Università degli Studi di Padova

Corso di Laurea Magistrale
in Scienze Statistiche

TESI DI LAUREA

STIMA DI MASSIMA VEROSIMIGLIANZA PENALIZZATA PER IL MODELLO DI RASCH

Relatore: Prof. Nicola Sartori
Dipartimento di Scienze Statistiche

Laureando: Alessandro De Bettin
Matricola N° 1097542

ANNO ACCADEMICO 2015/2016

Indice

Introduzione

Spesso, in un'analisi statistica, il modello scelto per spiegare un determinato fenomeno è composto di diversi parametri; l'obiettivo inferenziale dello studio, in questi casi, può essere circoscritto ad un sottoinsieme di essi, detti parametri di interesse. I parametri di disturbo, ovvero i parametri non di diretto interesse, giocano un ruolo chiave nel determinare la bontà delle analisi condotte sui parametri di interesse, poiché senza di essi parte della variabilità presente nei dati non verrebbe spiegata. Quando la dimensione della componente di disturbo è dipendente dalla numerosità campionaria, si presentano i cosiddetti problemi di Neyman e Scott (Neyman e Scott, 1948); ossia, poiché la verosimiglianza non rispetta le condizioni di regolarità, lo stimatore di massima verosimiglianza non gode delle usuali buone proprietà; in particolare, tale stimatore può non essere consistente. Questo tipo di problematica è presente, tipicamente, in modelli in cui ogni unità statistica ha un proprio parametro che ne spiega l'unicità; una situazione del genere è comune, ad esempio, in modelli per dati di panel.

Come riferimento in questa tesi è stato scelto un modello particolarmente esplicativo di queste problematiche: il modello di Rasch. Esso è stato sviluppato nell'ambito della psicometria, per misurare l'abilità dei soggetti e la difficoltà delle prove a cui essi vengono sottoposti in uno studio. Questo modello si presta molto bene da esempio per i fini di questa tesi perché i suoi parametri sono di due tipi: uno spiega l'abilità dei soggetti e l'altro la difficoltà delle domande. Quando l'obiettivo dell'analisi statistica è la quantificazione della difficoltà delle prove, ad esempio, aumentare la quantità dei soggetti facenti parte dello studio non porta, utilizzando lo stimatore di massima verosimiglianza, ad avere uno stimatore più preciso, poiché ogni soggetto

è legato ad un parametro diverso che necessita di una parte dell'informazione per essere stimato.

Nel corso degli anni sono stati proposti diversi metodi per superare questi problemi. Un metodo classico si basa su una particolare pseudo-verosimiglianza: la verosimiglianza condizionata. Questo metodo, dalle buone proprietà, non è sempre applicabile. Un metodo più moderno si basa su un modello ad effetti misti; benché abbia delle ottime proprietà, un'ipotesi su cui si basa non può essere verificata, per cui il suo utilizzo in taluni ambiti, come quello econometrico, non ha preso piede.

Il metodo che si vuole indagare in questa tesi consiste nell'utilizzare moderni metodi di regolarizzazione per ottenere degli stimatori dalle proprietà migliori di quello di massima verosimiglianza. In Tutz e Oelker (2016) viene presentata una panoramica di queste problematiche e uno dei metodi citati si basa su questo tipo di approccio. I tipi di penalizzazione della verosimiglianza presi in considerazione in questa tesi sono due. Il primo, basato sulla regressione ridge, consiste nell'utilizzare come penalità la somma quadratica dei parametri di disturbo. Le stime così ottenute saranno diverse fra loro, ed ognuna di esse sarà minore della corrispettiva stima di massima verosimiglianza. Il secondo, esposto anche in Tutz e Oelker (2016), è basato sul *geralized fused lasso*; la penalità introdotta è la somma dei valori assoluti di tutte le possibili differenze fra i parametri di disturbo. Grazie alle proprietà della penalizzazione basata sulla norma L_1, alcune delle differenze fra parametri vengono poste esattamente a zero, in modo tale che le relative stime dei parametri di disturbo siano uguali: nel caso del modello di Rasch, ciò equivale a raggruppare automaticamente i soggetti parte dello studio in base alla loro abilità. L'introduzione di una penalizzazione come quelle in esame ha l'effetto di imporre una distorsione nello stimatore sia dei parametri di disturbo che di quelli di interesse, diminuendone però la varianza.

L'utilizzo di una regolarizzazione per i termini relativi ai parametri non di interesse riduce la quantità di informazione necessaria per la loro stima, così che essa possa essere utilizzata dagli stimatori dei parametri di interesse. La quantità di informazione che si desidera assegnare allo stimatore dei parametri di disturbo può, in questo caso, essere quantificata da un indice

noto come numero effettivo di parametri; più esso è basso meno informazione viene utilizzata dallo stimatore dei parametri regolarizzati, più è alto più informazione è necessaria. Nel caso in cui non ci sia regolarizzazione, questo indice è uguale al numero p di parametri presenti nel modello; nel caso in cui solo una parte di dimensione J del parametro venga regolarizzata, tale indice è compreso fra $p - J$ e p. Nel caso di verosimiglianza regolare, il numero di parametri in un modello è costante; regolarizzando opportunamente il modello si può fare in modo che al variare della numerosità campionaria il numero effettivo di parametri sia approssimativamente costante. Lo scopo di questa tesi è indagare come l'utilizzo di questo approccio modifichi distorsione e varianza dello stimatore dei parametri di interesse al variare della numerosità campionaria. Per fare ciò si svolge uno studio di simulazione.

Nel Capitolo 1 vengono presentate alcune nozioni di base dell'inferenza frequentista basata sulla verosimiglianza, del modello lineare e del modello lineare generalizzato. In particolare, l'attenzione è rivolta verso lo stimatore di massima verosimiglianza e alcune delle sue proprietà, come la consistenza e la distribuzione asintotica normale; verso il metodo di stima dei minimi quadrati e le caratteristiche degli stimatori nel modello lineare; verso la procedura di stima dei modelli lineari generalizzati nota come *scoring di Fisher* e le proprietà inferenziali dello stimatore che se ne ricava.

Nel Capitolo 2 viene presentato il nucleo teorico di questa tesi. Innanzitutto viene esposto il problema dei parametri incidentali, facendo riferimento ai dati di panel e al modello di Rasch. In seguito, vengono descritte due soluzioni note in letteratura: il modello ad effetti misti e la verosimiglianza condizionata; di entrambi si riportano le principali obiezioni di cui sono oggetto. Infine, vengono esposti i metodi di regolarizzazione ridge, lasso e *generalized fused lasso*, nel caso in cui vengano regolarizzati tutti i parametri o solo una parte. Di ridge e *generalized fused lasso* vengono presentati gli algoritmi di stima e, dove possibile, alcuni risultati inferenziali. Viene presentato anche il concetto di numero effettivo di parametri, mostrando come possa essere stimato. I modelli regolarizzati vengono poi applicati al modello di Rasch.

Nel Capitolo 3 vengono riportati i risultati dello studio di simulazione atto a confrontare distorsione, varianza ed errore quadratico medio degli sti-

matori regolarizzati e non dei parametri di interesse. Le stime di distorsione e varianza vengono rappresentate graficamente al variare della numerosità campionaria, mentre le stime dell'errore quadratico medio vengono riportate in delle tabelle. Gli stimatori basati sui modelli con regolarizzazione, nei casi presi in esame, hanno sempre errore quadratico medio inferiore di quello dello stimatore di massima verosimiglianza, in alcuni casi anche di diversi ordini di grandezza; la distorsione, invece, pare essere paragonabile a quella dello stimatore di massima verosimiglianza senza mostrare una dipendenza dalla numerosità campionaria, mentre la varianza è spesso nettamente inferiore e mostra un andamento decrescente al crescere della numerosità campionaria.

Capitolo 1

Verosimiglianza e Problemi di Regressione

1.1 Introduzione

La verosimiglianza, per la bontà delle sue proprietà e per la vastità dei casi in cui può essere applicata, è uno degli strumenti per l'inferenza statistica frequentista più utilizzati. Essa permette di risolvere le tre problematiche di base dell'inferenza, ovvero la verifica di ipotesi, la stima intervallare e quella puntuale, a partire da un nucleo comune di risultati teorici, in maniera efficace. Una sua applicazione particolarmente interessante è nei modelli di regressione; essi permettono di schematizzare efficacemente fenomeni in cui la media della variabile aleatoria di interesse è una funzione di variabili esplicative.

Questo capitolo si divide in tre sezioni. La prima affronta brevemente l'inferenza di verosimiglianza; dopo una sintetica trattazione teorica, l'attenzione viene rivolta verso due proprietà dello stimatore di massima verosimiglianza: distribuzione asintotica e consistenza. La seconda parte riguarda il modello lineare; vengono presentati il metodo di stima dei minimi quadrati e quello basato sulla verosimiglianza, con le relative proprietà. L'ultima sezione presenta il modello lineare generalizzato, con attenzione verso le sue caratteristiche inferenziali e il suo algoritmo di stima.

Questo capitolo non intende essere esaustivo dei temi trattati, ma ha come scopo quello di inquadrare l'ambito della statistica inferenziale entro cui si muove il nucleo di questa tesi. Infatti, per la comprensione dei problemi connessi ai dati di panel, la teoria della verosimiglianza relativa ai modelli di regressione è imprescindibile. Per una spiegazione più completa si rimanda, ad esempio, a Pace e Salvan (2001).

1.2 Inferenza di Verosimiglianza

Sia $\mathbf{y} = (y_1, \ldots, y_n)$ un campione di osservazioni. Un modello statistico parametrico $\mathcal{F}$ per il campione $\mathbf{y}$ è definito da:

- spazio campionario, $\mathbf{y} \in \mathcal{Y}$;
- funzione del modello, $p(\mathbf{y}; \boldsymbol{\theta})$;
- spazio parametrico, Θ.

La funzione del modello $p(\mathbf{y}; \boldsymbol{\theta})$ è la funzione di densità, se $\mathcal{Y}$ è uno spazio continuo, la funzione di probabilità, nel caso in cui sia uno spazio discreto. Il parametro del modello è $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Lo spazio parametrico, Θ, è lo spazio a cui appartiene il parametro $\boldsymbol{\theta}$. Quindi, nel caso in cui $\mathcal{Y} \subset \mathbb{R}^n$ e $\Theta \subset \mathbb{R}^p$, si ha

$$\mathcal{F} = \left\{ p(\mathbf{y}; \boldsymbol{\theta}),\ \mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^n,\ \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p \right\}.$$

1.2.1 Lo Stimatore di Massima Verosimiglianza

Una volta che il modello statistico è stato definito e che i dati $\mathbf{y} = (y_1, \ldots, y_n)$ sono stati osservati, è possibile ricavare la funzione di verosimiglianza $L(\boldsymbol{\theta}; \mathbf{y})$. Essa è proporzionale alla funzione del modello, dove però $\mathbf{y}$ è fissato (sono i valori osservati) e $\boldsymbol{\theta}$ è l'argomento. Dominio e codominio sono, quindi, $L : \Theta \to \mathbb{R}$. In particolare, la funzione di verosimiglianza è

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{y}) = c(\mathbf{y}) p(\mathbf{y}; \boldsymbol{\theta}).$$

Essa è una sintesi dell'informazione su $\boldsymbol{\theta}$ contenuta nei dati alla luce del modello statistico ipotizzato. La costante $c(\mathbf{y})$ non ha alcuna rilevanza pratica nel processo dell'inferenza di verosimiglianza, ma viene introdotta al fine di rendere $L(\boldsymbol{\theta}; \mathbf{y})$ più facile da studiare, semplificando tutti i termini moltiplicativi di $p(\mathbf{y}; \boldsymbol{\theta})$ che non dipendono da $\boldsymbol{\theta}$.

Nel caso in cui il campione sia formato da osservazioni i.i.d. (indipendenti ed identicamente distribuite) della variabile casuale Y con funzione del modello $p_Y(y_i; \boldsymbol{\theta})$, la funzione di verosimiglianza può essere scritta come

$$L(\theta) = c(\mathbf{y}) \prod_{i=1}^{n} p_Y(y_i; \boldsymbol{\theta}).$$

Per rendere più semplice l'analisi delle proprietà della funzione di verosimiglianza, è uso comune studiare la sua trasformazione logaritmica, la funzione di log-verosimiglianza

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}; \mathbf{y}) = \log L(\boldsymbol{\theta}; \mathbf{y}) = \log c(\mathbf{y}) + \log p(\mathbf{y}; \boldsymbol{\theta}).$$

Nel caso di osservazioni i.i.d., la log-verosimiglianza può essere scritta come

$$l(\boldsymbol{\theta}) = \log c(\mathbf{y}) + \sum_{i=1}^{n} \log p_Y(y_i; \boldsymbol{\theta}).$$

Il valore di $\boldsymbol{\theta}$ che massimizza la funzione di verosimiglianza è detto stima di massima verosimiglianza (`SMV`). Si può quindi definire lo stimatore di massima verosimiglianza

$$\hat{\boldsymbol{\theta}} = \operatorname*{argsup}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) = \operatorname*{argsup}_{\boldsymbol{\theta} \in \Theta} l(\boldsymbol{\theta}). \tag{1.1}$$

1.2.2 Quantità di Verosimiglianza

Nel caso in cui la verosimiglianza di un modello soddisfi opportune condizioni di regolarità, la verosimiglianza gode di proprietà generali di grande interesse statistico. Le condizioni di regolarità (Azzalini, 2001, Paragrafo 3.2.3) sono:

1. Il modello è identificato, ovvero valori diversi di $\boldsymbol{\theta}$ definiscono diversi modelli probabilistici e viceversa.

2. Lo spazio parametrico Θ è un sottoinsieme aperto di $\mathbb{R}^p$, con p non dipendente dalla numerosità campionaria.

3. Il supporto dei modelli probabilistici implicati dal modello statistico $\mathcal{F}$ è il medesimo e non dipende quindi da $\boldsymbol{\theta}$.

4. La log-verosimiglianza è derivabile almeno tre volte, con derivate parziali continue in Θ.

Per l'esposizione delle proprietà dello `SMV` è necessario fare riferimento alle seguenti quantità di verosimiglianza:

- la funzione punteggio (o funzione *score*) è lo jacobiano della log-verosimiglianza $l_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. Essa, quindi, è una funzione vettoriale di dimensione p, il cui r-esimo elemento è $l_{\theta_r}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_r}$, con $r = 1, \ldots, p$;

- la matrice di informazione osservata $j(\boldsymbol{\theta})$ è la matrice Hessiana di $l(\boldsymbol{\theta})$ cambiata di segno. Per esteso: $j(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$; l'elemento di posizione (r, s) di $j(\boldsymbol{\theta})$, $j_{rs}(\boldsymbol{\theta})$, è $\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s}$, con $r, s = 1, \ldots, p$;

- la matrice di informazione attesa di Fisher è il valore atteso della matrice di informazione osservata ed è indicata con $i(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \{j(\boldsymbol{\theta})\}$.

Nel caso di verosimiglianza regolare valgono le seguenti due proprietà:

- prima identità di Bartlett: $\mathbb{E}_{\boldsymbol{\theta}}\{l_{\boldsymbol{\theta}}(\boldsymbol{\theta})\} = \mathbf{0}$;

- seconda identità di Bartlett: $i(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\{l_{\boldsymbol{\theta}}(\boldsymbol{\theta}) l_{\boldsymbol{\theta}}(\boldsymbol{\theta})^T\} = \mathrm{Var}\{l_{\boldsymbol{\theta}}(\boldsymbol{\theta})\}$.

L'informazione attesa è uguale alla varianza della funzione punteggio alla luce del fatto che $i(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\{l_{\boldsymbol{\theta}}(\boldsymbol{\theta}) l_{\boldsymbol{\theta}}(\boldsymbol{\theta})^T\}$ e del fatto che la funzione punteggio ha valore atteso nullo (prima identità di Bartlett).

1.2.3 Proprietà dello Stimatore di Massima Verosimiglianza

Lo stimatore di massima verosimiglianza, definito in (1.1), beneficia di importanti proprietà quando la verosimiglianza è regolare. Lo SMV è, almeno asintoticamente, non distorto: $\mathbb{E}_{\boldsymbol{\theta}}\{\hat{\boldsymbol{\theta}}\} \to \boldsymbol{\theta}$. Un altro risultato rilevante riguarda la varianza dello SMV al crescere della numerosità campionaria: $\mathrm{Var}\{\hat{\boldsymbol{\theta}}\} = i(\boldsymbol{\theta})^{-1} \to 0$. Non distorsione asintotica e convergenza della varianza a zero implicano la consistenza in senso debole di questa classe di stimatori:

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}.$$

Il risultato più importante per l'inferenza riguarda la distribuzione asintotica dello SMV quando la verosimiglianza soddisfa le condizioni di regolarità:

$$\hat{\boldsymbol{\theta}} \dot{\sim} N_p(\boldsymbol{\theta}, i(\boldsymbol{\theta})^{-1}).$$

Poiché il vero valore di $\boldsymbol{\theta}$ è sconosciuto, la quantità $i(\boldsymbol{\theta})$ non si può sempre utilizzare in pratica. Tuttavia, si può dimostrare che vale $j(\hat{\boldsymbol{\theta}})^{-1} i(\boldsymbol{\theta}) \xrightarrow{p} I_p$, per cui nel momento in cui si vogliono effettuare alcune verifiche di ipotesi o stime intervallari, si usa abitualmente la distribuzione asintotica $\hat{\boldsymbol{\theta}} \dot{\sim} N_p(\boldsymbol{\theta}, j(\hat{\boldsymbol{\theta}})^{-1})$.

1.3 Il Modello Lineare

Il modello lineare, per la sua facile interpretabilità e applicazione, è uno strumento molto utilizzato in diversi ambiti. Esso è un modello in cui ogni realizzazione di una variabile di interesse, la variabile risposta, viene ipotizzata essere la somma di una combinazione lineare di altre variabili, dette esplicative, e di un termine di errore. La variabile risposta è numerica, mentre le esplicative possono essere sia quantitative che qualitative. Solitamente, la i-esima osservazione della variabile risposta è indicata con y_i, e le relative p esplicative con $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$. Viene dunque ipotizzato $y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i$, con $i = 1, \ldots, n$, dove $\boldsymbol{\beta}$ è il vettore contenente i p parametri del modello, e ε_i è l'i-esimo termine d'errore. Siano

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

In forma matriciale, il modello può quindi essere scritto come $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, dove $\mathbf{X}$ è detta matrice di disegno.

Alcuni importanti risultati sono ottenibili senza assumere una distribuzione per la variabile risposta, ma a partire dalle cosiddette *ipotesi del secondo ordine*; esse sono:

1. la matrice di disegno $\mathbf{X}$ è non stocastica e di rango pieno $p \leq n$;
2. non ci sono errori sistematici, quindi $\mathbb{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0}$;
3. gli errori sono incorrelati, e ognuno ha la medesima varianza: $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$.

La prima condizione stabilisce che non ci può essere collinearità fra le colonne di $\mathbf{X}$; la seconda condizione implica che $\mathbb{E}\{\mathbf{Y}\} = \mathbb{E}\{\mathbf{X}\boldsymbol{\beta} + \varepsilon\} = \mathbf{X}\boldsymbol{\beta}$; dalla terza, anche conosciuta come ipotesi di *omoschedasticità*, e dalla prima, consegue che $\mathrm{Var}(\mathbf{Y}) = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta} + \varepsilon) = \sigma^2\mathbf{I}_n$.

1.3.1 Il Metodo dei Minimi Quadrati

Con le sole ipotesi del secondo ordine, è possibile ottenere un criterio di stima basato su considerazioni geometriche. Stabilita la forma parametrica di $\mathbb{E}\{\mathbf{Y}\} = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, il criterio dei *minimi quadrati* propone di stimare i parametri $\boldsymbol{\beta}$ minimizzando la distanza euclidea fra $\mathbf{y}$ e $\boldsymbol{\mu}$, $Q(\boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{\mu}\|^2 = (\mathbf{y} - \boldsymbol{\mu})^T(\mathbf{y} - \boldsymbol{\mu}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i - (\beta_1 x_{i1} + \cdots + \beta_p x_{ip})\}^2$. Lo stimatore ai minimi quadrati è

$$\hat{\boldsymbol{\beta}}_{MQ} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, Q(\boldsymbol{\beta}).$$

Poiché $Q(\boldsymbol{\beta})$ è una funzione convessa continua di $\boldsymbol{\beta}$, per trovarne il punto di minimo assoluto basta eguagliare a zero le sue derivate prime parziali, e

risolvere il sistema in funzione dei parametri. Si ottiene quindi lo stimatore

$$\hat{\boldsymbol{\beta}}_{MQ} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{1.2}$$

I valori predetti dal modello sono $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{MQ} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{P}\mathbf{y}$. La matrice $\mathbf{P}$ è la matrice di proiezione sullo spazio delle colonne di $\mathbf{X}$, $\mathcal{C}(\mathbf{X})$ (le colonne di $\mathbf{X}$ sono linearmente indipendenti fra loro); di conseguenza, $\hat{\boldsymbol{\mu}}$ è la proiezione ortogonale di $\mathbf{y}$ su $\mathcal{C}(\mathbf{X})$, e $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\boldsymbol{\mu}} = (\mathbf{I}_p - \mathbf{P})\mathbf{y}$, il vettore dei residui, è ortogonale a $\mathcal{C}(\mathbf{X})$.

Per lo stimatore (1.2) è immediato dimostrare le proprietà

- $\mathbb{E}\left\{\hat{\boldsymbol{\beta}}_{MQ}\right\} = \boldsymbol{\beta}$;
- $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{MQ}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

Un importante risultato, il *teorema di Guass-Markov*, garantisce che $\hat{\boldsymbol{\beta}}_{MQ}$, nel caso di modello correttamente specificato, è lo stimatore lineare non distorto con varianza minima.

