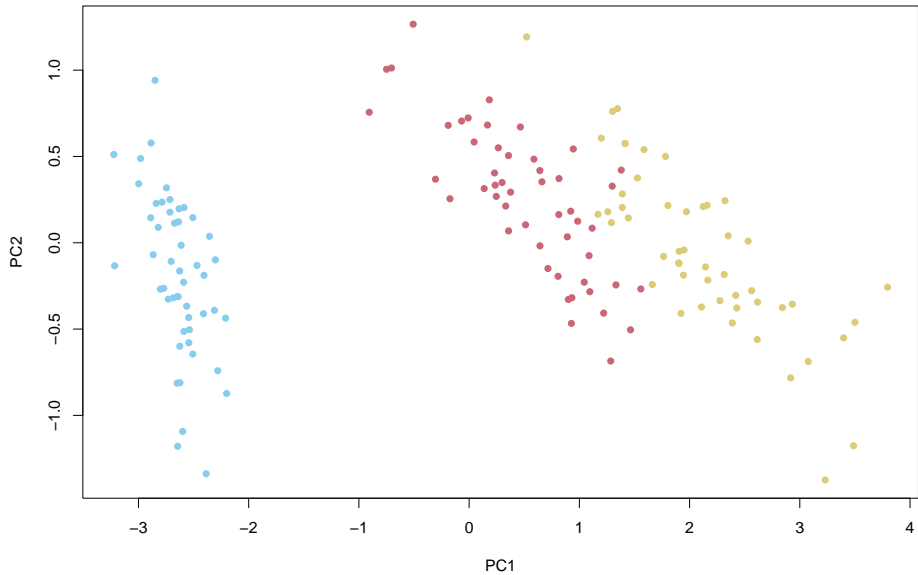
L'oggetto risultante contiene diverse informazioni sulla PCA, come ad esempio la base di vettori U (una matrice 4×4) e le deviazioni standard delle varie componenti.

```
##              PC1              PC2              PC3              PC4
## Sepal.Length  0.36138659 -0.65658877  0.58202985  0.3154872
## Sepal.Width   -0.08452251 -0.73016143 -0.59791083 -0.3197231
## Petal.Length  0.85667061  0.17337266 -0.07623608 -0.4798390
## Petal.Width   0.35828920  0.07548102 -0.54583143  0.7536574

## [1] 2.0562689 0.4926162 0.2796596 0.1543862
```


La PCA ha l'effetto in questo caso di ben separare le tre specie, o quanto meno la prima dalle altre due, come evidenziamo con la diversa colorazione.





Una giustificazione tramite MLE della PCA

È possibile giustificare il metodo di PCA in termini di una stima di massima verosimiglianza per un opportuno modello gaussiano.

- L'idea è che con la PCA stiamo recuperando un “segnale” (la X) osservandone una versione “rumorosa” e disposta su un sottospazio non noto.

- fissata (e nota) la dimensione k , introduciamo una variabile standardizzata $Z \in \mathbb{R}^k$ e imponiamo che valga

$$Y = AZ + W,$$

dove $A \in \mathbb{R}^{d \times k}$ è una matrice non nota (rispetto all'informazione priori).

- fissata (e nota) la dimensione k , introduciamo una variabile standardizzata $Z \in \mathbb{R}^k$ e imponiamo che valga

$$Y = AZ + W,$$

dove $A \in \mathbb{R}^{d \times k}$ è una matrice non nota (rispetto all'informazione priori).

- Il “segnale” da ricostruire è quindi AZ (quello che nella PCA abbiamo chiamato X) e W è una variabile che rappresenta il “rumore” aggiunto.

- fissata (e nota) la dimensione k , introduciamo una variabile standardizzata $Z \in \mathbb{R}^k$ e imponiamo che valga

$$Y = AZ + W,$$

dove $A \in \mathbb{R}^{d \times k}$ è una matrice non nota (rispetto all'informazione priori).

- Il “segnale” da ricostruire è quindi AZ (quello che nella PCA abbiamo chiamato X) e W è una variabile che rappresenta il “rumore” aggiunto.
- Supponiamo che Z , W siano indipendenti con densità gaussiane centrate e, oltre a $\Sigma_Z = Id$, supponiamo che $\Sigma_W = \sigma_0^2 Id$, per una costante opportuna (nota a priori e sufficientemente piccola).

Supponendo nota la matrice A , allora la densità di Y , è anch'essa una gaussiana centrata, con covarianza $\Sigma_Y = AA^T + \sigma_0^2 Id$, ch     una funzione di A . Pertanto la verosimiglianza di A associata ad Y si scrive

$$L(A; y) = p(Y = y|A) \propto \exp\left(-\frac{1}{2} \left(y^T \Sigma_Y^{-1} y + \log(\det(\Sigma_Y))\right)\right).$$

- se supponiamo di avere n osservazioni indipendenti $Y_i = y_i$, tutte gaussiane con gli stessi parametri – in particolare con la stessa matrice A , la verosimiglianza si ottiene come prodotto della funzione sopra (cambiando i valori osservati)

$$L(A; y_1, \dots, y_n) \propto \exp\left(-\frac{n}{2} \left(\frac{1}{n} \sum_{i=1}^n y_i^T \Sigma_Y^{-1} y_i + \log(\det(\Sigma_Y))\right)\right).$$

