PCA

Richiami

- **Teoria**: $Y \in \mathbb{R}^d$, $X = \Pi_Y Y \in \mathbb{R}^k$, dove Π_Y è costruita diagonalizzando Σ_Y.

Richiami

- **Teoria**: $Y \in \mathbb{R}^d$, $X = \Pi_Y Y \in \mathbb{R}^k$, dove Π_Y è costruita diagonalizzando Σ_Y.
- **Pratica**: Σ_Y non è nota $\rightarrow$

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

Richiami

- **Teoria**: $Y \in \mathbb{R}^d$, $X = \Pi_Y Y \in \mathbb{R}^k$, dove Π_Y è costruita diagonalizzando Σ_Y.
- **Pratica**: Σ_Y non è nota $\rightarrow$

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

- Definiamo $\Pi_y \in \mathbb{R}^{k \times d}$ come la proiezione nel sottospazio k-dim degli autovettori con autovalori il più grande possibile.

Richiami

- **Teoria**: $Y \in \mathbb{R}^d$, $X = \Pi_Y Y \in \mathbb{R}^k$, dove Π_Y è costruita diagonalizzando Σ_Y.
- **Pratica**: Σ_Y non è nota $\rightarrow$

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

- Definiamo $\Pi_y \in \mathbb{R}^{k \times d}$ come la proiezione nel sottospazio k-dim degli autovettori con autovalori il più grande possibile.
- Proiezioni: $x_i = \Pi_y y_i$.

Una giustificazione tramite MLE della PCA

PCA è una stima di massima verosimiglianza per un opportuno modello gaussiano.

- **Idea:** con la PCA recuperiamo "segnale" (X) osservandone una versione "rumorosa" (Y) disposta su un sottospazio non noto.

- fissata k, introduciamo una variabile gaussiana standard $Z \in \mathbb{R}^k$ e imponiamo

$$Y = AZ + W,$$

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

- fissata k, introduciamo una variabile gaussiana standard $Z \in \mathbb{R}^k$ e imponiamo

$$Y = AZ + W,$$

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

- Il "segnale" da ricostruire è quindi $X = AZ$ e W è una variabile che rappresenta il "rumore" residuo.

- fissata k, introduciamo una variabile gaussiana standard $Z \in \mathbb{R}^k$ e imponiamo

$$Y = AZ + W,$$

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

- Il "segnale" da ricostruire è quindi $X = AZ$ e W è una variabile che rappresenta il "rumore" residuo.
- *Supponiamo*: Z, W indipendenti con densità gaussiane centrate, $\Sigma_Z = Id$ e $\Sigma_W = \sigma_0^2 Id$, per una costante opportuna (sufficientemente piccola).

- Nota la matrice A, allora la densità di Y, è gaussiana centrata, con covarianza $\Sigma_Y = AA^T + \sigma_0^2 Id$

- Nota la matrice A, allora la densità di Y, è gaussiana centrata, con covarianza $\Sigma_Y = AA^T + \sigma_0^2 Id$
- Pertanto la verosimiglianza di A associata ad Y si scrive

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

- Nota la matrice A, allora la densità di Y, è gaussiana centrata, con covarianza $\Sigma_Y = AA^T + \sigma_0^2 Id$
- Pertanto la verosimiglianza di A associata ad Y si scrive

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

- date n osservazioni indipendenti $Y_i = y_i$, tutte gaussiane con gli stessi parametri – la stessa matrice A, la verosimiglianza è il prodotto della funzione sopra (cambiando i valori osservati)

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

- Nota la matrice A, allora la densità di Y, è gaussiana centrata, con covarianza $\Sigma_Y = AA^T + \sigma_0^2 Id$
- Pertanto la verosimiglianza di A associata ad Y si scrive

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

- date n osservazioni indipendenti $Y_i = y_i$, tutte gaussiane con gli stessi parametri – la stessa matrice A, la verosimiglianza è il prodotto della funzione sopra (cambiando i valori osservati)

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

- la massima verosimiglianza si ottiene quindi 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

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

dove:

- La stima di massima verosimiglianza per A è data da

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

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

- La stima di massima verosimiglianza per A è data da

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

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
- $D_{y|k} \in \mathbb{R}^{k \times k}$ indica la matrice diagonale contenente tali autovalori nell'ordine corrispondente.