La varianza σ^2 può essere stimata a partire dai residui $\hat{\boldsymbol{\varepsilon}}$. Poiché per ipotesi $\mathbb{E}\{\varepsilon_i\} = 0$ e $\mathrm{Var}(\varepsilon_i) = \sigma^2$, uno stimatore ragionevole è $\hat{\sigma}^2 = \sum_{i=1}^n \frac{\hat{\varepsilon}_i^2}{n}$. Siccome questo stimatore è distorto, spesso si preferisce utilizzare la varianza residua corretta $s^2 = \sum_{i=1}^n \frac{\varepsilon_i^2}{n-p}$, che è non distorta.

1.3.2 Il Modello Lineare Normale

I risultati ottenuti a partire dalle ipotesi del secondo ordine sono fondamentali; un'ulteriore assunzione sulla distribuzione del termine d'errore del modello permette di svilupparli in modo tale da formare una solida teoria statistica tramite cui fare inferenza parametrica frequentista.

Si assuma $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2\mathbf{I}_n)$. Grazie alle proprietà della distribuzione normale multivariata, ciò equivale ad assumere $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. Siano $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ e $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$, la funzione di log-verosimiglianza è

$$l(\boldsymbol{\theta}; \mathbf{y}) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}).$$

Risolvendo le equazioni di verosimiglianza, si ottengono

$$\hat{\boldsymbol{\beta}}_{MV} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y};$$

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^2}{n}.$$

Per il modello lineare normale, le stime ai minimi quadrati e di massima verosimiglianza coincidono. Inoltre, utilizzando i risultati derivanti dalla teoria della verosimiglianza, si ottengono gli stessi valori ottenuti dalle ipotesi del secondo ordine per valore atteso e varianza di $\hat{\boldsymbol{\beta}}_{MV}$; l'elemento di novità più rilevante riguarda la distribuzione, in questo caso esatta, dello stimatore, infatti vale

$$\hat{\boldsymbol{\beta}}_{MV} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}).$$

A partire da questo, è possibile effettuare verifiche di ipotesi e stime di regioni di confidenza.

1.4 Il Modello Lineare Generalizzato

Nonostante la versatilità del modello lineare normale, le assunzioni su cui si basa la sua validità possono non rispecchiare la realtà. Ad esempio, in presenza di eteroschedasticità o di risposta discreta, il modello lineare porta a risultati non affidabili. Per risolvere questi problemi, sono stati sviluppati i *modelli lineari generalizzati* (`GLM`); le assunzioni di questa classe di modelli sono meno restrittive di quelle del modello lineare, e permettono, quindi, una più vasta possibilità di applicazione. In particolare, i `GLM` sono applicabili quando la distribuzione della variabile risposta appartiene ad una famiglia di dispersione esponenziale (Pace e Salvan, 1997, Paragrafo 6.4). Ne consegue che il modello lineare normale ne è un caso specifico.

1.4.1 Le Famiglie di Dispersione Esponenziale

La distribuzione di probabilità della variabile aleatoria Y appartiene ad una famiglia di dispersione esponenziale, $Y \sim EF(b(\theta), \psi)$, se la sua funzione

di densità (o di probabilità) può essere scritta come

$$p(y;\theta,\psi) = \exp\left\{\frac{1}{\psi}(y\theta - b(\theta)) + c(y,\psi)\right\},$$

dove

- θ e ψ sono parametri scalari ignoti, definiti rispettivamente parametro naturale e di dispersione o scala;
- $b(\cdot)$ e $c(\cdot)$ sono funzioni note;
- il supporto di Y non dipende da θ e ψ.

Il parametro di dispersione ψ spesso è da considerarsi noto, e in ogni caso la sua stima non influenza quella di θ. La funzione di log-verosimiglianza associata ad una distribuzione appartenente ad una famiglia di dispersione esponenziale con ψ noto è quindi

$$l(\theta) = \frac{1}{\psi}(y\theta - b(\theta)) + c(y,\psi).$$

La funzione punteggio e l'informazione osservata sono rispettivamente

$$l_\theta(\theta) = \frac{1}{\psi}(y - b'(\theta)),\ j(\theta) = b''(\theta)\frac{1}{\psi}.$$

A partire dalle identità di Bartlett si ottengono

$$\mathbb{E}\{Y\} = \mu = b'(\theta),\ \mathrm{Var}\{Y\} = \psi b''(\theta) = \psi V(\mu).$$

La funzione $V(\mu)$ è chiamata funzione di varianza; essa è una funzione della media, ragione per cui nei `GLM` è possibile modellare determinate forme di eteroschedasticità.

1.4.2 Formulazione di un Modello Lineare Generalizzato

In un modello lineare generalizzato il legame tra i parametri $\boldsymbol{\beta}$ e la media è dato da

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta},$$

dove μ_i è la media dell'i-esima osservazione, $\mathbf{x}_i$ è il relativo vettore delle esplicative e $\boldsymbol{\beta}$ è il vettore di dimensione p dei parametri. Quindi, si ipotizza che una funzione della media sia uguale ad una combinazione lineare delle esplicative. Un `GLM` è composto da tre componenti:

- componente casuale,
- componente sistematica,
- funzione legame.

La componente casuale è determinata dalla distribuzione della variabile risposta. La componente sistematica, $\boldsymbol{\eta}$, è, come nel modello lineare, una combinazione lineare delle esplicative, $\eta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}$; vale $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, dove la matrice di disegno $\mathbf{X}$ ha le medesime caratteristiche di quella del modello lineare. La funzione legame $g(\cdot)$ ha il compito di mettere in relazione la media della variabile aleatoria con la sua componente sistematica. Essa deve essere nota, monotona e derivabile, cosicché $\mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$. L'utilizzo della funzione legame permette di mettere in relazione la media della variabile casuale, il cui supporto può non essere $\mathbb{R}$, con la componente sistematica $\eta_i \in \mathbb{R}$. La scelta di $g(\cdot)$ è libera, purché abbia le proprietà citate e il suo dominio sia il supporto della risposta e il suo codominio sia $\mathbb{R}$. Una funzione legame molto utilizzata per le sue proprietà statistiche desiderabili è la *funzione di legame canonico*. Essa dipende dalla distribuzione della risposta e si ottiene da $g(\mu_i) = \theta_i$, dove θ_i è il parametro naturale dell'i-esima osservazione. Poiché si vuole che $g(\mu_i) = g(b'(\theta_i)) = \theta_i$, ne consegue che

$$g(\cdot) = b'^{-1}(\cdot).$$

1.4.3 Inferenza e Stima nel Modello Lineare Generalizzato

Siano $\mathbf{y} = (y_1, \ldots, y_n)$ un campione di osservazioni indipendenti provenienti da $Y_i \sim EF(b(\theta_i), \psi)$ e $\mathbf{X}$ matrice di disegno la cui i-esima riga è $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$; la funzione di log-verosimiglianza è

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \frac{1}{\psi} \{y_i\theta_i - b(\theta_i)\} + c(y_i, \psi) \right\} = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}),$$

con $\theta_i = b'^{-1}(g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta}))$. Per ottenere le stime di massima verosimiglianza e fare inferenza è necessario ricavare le quantità di verosimiglianza

$$l_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \mathbf{X}^T\tilde{\mathbf{W}}\mathbf{u}, \;\; i(\boldsymbol{\beta}) = \mathbf{X}^T\tilde{\mathbf{W}}\mathbf{X},$$

con $\tilde{\mathbf{W}}$ matrice diagonale di elementi $w_i = \frac{1}{\psi V(\mu_i)(g'(\mu_i))^2}$, e $\mathbf{u}$ vettore di elementi $u_i = (y_i - \mu_i)g'(\mu_i)$. Nel caso un cui venga utilizzato il legame canonico, valgono i seguenti risultati:

- $l(\boldsymbol{\beta}) = \frac{1}{\psi}\left[\sum_{i=1}^n y_i\mathbf{x}_i^T\boldsymbol{\beta} - \sum_{i=1}^n b(\mathbf{x}_i^T\boldsymbol{\beta}))\right] + c(y_i, \psi)$;
- $l_{\beta_j}(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{y_i - \mu_i}{\psi} x_{ij}$;
- la derivata seconda della log-verosimiglianza non dipende dalle osservazioni: $j(\boldsymbol{\beta}) = i(\boldsymbol{\beta})$.

Per lo stimatore di massima verosimiglianza $\hat{\boldsymbol{\beta}}$ vale quanto detto nel Paragrafo 1.2.3:

$$\hat{\boldsymbol{\beta}} \dot{\sim} N_p(\boldsymbol{\beta}, i(\boldsymbol{\beta})^{-1}).$$

Poiché $l_{\beta_j}(\boldsymbol{\beta})$ è la j-esima componente della funzione punteggio, per ottenere le stima di massima verosimiglianza, bisogna risolvere il sistema delle equazioni di verosimiglianza

$$l_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \mathbf{0} \implies \mathbf{X}^T\tilde{\mathbf{W}}\mathbf{u} = \mathbf{0}.$$

A meno di casi particolari, la soluzione non è esplicita. Al fine di ottenere le stime, si utilizza l'algoritmo *scoring di Fisher*, una variante dell'algoritmo

di *Newton-Raphson*. Questo algoritmo è iterativo e consiste nel trovare ripetutamente lo zero dell'approssimazione lineare della funzione punteggio nella stima corrente fino a raggiungimento della convergenza. Il punto di partenza è

$$l_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \approx l_{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\beta}}^{(0)}\right) - i(\boldsymbol{\beta})\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(0)}\right),$$

l'approssimazione lineare della funzione punteggio nel punto $\hat{\boldsymbol{\beta}}^{(0)}$, con $i(\boldsymbol{\beta})$ che sostituisce $j(\boldsymbol{\beta})$. Risolvendo opportunamente, si ottiene il primo passo della procedura di stima in

$$\hat{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}^{(0)} + i^{-1}\left(\hat{\boldsymbol{\beta}}^{(0)}\right) l_{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\beta}}^{(0)}\right).$$

Il passo t-esimo dell'algoritmo è quindi

$$\hat{\boldsymbol{\beta}}^{(t)} = \hat{\boldsymbol{\beta}}^{(t-1)} + i^{-1}\left(\hat{\boldsymbol{\beta}}^{(t-1)}\right) l_{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\beta}}^{(t-1)}\right).$$

Ogni iterazione è equivalente a

$$\mathbf{X}^T\tilde{\mathbf{W}}^{(t-1)}\mathbf{X}\hat{\boldsymbol{\beta}}^{(t)} = \mathbf{X}^T\tilde{\mathbf{W}}^{(t-1)}\mathbf{z}^{(t-1)},$$

dove $\mathbf{z}^{(t-1)}$ è un vettore di componenti $z_i^{(t-1)} = \eta_i + (y_i - \mu_i)\frac{\partial \eta_i^{(t-1)}}{\partial \mu_i^{(t-1)}}$. L'iterazione dell'algoritmo può quindi essere scritta come

$$\hat{\boldsymbol{\beta}}^{(t)} = (\mathbf{X}^T\tilde{\mathbf{W}}^{(t-1)}\mathbf{X})^{-1}\mathbf{X}^T\tilde{\mathbf{W}}^{(t-1)}\mathbf{z}^{(t-1)},$$

ogni passo quindi consiste in una stima ai minimi quadrati pesati; perciò, la procedura prende il nome di *algoritmo dei minimi quadrati ponderati iterati* (**IWLS**). Dei buoni valori di partenza per l'algoritmo sono $z_i^{(0)} = g(y_i)$ e $\tilde{\mathbf{W}}^{(0)} = \mathbf{I}_p$.

Capitolo 2

Stima Regolarizzata del Modello di Rasch

2.1 Introduzione

Questo capitolo è diviso in tre parti; nella prima viene presentato il problema dei parametri incidentali, con riferimento ai dati di panel e al modello di Rasch; nella seconda vengono presentate due soluzioni presenti in letteratura: il modello ad effetti misti e la verosimiglianza condizionata; di entrambe si riportano le proprietà e si mostra quali sono le principali obiezioni al loro utilizzo; la terza sezione, nucleo di questa tesi, riguarda la regressione regolarizzata, in particolare la regressione ridge, il lasso, e una sua variante denominata *generalized fused lasso*, di interesse nel modello di Rasch perché permette un raggruppamento automatico dei soggetti in base alla loro abilità. Di questi metodi vengono indagati, dove possibile, distorsione e varianza, gli algoritmi di stima e l'aspetto interpretativo. Il modello di riferimento è il modello di Rasch, in quanto esso è un esempio chiave del problema dei parametri incidentali e la sua semplicità ne facilita l'esposizione; le considerazioni fatte per questo modello valgono anche per modelli più complessi. In Tutz e Oelker (2016) si trovano alcuni dei modelli presentati.

2.2 I Problemi di Neyman e Scott

Le buone proprietà dello SMV, presentate nel Capitolo 1, si basano su dei presupposti; qualora questi non siano rispettati, l'usuale teoria asintotica della verosimiglianza non è affidabile. Quando il modello statistico utilizzato ha specifiche caratteristiche, sorgono nello studio della verosimiglianza quelli che in letteratura prendono il nome di problemi di Neyman e Scott, i due autori di Neyman e Scott (1948); essi per primi, infatti, presentarono una soluzione inferenziale valida per questi modelli.

Più nello specifico, i modelli che portano a questi problemi sono quelli per cui la dimensione dello spazio parametrico dipende dalla numerosità campionaria (Pace e Salvan, 1997, Capitolo 4); in questo modo la condizione di regolarità 2 di pagina 12 non viene rispettata. Intuitivamente, quando lo spazio parametrico è fissato, più numeroso è l'insieme di dati utilizzato per l'analisi, più precise saranno le stime ottenute da questi, perché avere più dati vuol dire avere più informazione; quando invece la dimensione dello spazio parametrico cresce con la dimensione del campione, più dati vogliono dire sia più informazione, ma anche più parametri che necessitano di questa per essere stimati. Di conseguenza, aumentare il numero di unità che fanno parte dello studio non vuol più dire ottenere stime più precise. Per formalizzare il concetto appena esposto è necessario introdurre il concetto di parametri incidentali.

2.2.1 Parametri Incidentali

Spesso, in un'analisi statistica, non tutti i parametri sono interessanti agli occhi del ricercatore. Infatti, l'obiettivo inferenziale può essere lo studio di solo una parte dei parametri, detti *parametri di interesse* e indicati con $\boldsymbol{\psi}$; i restanti parametri sono detti *parametri di disturbo* e vengono indicati con $\boldsymbol{\zeta}$; la loro presenza è necessaria per l'adeguatezza del modello: senza di essi parte della variabilità presente nei dati non verrebbe spiegata, e tutta l'inferenza su $\boldsymbol{\psi}$ ne sarebbe inficiata. Essi determinano lo spazio parametrico del modello e, inoltre, lo stimatore di $\boldsymbol{\psi}$ spesso dipende da $\boldsymbol{\zeta}$. Vale, quindi, $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\zeta})$, con

- $\boldsymbol{\psi}$ vettore di lunghezza k dei parametri di interesse;
- $\boldsymbol{\zeta}$ vettore di lunghezza $p - k$ dei parametri di disturbo.

Ogni osservazione, solitamente, è la realizzazione della medesima variabile casuale, i cui parametri sono $\boldsymbol{\theta}_i = (\boldsymbol{\psi}, \boldsymbol{\zeta})$. Nei modelli, oggetto di questa tesi, in cui ogni osservazione, o gruppo di osservazioni, ha dei parametri individuali, vale $\boldsymbol{\theta}_i = (\boldsymbol{\psi}, \boldsymbol{\zeta}_i)$; in questi casi, la dimensione di Θ dipende dal numero di osservazioni presenti nel campione, per cui la condizione di regolarità 2 non è rispettata. Secondo la terminologia utilizzata in Neyman e Scott (1948), in questo caso, i parametri si dividono in

- *parametri strutturali*, indicati con $\boldsymbol{\psi}$;
- *parametri incidentali*, indicati con $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \ldots, \boldsymbol{\zeta}_n)$.

Una delle conseguenze della presenza dei parametri incidentali è che lo `SMV` non è più consistente: al crescere del numero di soggetti che fanno parte dello studio, la varianza dello stimatore dei parametri strutturali non decresce. In questa tesi ci si concentra sui metodi utilizzabili per ottenere stimatori consistenti quando sono presenti i problemi di Neyman e Scott. Per approfondimenti si veda, ad esempio, Arellano e Hahn (2007).

2.2.2 I Dati di Panel

Una tipologia di dati in cui molto spesso si incontrano i problemi oggetto di questo capitolo è quella dei dati di panel. Con dati di panel, o dati longitudinali, si intendono quei dati per cui, per ogni unità facente parte dello studio, si hanno misure ripetute delle sue caratteristiche nel tempo. Questi dati sono molto comuni in ambito econometrico e sociale. La ragione per cui i dati di panel sono legati ai problemi di Neyman e Scott è che, nel momento di stesura di un modello adeguato, per tenere conto delle caratteristiche uniche di ogni unità, si aggiunge un parametro per ognuna di esse. In questa tesi vengono considerati dati di questo tipo per modelli non dinamici, in cui ogni osservazione è considerata indipendente dalle altre; nonostante questo, le conclusioni ottenute sono valide anche per modelli con dipendenza temporale.

2.2.3 Il Modello di Rasch

Il modello di riferimento di questa tesi è il modello di Rasch. Questo modello è stato sviluppato nell'ambito della psicometria con l'obiettivo di stimare l'abilità di delle unità nel superare determinate prove, più o meno difficili. Un caso tipico in cui viene utilizzato è quello dell'analisi dei questionari per misurare l'abilità degli studenti. Questo modello è un caso particolare di modelli della *item-response theory*, e, di fatto, è un modello lineare generalizzato. Il modello di Rasch, nella sua forma più semplice, è adatto alla esemplificazione delle problematiche di questa tesi. Supponendo di sottoporre n soggetti a J diverse prove, i parametri da stimare sono $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_J)$, a rappresentare la difficoltà di ognuna delle prove, e $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$ a rappresentare l'abilità di ogni soggetto. Come è evidente, più unità fanno parte del campione, più aumenta la dimensione dello spazio parametrico. Alcuni esempi basati su questo modello sono descritti in Pace e Salvan (1997, Capitolo 4).

Più formalmente, la versione più semplice del modello di Rasch è un `GLM` binomiale con funzione di legame canonico il cui obiettivo è modellare le probabilità π_{ij}, dove $i = 1, 2, \ldots, n$ è l'indice relativo al soggetto e $j = 1, 2, \ldots, J$ quello relativo alla prova; π_{ij} è la probabilità che il soggetto i-esimo superi la prova j-esima. Il modello è

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \delta_j + \gamma_i.$$

Il modello così descritto è un `GLM` con due esplicative categoriali non ordinali, la domanda e il soggetto. Il risultato di ogni prova per ogni soggetto rappresenta un dato, per cui, se ognuno affronta tutte le prove (se non ci sono dati mancanti) il modello verrà stimato su $n \cdot J$ dati. I vettori della variabile risposta e dei parametri sono

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1J} \\ y_{21} \\ \vdots \\ y_{nJ} \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_J \\ \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix},$$

dove $y_{ij} = 0$ se il soggetto i-esimo fallisce la j-esima prova, $y_{ij} = 1$ se la completa con successo (y_{ij} è realizzazione della variabile casuale $Y_{ij} \sim$ Binomiale$(1, \pi_{ij})$). La matrice di disegno è

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Il modello presentato, sebbene chiaro nella sua interpretazione, non è identificato. Infatti, sommando le prime J colonne fra di loro si ottiene il vettore $\mathbf{1}_{n*J}$, medesimo vettore ottenibile sommando le ultime n colonne. La matrice di disegno, pertanto, soffre di collinearità. Per risolvere questo problema, tipico delle variabili esplicative categoriali, basta modificare la parametrizzazione del modello. Si definisce $\boldsymbol{\alpha} = \boldsymbol{\delta} + \gamma_1 \mathbf{1}_J$; i parametri $\boldsymbol{\alpha}$ rappresentano la difficoltà delle prove per il primo soggetto. Sia, inoltre, $\boldsymbol{\beta} = \boldsymbol{\gamma}_{-1} - \gamma_1 \mathbf{1}_{n-1}$, dove $\boldsymbol{\gamma}_{-1}$ è il vettore $\boldsymbol{\gamma}$ senza il parametro γ_1; i $\boldsymbol{\beta}$ rappresentano la differenza in abilità fra i soggetti e il primo soggetto. Si noti che questo modo di parametrizzare il modello non è l'unico possibile.

2.3 Soluzioni nella Letteratura

Nel corso degli anni, sono stati sviluppati diversi metodi utilizzabili per dati con le caratteristiche presentate nella sezione precedente; due dei metodi più conosciuti sono quello basato sul modello ad effetti misti e quello costruito sulla verosimiglianza condizionata. Entrambi generalmente porta-

no a dei buoni risultati, ma ognuno di essi ha dei difetti che non li rendono applicabili in ogni contesto. Il modello ad effetti misti è simile ad un metodo bayesiano empirico: l'eterogeneità degli individui viene spiegata assumendo una distribuzione a priori per i parametri ad essi associata; un'ipotesi del genere, purtroppo, non può essere verificata, motivo per cui in alcuni ambiti, come quello econometrico, questo tipo di modello non ha mai preso piede.

L'approccio basato su verosimiglianza condizionata porta a degli stimatori consistenti dei parametri strutturali, ma può essere utilizzato solo in determinate situazioni; ad esempio, nei `GLM` solo quando la funzione legame è di legame canonico. Ciò ne restringe le possibilità di applicazione. Nel resto di questa sezione verranno presentate queste soluzioni più nel dettaglio.