Supponendo nota la matrice A , allora la densità di Y , è anch'essa una gaussiana centrata, con covarianza $\Sigma_Y = AA^T + \sigma_0^2 Id$, ch  è una funzione di A . Pertanto la verosimiglianza di A associata ad Y si scrive

$$L(A; y) = p(Y = y|A) \propto \exp\left(-\frac{1}{2} \left(y^T \Sigma_Y^{-1} y + \log(\det(\Sigma_Y))\right)\right).$$

- se supponiamo di avere n osservazioni indipendenti $Y_i = y_i$, tutte gaussiane con gli stessi parametri – in particolare con la stessa matrice A , la verosimiglianza si ottiene come prodotto della funzione sopra (cambiando i valori osservati)

$$L(A; y_1, \dots, y_n) \propto \exp\left(-\frac{n}{2} \left(\frac{1}{n} \sum_{i=1}^n y_i^T \Sigma_Y^{-1} y_i + \log(\det(\Sigma_Y))\right)\right).$$

- la massima verosimiglianza si ottiene minimizzando

$$A \mapsto \frac{1}{n} \sum_{i=1}^n y_i^T \Sigma_Y^{-1} y_i + \log(\det(\Sigma_Y))$$

- La stima di massima verosimiglianza per A è data da

$$A_{\text{MLE}} = U_{y|k}(D_{y|k} - \sigma_0^2 \text{Id})^{1/2},$$

dove $U_{y|k} \in \mathbb{R}^{d \times k}$ indica la matrice corrispondente ai k autovettori della covarianza campionaria $\Sigma_y = \sum_{i=1}^n y_i y_i^T$ con autovalori più grandi, e $D_{y|k} \in \mathbb{R}^{k \times k}$ indica la matrice diagonale contenente tali autovalori nell'ordine corrispondente.

- La stima di massima verosimiglianza per A è data da

$$A_{\text{MLE}} = U_{y|k}(D_{y|k} - \sigma_0^2 \text{Id})^{1/2},$$

dove $U_{y|k} \in \mathbb{R}^{d \times k}$ indica la matrice corrispondente ai k autovettori della covarianza campionaria $\Sigma_y = \sum_{i=1}^n y_i y_i^T$ con autovalori più grandi, e $D_{y|k} \in \mathbb{R}^{k \times k}$ indica la matrice diagonale contenente tali autovalori nell'ordine corrispondente.

- Tutto questo purché σ_0^2 sia sufficientemente piccolo. Nel limite $\sigma_0 \rightarrow 0$ si ottiene che $A_{\text{MLE}} = U_{y|k} D_{y|k}^{1/2}$ e la variabile $A_{\text{MLE}} X$ si identifica con $\Pi_y Y$.

Section 3

Regressione

Regressione

Il problema della regressione è il seguente: date due variabili aleatorie $X \in E$, $Y \in E'$ determinare una funzione $g : E \rightarrow E'$ tale che Y sia “molto vicina” a $g(X)$, ossia

$$Y \sim g(X),$$

a partire dall'osservazione congiunta di (X, Y) (sottoforma di una o più copie, solitamente indipendenti).

- La variabile X è detta **predittore** (in inglese *predictor*) o **variabile esplicativa** (*explanatory variable*),

Regressione

Il problema della regressione è il seguente: date due variabili aleatorie $X \in E$, $Y \in E'$ determinare una funzione $g : E \rightarrow E'$ tale che Y sia “molto vicina” a $g(X)$, ossia

$$Y \sim g(X),$$

a partire dall'osservazione congiunta di (X, Y) (sottoforma di una o più copie, solitamente indipendenti).

- La variabile X è detta **predittore** (in inglese *predictor*) o **variabile esplicativa** (*explanatory variable*),
- La variabile Y è detta **risposta** (*response*) oppure **esito** (*outcome*).

Regressione

Il problema della regressione è il seguente: date due variabili aleatorie $X \in E$, $Y \in E'$ determinare una funzione $g : E \rightarrow E'$ tale che Y sia “molto vicina” a $g(X)$, ossia

$$Y \sim g(X),$$

a partire dall'osservazione congiunta di (X, Y) (sottoforma di una o più copie, solitamente indipendenti).

- La variabile X è detta **predittore** (in inglese *predictor*) o **variabile esplicativa** (*explanatory variable*),
- La variabile Y è detta **risposta** (*response*) oppure **esito** (*outcome*).
- Evitiamo appositamente il linguaggio trazionale di variabile “indipendente” (che sarebbe la X) e “dipendente” (la Y) per evitare di confonderlo con il concetto probabilistico di indipendenza.

In molte applicazioni, X è una variabile vettoriale, ossia $E = \mathbb{R}^d$ (se $d > 1$ si parla di regressione multipla),

- a seconda dell'obiettivo che ci si pone, Y potrebbe anche essere una variabile discreta (noi ci concentreremo al caso in cui sia una variabile continua).

In molte applicazioni, X è una variabile vettoriale, ossia $E = \mathbb{R}^d$ (se $d > 1$ si parla di regressione multipla),

- a seconda dell'obiettivo che ci si pone, Y potrebbe anche essere una variabile discreta (noi ci concentreremo al caso in cui sia una variabile continua).
- La regressione si usa anche per problemi di *classificazione*, in cui ad esempio bisogna “etichettare” i possibili valori di X per determinare due (o più) classi disgiunte e quindi la variabile di risposta $g(X)$ è discreta a valori nell'insieme delle possibili etichette.

Formulazione generale

- L'incognita del problema g di solito non è completamente determinata dall'osservazione di (X, Y)

Formulazione generale

- L'incognita del problema g di solito non è completamente determinata dall'osservazione di (X, Y)
- Si introduce una variabile aleatoria G a valori nell'insieme delle possibili funzioni da E in E' .

