Test delle Ipotesi applicato al modello di Regressione

Riprendiamo il modello di regressione lineare si assume una relazione di tipo lineare tra il valore medio della variabile dipendente Y da quello indipendente per cui

$$E(Y \mid X) = \eta_{Y|X} = a + b * x$$

Il modello si scrive come

$$y = a + b * x + \varepsilon$$

Dove ε è l'errore gaussiano con varianza σ^2 e valore medio nullo.

In Matlab tale matrice può essere stimata come (vedi dispensa *Corr_Cov.pdf*)

$$\beta = \left[\hat{a}\ \hat{b}\right] = X \setminus y$$

dove gli elementi di β sono tali da minimizzare l'errore quadratico medio $\sum_{k=1}^{n}\left(y(k) - \left(\hat{a} + \hat{b} * x(k)\right)\right)^2$

In particolare $\hat{b} = \dfrac{\sum_{k=1}^{n}\left(x(k) - \bar{x}\right)y(k)}{\sum_{k=1}^{n}\left(x(k) - \bar{x}\right)^2} = \sum_{k=1}^{n} h(k)y(k)$ con $h(k) = \dfrac{\left(x(k) - \bar{x}\right)}{\sum_{k=1}^{n}\left(x(k) - \bar{x}\right)^2}$

$\hat{b}$ è una variabile aleatoria il cui valore atteso è $E\{\hat{b}\} = b$ e deviazione standard $\sigma_{\hat{b}}$

Derivazione di $\sigma_{\hat{b}}$

Nel seguito otteremo una stima di $\sigma_{\hat{b}}$. Definita la somma dei quadrati degli errori (error sum of squares) $SSE = \sum_{k=1}^{n}\varepsilon(k)^2 = \sum_{k=1}^{n}\left(y(k) - \left(\hat{a} + \hat{b} * x(k)\right)\right)^2$.

Si definisce l'errore quadratico medio come (error mean square) $MSE = \sum_{k=1}^{n}\dfrac{\varepsilon(k)^2}{n-2}$ dove i gradi di libertà sono i *n-2* (2 gradi di libertà sono stati usati per stimare $\hat{a}$ e $\hat{b}$).

La varianza di $\hat{b}$ si può stimare come $\sigma_{\hat{b}}^2 = \dfrac{MSE}{\sum_{k=1}^{n}\left(x(k) - \bar{x}\right)^2}$

<u>*Dim.*</u>

Dato che $\sigma_{y(k)}^2 = \sigma^2$, ovvero la varianza delle misure è pari alla varianza dell'errore, si ha

$$\sigma_{\hat{b}}^2 = \operatorname{var}\left\{\sum_{k=1}^{n} h(k)y(k)\right\} = \sum_{k=1}^{n} h(k)^2 \operatorname{var}\{y(k)\} = \sigma^2 \sum_{k=1}^{n} h(k)^2 = \sigma^2 \frac{1}{\sum_{k=1}^{n}\left(x(k) - \bar{x}\right)^2}$$

visto che MSE è uno stimatore di σ^2 (infatti $E\{MSE\} = \sigma^2$) si ha che lo stimatore di $\sigma_{\hat{b}}^2 = \dfrac{MSE}{\sum_{k=1}^{n}\left(x(k) - \bar{x}\right)^2}$

Supponiamo di volere testare l'ipotesi nulla, con un livello di significatività pari ad α

$H_0 : b = 0$

con ipotesi alternativa

$H_a : b \neq 0$

Per testare questa ipotesi possiamo considerare la variabile $\frac{\hat{b} - b}{\sigma_{\hat{b}}}$ dove $\sigma_{\hat{b}} = \sqrt{\frac{MSE}{\sum_{k=1}^{n} \left(x(k) - \bar{x} \right)^2}}$.

La funzione di distribuzione di questa variabile è quella relativa ad una *t* di Student con *n-2* gradi di libertà.

In questo caso il test delle ipotesi consiste nel valutare la grandezza $t = \frac{\hat{b}}{\sigma_{\hat{b}}}$. Nel caso in cui *t* ricada fuori dalla regione di accettazione non si potrà accettare l'ipotesi nulla, che verrà accettata altrimenti.

Detto il *t* critico il valore $t\left(1 - \frac{\alpha}{2}; n-2\right)$ si ha

$$|t| \leq t\left(1 - \tfrac{\alpha}{2}; n-2\right) \Rightarrow H_0$$
$$|t| > t\left(1 - \tfrac{\alpha}{2}; n-2\right) \Rightarrow H_a$$

Esercitazione dati fMRI

Si considerei la matrice contenuta nel file *fetta_20_007_V6.mat*. La matrice ha dimensioni 100x4096. Ogni riga della matrice è una immagine da risonanza magnetica (MR) trasformata in un vettore riga. La matrice in totale rappresenta una sequenza temporale di 100 immagini. Le 100 immagini sono state acquisite su un soggetto che stava svolgendo un compito. Le immagini sono sensibili all'effetto BOLD (Blood Oxygenation Level Dependent) e presentano un aumento del segnale, ovvero del valore assunto in quel punto, se è avvenuto localmente un aumento della emoglobina ossigenata rispetto a quella deossigenata. Questo aumento potrebbe essere legato ad un aumento della attività cerebrale in quella zona.

In una serie temporale di 100 elementi, *paradigma_V6.mat*, assegnata a parte, è descritto l'andamento temporale del compito svolto: tale andamento consiste in una serie di 0 e 1, dove il valore "0" indica "nessun compito svolto da parte del soggetto", il valore "1" indica "svolgimento del compito da parte del soggetto".

Nell'immagine che consideriamo è possibile osservare una variazione del valore assunto dal voxel in corrispondenza dello svolgimento del compito (con una certa approssimazione possiamo pensare ad esempio ad un aumento del valore in corrispondenza di un 1 e una sua diminuzione in corrispondenza di 0).

Ogni colonna della matrice rappresenta l'andamento temporale di un voxel (volume element) dell'immagine di risonanza.

E' possibile stimare un modello di regressione per ogni voxel dell'immagine dove, la colonna corrispondente rappresenta la variabile dipendente mentre il paradigma sperimentale (la serie di 0 e 1 la variabile indipendente).

Nella esercitazione si chiede appunto di:

- stimare un modello di regressione per ogni colonna della matrice, ottenendo in questo modo 4096 valori di $\hat{a}$ e $\hat{b}$, una coppia per ogni voxel.
- creare una immagine 64x64 dove ogni pixel rappresenta il valore di $\hat{b}$
- stimare per ogni voxel il valore del *MSE*
- stimare per ogni voxel il valore $t = \frac{\hat{b}}{\sigma_{\hat{b}}}$, ovvero il valore di $t = \frac{\hat{b} - b}{\sigma_{\hat{b}}}$ sotto l'ipotesi nulla $H_0 : b = 0$
- creare una matrice T di dimensioni 64x64 dove ogni pixel rappresenta il valore di $t = \frac{\hat{b}}{\sigma_{\hat{b}}}$
- dato un valore di significatività dell'ipotesi nulla pari a $\alpha = 10^{-4}$ trovare i valori di t critici $t\left(1 - \frac{\alpha}{2}; n - 2\right)$ (si veda per questo la dispensa *Statistica_Matlab.pdf*
- creare una matrice H0 di dimensioni 64x64 dove ogni pixel vale 1 se l'ipotesi nulla può essere accettata, 0 altrimenti
- creare una matrice T_H0 di dimensioni 64x64 dove in corrispondenza dei pixel di H0 pari a 1, è presente il valore corrispondente di T, e vale 0 altrove