2.3.1 Modello ad Effetti Misti

Il modello ad effetti misti, in problemi di questo tipo, sarebbe la prima scelta per molti ricercatori: ha buone proprietà inferenziali, è flessibile, abbastanza robusto ad errori di specificazione e la sua popolarità garantisce che più persone siano in grado di interpretarne i risultati. Nonostante sia stato provato empiricamente che questa classe di modelli ha spesso un comportamento accettabile anche quando le ipotesi di partenza non sono realistiche, questi modelli possono essere sconsigliati poiché esse non possono essere verificate. Per lo scopo di questa tesi ci si limiterà a presentare i modelli ad intercetta casuale. Una breve spiegazione di questo modello è presente in Tutz e Oelker (2016).

La specificazione parametrica di un modello a intercetta casuale consiste in due parti fondamentali: gli effetti fissi e gli effetti casuali. I primi corrispondono ai parametri come li si intende nella statistica frequentista: valori fissati, oggetto dello studio dello statistico, che determinano il processo generatore dei dati; nell'ambito di questa tesi essi coincidono con quelli che nel Paragrafo 2.2.1 sono stati definiti come parametri strutturali. I secondi non sono dei parametri come li si intende nella statistica frequentista, ma sono più vicini alla loro accezione nella statistica bayesiana; un effetto casuale, come un effetto fisso, determina il processo generatore dei dati, tuttavia invece

di essere una quantità fissata, esso è la realizzazione di una variabile casuale non osservabile. In questo modo, si possono sfruttare alcune caratteristiche della statistica bayesiana, ma facendo affidamento esclusivamente sul principio del campionamento ripetuto. Infatti, in questo caso, si ipotizza che la distribuzione degli effetti casuali sia realistica, ovvero, che il vero processo generatore dei dati si appoggi su una variabile casuale per determinati parametri; secondo un approccio bayesiano puro, invece, le distribuzioni a priori dei parametri non rappresentano la struttura reale del fenomeno, ma sono una sintesi dell'informazione pre-sperimentale a disposizione su di essi.

Le caratteristiche della variabile casuale latente vengono stimate attraverso i dati; questo approccio, quindi, è simile a quello bayesiano empirico. Da queste considerazioni è evidente che utilizzare una distribuzione degli effetti casuali che non rispecchia la realtà può avere un impatto sull'inferenza, poiché coinciderebbe con una errata specificazione del modello. Da qui nascono le critiche a questo tipo di modellazione: verificare se le assunzioni sulla distribuzione degli effetti casuali siano giuste non è possibile, poiché essi non sono osservabili, e quindi l'inferenza su questo tipo di modello può potenzialmente essere non attendibile.

Nel caso del modello di Rasch, il modello a intercetta casuale è un modello lineare generalizzato a effetti misti (`GLMM`, *Generalized Linear Mixed Model*). La sua formulazione è

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \delta_j + b_i,$$

dove b_i è l'intercetta casuale. Solitamente si ipotizza che b_i sia la realizzazione della variabile casuale $B \sim N(0, \sigma_B^2)$. Per stimare il modello, occorre ricavare la funzione di verosimiglianza e modificarla in modo che non dipenda dai b_i, $i = 1, \ldots, n$, poiché essi non sono osservati. La probabilità congiunta del modello per $(\mathbf{y}_i, b_i)$ è

$$p_{(\mathbf{Y}_i,B)}(\mathbf{y}_i, b_i; \boldsymbol{\delta}, \sigma_B^2) = \prod_{j=1}^{J} p_{(Y_{ij}|b_i)}(y_{ij}|b_i; \boldsymbol{\delta}, \sigma_B^2) p_B(b_i; \sigma_B^2),$$

dove $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})$, $p_{(Y_{ij}|b_i)}(y_{ij}|b_i; \boldsymbol{\delta}, \sigma_B^2)$ è la funzione di densità (o di probabilità) di $Y_{ij}|b_i$ e $p_B(b_i; \sigma_B^2)$ quella della variabile casuale B. Nel caso del modello di Rasch, la formulazione ad effetti casuali prevede che $Y_{ij}|b_i \sim$ Binomiale$(1, \pi_{ij})$ e che B abbia distribuzione normale, come specificato in precedenza (questa è la scelta più comune, ma in teoria sono disponibili anche altre opzioni). A partire da questa funzione è possibile ottenere la verosimiglianza per una singola osservazione $(\mathbf{y}_i, b_i)$

$$L(\boldsymbol{\delta}, \sigma_B^2; y_{ij}, b_i) = \prod_{j=1}^{J} p_{(Y_{ij}|b_i)}(y_{ij}|b_i; \boldsymbol{\delta}, \sigma_B^2) p_B(b_i; \sigma_B^2).$$

La verosimiglianza per tutte le osservazioni $(\mathbf{y}, \mathbf{b})$ è

$$L(\boldsymbol{\delta}, \sigma_B^2; \mathbf{y}, \mathbf{b}) = \prod_{i=1}^{n} L(\boldsymbol{\delta}, \sigma_B^2; y_{ij}, b_i).$$

Poiché la verosimiglianza dipende dalle quantità non osservate $\mathbf{b}$, essa non può essere utilizzata. Per stimare i parametri strutturali si utilizza la *verosimiglianza marginale* (o *verosimiglianza integrata*), in cui si utilizza la funzione di densità (o probabilità) marginale degli $\mathbf{Y}_i$; essa viene ricavata a partire dalla densità congiunta di $(\mathbf{Y}_i, B)$. Solitamente la distribuzione di B viene scelta in modo tale che essa abbia come supporto $\mathbb{R}$ (come nel caso della normale), per cui per ottenere la distribuzione marginale di $\mathbf{Y}_i$ bisogna integrare $p_{(\mathbf{Y}_i, B)}(\mathbf{y}_i, b_i; \boldsymbol{\delta}, \sigma_B^2)$ rispetto alla quantità b_i. Per cui

$$p_{\mathbf{Y}_i}(\mathbf{y}_i; \boldsymbol{\delta}, \sigma_B^2) = \int_{-\infty}^{\infty} p_{(\mathbf{Y}_i, B)}(\mathbf{y}_i, b_i; \boldsymbol{\delta}, \sigma_B^2) \, \mathrm{d}b_i.$$

In questo modo, la funzione di densità (o di probabilità) non dipende più dai valori $\mathbf{b}$, per cui è ottenibile la verosimiglianza marginale

$$L(\boldsymbol{\delta}, \sigma_B^2; \mathbf{y}) = \prod_{i=1}^{n} p_{\mathbf{Y}_i}(\mathbf{y}_i; \boldsymbol{\delta}, \sigma_B^2).$$

A questo punto, la verosimiglianza marginale può essere utilizzata per ottenere la stima dei parametri strutturali e di σ_B^2; ora, la verosimiglianza da

studiare è regolare, infatti, all'aumentare della numerosità campionaria, lo spazio parametrico è il medesimo, e lo SMV ha le usuali buone proprietà.

Questo approccio è simile a quello bayesiano empirico perché gli iperparametri delle distribuzioni a priori vengono stimati a partire dai dati, e non sono fissati come in un approccio bayesiano puro. Questo modo di operare permette di ottenere delle stime bayesiane empiriche per i **b**. Infatti, per il teorema di Bayes vale

$$p_{B_i|\mathbf{y}_i}(b_i|\mathbf{y}_i;\boldsymbol{\delta},\sigma_B^2) \propto \prod_{j=1}^{J} p_{(Y_{ij}|b_i)}(y_{ij}|b_i;\boldsymbol{\delta},\sigma_B^2)p_B(b_i;\sigma_B^2).$$

Utilizzando la densità di $B_i|\mathbf{y}_i$ è possibile ottenere una stima puntuale, in genere chiamata previsione, per i parametri incidentali latenti del modello. Se, ad esempio, si decide di utilizzare il valore medio della distribuzione a posteriori di B_i a questo scopo,

$$\hat{b}_i = \mathbb{E}\{B_i|\mathbf{y}_i\}.$$

Qualora la distribuzione a posteriori non porti ad una soluzione analitica esplicita, può essere conveniente ottenere una stima utilizzando opportune approssimazioni, sia di tipo analitico che di tipo Monte Carlo.

2.3.2 Verosimiglianza Condizionata

Un'altra strategia utilizzata per risolvere il problema dei parametri incidentali è basata sulla verosimiglianza condizionata; questa è una pseudo-verosimiglianza che, qualora esista, non dipende dai parametri di disturbo, e che fornisce degli stimatori dalle proprietà desiderabili per i parametri di interesse. Rispetto al modello ad effetti misti, il metodo basato sulla verosimiglianza condizionata è meno moderno e in genere meno utilizzato, ma le sue valide caratteristiche ne rendono la spiegazione obbligatoria. Nei GLM, e quindi nel modello di Rasch, questo metodo è utilizzabile nel caso in cui la funzione legame sia di tipo canonico. Una spiegazione più esaustiva si trova in Davison (2003, Paragrafo 12.3).

Si supponga di essere nello scenario descritto nel Paragrafo 2.2.1. Sia u una statistica tale che

$$p_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\psi}, \boldsymbol{\zeta}) = p_U(u; \boldsymbol{\psi}, \boldsymbol{\zeta}) p_{\mathbf{Y}|u}(\mathbf{y}; u, \boldsymbol{\psi}),$$

dove $p_U(u; \boldsymbol{\psi})$ è la densità della statistica U, funzione del vettore casuale $\mathbf{Y}$. Una scelta utile, se possibile, per la statistica u è la statistica sufficiente minimale per $\boldsymbol{\zeta}$. Se si ammette che la quantità data da $p_U(u; \boldsymbol{\psi}, \boldsymbol{\zeta})$ sia trascurabile nella costruzione della verosimiglianza, si ottiene

$$L_C(\boldsymbol{\psi}) = L_C(\boldsymbol{\psi}; \mathbf{y}|u) = p_{\mathbf{Y}|u}(\mathbf{y}; u, \boldsymbol{\psi}).$$

Questa viene detta *verosimiglianza condizionata*, ed è una pseudo-verosimiglianza. Eliminare dalla verosimiglianza il termine $p_U(u; \boldsymbol{\psi}, \boldsymbol{\zeta})$ vuol dire perdere dell'informazione sui parametri strutturali, poiché esso dipende da $\boldsymbol{\psi}$; questa perdita solitamente è accettabile, perché permette anche di eliminare i parametri incidentali dalla verosimiglianza. In questo modo, si ottiene un problema regolare di stima, e lo stimatore basato sulla verosimiglianza condizionata è consistente.

Nel `GLM` binomiale con funzione legame canonica, e quindi anche nel modello di Rasch, la statistica sufficiente minimale è data da $S = \mathbf{X}^T\mathbf{y}$, dove $\mathbf{X}$ è la matrice di disegno; essa, quindi, ha dimensione pari al numero di parametri del modello, e in particolare il parametro j-esimo ha la sua statistica sufficiente in $\mathbf{x}_j^T y$, dove $\mathbf{x}_j$ è la j-esima colonna di $\mathbf{X}$. Per il modello di Rasch, in cui i parametri incidentali sono quelli relativi all'abilità di ogni soggetto, la statistica a cui condizionarsi è S_n; a secondo della parametrizzazione, la dimensione di S_n può essere n o $n-1$: nel caso in cui ogni soggetto abbia il suo parametro, l'i-esimo elemento di S_n è il numero di prove superate dall'i-esimo soggetto; nel caso in cui l'abilità di un soggetto di riferimento venga spiegata da un intercetta o faccia parte dei parametri strutturali, l'i-esimo elemento di S_n è il numero di prove superate dal soggetto $(i+1)$-esimo. Per cui, una volta ottenuta la verosimiglianza condizionata, è possibile stimare i parametri strutturali. Sebbene questo approccio porti a degli stimatori dalle

caratteristiche vicine a quelle dello `SMV` in un problema regolare di stima, la verosimiglianza condizionata può essere utilizzata per una ristretta gamma di modelli, e quindi è una soluzione non universale al problema dei parametri incidentali.

2.4 Soluzione tramite Regressione Regolarizzata

Negli ultimi anni, sempre di più, i metodi di regolarizzazione sono stati oggetto di interesse nel mondo della ricerca. Questi, infatti, permettono di risolvere problematiche che si sono presentate solo di recente nell'analisi dei dati. Quando l'obiettivo dello studio è ottenere un modello in grado di predire al meglio un determinato fenomeno, un modello regolarizzato può restituire predizioni con errore quadratico medio inferiore; quando, in un modello di regressione, si hanno molte variabili esplicative, può essere fondamentale una buona selezione delle variabili, cosa che alcuni metodi di regolarizzazione è stato dimostrato svolgono egregiamente. Problemi del genere si pongono spesso nel mondo d'oggi, in cui molti servizi si basano sull'accuratezza delle predizioni dei loro modelli, e in cui il basso costo della raccolta di informazione permette di avere insiemi di dati da analizzare di dimensione anche molto elevata. In questa tesi, i metodi di regolarizzazione vengono utilizzati per le loro proprietà di *shrinkage*, ovvero la capacità di comprimere le stime di determinati parametri verso lo zero; i parametri incidentali saranno oggetto della compressione, in modo che i parametri strutturali possano essere stimati in maniera più efficace. Questo approccio, a differenza di quello basato sulla verosimiglianza condizionata, può essere applicato a qualsiasi tipo di `GLM` e, basandosi su un modello ad effetti fissi, non soffre dei problemi del modello ad intercetta casuale.

2.4.1 La Regressione Ridge

La *regressione ridge* (Hoerl e Kennard, 1970) fu originariamente proposta per risolvere un problema per cui oggi viene utilizzata poco: la collinearità.

In un modello lineare, quando alcune esplicative sono, anche solo approssimativamente, la combinazione lineare di altre, il modello non è stimabile. Infatti, la matrice $\mathbf{X}^T\mathbf{X}$, sotto queste condizioni, non è a rango pieno e non può essere invertita, per cui lo stimatore $\hat{\boldsymbol{\beta}}_{MQ} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, riportato in (1.2), non è utilizzabile. Questo tipo di regolarizzazione, come molti altri, risolve questo problema. Poiché questo metodo permette di ottenere diverse quantità di interesse in maniera esplicita, è stato considerato come punto di partenza per questa tesi.

La regressione ridge, al fine di ottenere la compressione dei parametri, modifica il problema di ottimizzazione dei minimi quadrati aggiungendo un vincolo sulla somma quadratica dei parametri. Nel caso si vogliano comprimere tutti i parametri e il modello abbia intercetta, per ottenere le stime ridge si desidera minimizzare

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$
$$\text{soggetto a } \sum_{j=2}^{p} \beta_j^2 \leq t.$$

L'intercetta β_1 non viene penalizzata, in modo che cambiamenti di posizione nelle $\mathbf{y}$ vengano colti dal modello (se venisse regolarizzata anche l'intercetta, per $t \to 0$ la media stimata dal modello sarebbe sempre nulla). Utilizzando i moltiplicatori di Lagrange, la funzione obiettivo può essere riscritta come

$$T(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{2}\sum_{j=2}^{p} \beta_j^2.$$

Il parametro λ viene detto parametro di regolazione, e determina quanto i parametri vengano compressi verso lo zero. Esiste una corrispondenza uno a uno fra λ e t. Quando λ è vicino a zero (t grande) la compressione è praticamente nulla, quando λ è grande (t piccolo) la compressione è rilevante. A definire quando λ è grande o piccolo è la grandezza delle stime dei parametri in assenza di regolarizzazione. Il problema di ottimizzazione è di semplice risoluzione perché la funzione $T(\boldsymbol{\beta})$ è convessa e differenziabile e vale, nel

caso in cui i dati siano stati centrati,

$$\hat{\boldsymbol{\beta}}_{ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, T(\boldsymbol{\beta}) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}.$$

Lo stimatore $\hat{\boldsymbol{\beta}}_{ridge}$ è esplicito e non può soffrire del problema della collinearità, infatti se $\mathbf{X}^T\mathbf{X}$ non è a rango pieno, $\mathbf{X}^T\mathbf{X} + \lambda I_p$ lo è sicuramente, poiché a ogni elemento della diagonale di $\mathbf{X}^T\mathbf{X}$ viene aggiunta la quantità λ.

Lo stimatore $\hat{\boldsymbol{\beta}}_{ridge}$ ha delle proprietà diverse da quello ai minimi quadrati. L'introduzione della penalità comporta la presenza di distorsione nello stimatore, e ne modifica anche la forma della varianza. Si può dimostrare che rispetto dallo stimatore ai minimi quadrati, una scelta opportuna del parametro di regolazione porta a delle previsioni con errore quadratico medio inferiore, permettendo la presenza di distorsione nello stimatore e diminuendone la varianza. Sia $\mathbf{W} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}$, la distorsione di $\hat{\boldsymbol{\beta}}_{ridge}$ è

$$\begin{aligned}\operatorname{bias}(\hat{\boldsymbol{\beta}}_{ridge}) &= \mathbb{E}\{\hat{\boldsymbol{\beta}}_{ridge}\} - \boldsymbol{\beta} = \mathbf{W}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta} \\ &= (\mathbf{W}\mathbf{X}^T\mathbf{X} - \mathbf{I}_p)\boldsymbol{\beta} = \mathbf{W}\mathbf{W}^{-1}(\mathbf{W}\mathbf{X}^T\mathbf{X} - \mathbf{I}_p)\boldsymbol{\beta} \\ &= \mathbf{W}(\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{X} - \lambda\mathbf{I}_p)\boldsymbol{\beta} = -\lambda\mathbf{W}\boldsymbol{\beta}.\end{aligned}$$

La sua varianza è

$$\operatorname{Var}(\hat{\boldsymbol{\beta}}_{ridge}) = \sigma^2\mathbf{W}\mathbf{X}^T\mathbf{X}\mathbf{W}.$$

Se si pone $\lambda = 0$ si ottiene lo stimatore ai minimi quadrati, e distorsione e varianza tornano a essere quelle usuali.

In determinate situazioni, è preferibile regolarizzare solo una parte dei parametri. Sia $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, con $\boldsymbol{\alpha}$ di dimensione J e $\boldsymbol{\beta}$ di dimensione n; ipotizzando di volere comprimere solo i parametri $\boldsymbol{\beta}$, la funzione obiettivo diventa

$$T(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \frac{\lambda}{2}\sum_{i=1}^{n}\beta_i^2,$$

e lo stimatore ridge diventa

$$\hat{\boldsymbol{\theta}}_{ridge} = \left(\mathbf{X}^T\mathbf{X} + \lambda\begin{pmatrix}\mathbf{0}_{J\times J} & \mathbf{0}_{J\times n}\\ \mathbf{0}_{n\times J} & \mathbf{I}_n\end{pmatrix}\right)^{-1}\mathbf{X}^T\mathbf{y}.$$

In questo caso, $\mathbf{W} = \left(\mathbf{X}^T\mathbf{X} + \lambda\begin{pmatrix}\mathbf{0}_{J\times J} & \mathbf{0}_{J\times n}\\ \mathbf{0}_{n\times J} & \mathbf{I}_n\end{pmatrix}\right)^{-1}$. Distorsione e varianza sono

$$\text{bias}(\hat{\boldsymbol{\theta}}_{ridge}) = -\lambda\mathbf{W}\begin{pmatrix}\mathbf{0}_J\\ \boldsymbol{\beta}\end{pmatrix},\ \text{Var}(\hat{\boldsymbol{\theta}}_{ridge}) = \sigma^2\mathbf{W}\mathbf{X}^T\mathbf{X}\mathbf{W}.$$

Sebbene lo stimatore dei parametri $\boldsymbol{\alpha}$ non venga compresso, poiché esso è correlato con lo stimatore dei $\boldsymbol{\beta}$, esso risulta essere distorto. Si considerino le seguenti matrici nella loro formulazione a blocchi:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix}\mathbf{A}_{J\times J} & \mathbf{C}_{J\times n}\\ \mathbf{D}_{n\times J} & \mathbf{E}_{n\times n}\end{pmatrix},\ \mathbf{W}^{-1} = \begin{pmatrix}\mathbf{A}_{J\times J} & \mathbf{C}_{J\times n}\\ \mathbf{D}_{n\times J} & \mathbf{B}_{n\times n}\end{pmatrix},\ \mathbf{W} = \begin{pmatrix}\mathcal{A}_{J\times J} & \mathcal{C}_{J\times n}\\ \mathcal{D}_{n\times J} & \mathcal{B}_{n\times n}\end{pmatrix}.$$

Ne consegue che $\text{bias}(\hat{\boldsymbol{\alpha}}_{ridge}) = -\lambda\mathcal{C}\boldsymbol{\beta}$. Quando le colonne della matrice $\mathbf{X}$ sono fra loro ortogonali, la matrice $\mathcal{C}$ è composta esclusivamente da zeri, per cui la distorsione di $\hat{\boldsymbol{\alpha}}_{ridge}$ è nulla; quando le colonne di $\mathbf{X}$ sono approssimativamente ortogonali, situazione più comune, i valori che compongono $\mathcal{C}$ sono vicini a zero, e lo stimatore ha distorsione trascurabile.
Poiché le matrici in esame sono simmetriche per costruzione, valgono $\mathbf{D} = \mathbf{C}^T$ e $\mathcal{D} = \mathcal{C}^T$. Sfruttando le regole di inversione di matrici a blocchi, si ottiene

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\alpha}}_{ridge}) =&(\mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^T)^{-1}(\mathbf{A} - 2\mathbf{C}\mathbf{B}^{-1}\mathbf{C}^T + \mathbf{C}\mathbf{B}^{-1}\mathbf{E}\mathbf{B}^{-1}\mathbf{C}^T)\cdot\\ &\cdot(\mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^T)^{-1}.\end{aligned}$$

Poiché $\mathbf{B} = \mathbf{E} + \lambda\mathbf{I}_n$ per costruzione, è evidente che la varianza dello stimatore dipenda dal parametro di regolazione, e che quando $\lambda = 0$ la varianza dello stimatore ridge coincida con quella dello stimatore ai minimi quadrati.