Formulazione generale

- L'incognita del problema g di solito non è completamente determinata dall'osservazione di (X, Y)
- Si introduce una variabile aleatoria G a valori nell'insieme delle possibili funzioni da E in E' .
- Il problema diventa quindi determinare la legge di G sulla base dell'informazione a priori I e dei dati osservati, ossia (X, Y) .

Regressione e curve interpolanti

La regressione generalizza quindi il concetto di “curva interpolante”, o più in generale il problema di determinare una funzione il cui grafico passi per determinati punti (x, y) . Questa generalizzazione avviene almeno su due fronti:

- 1 G non è una singola funzione ma una densità di probabilità sulle funzioni (ovviamente poi si dovrà scegliere una stima, ad esempio tramite massima verosimiglianza)

Regressione e curve interpolanti

La regressione generalizza quindi il concetto di “curva interpolante”, o più in generale il problema di determinare una funzione il cui grafico passi per determinati punti (x, y) . Questa generalizzazione avviene almeno su due fronti:

- 1 G non è una singola funzione ma una densità di probabilità sulle funzioni (ovviamente poi si dovrà scegliere una stima, ad esempio tramite massima verosimiglianza)
- 2 non si richiede che la curva interpoli esattamente i punti osservati, ma introducendo un certo “residuo” (o errore), definito spesso come la differenza tra Y e $G(X)$ ossia $Y - G(X)$.

Modelli

- 1 In teoria per affrontare la regressione basta ragionare con il solito schema:

Modelli

- 1 In teoria per affrontare la regressione basta ragionare con il solito schema:
 - specificare una densità a priori per G

Modelli

- 1 In teoria per affrontare la regressione basta ragionare con il solito schema:
 - specificare una densità a priori per G
 - usare Bayes per stimare la densità a posteriori date le osservazioni (X, Y)

Modelli

- 1 In teoria per affrontare la regressione basta ragionare con il solito schema:
 - specificare una densità a priori per G
 - usare Bayes per stimare la densità a posteriori date le osservazioni (X, Y)
 - in alternativa usare la stima di massima verosimiglianza.

Modelli

- 1 In teoria per affrontare la regressione basta ragionare con il solito schema:
 - specificare una densità a priori per G
 - usare Bayes per stimare la densità a posteriori date le osservazioni (X, Y)
 - in alternativa usare la stima di massima verosimiglianza.
- 2 Il problema è che l'insieme delle funzioni è troppo grande per essere trattato. Soluzione:

Modelli

- 1 In teoria per affrontare la regressione basta ragionare con il solito schema:
 - specificare una densità a priori per G
 - usare Bayes per stimare la densità a posteriori date le osservazioni (X, Y)
 - in alternativa usare la stima di massima verosimiglianza.
- 2 Il problema è che l'insieme delle funzioni è troppo grande per essere trattato. Soluzione:
 - specificare un *modello*, una famiglia (parametrizzata) di funzioni:

Modelli

- 1 In teoria per affrontare la regressione basta ragionare con il solito schema:
 - specificare una densità a priori per G
 - usare Bayes per stimare la densità a posteriori date le osservazioni (X, Y)
 - in alternativa usare la stima di massima verosimiglianza.
- 2 Il problema è che l'insieme delle funzioni è troppo grande per essere trattato. Soluzione:
 - specificare un *modello*, una famiglia (parametrizzata) di funzioni:
 - introdurre un parametro U a valori in \mathbb{R}^k e per ogni valore $U = u$ definire

$$g(\cdot; u) : E \rightarrow E', \quad \text{ad } x \in E \text{ associa } g(x; u) \in E'.$$

Modelli lineari

Sia $E' = \mathbb{R}^{d'}$ e la parametrizzazione sia *lineare* nel parametro $U \in \mathbb{R}^k$, cioè

$$u \in \mathbb{R}^k \mapsto g(\cdot; u)$$

sia della forma

$$g(x; u) = \sum_{i=1}^k g_i(x) u_i$$

per opportune funzioni (note e fissate a priori) $g_i : E \rightarrow E' = \mathbb{R}^{d'}$.

- Ciascuna g_i , $x \mapsto g_i(x)$, può anche essere *non* lineare, ad esempio $g(x) = x^2$,

Modelli lineari

Sia $E' = \mathbb{R}^{d'}$ e la parametrizzazione sia *lineare* nel parametro $U \in \mathbb{R}^k$, cioè

$$u \in \mathbb{R}^k \mapsto g(\cdot; u)$$

sia della forma

$$g(x; u) = \sum_{i=1}^k g_i(x) u_i$$

per opportune funzioni (note e fissate a priori) $g_i : E \rightarrow E' = \mathbb{R}^{d'}$.

- Ciascuna g_i , $x \mapsto g_i(x)$, può anche essere *non* lineare, ad esempio $g(x) = x^2$,
- si può considerare, al posto della X , la variabile $X' = (g_i(X))_{i=1}^k$, così

$$g(x'; u) = \sum_{i=1}^k x'_i u_i.$$

- Esempio:

$$g(x; (u_1, u_2)) = u_1x + u_2x^2$$

- Esempio:

$$g(x; (u_1, u_2)) = u_1x + u_2x^2$$

- anche un modello “affine”

$$g(x; u) = u_0 + \sum_{i=1}^k x_i u_i$$

è in realtà lineare (nella u)