- La stima di massima verosimiglianza per A è data da

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

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
  - $D_{y|k} \in \mathbb{R}^{k \times k}$ indica la matrice diagonale contenente tali autovalori nell'ordine corrispondente.

- **Purché** σ_0^2 sia sufficientemente piccolo.

- La stima di massima verosimiglianza per A è data da

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

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
  - $D_{y|k} \in \mathbb{R}^{k \times k}$ indica la matrice diagonale contenente tali autovalori nell'ordine corrispondente.

- **Purché** σ_0^2 sia sufficientemente piccolo.
- Nel limite $\sigma_0 \to 0$ si ottiene che $A_{\text{MLE}} = U_{y|k} D_{y|k}^{1/2} \to A_{\text{MLE}} Z$ recupera $\Pi_y Y$.

Regressione

Regressione

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

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

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 $g : E \to E'$ tale che Y sia "molto vicina" a $g(X)$, ossia

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

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 $g : E \to E'$ tale che Y sia "molto vicina" a $g(X)$, ossia

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

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 varibile 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 varibile 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

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

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) $g_i : E \to E' = \mathbb{R}^{d'}$.

- 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) $g_i : E \to E' = \mathbb{R}^{d'}$.

- 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:
$$g(x;(u_1,u_2)) = u_1 x + u_2 x^2$$
- 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