Modello Lineare Generalizzato con regolarizzazione L_2

Un approccio analogo si può adottare per il modello lineare generalizzato; modificando la verosimiglianza con una penalità di tipo ridge (basata sulla norma L_2) è possibile ottenere stime compresse dei parametri. Sia $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ il parametro, di cui si vuole regolarizzare $\boldsymbol{\beta}$. Sia $l(\cdot)$ la funzione di log-verosimiglianza, $l_{\boldsymbol{\theta}}(\cdot)$ la funzione punteggio e $i(\cdot)$ la matrice di informazione attesa di $l(\cdot)$. La funzione obiettivo, da minimizzare per ottenere le stime, è

$$T(\boldsymbol{\theta}) = -l(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \frac{\lambda}{2} \sum_{i=1}^{n} \beta_i^2. \tag{2.1}$$

Nel caso di verosimiglianze di famiglie esponenziali, come nel caso di `GLM`, la funzione obiettivo è convessa e differenziabile, quindi per la sua ottimizzazione sono utili le quantità

$$T_{\boldsymbol{\theta}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -l_{\boldsymbol{\theta}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda \begin{pmatrix} \mathbf{0}_J \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix},$$

$$T_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = i(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda \begin{pmatrix} \mathbf{0}_{J\times J} & \mathbf{0}_{J\times n} \\ \mathbf{0}_{n\times J} & \mathbf{I}_n \end{pmatrix}.$$

Utilizzando queste, è possibile sviluppare un algoritmo iterativo di stima, molto simile all'algoritmo di *scoring di Fisher*. Una sua spiegazione è presente in Park (2006, Paragrafo 2.3). Partendo dalla approssimazione lineare di $T_{\boldsymbol{\theta}}(\cdot)$ nel punto $\hat{\boldsymbol{\theta}}^{(0)}$

$$T_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \approx T_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}^{(0)}\right) + T_{\boldsymbol{\theta}\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}^{(0)}\right)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(0)}\right),$$

e risolvendo rispetto a $\boldsymbol{\theta}$ si ottiene

$$\hat{\boldsymbol{\theta}}^{(1)} = \hat{\boldsymbol{\theta}}^{(0)} - T_{\boldsymbol{\theta}\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}^{(0)}\right)^{-1} T_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}^{(0)}\right),$$

il primo passo dell'algoritmo. Il passo t-esimo della procedura è

$$\hat{\boldsymbol{\theta}}^{(t)} = \hat{\boldsymbol{\theta}}^{(t-1)} - T_{\boldsymbol{\theta}\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}^{(t-1)}\right)^{-1} T_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}^{(t-1)}\right).$$

Siano $\tilde{\mathbf{W}}^{(t)}$ la matrice definita nel Paragrafo 1.4.3 e $\mathbf{\Lambda}$ la matrice $\lambda \begin{pmatrix} \mathbf{0}_{J\times J} & \mathbf{0}_{J\times n} \\ \mathbf{0}_{n\times J} & \mathbf{I}_n \end{pmatrix}$, $\mathbf{z}^{(t)} = \mathbf{X}\boldsymbol{\theta}^{(t)} + \tilde{\mathbf{W}^{(t)}}^{-1}(\mathbf{y} - \mathbf{p}^{(\mathbf{t})})$ con $\mathbf{p}^{(\mathbf{t})}$ vettore delle stime correnti delle medie di $\mathbf{y}$; il passo t-esimo può essere riscritto come

$$\hat{\boldsymbol{\theta}}^{(t)} = (\mathbf{X}^T\tilde{\mathbf{W}}^{(t-1)}\mathbf{X} + \mathbf{\Lambda})^{-1}\mathbf{X}^T\tilde{\mathbf{W}}^{(t-1)}\mathbf{z}^{(t-1)}.$$

Ogni iterazione coincide con una regressione ridge, motivo per cui questo algoritmo ha acronimo *IRRR*, *iteratively reweighted ridge regressions*. Quando l'algoritmo raggiunge la convergenza si ottengono le stime $\hat{\boldsymbol{\theta}}_{ridge} = (\hat{\boldsymbol{\alpha}}_{ridge}, \hat{\boldsymbol{\beta}}_{ridge})$. Per le proprietà della funzione punteggio, vale

$$\mathbb{E}\left\{T_{\boldsymbol{\theta}}(\boldsymbol{\alpha}, \boldsymbol{\beta})\right\} = \lambda \begin{pmatrix} \mathbf{0}_J \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}.$$

Le quantità $\tilde{\mathbf{W}}$, $\mathbf{p}$ e $\mathbf{z}$, ricavate dall'ultimo passo dell'algoritmo, possono essere sfruttate per descrivere alcune caratteristiche di $\hat{\boldsymbol{\theta}}_{ridge}$. Utilizzando l'approssimazione lineare della funzione $T_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ in $\hat{\boldsymbol{\theta}}_{ridge}$, si può ottenere la seguente forma per la distorsione:

$$\begin{aligned} \text{bias}(\hat{\boldsymbol{\theta}}_{ridge}) &\approx -\mathbb{E}\left\{T_{\boldsymbol{\theta}\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_{ridge}\right)^{-1} T_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_{ridge}\right)\right\} \\ &\approx -\lambda(\mathbf{X}^T\tilde{\mathbf{W}}\mathbf{X} + \mathbf{\Lambda})^{-1} \begin{pmatrix} \mathbf{0}_J \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}. \end{aligned}$$

Considerando il passo finale dell'algoritmo come una semplice regressione

ridge, si può ricavare

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\theta}}_{ridge}) &= \mathrm{Var}((\mathbf{X}^T\tilde{\mathbf{W}}\mathbf{X}+\boldsymbol{\Lambda})^{-1}\mathbf{X}^T\mathbf{z}) \\ &= (\mathbf{X}^T\tilde{\mathbf{W}}\mathbf{X}+\boldsymbol{\Lambda})^{-1}\mathrm{Var}(\tilde{\mathbf{W}}^{-1}(\mathbf{y}-\mathbf{p}))(\mathbf{X}^T\tilde{\mathbf{W}}\mathbf{X}+\boldsymbol{\Lambda})^{-1} \\ &= (\mathbf{X}^T\tilde{\mathbf{W}}\mathbf{X}+\boldsymbol{\Lambda})^{-1}\mathbf{X}^T\tilde{\mathbf{W}}\mathbf{X}(\mathbf{X}^T\tilde{\mathbf{W}}\mathbf{X}+\boldsymbol{\Lambda})^{-1}.\end{aligned}$$

2.4.2 Il Lasso

Il *lasso* (Tibshirani, 1996), *least absolute shrinkage and selection operator*, è un metodo di regolarizzazione molto apprezzato per le sue buone proprietà; infatti, oltre a comprimere le stime dei parametri verso lo zero, esso porta naturalmente alcune di esse esattamente a zero, effettuando di fatto una selezione delle variabili. Esistono numerose versioni di questo operatore, e una sua variante specifica è di particolare interesse per lo scopo di questa tesi; per una sua spiegazione occorre inizialmente presentare il lasso nella sua forma base.

Il lasso, in maniera non dissimile dalla regressione ridge, si basa sulla risoluzione del problema dei minimi quadrati ponendo un vincolo sui parametri. Supponendo che la variabile risposta abbia media pari a zero, la stima lasso è tale da minimizzare

$$\begin{gathered}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \\ \text{soggetto a } \sum_{j=1}^{p}|\beta_j| \leq t.\end{gathered}$$

Sfruttando i moltiplicatori di Lagrange, la funzione obiettivo è

$$T_{lm}(\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|.$$

Nel caso in cui si voglia regolarizzare un `GLM` è possibile utilizzare una penalità

di tipo lasso, basata sulla norma L_1. In questo caso la funzione obiettivo è

$$T_{glm}(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Il parametro di regolazione λ ha la stessa interpretazione che nella regressione ridge: più esso è grande più le stime vengono compresse, più è vicino a zero più le stime sono vicine a quelle usuali.

Gli stimatori

$$\hat{\boldsymbol{\beta}}^{lm}_{lasso} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, T_{lm}(\boldsymbol{\beta}),\ \hat{\boldsymbol{\beta}}^{glm}_{lasso} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, T_{glm}(\boldsymbol{\beta})$$

non hanno forma esplicita, a causa del tipo di penalizzazione adottata; infatti, sebbene le funzioni obiettivo siano convesse, esse non sono differenziabili, poiché la derivata del valore assoluto non esiste in zero. La minimizzazione della funzione obiettivo è un problema di programmazione quadratica, e quindi computazionalmente oneroso. Negli anni sono state proposte alcune soluzioni che permettono di ottenere le stime in maniera efficiente, con algoritmi dalla complessità paragonabile a quella della regressione ridge; nel caso del modello lineare, l'algoritmo più utilizzato è una variante del *lar* (*least angle regression*), mentre nel caso di un `GLM` è il *cyclical coordinate descent* (per entrambi gli algoritmi di veda Friedman *et al.*, 2001). Questi algoritmi, usando opportuni accorgimenti, permettono di raggiungere la soluzione esatta del problema di ottimizzazione.

La differenza fondamentale fra la regressione lasso e quella ridge è che la prima effettua anche una selezione delle variabili; un'interpretazione geometrica del processo di stima può essere d'aiuto alla comprensione del perché ciò accada. In Figura 2.1 viene rappresentato il caso in cui siano presenti solo due covariate. Le curve di livello sono quelle della log-verosimiglianza, e i punti denotati come $\hat{\boldsymbol{\beta}}$ sono le stime di massima verosimiglianza; le forme evidenziate specificano i vincoli imposti quando si utilizza il lasso (a sinistra) e la ridge (destra): la soluzione regolarizzata deve soddisfare i vincoli posti e essere entro queste figure. Poiché il lasso ha una penalità basata sul valore

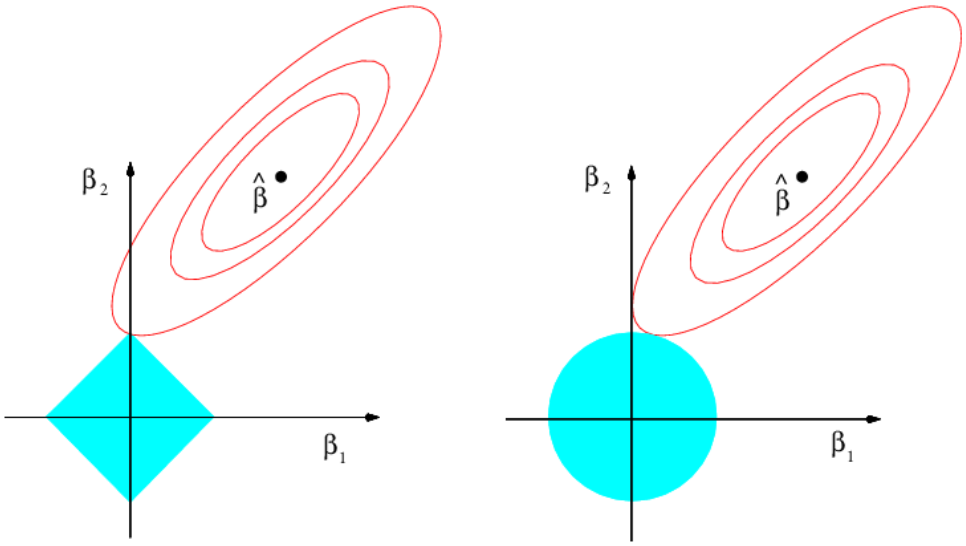

Figura 2.1: Interpretazione geometrica di lasso e ridge. Il grafico proviene da James *et al.* (2013).

assoluto, l'area entro cui devono essere le stime è a diamante; per questo, è facile che il punto di ottimo coincida con un vertice di essa, e che quindi alcune delle stime siano esattamente zero. Per la regressione ridge, invece, è molto meno facile che alcune siano pari a zero, perché la sua penalizzazione porta ad un'area entro cui devono essere le stime circolare.

2.4.3 Interpretazione Bayesiana di Regressione Ridge e Lasso

I due metodi di regolarizzazione presentati hanno un'interpretazione bayesiana. Si ipotizza un modello senza intercetta, in cui tutte le stime dei parametri vengono compresse. Le stime ottenute con i metodi esposti per un tale modello sono equivalenti a quelle ottenibili da un opportuno modello bayesiano stimando le mode a posteriori dei parametri. Il modello bayesiano equivalente è composto dalle seguenti quantità:

- $\pi(\boldsymbol{\beta})$, la distribuzione a priori dei parametri;

- $L(\boldsymbol{\beta}; \mathbf{y}) = p_{\mathbf{Y}|\boldsymbol{\beta}}(\mathbf{y}|\boldsymbol{\beta})$, la funzione di verosimiglianza.

Esse permettono di ottenere la distribuzione a posteriori di $\boldsymbol{\beta}$, sintesi dell'informazione a priori e sperimentale sul parametro. Essa è

$$\pi(\boldsymbol{\beta}|\mathbf{y}) \propto L(\boldsymbol{\beta}; \mathbf{y})\pi(\boldsymbol{\beta}).$$

Per semplificare alcune operazioni, viene spesso utilizzata la log-posteriori

$$\log(\pi(\boldsymbol{\beta}|\mathbf{y})) = l(\boldsymbol{\beta}; \mathbf{y}) + \log(\pi(\boldsymbol{\beta})) + c,$$

dove c è una costante non dipendente dai parametri.

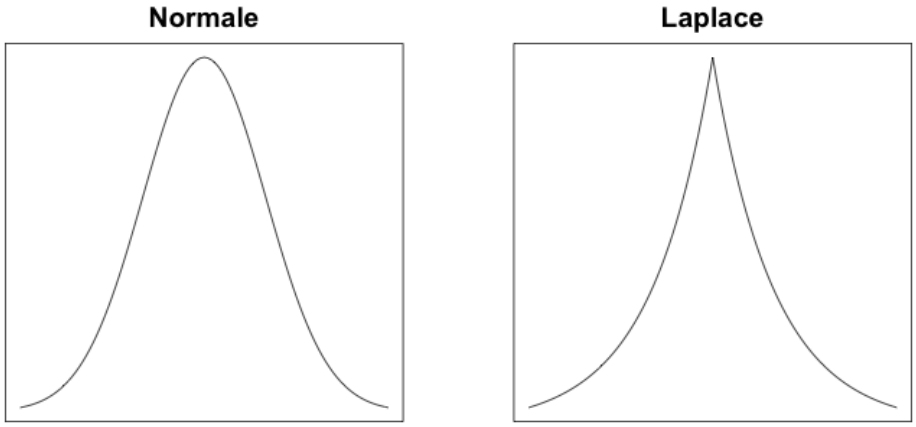

Figura 2.2: Densità normale e Laplace.

Nella regressione ridge si pone come distribuzione a priori per i parametri $\boldsymbol{\beta} \sim N_p(0, \frac{1}{\lambda}\mathbf{I}_p)$, per cui $\pi(\boldsymbol{\beta}) = \frac{\lambda}{\sqrt{(2\pi)^p|\mathbf{I}_p|}} \exp\left(-\frac{\lambda}{2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right)$. Un grafico della densità della distribuzione normale è riportato in Figura 2.2. La log-posteriori risulta essere

$$\log(\pi(\boldsymbol{\beta}|\mathbf{y})) = l(\boldsymbol{\beta}; \mathbf{y}) - \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2 + c,$$

che coincide con la funzione obiettivo cambiata di segno a cui viene aggiunta la costante c. Massimizzare la log-posteriori rispetto a $\boldsymbol{\beta}$ permette di trovare le mode a posteriori dei parametri, coincidenti con le stime ridge.

Nel lasso bayesiano, ogni parametro ha distribuzione a priori Laplace, $\beta_j \sim \text{Laplace}(0, \lambda)$ e $\pi(\beta_j) = \frac{\lambda}{2} \exp(-\lambda|\beta_j|)$. Un grafico della densità della distribuzione di Laplace è rappresentato in Figura 2.2. La priori di $\boldsymbol{\beta}$ è quindi $\pi(\boldsymbol{\beta}) = \prod_{j=1}^{p} \pi(\beta_j) = \left(\frac{\lambda}{2}\right)^p \exp\left(-\lambda \sum_{j=1}^{p} |\beta_j|\right)$. La log-posteriori dei parametri è

$$\log(\pi(\boldsymbol{\beta}|\mathbf{y})) = l(\boldsymbol{\beta}; \mathbf{y}) - \lambda \sum_{j=1}^{p} |\beta_j| + c,$$

di nuovo coincidente con la funzione obiettivo cambiata di segno a cui viene sommata la costante c, per cui le mode a posteriori sono uguali alle stime lasso.

Sia nella regressione ridge che nel lasso il parametro di regolazione λ è l'inverso del parametro di scala della distribuzione a priori dei parametri; poiché in entrambi i casi la varianza delle distribuzioni a priori è una funzione monotona del parametro di scala, il parametro di regolazione è legato alla varianza dei parametri a priori. In particolare, se $\lambda \to 0$, la varianza delle a priori va a infinito (a priori non informativa), se invece $\lambda \to \infty$ la varianza delle a priori va a zero (a priori molto informativa).

L'interpretazione bayesiana proposta rivela quanto il modello con regolarizzazione degli effetti fissi sia simile al modello ad effetti misti. Le differenze fondamentali sono due: innanzitutto, nel primo modello la distribuzione a priori ha il solo compito di comprimere i parametri verso lo zero, mentre nel secondo essa descrive come il fenomeno oggetto dello studio funzioni; poi, nel modello regolarizzato, il parametro λ è fisso oppure stimato in modo da minimizzare l'errore di previsione, mentre nel modello ad effetti misti viene stimato dai dati tramite la verosimiglianza.

2.4.4 Il Generalized Fused Lasso

Spesso, in un'analisi statistica, il problema da studiare richiede di raggruppare, secondo un criterio, le osservazioni. Solitamente, in questi casi si fa affidamento sui metodi di apprendimento non supervisionato. Una modifica nella penalizzazione del lasso permette di effettuare un raggruppamento

delle categorie di una variabile categoriale in base alla relazione presente fra le osservazioni e una variabile di interesse, in maniera supervisionata; nel caso in cui questa variabile identifichi i soggetti parte dello studio, questo metodo risulta in un raggruppamento delle unità.

Il *fused lasso* (Tibshirani *et al.*, 2005), nel caso la variabile categoriale sia ordinale, permette di raggrupparne le modalità. Oltre a comprimere le stime dei parametri, esso comprime anche le differenze fra parametri di modalità adiacenti della variabile categoriale, in modo tale che una parte di esse vengano portate a zero, rendendo le stime dei relativi parametri esattamente uguali. Sia $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ il vettore dei parametri, di cui $\boldsymbol{\beta}$ i k parametri relativi alla variabile categoriale ordinale le cui categorie si vogliono raggruppare. La stima dei parametri si ottiene da

$$\hat{\boldsymbol{\theta}}_{fl} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left(-l(\boldsymbol{\theta}) + \lambda_1 \sum_{j=1}^{k} |\beta_j| + \lambda_2 \sum_{j=2}^{k} |\beta_j - \beta_{j-1}| \right).$$

Questa tipologia di lasso può essere generalizzata per funzionare anche per variabili categoriali non ordinali (Bondell e Reich, 2009). Questo metodo può essere molto utile, oltre che per quando si desidera fare un'analisi di raggruppamento, quando si ha a disposizione una esplicativa qualitativa dalle numerose modalità, alcune delle quali poco rappresentate nel campione in esame. La funzione di perdita del *generalized fused lasso* è

$$T(\boldsymbol{\theta}) = -l(\boldsymbol{\theta}) + P_\lambda(\boldsymbol{\theta}) = -l(\boldsymbol{\theta}) + \lambda \sum_{r>m} |\beta_r - \beta_m|,$$

dove $\boldsymbol{\beta}$ è il vettore dei parametri relativi alla variabile categoriale, e $P_\lambda(\boldsymbol{\theta})$ è detta funzione di penalizzazione. Un modello di questo tipo è utilizzato in She (2010) per l'analisi di microarray e in Masarotto e Varin (2012) per effettuare delle classifiche sportive. La penalizzazione consiste nella somma dei valori assoluti di tutte le possibili differenze fra i parametri. Di conseguenza, ognuno dei k parametri compare $k-1$ volte in $P_\lambda(\boldsymbol{\theta})$, poiché viene confrontato con $k-1$ parametri. Il numero di termini nella penalizzazione è quindi $\frac{k(k-1)}{2}$, la somma dei primi $k-1$ numeri naturali. Infatti, il primo parametro è oggetto

di $k-1$ confronti unici, il secondo di $k-2$, perché il suo confronto col primo parametro è già stato contato, il terzo con $k-3$, e così via fino all'ultimo parametro, che non aggiunge nuovi termini di confronto poiché esso compare già nei confronti dei $k-1$ parametri precedenti. I termini oggetto dei valori assoluti della penalizzazione possono essere scritti in forma vettoriale come $\mathbf{A}\boldsymbol{\theta}$. Ad esempio, nel caso $k=4$ e ipotizzando $\boldsymbol{\theta}=\boldsymbol{\beta}$

$$\mathbf{A}\boldsymbol{\beta}=\begin{pmatrix}-1 & 1 & 0 & 0\\ -1 & 0 & 1 & 0\\ -1 & 0 & 0 & 1\\ 0 & -1 & 1 & 0\\ 0 & -1 & 0 & 1\\ 0 & 0 & -1 & 1\end{pmatrix}\boldsymbol{\beta}=\begin{pmatrix}\beta_2-\beta_1\\ \beta_3-\beta_1\\ \beta_4-\beta_1\\ \beta_3-\beta_2\\ \beta_4-\beta_2\\ \beta_4-\beta_3\end{pmatrix}.$$

Siano L il numero di righe della matrice $\mathbf{A}$ (ovvero il numero di termini nella penalizzazione) e $\mathbf{a}_l$ la riga l-esima di $\mathbf{A}$, la penalizzazione può essere scritta come

$$P_\lambda(\boldsymbol{\theta})=\sum_{l=1}^{L}\lambda|\mathbf{a}_l^T\boldsymbol{\theta}|.$$