Modello logistico

Un esempio di modello non lineare è ottenuto tramite composizione del modello lineare con una funzione logistica (sigmoide)

$$\ell(z) = \frac{1}{1 + e^{-z}}.$$

Si ottiene pertanto un modello della forma

$$g(x; u) = \ell \left(\sum_{i=1}^k g_i(x) u_i \right) = \frac{1}{1 + \exp \left(- \sum_{i=1}^k g_i(x) u_i \right)}.$$

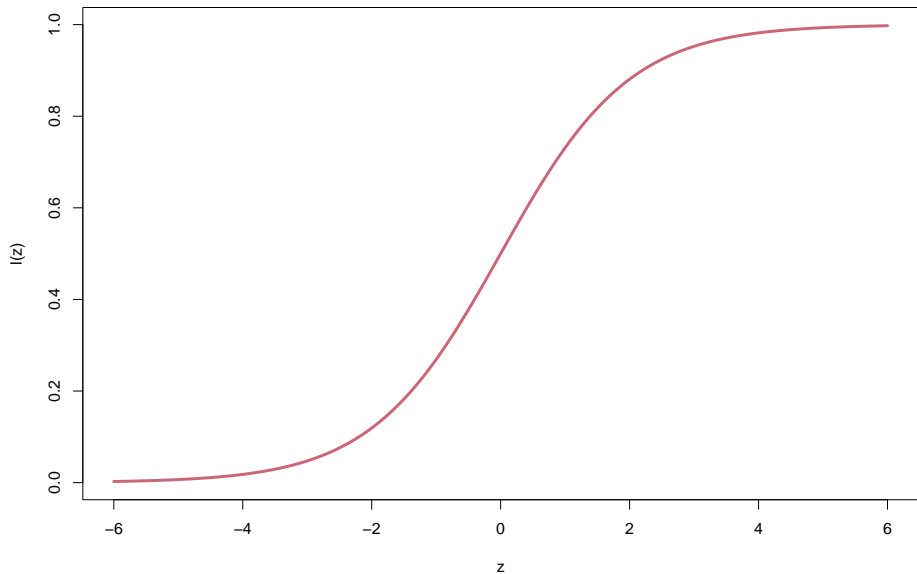


Figure 2: grafico della funzione logistica l

Il metodo dei minimi quadrati

Si suppone $Y \in E' = \mathbb{R}^{d'}$ e $U \in \mathbb{R}^k$, e si pone il **residuo**

$$Y - g(X; U).$$

- Il metodo prescrive, avendo osservato $X = x$, $Y = y$, di determinare un valore del parametro u che minimizzi il “residuo quadratico”, ossia

$$u_{\text{OLS}} \in \arg \min_{u \in \mathbb{R}^k} |y - g(x; u)|^2,$$

Il metodo dei minimi quadrati

Si suppone $Y \in E' = \mathbb{R}^{d'}$ e $U \in \mathbb{R}^k$, e si pone il **residuo**

$$Y - g(X; U).$$

- Il metodo prescrive, avendo osservato $X = x$, $Y = y$, di determinare un valore del parametro u che minimizzi il “residuo quadratico”, ossia

$$u_{\text{OLS}} \in \arg \min_{u \in \mathbb{R}^k} |y - g(x; u)|^2,$$

- Avendo n osservazioni indipendenti di coppie di variabili $(X_i, Y_i) = (x_i, y_i)$ per cui si suppone che il parametro U sia lo stesso, ossia $Y_i \sim g(X_i, U)$, si minimizza la somma dei residui quadratici

$$u_{\text{OLS}} \in \arg \min_{u \in \mathbb{R}^k} \sum_{i=1}^n |y_i - g(x_i; u)|^2.$$

- La stima della “varianza” del residuo tipico $Y - g(X; u_{OLS})$, è l'**errore quadratico medio** (in inglese *mean squared error*, MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n |y_i - g(x_i; u_{OLS})|^2,$$

Regressione lineare semplice

Siano X, Y a valori reali e un modello lineare con $u = (a, b) \in \mathbb{R}^2$

$$g(x; u) = ax + b$$

- partendo da n osservazioni (x_i, y_i) , si deve minimizzare

$$(a, b) \mapsto \sum_{i=1}^n (y_i - ax_i - b)^2.$$

Regressione lineare semplice

Siano X, Y a valori reali e un modello lineare con $u = (a, b) \in \mathbb{R}^2$

$$g(x; u) = ax + b$$

- partendo da n osservazioni (x_i, y_i) , si deve minimizzare

$$(a, b) \mapsto \sum_{i=1}^n (y_i - ax_i - b)^2.$$

- Imponendo che le derivate si annullino si trova un semplice sistema (lineare) nelle incognite a, b , che risolto permette di determinare

$$a_{\text{OLS}} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\Sigma_{xy}}{\Sigma_{xx}}$$

avendo indicato con Σ le covarianze campionarie, e

$$b_{\text{OLS}} = \bar{y} - \frac{\Sigma_{xy}}{\Sigma_{xx}} \bar{x}.$$

Il segno di a_{OLS} coincide con quello della covarianza campionaria Σ_{xy} (essendo la varianza a denominatore sempre positiva).

- Recuperiamo quindi il significato di positiva (o negativa) correlazione in termini della “concentrazione” della densità della variabile congiunta (X, Y) intorno ad una retta con coefficiente angolare positivo (o negativo).

Il segno di a_{OLS} coincide con quello della covarianza campionaria Σ_{xy} (essendo la varianza a denominatore sempre positiva).