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

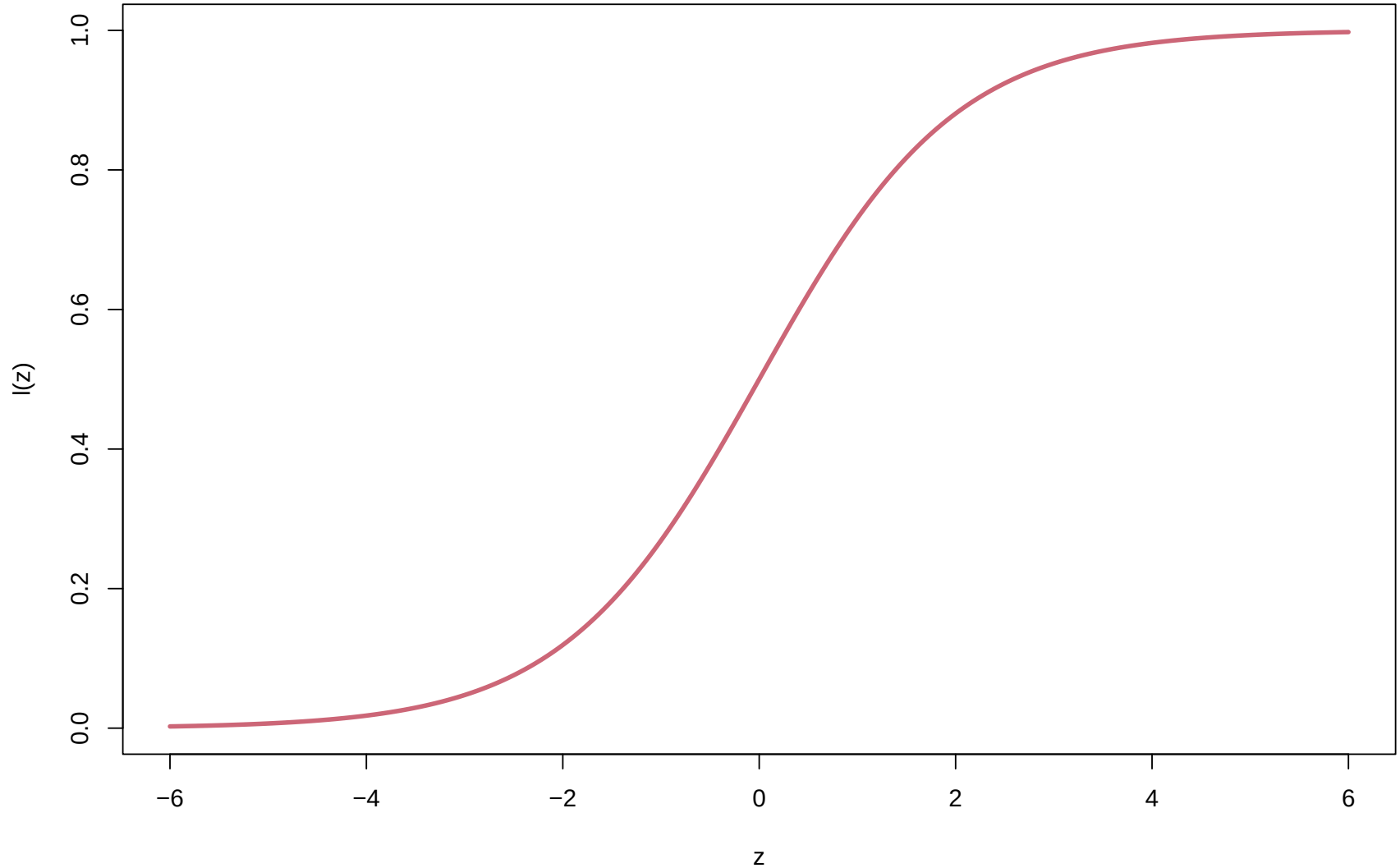


Figure 1: grafico della funzione logistica ℓ

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

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

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 $Y - g(X; u_{\mathrm{OLS}})$, è l'**errore quadratico medio** (in inglese *mean squared error*, MSE)