Nel caso in cui una modalità della variabile categoriale sia spiegata da una intercetta, i parametri $\boldsymbol{\beta}$ rappresentano la differenza fra il valore di una modalità e quella di riferimento, e quindi anch'essi vanno regolarizzati. La funzione obiettivo diventa

$$T(\boldsymbol{\theta})=-l(\boldsymbol{\theta})+\lambda\sum_i|\beta_i|+\lambda\sum_{r>m}|\beta_r-\beta_m|. \tag{2.2}$$

Supponendo che la variabile categoriale abbia k modalità, $\boldsymbol{\beta}$ è un vettore di $k-1$ parametri. Il termine di penalizzazione è composto da $(k-1)+\frac{(k-1)(k-2)}{2}=\frac{k(k-1)}{2}$ elementi. Di nuovo, i termini i cui valori assoluti formano la penalizzazione possono essere scritti come $\mathbf{A}\boldsymbol{\theta}$; ad esempio, con $k=4$ e

$\boldsymbol{\theta} = (\alpha_0, \boldsymbol{\beta})$

$$\mathbf{A}\boldsymbol{\theta} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_2 - \beta_1 \\ \beta_3 - \beta_1 \\ \beta_3 - \beta_2 \end{pmatrix}.$$

Il problema di ottimizzazione di $T(\boldsymbol{\theta})$ è di difficile risoluzione. Un algoritmo che porta ad una soluzione approssimata del problema è l'algoritmo *PIRLS*, *penalized iterated reweighted least squares*; esso, attraverso un'approssimazione differenziabile della penalizzazione, permette di ottenere una soluzione approssimata. Una spiegazione nel dettaglio di questo è presente in Oelker (2015b). Come nel lasso, il problema nella stima è dato dalla presenza di diversi valori assoluti nel termine di penalizzazione; poiché questa funzione non è derivabile in zero, infatti, non è possibile utilizzare gli usuali metodi analitici per ottenere le stime. La strategia adottata in questa procedura consiste nell'utilizzare un'approssimazione differenziabile del valore assoluto. In particolare

$$|\epsilon| \approx \sqrt{\epsilon^2 + c},$$

dove c è una costante; più il suo valore è vicino a zero, migliore è l'approssimazione (in Oelker (2015b) viene suggerito di utilizzare $c = 10^{-5}$). Una rappresentazione grafica della funzione valore assoluto e della sua approssimazione è riportata in Figura 2.3. Siano, quindi,

$$\mathcal{N}_l(\epsilon) = \sqrt{\epsilon^2 + c},\ \mathcal{D}_l(\epsilon) = \mathcal{N}_l(\epsilon)' = (\epsilon^2 + c)^{-\frac{1}{2}} \cdot \epsilon.$$

L'algoritmo PIRLS si basa sulla approssimazione lineare della funzione di penalizzazione nel punto $\boldsymbol{\theta}^{(k)}$

$$P_\lambda(\boldsymbol{\theta}) \approx P_\lambda\left(\boldsymbol{\theta}^{(k)}\right) + \nabla P_\lambda\left(\boldsymbol{\theta}^{(k)}\right)^T \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}\right),$$

Figura 2.3: Approssimazione del valore assoluto; la linea tratteggiata è la funzione valore assoluto, la linea continua una sua approssimazione.

con

$$\nabla P_\lambda \left(\boldsymbol{\theta}^{(k)}\right)^T \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}\right) = \sum_{l=1}^{L} \lambda \left(\nabla \left|\mathbf{a}_l^T \boldsymbol{\theta}^{(k)}\right|\right)^T \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}\right).$$

Si può dimostrare (Oelker, 2015b, Paragrafo 2.2) che

$$\left(\nabla \left|\mathbf{a}_l^T \boldsymbol{\theta}^{(k)}\right|\right)^T \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}\right) \approx \frac{1}{2}\left(\boldsymbol{\theta}^T \mathbf{A}_l \boldsymbol{\theta} + \boldsymbol{\theta}^{(k)^T} \mathbf{A}_l \boldsymbol{\theta}^{(k)}\right),$$

dove

$$\mathbf{A}_l = \frac{\mathcal{D}_l\left(\mathbf{a}_l^T \boldsymbol{\theta}^{(k)}\right)}{\mathbf{a}_l^T \boldsymbol{\theta}^{(k)}} \mathbf{a}_l \mathbf{a}_l^T.$$

In questo modo la penalità può essere approssimata da

$$P_\lambda(\boldsymbol{\theta}) \approx P_\lambda\left(\boldsymbol{\theta}^{(k)}\right) + \frac{1}{2}\left(\boldsymbol{\theta}^T \mathbf{A}_\lambda \boldsymbol{\theta} + \boldsymbol{\theta}^{(k)^T} \mathbf{A}_\lambda \boldsymbol{\theta}^{(k)}\right),$$

dove $\mathbf{A}_\lambda = \lambda \sum_{l=1}^{L} \mathbf{A}_l$. A questo punto, è possibile ottenere un algoritmo di stima basato sullo *scoring di Fisher*. Utilizzando le funzioni

$$T(\boldsymbol{\theta}) = -l(\boldsymbol{\theta}) + P_\lambda(\boldsymbol{\theta}),\ T_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = -l_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \mathbf{A}_\lambda \boldsymbol{\theta},\ T_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}) = i(\boldsymbol{\theta}) + \mathbf{A}_\lambda,$$

in maniera del tutto analoga alla procedura adottata per i GLM, si ottiene un algoritmo iterativo il cui t-esimo passo è

$$\hat{\boldsymbol{\theta}}^{(t)} = (\mathbf{X}^T\tilde{\mathbf{W}}^{(t-1)}\mathbf{X} + \mathbf{A}_\lambda)^{-1}\mathbf{X}^T\tilde{\mathbf{W}}^{(t-1)}\mathbf{z}^{(t-1)},$$

dove le quantità $\tilde{\mathbf{W}}^{(t-1)}$ e $\mathbf{z}^{(t-1)}$ sono le medesime dello *scoring di Fisher*. L'algoritmo viene fermato quando le stime raggiungono convergenza.

2.4.5 Numero Effettivo di Parametri

Uno dei motivi per cui si decide di regolarizzare le stime dei parametri di un modello è per ridurne la complessità. Un modello troppo complesso rischia di soffrire di *sovra-adattamento*, oppure, come nei modelli oggetto di questa tesi, di non produrre stime affidabili, anche se correttamente specificato. Una misura intuitiva della complessità di un modello è il numero di parametri che lo caratterizzano; tuttavia, nel caso si utilizzino metodi di compressione dei parametri, il numero di parametri non ne indica la complessità, poiché il valore qualitativo di un parametro non regolarizzato e di uno regolarizzato è diverso, in quanto il secondo, a causa dei vincoli imposti, ha una stima in valore assoluto minore. Hastie e Tibshirani (1990), per ottenere una misura equivalente della complessità di un modello, hanno proposto il *numero effettivo di parametri* (*effective number of parameters*); questa quantità è stimabile col metodo presentato solo nel caso in cui il processo di stima sia lineare nelle osservazioni della variabile risposta, ovvero se

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}.$$

Il numero effettivo di parametri è

$$\mathrm{d}(\mathbf{H}) = \mathrm{tr}(\mathbf{H}).$$

Per modelli non regolarizzati, si può dimostrare che $\mathrm{d}(\mathbf{H}) = p$, dove p è il numero di parametri. Nel caso il modello sia un GLM regolarizzato e l'algoritmo di stima sia una variante dei minimi quadrati ponderati iterati, è possibile

utilizzare le quantità dell'ultimo passo dell'algoritmo per ottenere

$$\hat{\mathbf{z}} = \mathbf{H}\mathbf{z},$$

dove $\mathbf{z}$ è la quantità descritta nel Paragrafo 1.4.3; in questo modo è possibile ottenere una stima di $\mathrm{d}(\mathbf{H})$. Questo indice di complessità, quindi, può essere agilmente utilizzato nel caso di modelli lineari e `GLM` con penalità di tipo ridge e basata sul *generalized fused lasso* (qualora venga utilizzato l'algoritmo PIRLS).

La quantità $\mathrm{d}(\mathbf{H})$, nel caso di regolarizzazione, dipende dal parametro di regolazione λ, in modo tale che, qualora si decida a priori il numero effettivo di parametri di un modello, è possibile ricavare il valore di λ adeguato. Chiaramente, $\mathrm{d}(\mathbf{H})$ deve essere maggiore o uguale al numero di parametri non regolarizzati.

Qualora non si volesse scegliere il parametro di penalizzazione fissando $\mathrm{d}(\mathbf{H})$, è possibile utilizzare un criterio basato sui dati, il cui obiettivo è minimizzare l'errore di previsione. Questo metodo prende il nome di *convalida incrociata generalizzata*, e sfrutta delle quantità che, dato un determinato λ, approssimano l'errore stimato tramite convalida incrociata *leave-one-out* senza dover stimare più volte il modello, e quindi riducendo considerevolmente il carico computazionale. Tale quantità è

$$\mathrm{GCV}(\hat{\boldsymbol{\theta}}_\lambda) = \frac{N \ \mathrm{dev}(\hat{\boldsymbol{\theta}}_\lambda)}{(N - \mathrm{d}(\mathbf{H}_\lambda))^2},$$

dove N è il numero di osservazioni, e $\mathrm{dev}(\hat{\boldsymbol{\theta}}_\lambda)$ è la devianza del `GLM` calcolata in $\hat{\boldsymbol{\theta}}_\lambda$. La stima del parametro di regolazione è data da

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}} \, \mathrm{GCV}(\hat{\boldsymbol{\theta}}_\lambda).$$

2.4.6 Applicazione al Modello di Rasch

Il modello di Rasch, descritto nel Paragrafo 2.2.3, è un `GLM` binomiale con funzione di legame canonico. Supponendo che l'obiettivo dello studio

sia stimare adeguatamente i parametri associati alla difficoltà delle prove a cui i soggetti sono sottoposti, il metodo proposto in questa tesi consiste nel regolarizzare le stime dei parametri incidentali, lasciando libere le stime dei parametri di interesse. Modificando opportunamente di volta in volta il parametro di regolazione λ, è possibile ottenere un modello la cui complessità è approssimativamente la stessa all'aumentare della numerosità campionaria. Le penalizzazioni adottate sono due: quella di tipo ridge, per la sua interpretazione simile a quella di un modello ad effetti misti e perché, a differenza del lasso, mantiene un effetto specifico per ogni soggetto; quella basata sul *generalized fused lasso*, in quanto spesso è legittimo aspettarsi che i soggetti siano divisibili in gruppi in base alla loro abilità.

Siano $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ i parametri, con $\boldsymbol{\alpha}$ i J parametri strutturali e $\boldsymbol{\beta}$ n parametri incidentali. Se si opta per una penalità di tipo ridge le stime del modello sono

$$\hat{\boldsymbol{\theta}}_{ridge} = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} -l(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \sum_{i=1}^{n} \beta_i^2.$$

Lo stimatore dei parametri strutturali, come mostrato nel Paragrafo 2.4.1, è distorto e la sua varianza può essere ricavata analiticamente; in un contesto del genere lo stimatore di massima verosimiglianza soffre anch'esso di distorsione, e la sua varianza è elevata; le stime dei parametri incidentali sono in valore assoluto minori a quelle di massima verosimiglianza, e la loro distorsione è potenzialmente anche elevata.

Qualora la penalità fosse di tipo *generalized fused lasso*, le stime si ottengono da

$$\hat{\boldsymbol{\theta}}_{gfl} = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} -l(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda \sum_{i=1}^{n} |\beta_i| + \lambda \sum_{r>m} |\beta_r - \beta_m|,$$

nel caso in cui l'abilità del primo soggetto sia spiegata dai parametri $\boldsymbol{\alpha}$. Le stime dei parametri di interesse hanno proprietà simili a quelle di $\hat{\boldsymbol{\theta}}_{ridge}$. Le stime dei parametri relativi ai soggetti, in caso di reale presenza di gruppi di soggetti, hanno un'interpretazione diversa; anche in questo caso, se il parametro di regolazione non è scelto adeguatamente, lo stimatore è molto

distorto.

Nel prossimo capitolo si indagano, attraverso uno studio di simulazione, le proprietà di questi metodi di stima; in particolare, si confrontano distorsione, varianza ed errore quadratico medio di questi stimatori con quello di massima verosimiglianza al variare della numerosità campionaria.

Capitolo 3

Studi di simulazione

3.1 Introduzione

Lo scopo di questo capitolo è indagare, attraverso degli studi di simulazione, le proprietà frequentiste degli stimatori presentati nel Paragrafo 2.4. In particolare, nell'ambito del problema dei parametri incidentali, si vogliono confrontare distorsione, varianza ed errore quadratico medio dello stimatore basato sul modello regolarizzato e dello stimatore di massima verosimiglianza. Il modello di riferimento è un modello la cui variabile risposta in un caso è continua e di distribuzione normale, in un altro è binomiale; esso è composto di due esplicative categoriali: gli effetti di una di esse sono di interesse, dell'altra sono di disturbo. La struttura del problema impone che, all'aumentare della dimensione campionaria, aumentino anche le categorie della seconda esplicativa, in modo che sia presente il problema di Neyman e Scott. Il modello così descritto coincide con il modello di Rasch nel caso di variabile risposta binomiale; la sua variante con risposta normale serve come introduzione al problema, per la sua semplicità. Si ipotizza non ci siano dati mancanti, per cui si ha un'osservazione per ogni coppia di modalità delle due variabili categoriali (nel caso del modello di Rasch, ogni soggetto in esame prende parte a tutte le prove). In una situazione di questo tipo, ci si aspetta che lo stimatore di massima verosimiglianza abbia distorsione non decrescente all'aumentare del numero di soggetti, e che quindi non sia

consistente. L'interesse primario di questo studio è osservare come sia la distorsione dello stimatore regolarizzato rispetto a quella dello `SMV` e come si comporti la sua varianza all'aumentare della numerosità campionaria. L'idea di base consiste nell'utilizzare dei parametri di regolazione diversi per ogni numerosità campionaria, in modo tale che il numero effettivo di parametri sia approssimativamente uguale. I metodi di stima adottati sono quelli basati sulla regressione ridge e sul *generalized fused lasso*, sia per la risposta normale che per quella binomiale.

3.2 Aspetti Computazionali

I problemi di stima presenti in questa tesi potrebbero essere risolti tutti utilizzando il pacchetto `R gvcm.cat` (Oelker, 2015a). Al fine di avere un maggiore controllo, tutti gli algoritmi sono stati re-implementati in `R`. Questo approccio ha permesso di utilizzare in maniera più parsimoniosa le risorse computazionali, riducendo l'utilizzo della memoria e il numero di operazioni svolte.

La complessità degli algoritmi di stima dei due modelli regolarizzati è paragonabile a quella dello *scoring di Fisher*. Nel caso dell'algoritmo PIRLS, tuttavia, il calcolo della matrice $\mathbf{A}_\lambda$ viene aggiornato ad ogni iterazione, e questa operazione rallenta i tempi di esecuzione. Questo algoritmo, inoltre, richiede il calcolo della matrice $\mathbf{A}$, la cui dimensione può essere elevata (si veda il Paragrafo 2.4.4). Nello studio di simulazione, presentato nel resto del capitolo, sono stati stimati anche modelli composti da più di settecento parametri; per questo, l'operazione che in tutti gli algoritmi ha richiesto più risorse è l'inversione della matrice $(\mathbf{X}^T\tilde{\mathbf{W}}\mathbf{X} + \mathbf{\Lambda})^{-1}$, dove $\mathbf{\Lambda}$ è una penalizzazione generica. Più grande è il numero di soggetti, più tempo è richiesto per la stima del modello.

Al fine di svolgere le simulazioni in maniera efficiente, è stato fatto uso del calcolo parallelo grazie alla libreria `R snowfall` (Knaus, 2013).

IRRR	100	300	500	700
PIRLS	50	100	200	300

Tabella 3.1: Numerosità campionarie adottate per ogni algoritmo.

3.3 Struttura dello Studio di Simulazione

Le funzioni obiettivo sono quelle riportate in (2.1) e (2.2). A causa delle diversità nei tempi di calcolo degli algoritmi IRRR e PIRLS, per le due penalizzazioni sono state scelte alcune caratteristiche uniche nella simulazione. Per il primo algoritmo, distorsione e varianza per ogni diversa numerosità campionaria sono state calcolate su centocinquanta repliche, mentre per il secondo solamente su cento. Le diverse numerosità campionarie per le procedure IRRR e PIRLS sono riportate in Tabella 3.1.

A parte queste differenze, per entrambi i tipi di penalizzazione sono state usate le stesse caratteristiche, sia per il modello lineare che per il modello lineare generalizzato. Per ogni coppia di tipologia di modello e di regolarizzazione sono state svolte due simulazioni: nella prima i soggetti vengono sottoposti a cinque prove ($\boldsymbol{\alpha} = [-2, -1,\ 0,\ 1,\ 2]^T$), nella seconda a dieci ($\boldsymbol{\alpha} = [-4, -3, -2, -1,\ 0,\ 1,\ 2,\ 3,\ 4,\ 5]^T$). I parametri $\boldsymbol{\beta}$ vengono generati in maniera pseudo-casuale da una normale di media zero e varianza uno; all'interno della medesima simulazione questi parametri sono da considerarsi fissati. L'abilità del primo soggetto, spiegata dai parametri $\boldsymbol{\alpha}$, viene imposta essere pari a zero, in modo da poter facilmente studiare la distorsione delle stime di $\boldsymbol{\alpha}$. Nel caso del modello con risposta normale, la varianza degli errori è pari a uno.

Per i due tipi di penalizzazione vengono presi in esame il modello lineare e `GLM` binomiale; per ognuno di essi si studia il caso in cui la dimensione di $\boldsymbol{\alpha}$ sia pari a cinque e quello in cui sia pari a dieci; in totale, quindi, vengono presentati un totale di otto casi in cui vengono studiate varianza e distorsione al variare della numerosità campionaria. Utilizzando i valori calcolati per varianza e distorsione, poi, vengono stimati gli errori quadratici medi degli stimatori. Per ogni tipo di modello e per ogni numerosità campionaria

il parametro di regolazione è stato scelto in modo che il numero effettivo di parametri sia approssimativamente pari a venti. Se, nel caso del modello lineare ridge, è possibile trovare in maniera molto precisa il valore adatto di λ, per modelli più complessi l'approssimazione è più grande. L'approccio più conveniente è risultato essere quello di provare diversi valori su dati simulati, fino ad ottenere un numero effettivo di parametri abbastanza vicino all'obiettivo.

3.4 Risultati

I risultati delle simulazioni vengono riportati attraverso dei grafici e delle tabelle. Per ogni tipologia di penalizzazione, di distribuzione della variabile risposta e di dimensione del parametro $\boldsymbol{\alpha}$, vengono riportati due grafici. In ognuno di essi vengono rappresentate, rispettivamente, distorsione e varianza dello stimatore dei parametri strutturali $\boldsymbol{\alpha}$, messe in relazione con la numerosità campionaria. Le stime di queste quantità per i modelli con regolarizzazione e senza regolarizzazione vengono rappresentate con punti di colore e forma diversi. In seguito, vengono riportate in forma tabellare le stime dell'errore quadratico medio degli stimatori dei parametri strutturali per ognuno dei modelli in esame.

Dalla Figura 3.1 alla Figura 3.8 sono riportati i grafici riguardo alla penalità di tipo ridge. Si nota immediatamente che la distorsione è paragonabile a quella dello `SMV` e che la varianza è nettamente inferiore. Nel modello lineare, la distorsione dello stimatore è più bassa quando il numero dei parametri di interesse è più alto; a dimostrare che la distorsione nel modello regolarizzato ha una natura diversa rispetto a quella dello `SMV`, in Figura 3.1 esse hanno segni opposti, pur avendo grandezza comparabile. Non è evidente una dipendenza della distorsione dal numero di soggetti, mentre pare evidente che, in ogni caso, aumentare la dimensione di $\boldsymbol{\alpha}$ corrisponda ad una diminuzione nel valore assoluto della distorsione. Il comportamento della varianza dello stimatore al variare della numerosità campionaria, invece, è chiaro: quella dello `SMV` è all'incirca costante, mentre quella dello stimatore ridge ha un andamento decrescente. Questo effetto è ciò che si cercava imponendo la stessa

complessità ai diversi modelli: sembra che l'utilizzo di una regolarizzazione possa smorzare gli effetti spiacevoli dovuti alla presenza dei parametri incidentali. Nel `GLM` le differenze fra gli stimatori sono ancora più evidenti. La distorsione è paragonabile per il modello regolarizzato e non, anche se spesso lo stimatore regolarizzato ha distorsione inferiore. La differenza in varianza, invece, è palese. Infatti, lo `SMV` ha varianza anche molto elevata, mentre quella del modello con penalità ridge, oltre ad avere l'andamento decrescente già descritto, è nettamente inferiore.

Nelle Figure dalla 3.9 alla 3.16 sono rappresentati i risultati relativi ai modelli con penalità *generalized fused lasso*. Per come è stato impostato lo studio di simulazione, questo tipo di penalizzazione è svantaggiata, in quanto i modelli stimati dividono i soggetti in gruppi in base alla loro similarità, non in base ad una reale struttura. Nel modello lineare, la distorsione è spesso più grande per il modello regolarizzato, sia con cinque che con dieci prove. La varianza ha un comportamento analogo a quello osservato nei modelli con penalità ridge: decresce con l'aumentare della numerosità campionaria ed è sempre inferiore a quella dello stimatore di massima verosimiglianza. Nel `GLM` la distorsione del modello regolarizzato è spesso inferiore a quella dello `SMV`. Si noti che la distorsione può essere anche molto elevata: ciò è dovuto al valore di alcuni parametri $\boldsymbol{\alpha}$. Infatti, quando questi sono in valore assoluto elevati, può accadere che nessuno dei soggetti superi (o fallisca) una determinata prova. In questo caso le vere stime di massima verosimiglianza sono infinite. La varianza dello stimatore regolarizzato è di nuovo decisamente inferiore a quella dello stimatore di massima verosimiglianza, e mostra di nuovo l'andamento decrescente desiderato.