- Recuperiamo quindi il significato di positiva (o negativa) correlazione in termini della “concentrazione” della densità della variabile congiunta (X, Y) intorno ad una retta con coefficiente angolare positivo (o negativo).
- Si può anche esprimere in alternativa

$$a_{OLS} = \rho_{xy} \frac{\sigma_y}{\sigma_x}, \quad b_{OLS} = \bar{y} - \rho_{xy} \frac{\sigma_y}{\sigma_x} \bar{x}$$

usando il coefficiente di correlazione e le deviazioni standard campionarie

$$\sigma_x = \sqrt{\Sigma_{xx}}, \quad \sigma_y = \sqrt{\Sigma_{yy}}, \quad \rho_{xy} = \frac{\Sigma_{xy}}{\sigma_x \sigma_y}.$$

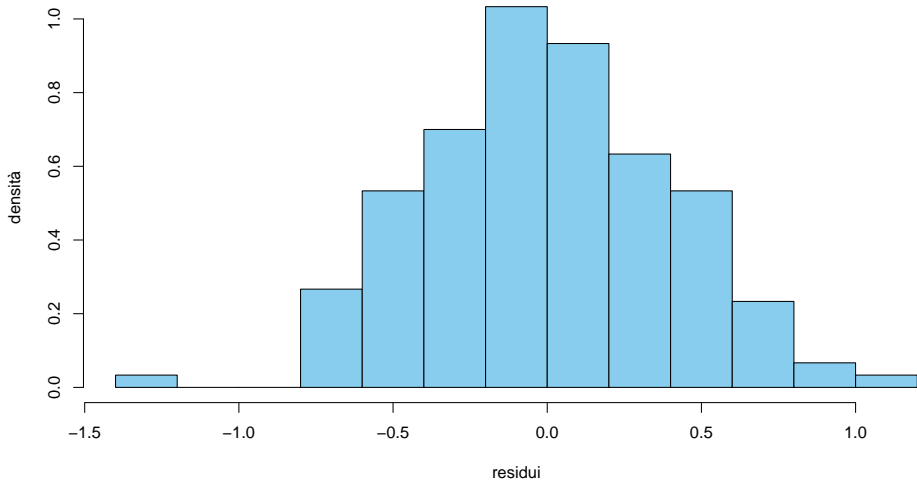
Un esempio in R

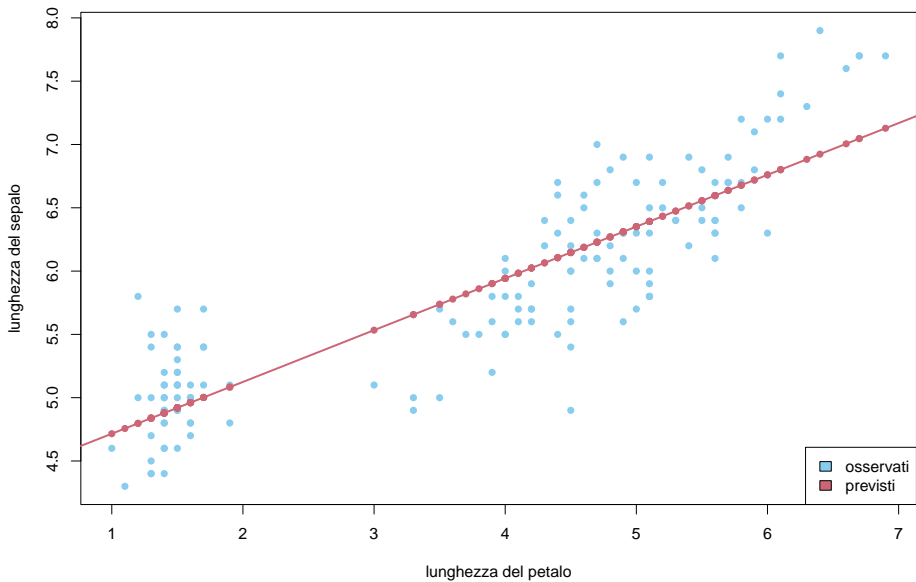
In R la regressione su un modello lineare è implementata tramite la funzione `lm()`.

- Nel dataset iris si vuole predire la lunghezza del sepalo (prima colonna) a partire da quella del petalo (terza colonna).

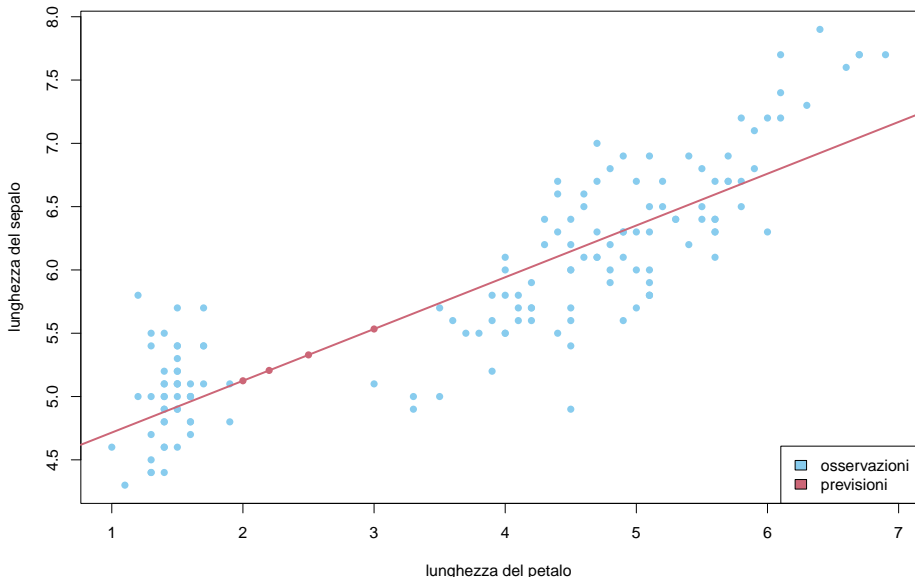
```
## (Intercept)          x  
## 4.3066034 0.4089223
```

Histogram of iris_reg_lin\$residuals





Possiamo effettuare semplici previsioni partendo dal modello. In R basta usare il comando `predict()`.