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

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

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

usando il coefficiente di correlazione e le deviazioni standard campionarie

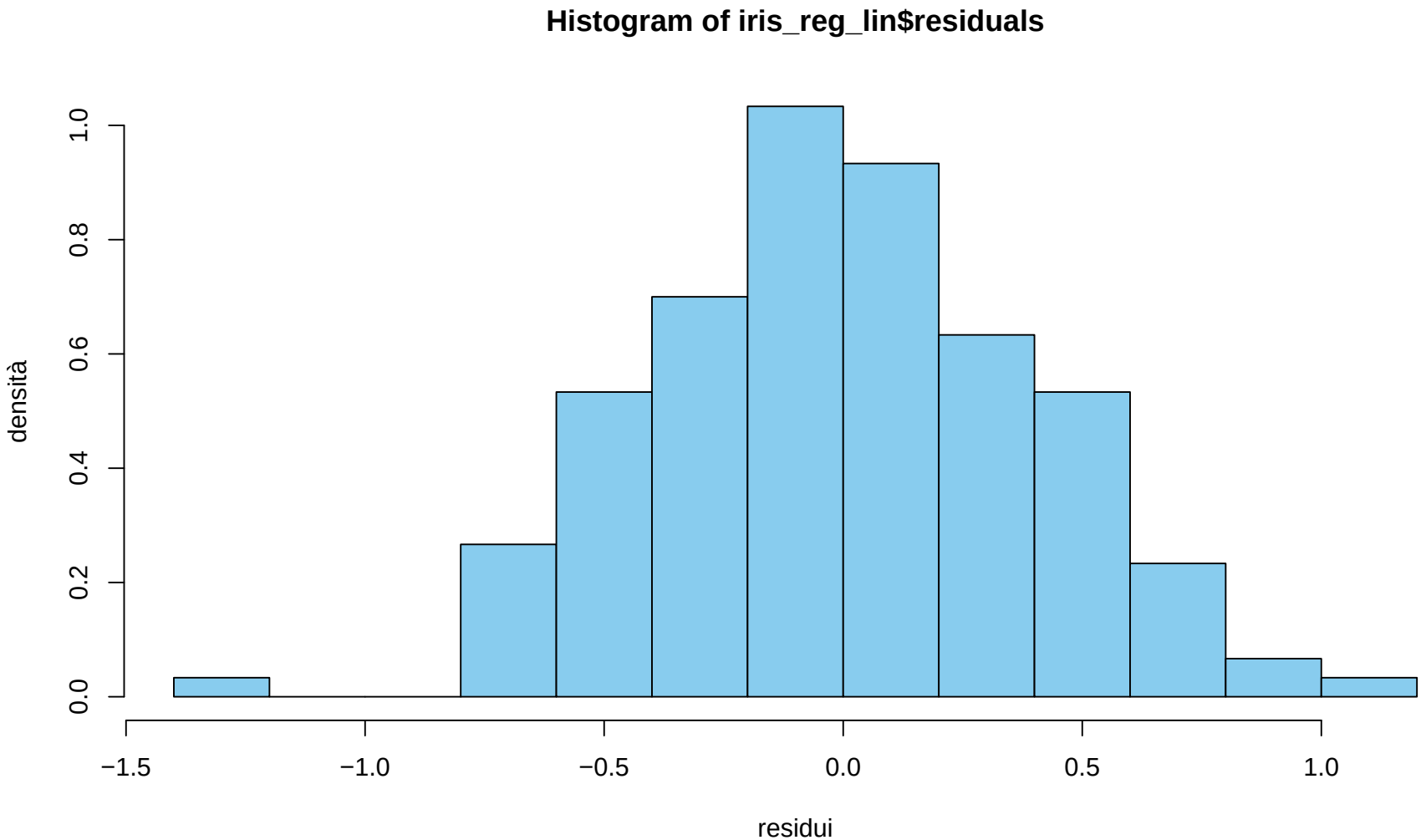
$$\sigma_x = \sqrt{\Sigma_{xx}}, \quad \sigma_y = \sqrt{\Sigma_{yy}}, \quad \rho_{xy} = \frac{\Sigma_{xy}}{\sigma_x\sigma_y}.$$