Le Tabelle 3.2, 3.3 e 3.4 riportano le stime degli errori quadratici medi degli stimatori dei parametri strutturali al variare della numerosità campionaria. In ogni caso, l'errore dello stimatore regolarizzato è inferiore a quello dello stimatore di massima verosimiglianza. In alcuni casi, la differenza può raggiungere i tre ordini di grandezza. Pare evidente, quindi, che una stima puntuale basata su un modello con opportuna regolarizzazione, possa essere anche molto più precisa di quella basata sulla massima verosimiglianza.

Le simulazioni, quindi, mostrano che utilizzare un modello con regolariz-

zazione come quelli presentati permette di abbassare anche notevolmente la varianza dello stimatore. Per quanto riguarda la distorsione, lo studio non è conclusivo, in quanto si nota che a volte essa è inferiore a quella dello `SMV` mentre altre volte è maggiore. In termini di errore quadratico medio, tutti i casi presi in considerazione mostrano un miglioramento, anche considerevole, nel momento in cui venga adottato un modello con regolarizzazione. Una stima intervallare effettuata senza tenere conto della distorsione nel caso dello stimatore regolarizzato, tuttavia, porterebbe a intervalli dal livello di copertura empirico più basso di quello dello `SMV`.

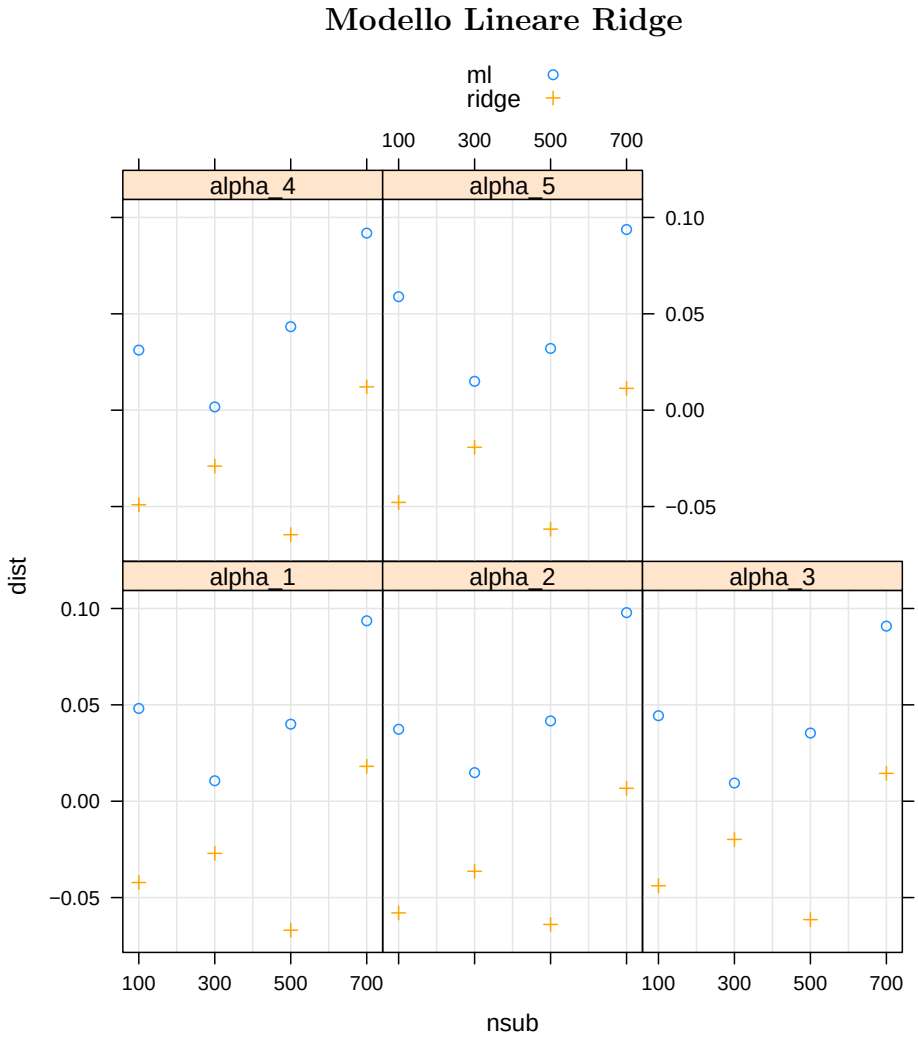

Figura 3.1: Distorsione, cinque prove.

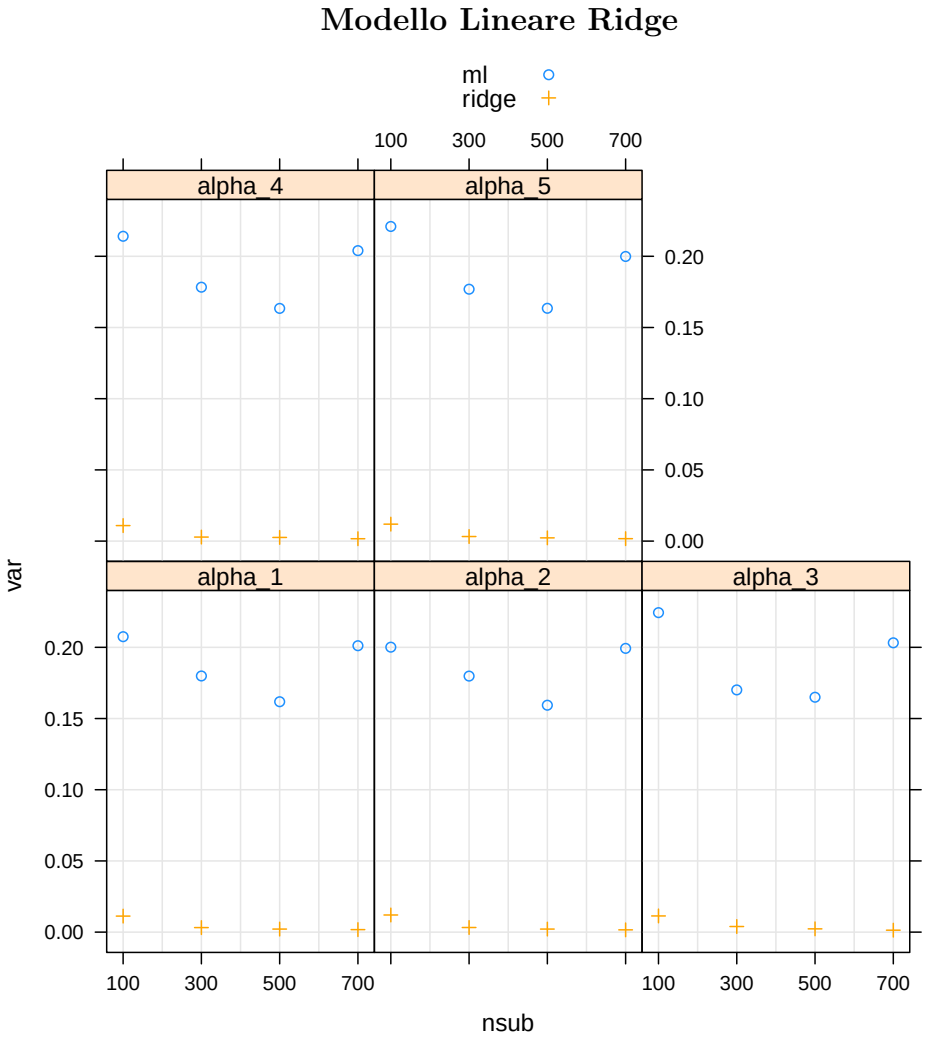

Figura 3.2: Varianza, cinque prove.

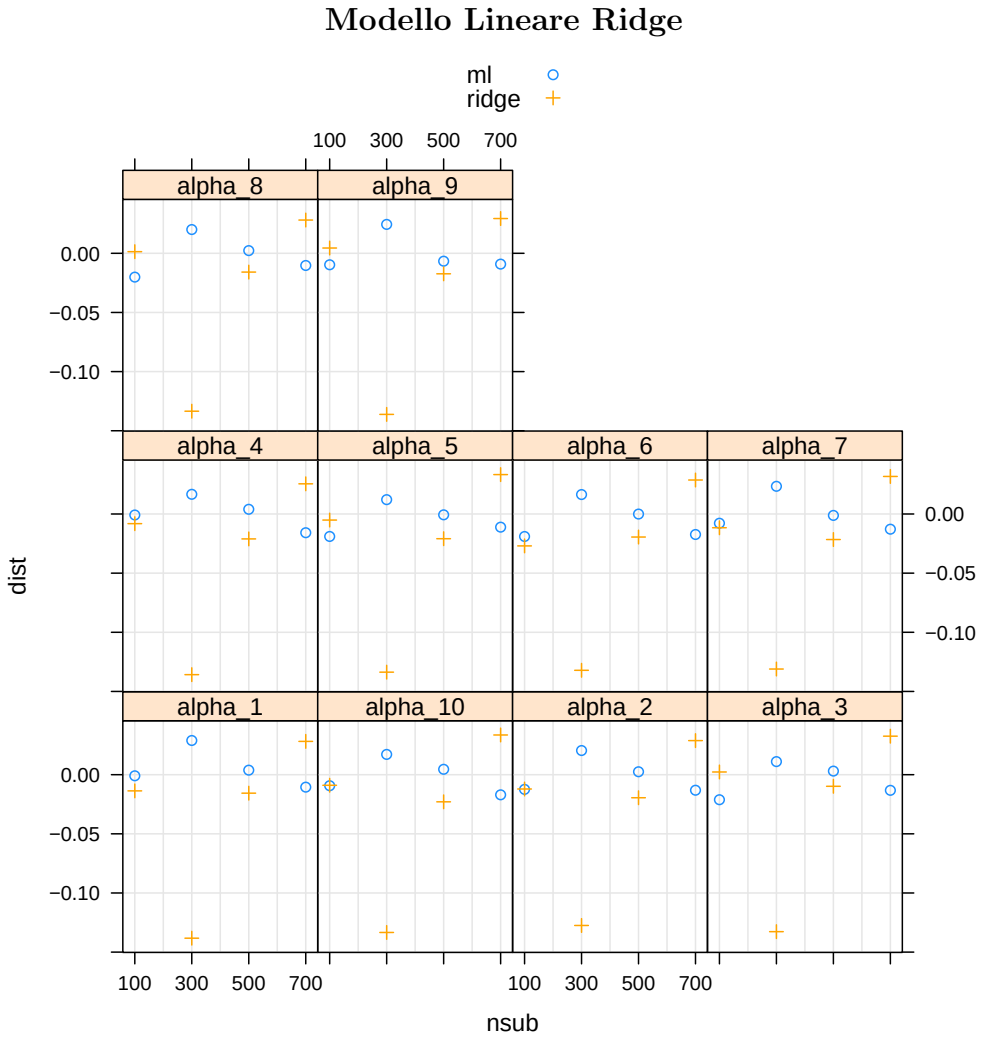

Figura 3.3: Distorsione, dieci prove.

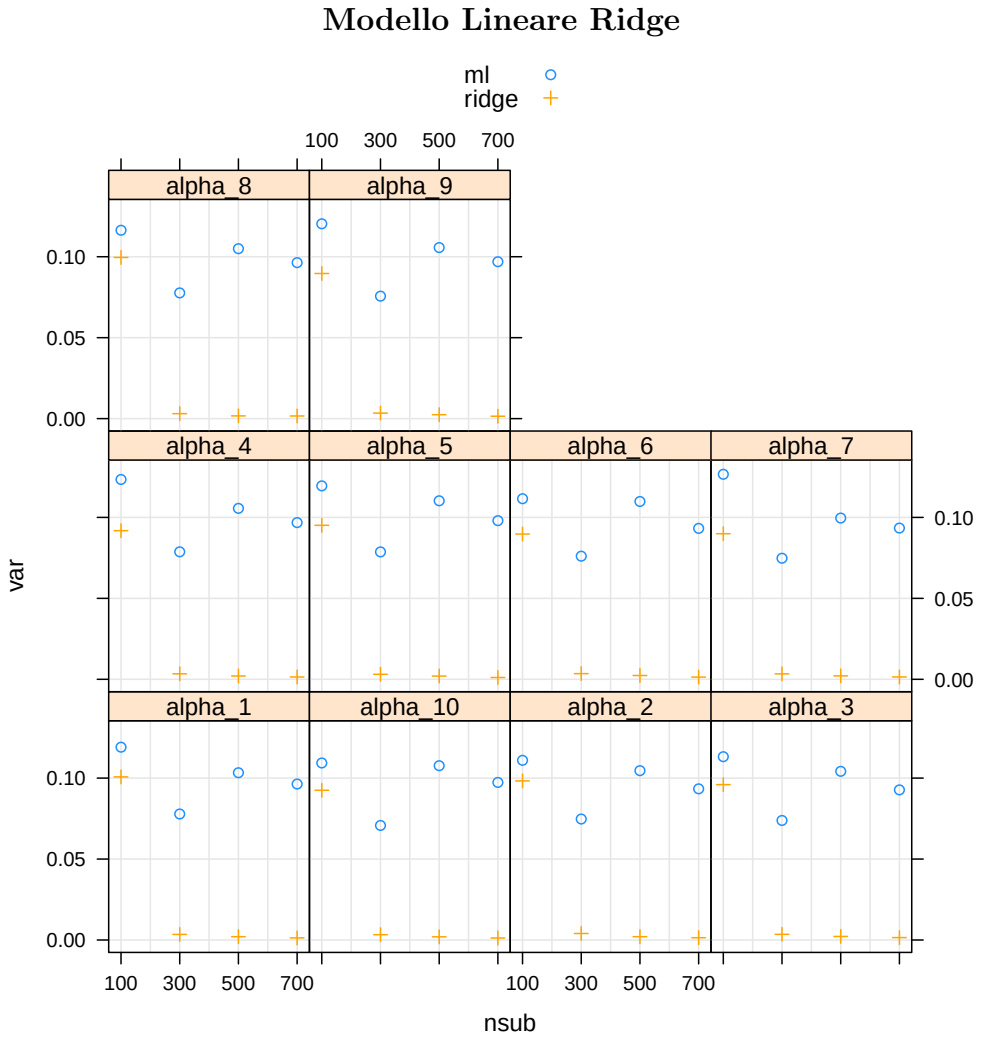

Figura 3.4: Varianza, dieci prove.

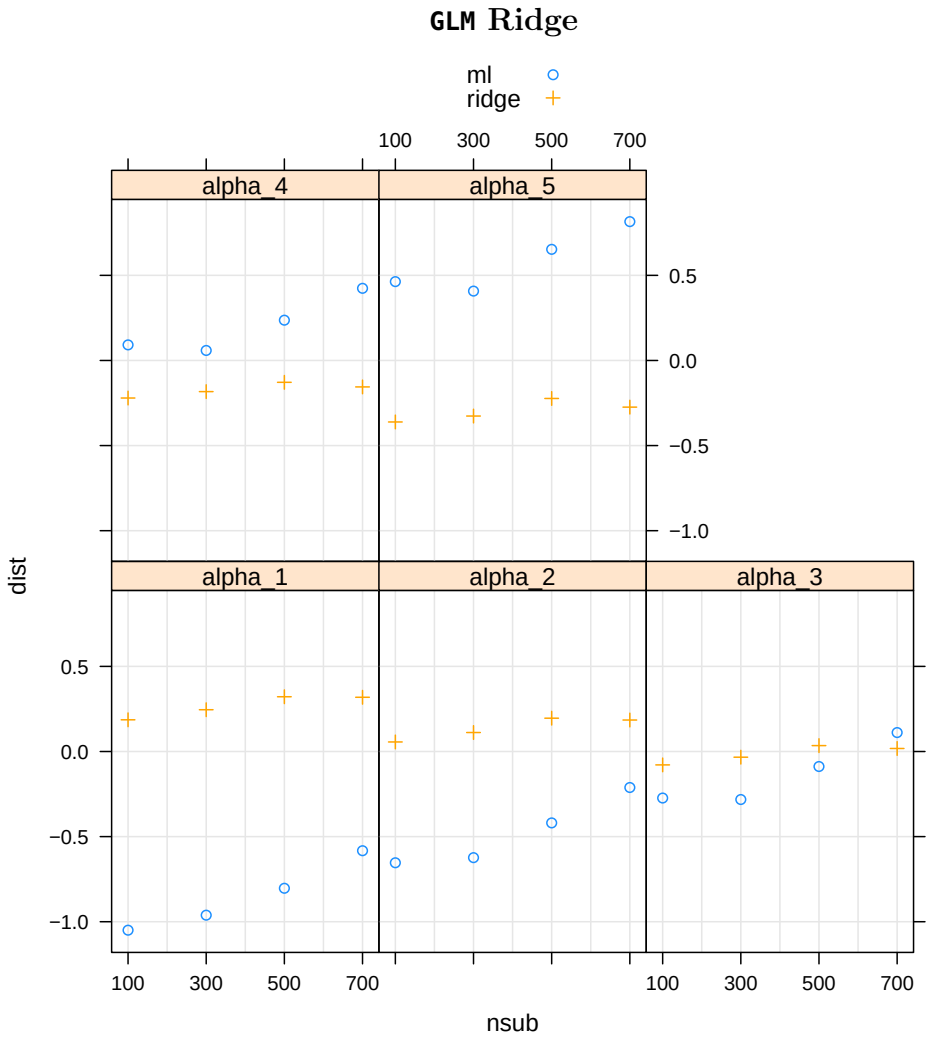

Figura 3.5: Distorsione, cinque prove.

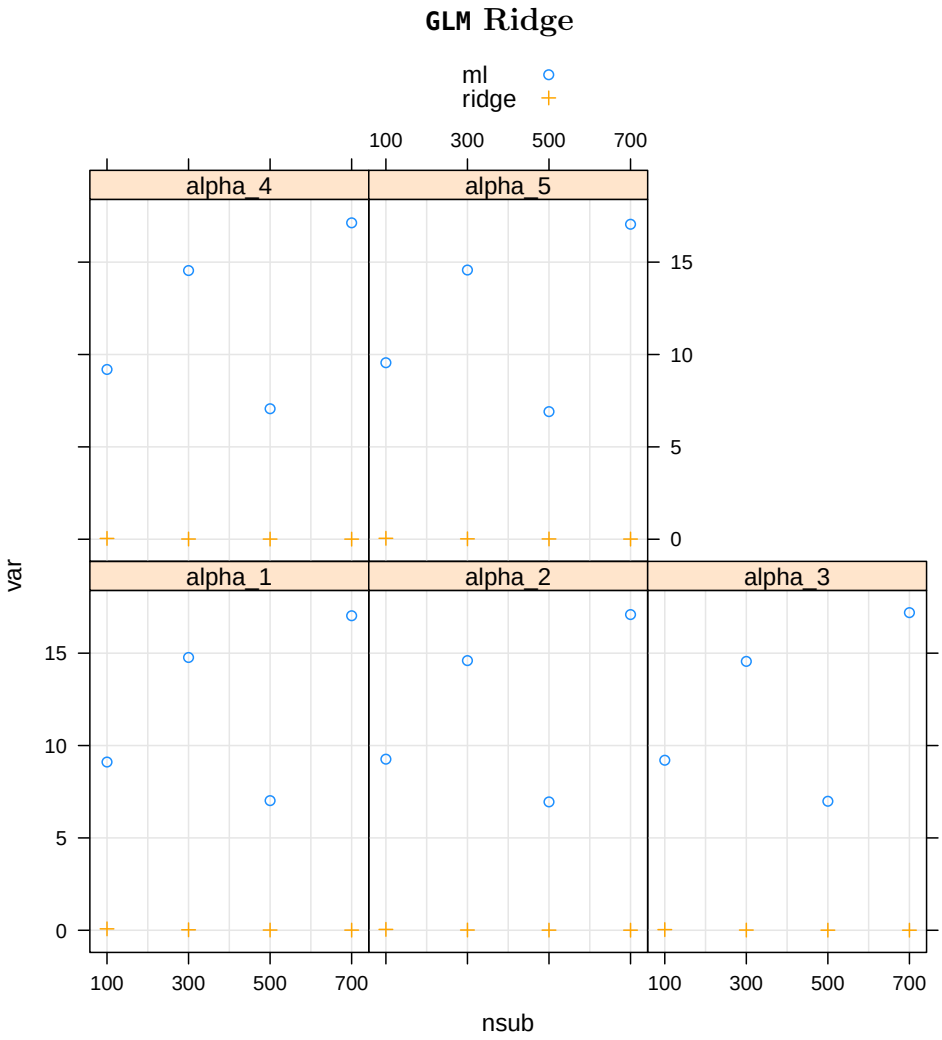

Figura 3.6: Varianza, cinque prove.