Regressione lineare multipla

Un modello lineare più generale a più variabili è il caso $X \in \mathbb{R}^d$, $U \in \mathbb{R}^k$, $Y \in \mathbb{R}$ e

$$g(x; u) = \sum_{j=1}^k x_j u_j = x \cdot u,$$

- Avendo osservato (x_i, y_i) , dobbiamo minimizzare la funzione *quadratica*

$$u \mapsto \sum_{i=1}^n (y_i - x_i \cdot u)^2,$$

Regressione lineare multipla

Un modello lineare più generale a più variabili è il caso $X \in \mathbb{R}^d$, $U \in \mathbb{R}^k$, $Y \in \mathbb{R}$ e

$$g(x; u) = \sum_{j=1}^k x_j u_j = x \cdot u,$$

- Avendo osservato (x_i, y_i) , dobbiamo minimizzare la funzione *quadratica*

$$u \mapsto \sum_{i=1}^n (y_i - x_i \cdot u)^2,$$

- Si ottiene un sistema lineare che ammette come soluzione esplicita

$$u_{\text{OLS}} = (x^T x)^{-1} x^T y,$$

dove $x \in \mathbb{R}^{n \times d}$ è intesa come matrice le cui righe sono le x_i^T .

Regressione lineare multipla

Un modello lineare più generale a più variabili è il caso $X \in \mathbb{R}^d$, $U \in \mathbb{R}^k$, $Y \in \mathbb{R}$ e

$$g(x; u) = \sum_{j=1}^k x_j u_j = x \cdot u,$$

- Avendo osservato (x_i, y_i) , dobbiamo minimizzare la funzione *quadratica*

$$u \mapsto \sum_{i=1}^n (y_i - x_i \cdot u)^2,$$

- Si ottiene un sistema lineare che ammette come soluzione esplicita

$$u_{\text{OLS}} = (x^T x)^{-1} x^T y,$$

dove $x \in \mathbb{R}^{n \times d}$ è intesa come matrice le cui righe sono le x_i^T .

- La previsione è quindi data dalla funzione

$$z \mapsto g(z, u_{\text{OLS}}) = z \cdot (x^T x)^{-1} x^T y.$$

Un esempio in R

Il comando `lm()` permette di effettuare regressione lineare multipla in dimensione arbitraria. Ad esempio, possiamo considerare come predittori della lunghezza del sepalo nel dataset Iris tutte le variabili (eccetto la specie).

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width,
##     data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82816 -0.21989  0.01875  0.19709  0.84570
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.85600    0.25078   7.401 9.85e-12 ***
## Sepal.Width   0.65084    0.06665   9.765 < 2e-16 ***
## Petal.Length  0.70913    0.05672  12.502 < 2e-16 ***
## Petal.Width  -0.55648    0.12755  -4.363 2.41e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con la funzione `summary()` si leggono anche altre informazioni rilevanti:

- L'**errore standard** dei residui (*residual standard error*)

$$s = \sqrt{\frac{1}{n-k} \sum_{i=1}^n |y_i - x_i \cdot u_{OLS}|^2}.$$

Con la funzione `summary()` si leggono anche altre informazioni rilevanti:

- L'**errore standard** dei residui (*residual standard error*)

$$s = \sqrt{\frac{1}{n-k} \sum_{i=1}^n |y_i - x_i \cdot u_{OLS}|^2}.$$

- Ogni parametro, ossia ogni componente u_j del vettore u_{OLS} è accompagnato da una stima della deviazione standard (visibile nella seconda colonna *Std. Error*, accanto a quella contenente la stima *Estimate*),

$$s_j = s \sqrt{(x^T x)^{-1}_{jj}}.$$

Con la funzione `summary()` si leggono anche altre informazioni rilevanti:

- L'**errore standard** dei residui (*residual standard error*)

$$s = \sqrt{\frac{1}{n-k} \sum_{i=1}^n |y_i - x_i \cdot u_{OLS}|^2}.$$

- Ogni parametro, ossia ogni componente u_j del vettore u_{OLS} è accompagnato da una stima della deviazione standard (visibile nella seconda colonna *Std. Error*, accanto a quella contenente la stima *Estimate*),

$$s_j = s \sqrt{(x^T x)^{-1}_{jj}}.$$

- Per valutare l'efficacia della regressione si usa spesso il **coefficiente di determinazione** definito come

$$R^2 = 1 - \frac{\sum_{i=1}^n |y_i - x_i \cdot u_{OLS}|^2}{\sum_{i=1}^n |y_i - \bar{y}|^2},$$

la “percentuale” di varianza (dei dati) spiegata dal modello lineare.

Section 4

Intepretazione probabilistica del metodo dei minimi quadrati

Metodo dei minimi quadrati e MLE

Intepretiamo il metodo come una stima di massima verosimiglianza sotto opportune ipotesi (residui gaussiani indipendenti). Supponiamo

- $X \in E$, $Y \in E = \mathbb{R}^{d'}$, $U \in \mathbb{R}^k$ tali che

$$Y = g(X; U) + W,$$

Metodo dei minimi quadrati e MLE

Intepretiamo il metodo come una stima di massima verosimiglianza sotto opportune ipotesi (residui gaussiani indipendenti). Supponiamo

- $X \in E$, $Y \in E = \mathbb{R}^{d'}$, $U \in \mathbb{R}^k$ tali che

$$Y = g(X; U) + W,$$

- W è una variabile (il residuo) con densità gaussiana vettoriale $\mathcal{N}(0, \nu Id)$ (dove $\nu > 0$ è un parametro).

