

Capitolo 1

Caso con covariate

1.1 Dimostrazione

Se si aggiunge il termine di covariate, la risposta sarà modellizzata come:

$$\underline{z} = W\underline{\beta} + \Pi\underline{c} + \epsilon$$

La funzione da minimizzare in questo caso sarà:

$$(\underline{z} - W\underline{\beta} - \Pi\underline{c})^T(\underline{z} - W\underline{\beta} - \Pi\underline{c}) + \underline{c}^T S \underline{c}$$

Eseguo le moltiplicazioni, e quindi si avrà:

$$\underline{z}^T \underline{z} - 2\underline{z}^T W\underline{\beta} - 2\underline{z}^T \Pi\underline{c} + 2\underline{\beta}^T W^T \Pi\underline{c} + \underline{\beta}^T W^T W \underline{\beta} + \underline{c}(\Pi^T \Pi + S)\underline{c}$$

Derivo rispetto a $\underline{\beta}$ e \underline{c} :

$$\frac{\partial}{\partial \underline{\beta}} = -2W^T \underline{z} + 2W^T \Pi\underline{c} + 2W^T W \underline{\beta}$$

$$\frac{\partial}{\partial \underline{c}} = -2\Pi^T \underline{z} + 2\Pi^T W \underline{\beta} + 2(\Pi^T \Pi + S)\underline{c}$$

Pongo le derivate uguali a zero per avere il punto di minimo:

$$\frac{\partial}{\partial \underline{\beta}} \Rightarrow W^T W \underline{\beta} = W^T (\underline{z} - \Pi\underline{c})$$

$$\frac{\partial}{\partial \underline{c}} \Rightarrow (\Pi^T \Pi + S)\underline{c} = \Pi^T (\underline{z} - W \underline{\beta})$$

che ricordano molto i modelli marginali (regressione classica per la stima di $\underline{\beta}$, e regressione penalizzata per la stima di \underline{c}), ma entrambi i casi tolgono a

\underline{z} la parte spiegata dall'altra metà di modello.

Di conseguenza

$$\underline{\beta} = (W^T W)^{-1} W^T (\underline{z} - \Pi \underline{c})$$

Sostituisco nell'altra equazione, e si trova (tralascio tutti i conti):

$$\hat{\underline{c}} = [\Pi^T \Pi + S + \Pi^T W (W^T W)^{-1} W^T \Pi]^{-1} \Pi^T [I - W (W^T W)^{-1} W^T] \underline{z}$$

Questo valore poi si sostituisce nell'equazione precedente per trovare anche $\hat{\underline{\beta}}$. La formulazione analitica è complessa, ma alcune parti del calcolo si ripetono (come $W(W^T W)^{-1} W^T$). Quindi è possibile riutilizzare alcune matrici temporanee nel calcolo computazionale della stima.

Inoltre, se si controllano le dimensioni delle matrici, i prodotti tornano. Infatti se si hanno n punti in spazio con dati (N contando anche quelli senza dati, di frontiera), m istanti di tempo (volendo, se si vuole avere una stima temporale più precisa, si può aumentare ad M il numero di basi in tempo), K covariate, si hanno le seguenti dimensioni

Π ha dimensione $nm \times NM$

S ha dimensione $NM \times NM$

W ha dimensione $nm \times K$

\underline{z} ha dimensione $nm \times 1$

\underline{c} ha dimensione $NM \times 1$

$\underline{\beta}$ ha dimensione $K \times 1$

e i prodotti matriciali sono coerenti.

Capitolo 2

GCV

2.1 Dimostrazione

Marra propone di minimizzare la seguente quantità per individuare il miglior $\underline{\lambda}$:

$$V(\underline{\lambda}) = \frac{nm}{(nm - \text{tr}(H))^2} D(\hat{\underline{c}})$$

dove nm è il numero totale di dati, H è la hat matrix che lega \underline{z} a $\hat{\underline{z}}$, D è la devianza, che si calcola come:

$$D(\underline{c}) = 2\phi(l_{\text{sat}} - l(\hat{\underline{c}}))$$

dove ϕ è il parametro di dispersione della distribuzione di Tweedie con cui si modellizzano i dati (quindi nel caso di distribuzione normale, $\phi = \sigma^2$), l è la logverosimiglianza dei dati. Viene usata in due modi:

- l_{sat} , cioè valore di saturazione (massimo della logverosimiglianza), con dati usati al posto di parametri
- $l(\underline{c})$, cioè calcolata con i valori stimati dal modello.

Di conseguenza, se si ha il modello (il caso con covariate è perfettamente analogo)

$$\underline{z} = \Pi \underline{c} + \underline{\epsilon}$$

con

$$\underline{\epsilon} \sim N(\underline{0}, \sigma^2 I)$$

Allora i dati hanno la seguente distribuzione normale multivariata (ricordo che si hanno nm dati):

$$\underline{z} \sim N(\Pi \underline{c}, \sigma^2 I)$$

e quindi

$$l_{\text{sat}} = \log\left(\frac{1}{\sqrt{(2\pi\sigma^2)^{nm}}}\right) - \frac{1}{2\sigma^2}(\underline{z} - \underline{\mu})^T(\underline{z} - \underline{\mu})$$

Marra indica di usare per il valore di saturazione i dati al posto dei parametri di distribuzione, quindi $\underline{\mu} = \underline{z}$, e infatti la logverosimiglianza assume il valore massimo:

$$l_{\text{sat}} = \log\left(\frac{1}{\sqrt{(2\pi\sigma^2)^{nm}}}\right)$$

Invece

$$l(\hat{\underline{c}}) = \log\left(\frac{1}{\sqrt{(2\pi\sigma^2)^{nm}}}\right) - \frac{1}{2\sigma^2}(\underline{z} - \Pi\hat{\underline{c}})^T(\underline{z} - \Pi\hat{\underline{c}})$$

ma $\Pi\hat{\underline{c}} = \hat{\underline{z}}$, quindi

$$l(\hat{\underline{c}}) = \log\left(\frac{1}{\sqrt{(2\pi\sigma^2)^{nm}}}\right) - \frac{1}{2\sigma^2}(\underline{z} - \hat{\underline{z}})^T(\underline{z} - \hat{\underline{z}})$$

Nella differenza $l_{\text{sat}} - l(\hat{\underline{c}})$ i termini $\log\left(\frac{1}{\sqrt{(2\pi\sigma^2)^{nm}}}\right)$ si semplificano.

Inserendo questi valori nella formula della devianza, si ha

$$D(\underline{c}) = (\underline{z} - \hat{\underline{z}})^T(\underline{z} - \hat{\underline{z}})$$

e quindi

$$V(\underline{\lambda}) = \frac{nm}{(nm - \text{tr}(H))^2}(\underline{z} - \hat{\underline{z}})^T(\underline{z} - \hat{\underline{z}})$$

Anche nel caso con covariate si può dimostrare che $V(\underline{\lambda})$ assume lo stesso valore.