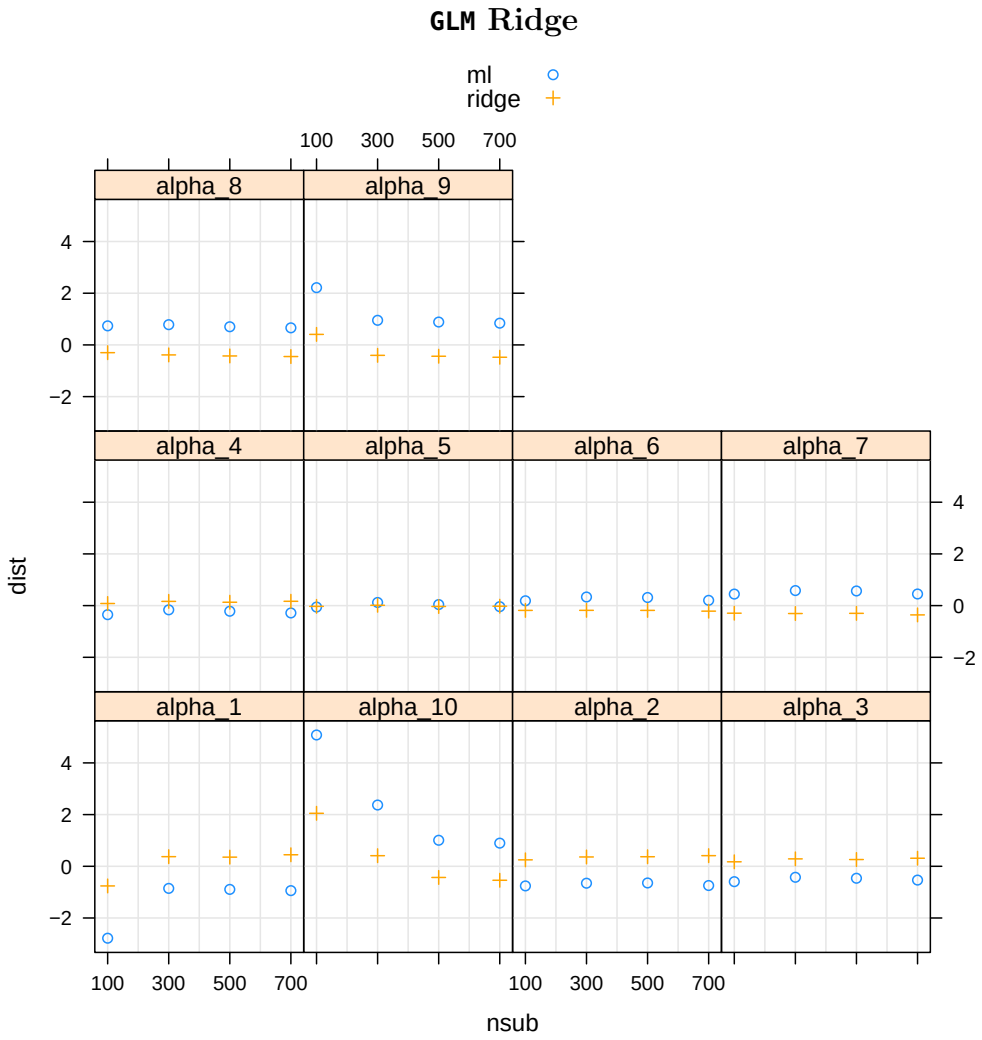

Figura 3.7: Distorsione, dieci prove.

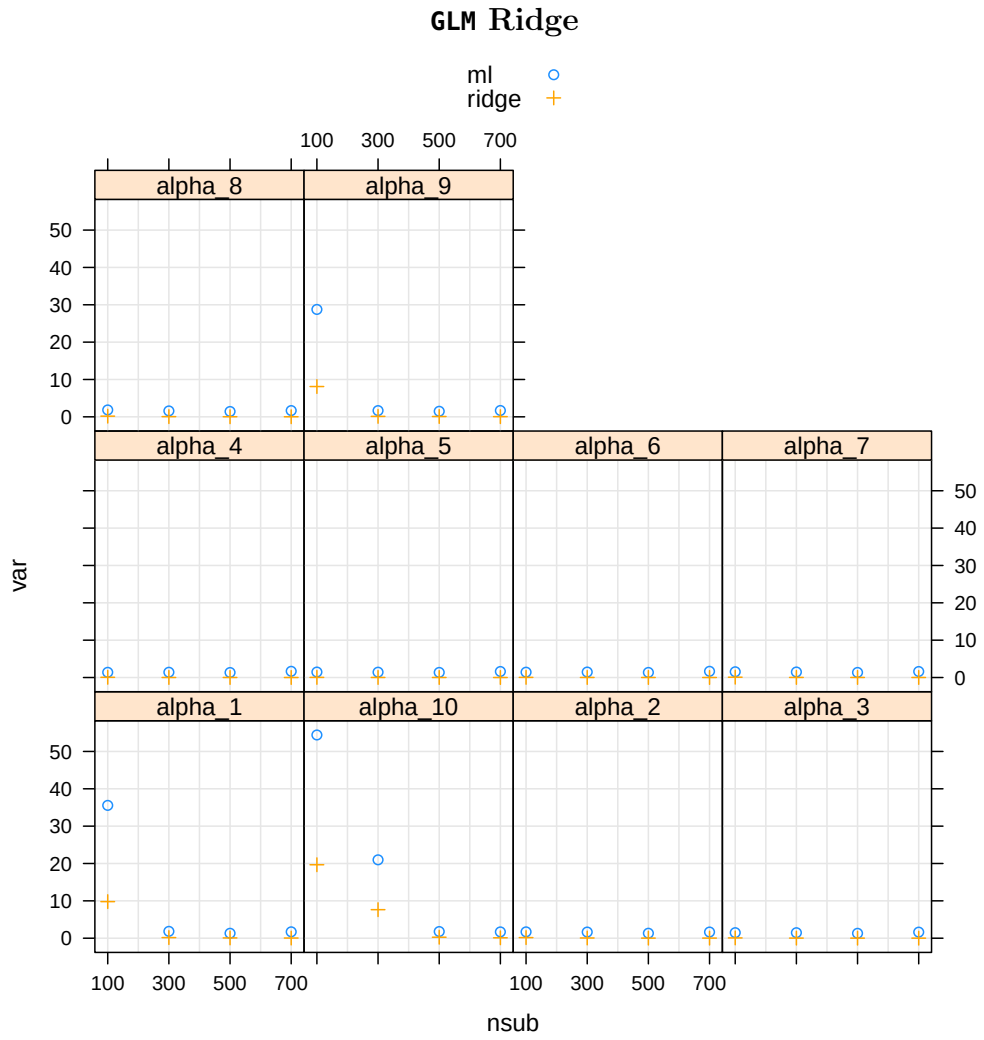

Figura 3.8: Varianza, dieci prove.

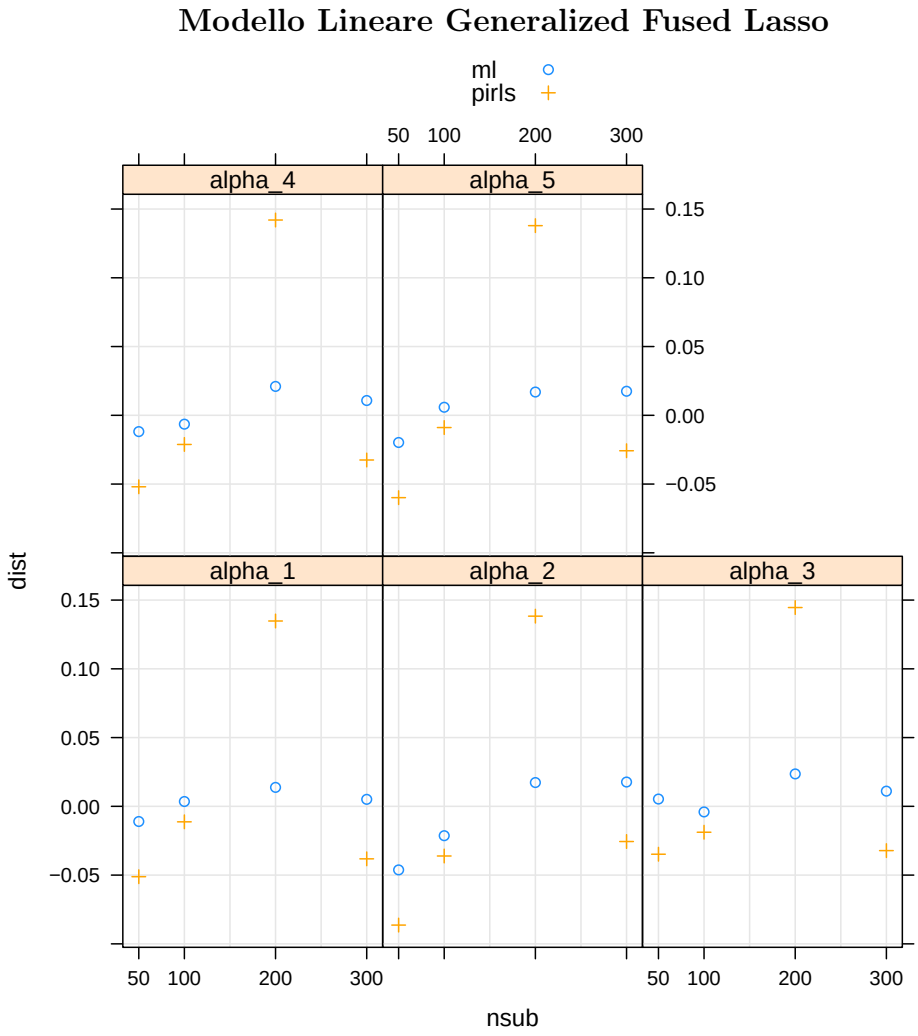

Figura 3.9: Distorsione, cinque prove.

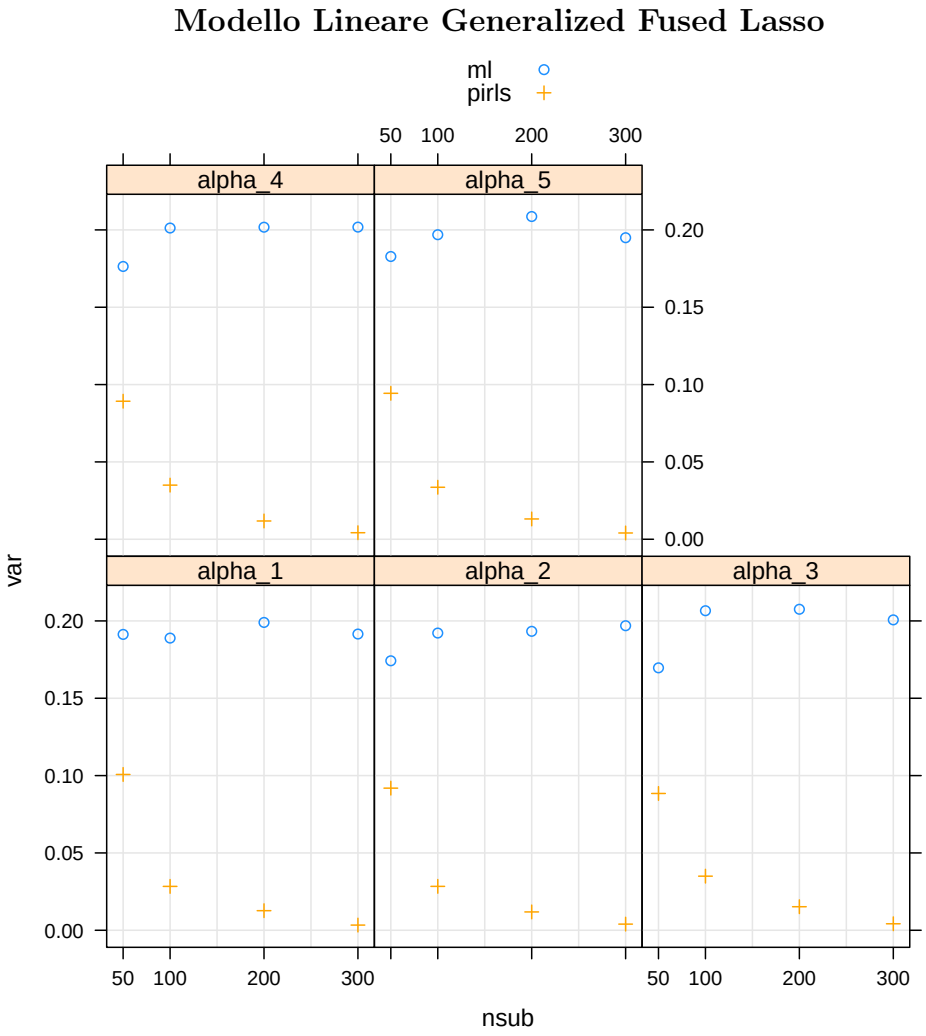

Figura 3.10: Varianza, cinque prove.

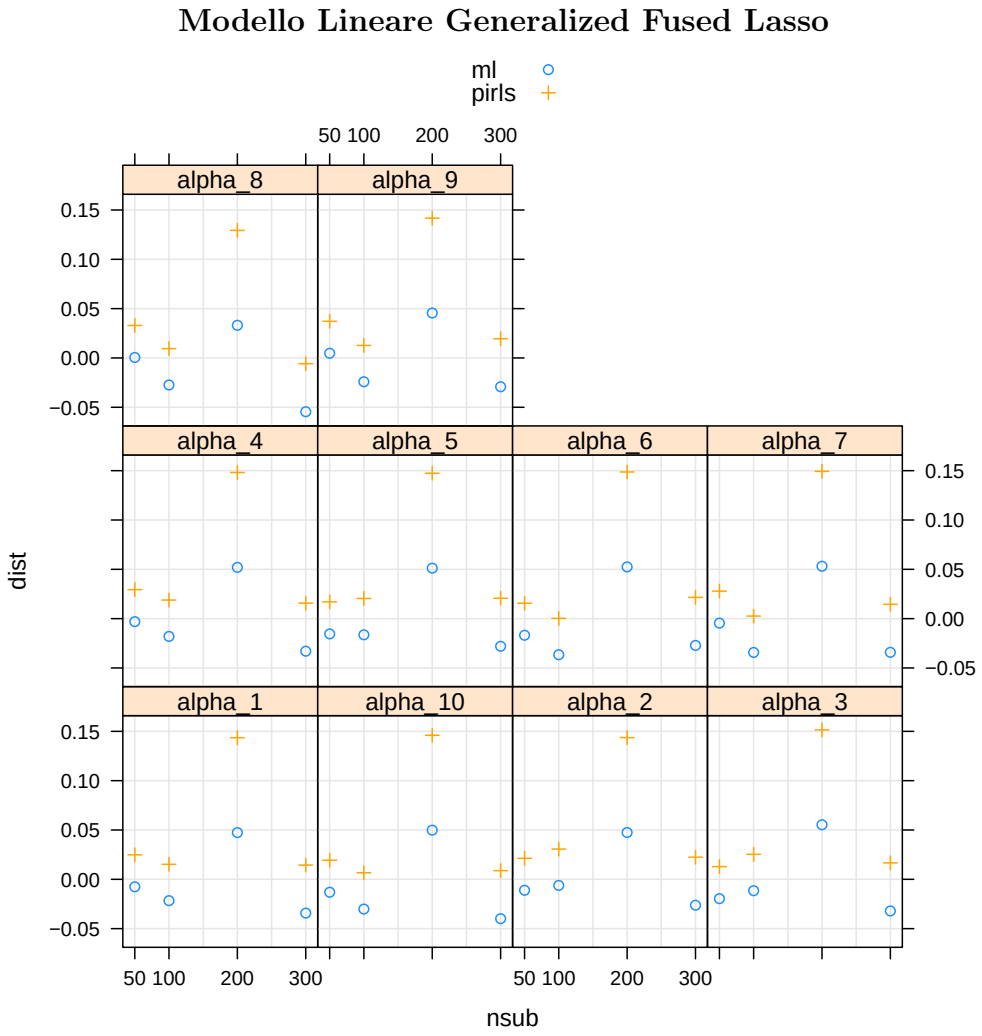

Figura 3.11: Distorsione, dieci prove.

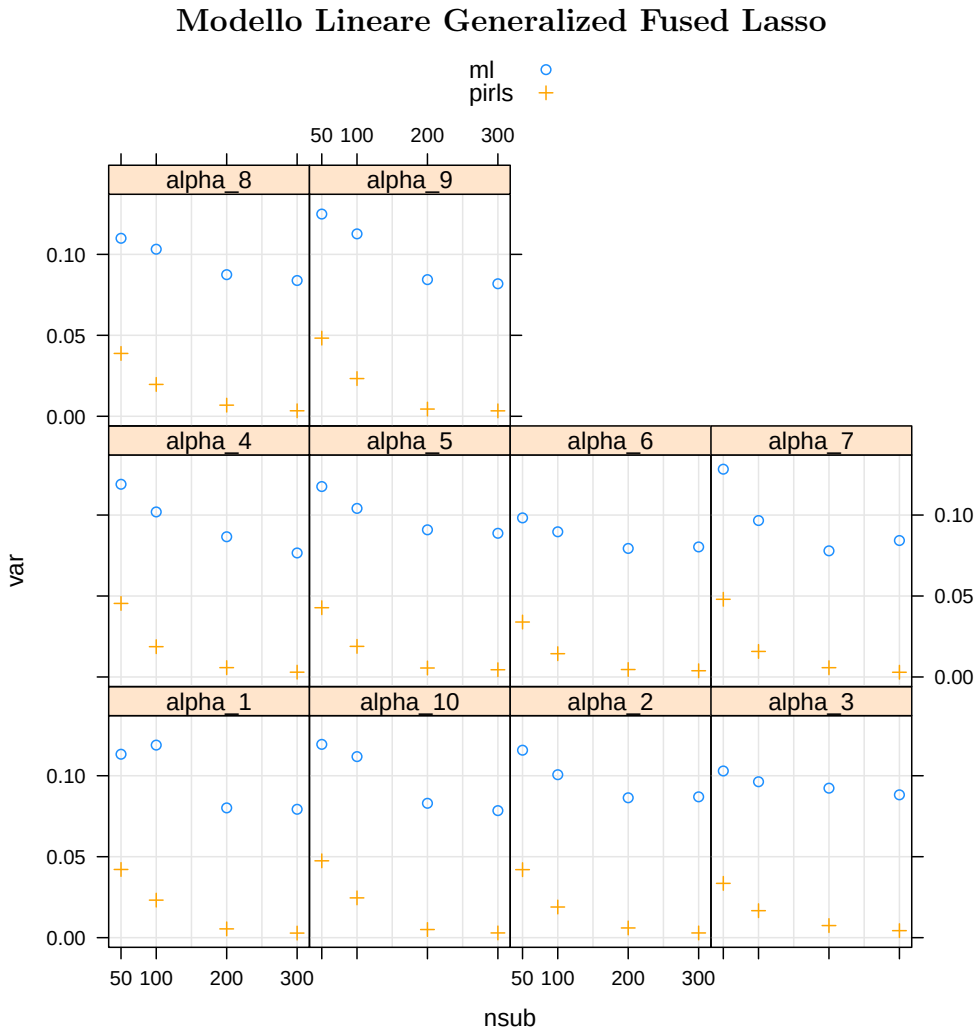

Figura 3.12: Varianza, dieci prove.

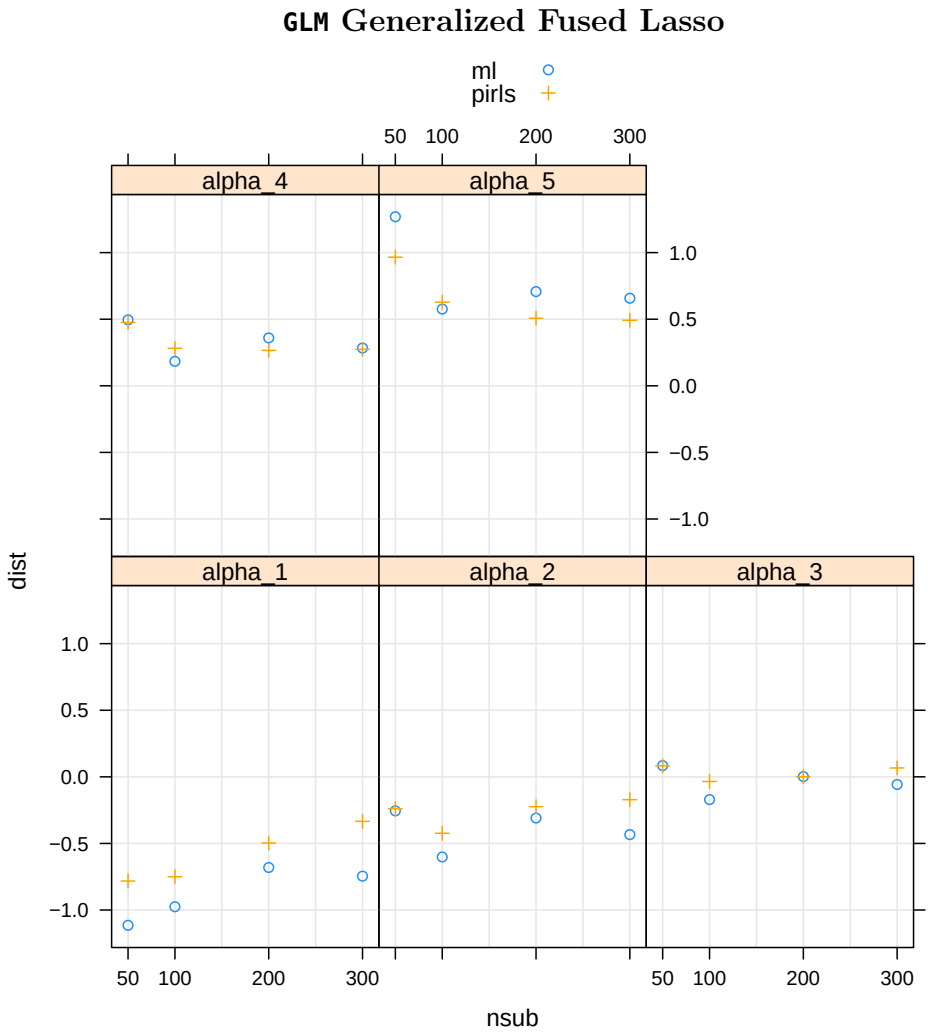

Figura 3.13: Distorsione, cinque prove.