Metodo dei minimi quadrati e MLE

Intepretiamo il metodo come una stima di massima verosimiglianza sotto opportune ipotesi (residui gaussiani indipendenti). Supponiamo

- $X \in E$, $Y \in E = \mathbb{R}^{d'}$, $U \in \mathbb{R}^k$ tali che

$$Y = g(X; U) + W,$$

- W è una variabile (il residuo) con densità gaussiana vettoriale $\mathcal{N}(0, \nu Id)$ (dove $\nu > 0$ è un parametro).
- Le variabili X , U e W siano tra loro indipendenti.

Metodo dei minimi quadrati e MLE

Intepretiamo il metodo come una stima di massima verosimiglianza sotto opportune ipotesi (residui gaussiani indipendenti). Supponiamo

- $X \in E$, $Y \in E = \mathbb{R}^{d'}$, $U \in \mathbb{R}^k$ tali che

$$Y = g(X; U) + W,$$

- W è una variabile (il residuo) con densità gaussiana vettoriale $\mathcal{N}(0, \nu I_d)$ (dove $\nu > 0$ è un parametro).
- Le variabili X , U e W siano tra loro indipendenti.
- Anche senza ipotesi sulla densità a priori di X , scriviamo la verosimiglianza come segue:

$$\begin{aligned}
L(u; x, y) &= p(X = x, Y = y | U = u) \\
&= p(Y = y | U = u, X = x) p(X = x | U = u) \\
&= p(Y - g(x; u) = y - g(x, u) | U = u, X = x) p(X = x) \\
&= p(W = y - g(x, u) | U = u, X = x) p(X = x) \\
&= \exp\left(-\frac{1}{2v} |y - g(x, u)|^2\right) \frac{1}{\sqrt{2\pi v}} p(X = x).
\end{aligned}$$

- La stima di massima verosimiglianza per U equivale a minimizzare

$$u \mapsto |y - g(x, u)|^2,$$

ossia il minimo residuo quadratico.

Se si dispone di n variabili (X_i, W_i) tutte indipendenti tra loro (e dalla U) tali che

$$Y_i = g(X_i; U) + W_i,$$

allora la verosimiglianza di U associata alle osservazioni $x = (x_i)_{i=1}^n$, $y = (y_i)_{i=1}^n$,

$$\begin{aligned} L(u; x, y) &= p(X = x, Y = y | U = u) \\ &= \exp\left(-\frac{1}{2\nu} \sum_{i=1}^n |y_i - g(x_i, u)|^2\right) \frac{1}{\sqrt{(2\pi\nu)^n}} \prod_{i=1}^n p(X_i = x_i). \end{aligned}$$

- La stima di massima verosimiglianza consiste nel minimizzare

$$u \mapsto \sum_{i=1}^n |y_i - g(x_i, u)|^2,$$

Se si dispone di n variabili (X_i, W_i) tutte indipendenti tra loro (e dalla U) tali che

$$Y_i = g(X_i; U) + W_i,$$

allora la verosimiglianza di U associata alle osservazioni $x = (x_i)_{i=1}^n$, $y = (y_i)_{i=1}^n$,

$$\begin{aligned} L(u; x, y) &= p(X = x, Y = y | U = u) \\ &= \exp\left(-\frac{1}{2v} \sum_{i=1}^n |y_i - g(x_i, u)|^2\right) \frac{1}{\sqrt{(2\pi v)^n}} \prod_{i=1}^n p(X_i = x_i). \end{aligned}$$

- La stima di massima verosimiglianza consiste nel minimizzare

$$u \mapsto \sum_{i=1}^n |y_i - g(x_i, u)|^2,$$

- Massimizzando anche in v si ottiene anche l'errore quadratico medio

$$v_{\text{MLE}} = \text{MSE} = \frac{1}{n} \sum_{i=1}^n |y_i - g(x_i; u)|^2.$$

Approccio bayesiano

La derivazione del metodo come MLE suggerisce un approccio bayesiano per raffinare il metodo.

- Supponiamo che sia noto a priori che U non si discosta troppo da un parametro noto u_0 , ad esempio con una variabilità dell'ordine di $\sigma_u > 0$ (lungo ciascuna componente) e che le componenti di U siano indipendenti tra loro.

Approccio bayesiano

La derivazione del metodo come MLE suggerisce un approccio bayesiano per raffinare il metodo.