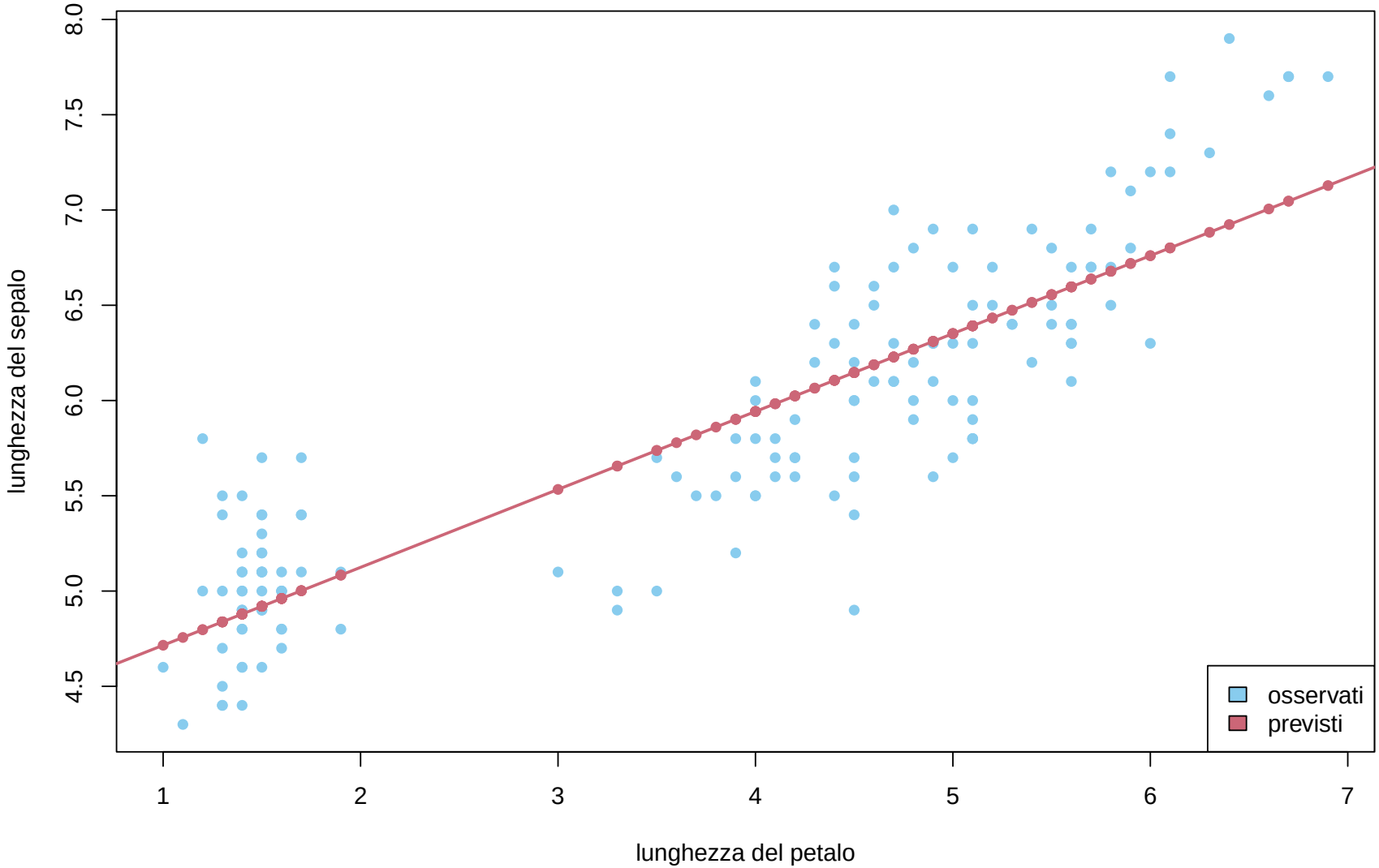
Un esempio in R

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

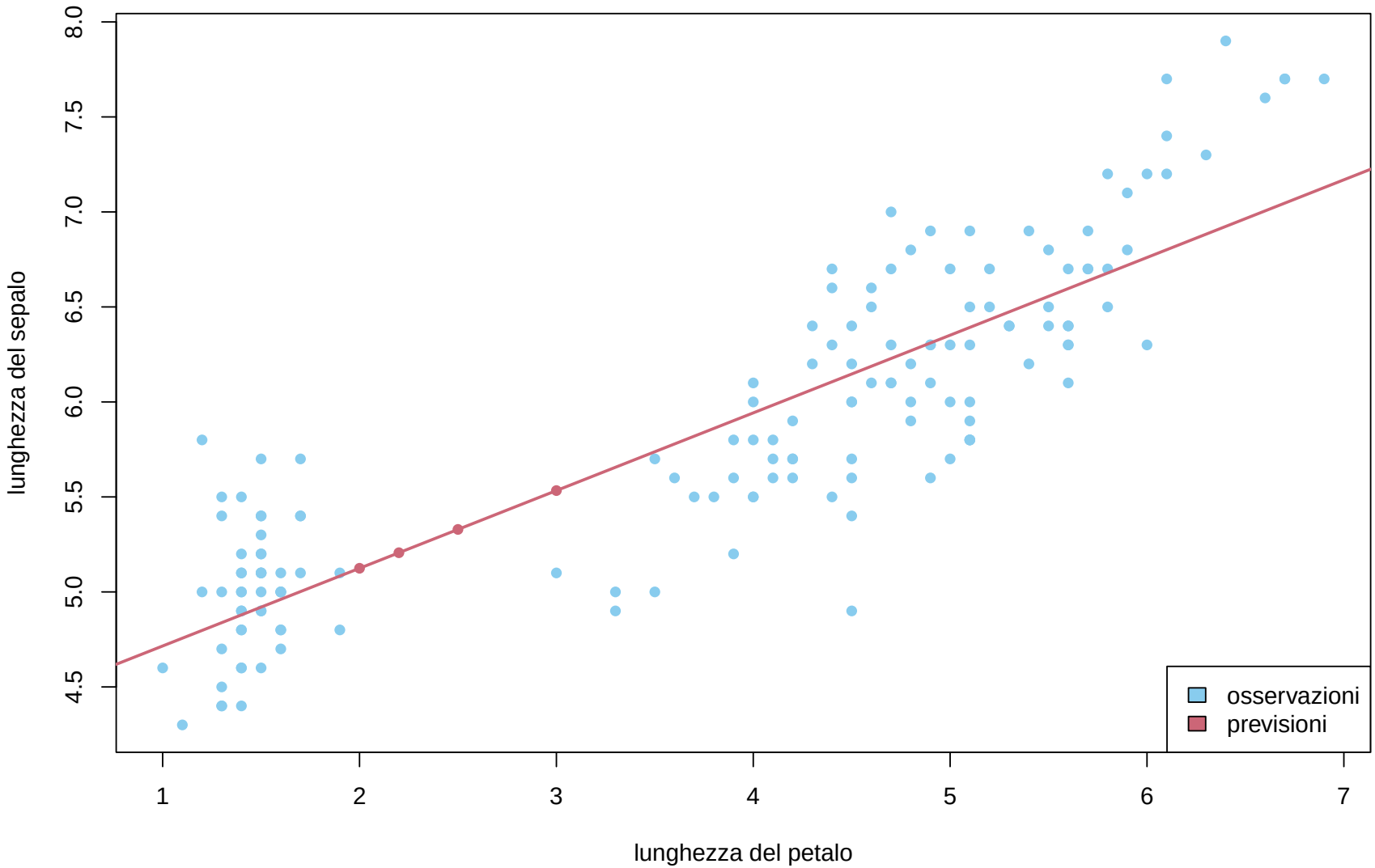
- Nel dataset iris si vuole predire la lunghezza del sepalo (prima colonna) a partire da quella del petalo (terza colonna).

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




lunghezza del sepalo
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
1
2
3
4
5
6
7
lunghezza del petalo
osservati
previsti

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 con 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 con 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 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 +
##     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
##
## Residual standard error: 0.3145 on 146 degrees of freedo
## Multiple R-squared:  0.8586, Adjusted R-squared:  0.8557
```

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

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

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

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

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

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

- 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 altre informazioni rilevanti:

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

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

- 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 può usare il **coefficiente di determinazione**:

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

la "percentuale" di varianza (dei dati) spiegata dal modello

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 $\mathcal{N}(0, vId)$ (dove $v > 0$ 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 $\mathcal{N}(0, vId)$ (dove $v > 0$ 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 $\mathcal{N}(0, vId)$ (dove $v > 0$ 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 $x = (x_i)_{i=1}^n$, $y = (y_i)_{i=1}^n$,

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

- La stima di massima verosimiglianza consiste nel minimizzare

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

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

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

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

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

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,

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

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

$$\sigma_{U_j|X=x,Y=y} = \sqrt{\mathrm{Var}\left(U_j|X=x, Y=y\right)}$$
$$= \sqrt{v\left(x^T x + (v/\sigma_u^2)Id\right)^{-1}_{jj}}.$$

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

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

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

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

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

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

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 resudui hanno densità detta di Laplace $p(W = w) \propto \exp\left(-\frac{|w|}{b}\right)$.