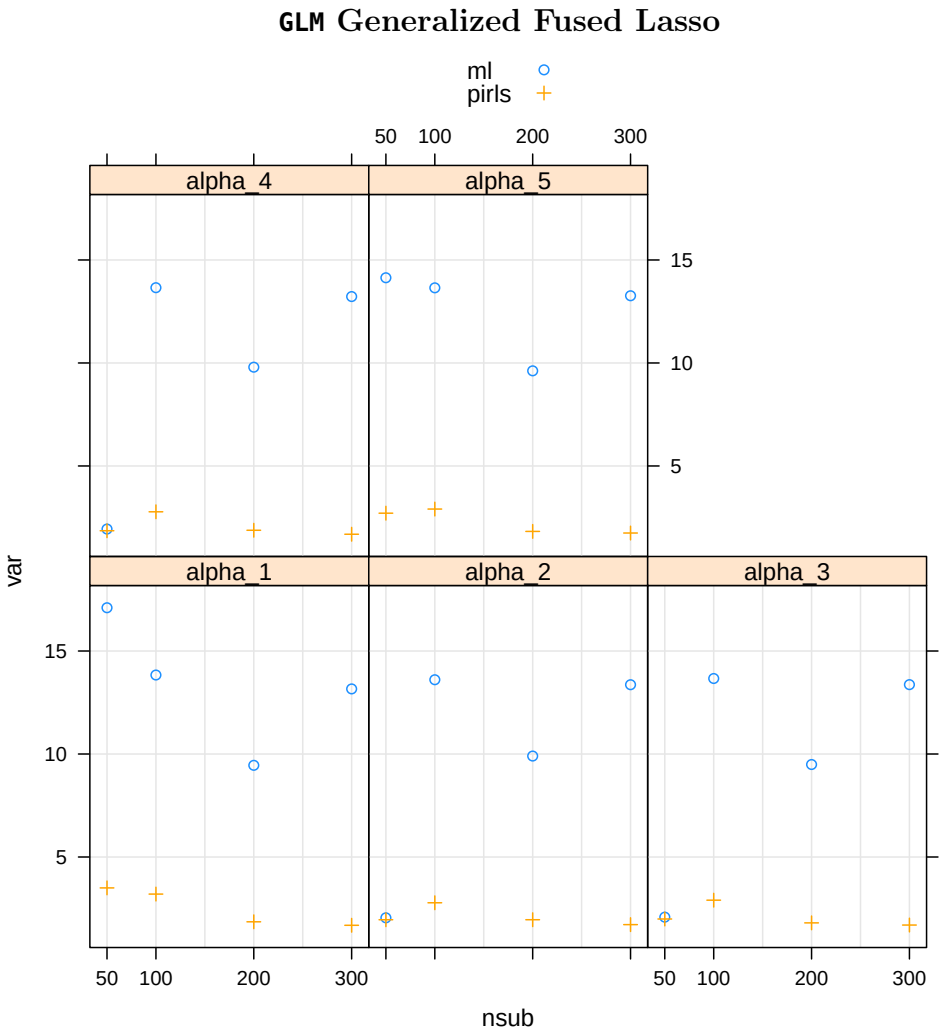

Figura 3.14: Varianza, cinque prove.

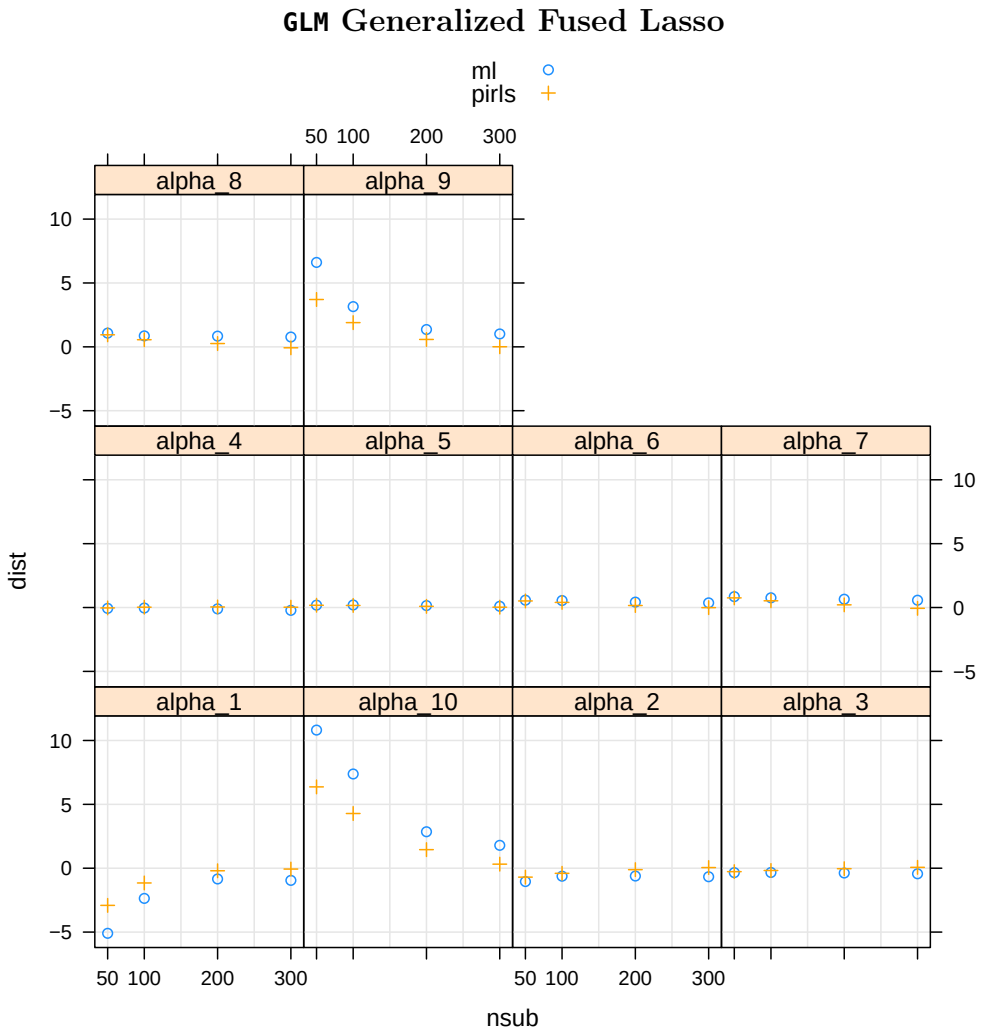

Figura 3.15: Distorsione, 10 prove.

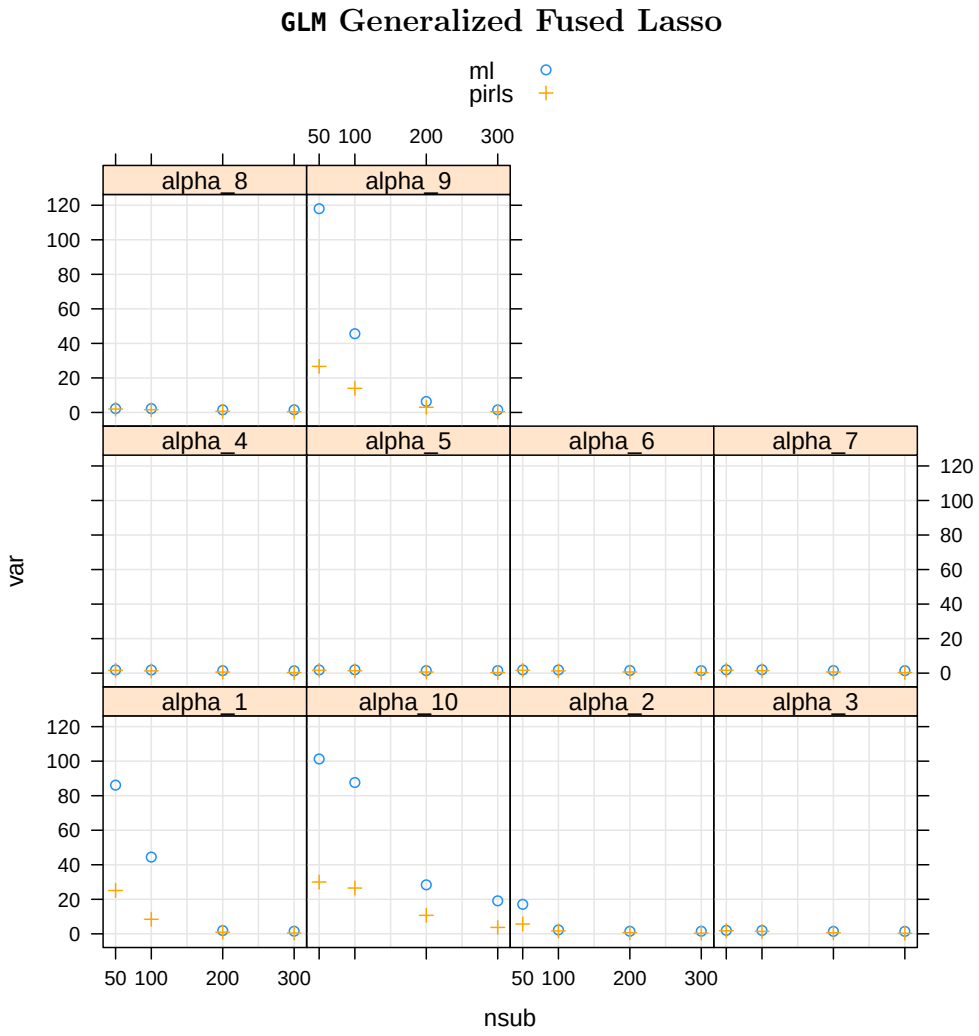

Figura 3.16: Varianza, 10 prove.

Cinque prove

	n	α_1	α_2	α_3	α_4	α_5
Modello Lineare Ridge	100	0.013	0.015	0.013	0.013	0.014
	300	0.004	0.005	0.004	0.004	0.004
	500	0.007	0.006	0.006	0.007	0.006
	700	0.002	0.002	0.002	0.002	0.002
Modello Lineare	100	0.210	0.202	0.226	0.215	0.224
	300	0.180	0.180	0.170	0.178	0.177
	500	0.163	0.161	0.166	0.165	0.165
	700	0.210	0.209	0.211	0.212	0.209
`GLM` Ridge	100	0.112	0.051	0.043	0.093	0.181
	300	0.086	0.026	0.012	0.045	0.128
	500	0.116	0.046	0.007	0.025	0.065
	700	0.110	0.040	0.005	0.030	0.084
`GLM`	100	10.210	9.690	9.280	9.200	9.770
	300	15.692	14.987	14.635	14.549	14.739
	500	7.665	7.126	6.992	7.120	7.335
	700	17.367	17.136	17.207	17.309	17.717
Modello Lineare Generalized	50	0.103	0.099	0.090	0.092	0.098
Fused Lasso	100	0.028	0.030	0.035	0.035	0.034
	200	0.031	0.031	0.036	0.032	0.032
	300	0.005	0.005	0.005	0.005	0.005
Modello Lineare	50	0.191	0.176	0.170	0.176	0.183
	100	0.189	0.193	0.207	0.201	0.197
	200	0.199	0.194	0.208	0.202	0.209
	300	0.192	0.197	0.201	0.202	0.195
`GLM` Generalized	50	4.112	2.011	1.997	2.081	3.641
Fused Lasso	100	3.762	2.957	2.905	2.854	3.302
	200	2.101	2.005	1.802	1.949	2.081
	300	1.797	1.746	1.701	1.761	1.986
`GLM`	50	18.343	2.114	2.088	2.184	15.750
	100	14.786	13.968	13.695	13.689	13.980
	200	9.915	9.997	9.491	9.926	10.120
	300	13.717	13.552	13.372	13.306	13.696

Tabella 3.2: Errore quadratico medio calcolato via simulazione in modelli regolarizzati e non per i cinque parametri strutturali.

Dieci prove, penalità ridge

	n	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}
Modello Lineare Ridge	100	0.101	0.098	0.096	0.092	0.095	0.090	0.090	0.100	0.090	0.093
	300	0.023	0.020	0.021	0.022	0.021	0.021	0.021	0.021	0.022	0.021
	500	0.002	0.002	0.002	0.002	0.002	0.003	0.003	0.002	0.003	0.002
	700	0.002	0.002	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002
Modello Lineare	100	0.119	0.111	0.114	0.123	0.120	0.112	0.127	0.117	0.120	0.109
	300	0.079	0.075	0.074	0.079	0.079	0.076	0.075	0.078	0.076	0.071
	500	0.103	0.105	0.104	0.106	0.110	0.110	0.100	0.105	0.106	0.108
	700	0.096	0.094	0.093	0.097	0.098	0.094	0.094	0.096	0.097	0.098
`GLM` Ridge	100	10.368	0.224	0.105	0.060	0.033	0.072	0.167	0.269	8.276	23.872
	300	0.304	0.185	0.105	0.038	0.013	0.045	0.114	0.204	0.282	7.810
	500	0.196	0.158	0.081	0.025	0.009	0.042	0.106	0.207	0.279	0.447
	700	0.242	0.192	0.108	0.033	0.004	0.052	0.139	0.223	0.278	0.402
`GLM`	100	43.299	2.277	1.841	1.518	1.474	1.466	1.742	2.390	33.660	80.223
	300	2.568	2.080	1.653	1.444	1.446	1.580	1.822	2.200	2.577	26.608
	500	2.120	1.749	1.524	1.387	1.367	1.462	1.682	1.930	2.271	2.804
	700	2.596	2.210	1.917	1.743	1.615	1.708	1.813	2.115	2.405	2.492

Tabella 3.3: Errore quadratico medio calcolato via simulazione in modelli con e senza regolarizzazione ridge per i dieci parametri strutturali.

Dieci prove, penalità generalized fused lasso

	n	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}
Modello Lineare Generalized	50	0.043	0.042	0.034	0.046	0.043	0.034	0.049	0.040	0.050	0.048
Fused Lasso	100	0.023	0.020	0.017	0.019	0.019	0.014	0.016	0.020	0.023	0.025
	200	0.026	0.027	0.030	0.028	0.027	0.027	0.028	0.024	0.025	0.026
	300	0.003	0.003	0.005	0.003	0.005	0.004	0.003	0.003	0.004	0.003
Modello Lineare	50	0.113	0.116	0.103	0.119	0.118	0.099	0.128	0.110	0.125	0.120
	100	0.119	0.101	0.096	0.102	0.104	0.091	0.098	0.104	0.113	0.113
	200	0.082	0.089	0.095	0.089	0.094	0.082	0.081	0.089	0.086	0.085
	300	0.080	0.088	0.089	0.078	0.090	0.081	0.085	0.087	0.083	0.080
`GLM` Generalized	50	33.554	6.202	1.848	1.640	1.664	1.926	2.254	2.868	40.443	70.552
Fused Lasso	100	9.762	1.848	1.443	1.323	1.428	1.475	1.706	1.953	17.566	44.870
	200	0.990	0.689	0.629	0.626	0.636	0.708	0.705	0.737	3.349	12.817
	300	0.455	0.406	0.348	0.318	0.346	0.329	0.341	0.390	0.404	3.819
`GLM`	50	112.078	18.178	2.137	1.874	1.896	2.222	2.648	3.395	161.705	218.264
	100	50.048	2.669	2.054	1.835	1.984	2.147	2.569	2.981	55.529	142.093
	200	2.593	1.857	1.563	1.444	1.465	1.720	1.924	2.210	8.200	36.549
	300	2.417	1.947	1.614	1.413	1.475	1.566	1.768	2.144	2.540	22.331

Tabella 3.4: Errore quadratico medio calcolato via simulazione in modelli con e senza regolarizzazione generalized fused lasso per i dieci parametri strutturali.

Conclusioni

In questa tesi, sono stati presentati alcuni metodi per ottenere stimatori con proprietà migliori di quello di massima verosimiglianza, in presenza di parametri incidentali. In questo caso, infatti, è risaputo che le usuali tecniche non producono risultati adeguati, in quanto la verosimiglianza non è regolare. I metodi più usati, sebbene abbiano delle buone proprietà, non sono sempre applicabili oppure si basano su ipotesi che non possono essere verificate; per questo, è stato proposto un metodo basato sulla verosimiglianza penalizzata, applicabile per qualsiasi tipo di `GLM` e dalle ipotesi equivalenti a quelle di un semplice modello ad effetti fissi. Negli studi di simulazione, questo approccio si è dimostrato essere migliore in termini di errore quadratico medio sia nel modello lineare che nel `GLM` binomiale.

Sebbene questo approccio abbia delle buone caratteristiche frequentiste, mantiene delle problematiche. Innanzitutto, la distorsione dello stimatore regolarizzato si è dimostrata essere simile a quella dello `SMV`, anche se in alcuni casi patologici sembra portare risultati nettamente migliori; la presenza della distorsione è comunque tollerabile a causa della considerevole diminuzione della varianza.

Poi, la scelta del parametro di regolazione λ è cruciale. La strategia utilizzata in questa tesi è stata di decidere a priori la complessità dei modelli stimati; sebbene questo approccio si sia dimostrato di successo nei casi particolari presi in esame, non essendo oggettivo i buoni risultati ottenuti non sono necessariamente raggiungibili in situazioni analoghe. Un approccio basato sui dati su cui indagare consiste nella selezione di λ tramite convalida incrociata generalizzata; nonostante questo metodo richieda tempi di calcolo decisamente inferiori ad una convalida incrociata semplice, le prove effettua-

te nella stesura di questa tesi hanno evidenziato che questa tecnica è esosa nelle richieste computazionali. Un modo per superare questo scoglio potrebbe essere sviluppare un'implementazione più performante degli algoritmi di stima, utilizzando un linguaggio di programmazione di basso livello come ad esempio `C`.

Se si vuole perseguire la strada del λ fissato per ogni numerosità campionaria, poi, metodi inferenziali validi non sono presenti per i modelli presentati. Sebbene nel caso della penalizzazione di tipo ridge varianza e distorsione siano ottenibili esplicitamente, utilizzare queste quantità per effettuare inferenza non è semplice. Una soluzione potrebbe essere l'utilizzo di metodi di ricampionamento come il *bootstrap*, ma le richieste computazionali sarebbero elevate. Una strada che potrebbe essere molto più fruttuosa è quella della *inferenza dopo la selezione* (*post-selection inference*); in questo filone della ricerca, di recente sono stati presentati in Lee *et al.* (2016) e in Taylor e Tibshirani (2016) dei metodi per fare inferenza per modelli lineari e lineari generalizzati basati su regolarizzazione di tipo lasso con λ fissato. Nel caso del *generalized fused lasso*, ricavare una regola per fare inferenza permetterebbe di poter eseguire stime intervallari e test di ipotesi in maniera analoga a quella dell'inferenza classica, sia per i parametri strutturali che per i parametri incidentali.

In conclusione, si ritiene che i modelli basati sulla verosimiglianza penalizzata, quando λ viene scelto adeguatamente, abbiano le proprietà desiderate, portando a una stimatore dei parametri strutturali più accurato di quello di massima verosimiglianza. Tuttavia, perché un approccio del genere possa venire preso in considerazione in un'analisi, sarebbe necessario lo sviluppo di risultati teorici e strumenti informatici adeguati che diano un ulteriore valore aggiunto a questo metodo. Infatti, gli usuali strumenti, benché non esenti da difetti, sono di veloce applicazione e permettono di fare inferenza.

Bibliografia

Arellano M.; Hahn J. (2007). Understanding bias in nonlinear panel models: Some recent developments. *Econometric Society Monographs*, **43**, 381.

Azzalini A. (2001). *Inferenza Statistica. Una Presentazione Basata sul Concetto di Verosimiglianza*. Springer-Verlag Italia, Milano.

Bondell H. D.; Reich B. J. (2009). Simultaneous factor selection and collapsing levels in anova. *Biometrics*, **65**(1), 169–177.

Davison A. C. (2003). *Statistical models*. Cambridge University Press, Cambridge.

Friedman J.; Hastie T.; Tibshirani R. (2001). *The elements of statistical learning*. Springer-Verlag, New York.

Hastie T. J.; Tibshirani R. J. (1990). *Generalized additive models*, volume 43. CRC Press, London.

Hoerl A. E.; Kennard R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

James G.; Witten D.; Hastie T.; Tibshirani R. (2013). *An introduction to statistical learning*. Springer-Verlag, New York.

Knaus J. (2013). *snowfall: Easier cluster computing (based on snow)*. R package version 1.84-6.1.

Lee J. D.; Sun D. L.; Sun Y.; Taylor J. E. *et al.* (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, **44**, 907–927.

Masarotto G.; Varin C. (2012). The ranking lasso and its application to sport tournaments. *The Annals of Applied Statistics*, **6**, 1949–1970.

Neyman J.; Scott E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16**, 33.

Oelker M.-R. (2015a). *gvcm.cat: Regularized Categorical Effects/Categorical Effect Modifiers/Continuous/Smooth Effects in GLMs*. R package version 1.9.

Oelker M.-R. (2015b). *Penalized regression for discrete structures*. Tesi di Dottorato di Ricerca, Ludwig-Maximilians-Universität München.

Pace L.; Salvan A. (1997). *Principles of statistical inference: from a Neo-Fisherian perspective*. World scientific, Singapore.

Pace L.; Salvan A. (2001). *Introduzione alla Statistica. II Inferenza, verosimiglianza, modelli.* CEDAM, Padova.

Park M. Y. (2006). *Generalized linear models with regularization*. Tesi di Dottorato di Ricerca, Stanford University.

She Y. (2010). Sparse regression with exact clustering. *Electron. J. Statist.*, **4**, 1055–1096.

Taylor J.; Tibshirani R. (2016). Post-selection inference for l1-penalized likelihood models. *https://arxiv.org/abs/1602.07358*.

Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58**, 267–288.

Tibshirani R.; Saunders M.; Rosset S.; Zhu J.; Knight K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society B*, **67**, 91–108.

Tutz G.; Oelker M.-R. (2016). Modelling clustered heterogeneity: Fixed effects, random effects and mixtures. *International Statistical Review*.