- Supponiamo che sia noto a priori che U non si discosta troppo da un parametro noto u_0 , ad esempio con una variabilità dell'ordine di $\sigma_u > 0$ (lungo ciascuna componente) e che le componenti di U siano indipendenti tra loro.
- Si pone U a priori $\mathcal{N}(u_0, \sigma_u^2 Id)$, e dalla formula di Bayes

$$\begin{aligned}
 p(U = u | X_i = x_i, Y_i = y_i, \forall i = 1, \dots, n) \\
 &\propto \exp\left(-\frac{1}{2\sigma_u^2} |u - u_0|^2\right) L(u; x, y) \\
 &\propto \exp\left(-\frac{1}{2} \left(\frac{1}{v} \sum_{i=1}^n |y_i - g(x_i; u)|^2 + \frac{1}{\sigma_u^2} |u - u_0|^2\right)\right)
 \end{aligned}$$

- Il massimo della densità a posteriori (stima MAP) si ottiene minimizzando

$$u \mapsto \frac{1}{v} \sum_{i=1}^n |y_i - g(x_i; u)|^2 + \frac{1}{\sigma_u^2} |u - u_0|^2.$$

- Il massimo della densità a posteriori (stima MAP) si ottiene minimizzando

$$u \mapsto \frac{1}{v} \sum_{i=1}^n |y_i - g(x_i; u)|^2 + \frac{1}{\sigma_u^2} |u - u_0|^2.$$

- È stato quindi introdotto un termine di *regolarizzazione* (o penalizzazione) alla somma dei residui, che diventa rilevante se u è troppo lontano dal parametro u_0 .

- Il massimo della densità a posteriori (stima MAP) si ottiene minimizzando

$$u \mapsto \frac{1}{v} \sum_{i=1}^n |y_i - g(x_i; u)|^2 + \frac{1}{\sigma_u^2} |u - u_0|^2.$$

- È stato quindi introdotto un termine di *regolarizzazione* (o penalizzazione) alla somma dei residui, che diventa rilevante se u è troppo lontano dal parametro u_0 .
- L'introduzione di questi ed altre funzioni è spesso utile per regolarizzare appunto la soluzione fornita dal semplice metodo dei minimi quadrati (queste tecniche hanno diversi nomi a seconda del tipo di termini che si aggiungono, ad esempio *ridge*, *weight decay*, *LASSO*, ecc.).

Approccio bayesiano e modelli lineari

- Supponendo ulteriormente che $g(x; u) = x \cdot u$, la densità a posteriori per U diventa

$$p(U = u | X_i = x_i, Y_i = y_i, \forall i = 1, \dots, n) \\ \propto \exp \left(-\frac{1}{2} \left(\frac{1}{v} \sum_{i=1}^n |y_i - x_i \cdot u|^2 + \frac{1}{\sigma_u^2} |u - u_0|^2 \right) \right),$$

che è una densità gaussiana vettoriale (essendo un esponenziale di polinomio di secondo grado rispetto alla variabile u).

Approccio bayesiano e modelli lineari

- Supponendo ulteriormente che $g(x; u) = x \cdot u$, la densità a posteriori per U diventa

$$p(U = u | X_i = x_i, Y_i = y_i, \forall i = 1, \dots, n) \\ \propto \exp \left(-\frac{1}{2} \left(\frac{1}{v} \sum_{i=1}^n |y_i - x_i \cdot u|^2 + \frac{1}{\sigma_u^2} |u - u_0|^2 \right) \right),$$

che è una densità gaussiana vettoriale (essendo un esponenziale di polinomio di secondo grado rispetto alla variabile u).

- Si trova che U ha come nuovi parametri, il vettore dei valor medi

$$u_{|X=x, Y=y} = \left(x^T x + (v/\sigma_u^2) Id \right)^{-1} \left(x^T y + (v/\sigma_u^2) u_0 \right)$$

e la matrice delle covarianze

$$\Sigma_{U|X=x, Y=y} = v \left(x^T x + (v/\sigma_u^2) Id \right)^{-1}.$$

- La deviazione standard della componente U_j del vettore dei parametri U , si ottiene dal termine diagonale della matrice,

$$\begin{aligned}\sigma_{U_j|X=x, Y=y} &= \sqrt{\text{Var}(U_j|X=x, Y=y)} \\ &= \sqrt{v(x^T x + (v/\sigma_u^2)Id)_{jj}^{-1}}.\end{aligned}$$

- La deviazione standard della componente U_j del vettore dei parametri U , si ottiene dal termine diagonale della matrice,

$$\begin{aligned}\sigma_{U_j|X=x, Y=y} &= \sqrt{\text{Var}(U_j|X=x, Y=y)} \\ &= \sqrt{v(x^T x + (v/\sigma_u^2)Id)_{jj}^{-1}}.\end{aligned}$$

- Nel limite $v \ll \sigma_u^2$ dalle formule sopra si recuperano la stima del metodo classico dei minimi quadrati per il modello lineare

$$u_{\text{OLS}} = (x^T x)^{-1} x^T y$$

e (avendo posto $v = v_{\text{MLE}} = \text{MSE}$ la stima di massima verosimiglianza) gli errori standard dei parametri, per $j \in \{1, \dots, k\}$,

$$\sigma_j = \sqrt{v(x^T x)_{jj}^{-1}}.$$

Altre funzioni obiettivo

Perché minimizzare i quadrati dei residui? in effetti ci sono altre scelte possibili e ragionevoli (ma non si trovano formule esplicite).

- Ad esempio il valore assoluto (*least absolute deviation* in inglese) darebbe

$$u_{\text{LAD}} \in \arg \min_{u \in \mathbb{R}^k} \sum_{i=1}^n |y_i - g(x_i; u)|.$$

Altre funzioni obiettivo

Perché minimizzare i quadrati dei residui? in effetti ci sono altre scelte possibili e ragionevoli (ma non si trovano formule esplicite).

- Ad esempio il valore assoluto (*least absolute deviation* in inglese) darebbe

$$u_{\text{LAD}} \in \arg \min_{u \in \mathbb{R}^k} \sum_{i=1}^n |y_i - g(x_i; u)|.$$

- L'interpretazione è che i residui hanno densità detta di Laplace $p(W = w) \propto \exp\left(-\frac{|w|}{b}\right)$.