Università
Ca'Foscari
Venezia

Corso di Laurea Specialistica
in Statistica e Sistemi Informativi Aziendali

Tesi di Laurea

Ca' Foscari
Dorsoduro 3246
30123 Venezia

Intervalli di previsione dal Lasso: uno studio di simulazione

Relatore
Chiar.ma Prof.ssa Federica Giummolè

Laureanda
Marilisa Chiesurin
Matricola 815018

Anno Accademico
2011/2012

A me stessa...

"L'amore non è una cosa che si può insegnare, ma è la cosa più importante da imparare."

Papa Giovanni Paolo II (Karol Wojtyla)

Ringraziamenti

Ringrazio di cuore Matteo per forza che mi ha trasmesso;

i miei genitori per avermi incoraggiata e aiutata anche nella gestione quotidiana della famiglia;

Giovanni e Pierpaolo per avermi sempre donato dei sorrisi carichi d'amore e di gioia;

la prof.ssa Federica Giummolè per la disponibilità e l'opportunità datami di lavorare su argomenti di mio interesse;

gli amici che mi sono sempre stati accanto;

Tea e Polly per avermi sempre coccolata;

le mie colleghe, soprattutto amiche, di lavoro per aver sempre creduto in me;

la vita che mi ha dato la possibilità di realizzare tutto questo...

Sommario

Nello studio di un esperimento statistico, generalmente si assume che i dati osservati siano generati da un meccanismo generatore dotato di una componente casuale. Allo stesso tempo, si assume che questo meccanismo rientri tra quelli previsti all'interno del modello statistico. Un esempio classico è quando si definiscono le osservazioni $x_1, ...x_n$ come delle realizzazioni di una legge normale $\mathrm{N}(\mu, \sigma^2)$: in questo modo si vincola soggettivamente il meccanismo generatore alla famiglia normale. In questo esempio rimane quindi solo da stimare i parametri della normale attraverso una valutazione probabilistica. A tal proposito entrano in gioco ben due assunzioni: l'esistenza di un vero meccanismo generatore e che tale meccanismo sia contenuto nel modello statistico da noi utilizzato. La seconda assunzione è al centro del dibattito metodologico: molti studiosi criticano l'appraccio parametrico all'inferenza statistica proprio perché si cerca di vincolare il meccanismo generatore ad appartenere ad una famiglia troppo ristretta solo per garantire una facilità di calcolo e di elaborazione. Tutti i modelli statistici parametrici rappresentano un costrutto matematico, teorico, il cui obiettivo è approssimare il vero meccanismo generatore. Si può quindi affermare che tutti i modelli statistici che si utilizzano sono sbagliati anche se di vitale importanza per l'interpretazione e la modellazione dei fenomeni reali. Compito dello statistico è individuare quello più adatto nella situazione specifica. Ricordiamo a

riguardo la citazione di un grande statistico dei nostri tempi, G.E.P. Box: "All models are wrong, but some are useful".

Tradizionalmente, l'inferenza statistica si occupa principalemte dei problemi di stima, di verifica di ipotesi sui parametri. In un approccio più moderno, molti analisti concordano sul fatto che il loro lavoro copre uno studio ben più ampio che concerne la scelta del modello, la criticità di tale modello e la previsione. Chatfield (1995) ha sostenuto che l'inferenza statistica dovrebbe essere ampliata per includere l'intero processo di costruzione del modello statistico. La costruzione del modello è solo una parte dello studio del problem-solving in cui la costruzione del modello è generalmente un processo iterattivo. Se pensiamo ad esempio alla maggior parte degli studi sulle serie temporali, i modelli vengono determinati dai dati attraverso un ciclo iterativo: formulazione del modello, stima del modello e la sua validazione. Di solito l'analista cerca su una gamma di modelli e seleziona quello che ritiene essere il migliore secondo alcuni criteri, per esempio il criterio minimo di Akaike (AIC). Fatto ciò l'analista procede con la stima dei parametri del modello scelto attraverso le stesse tecniche che sarebbero state utilizzate nella tradizionale inferenza statistica dove il modello si assume a priori. Questo ragionamento, che risulta essere fallace e ingannevole (Zhang, 1992), non è stato considerato perché non era chiaro cos'altro si poteva e doveva fare.

Tipicamente ci sono tre fonti di inceretezza:

- l'incertezza della struttura modello;
- l'incertezza sulle stime dei parametri del modello;
- l'incertezza sui dati dove si includono anche gli errori di misura e di registrazione.

La letteratura statistica ha molto da dire sul secondo e terzo punto, ma piuttosto poco sul primo. La scarsità della letteratura è sorprendente dato che gli errori derivanti dall'incertezza del modello sono suscettibili di essere di gran lunga peggiori di quelli derivanti da altre fonti.
Possiamo affermare che le procedure parametriche classiche inseguono l'efficienza e si contrappongono alle procedure basate sull'eliminazione delle osservazioni anomale per ricercare la stabilità. Vi è un ultimo approccio attraverso la statistica robusta che si colloca in una situazione intermedia. L'obiettivo della statistica robusta è di predisporre strumenti per valutare la bontà delle procedure statistiche in intorni di modelli stocastici, e quindi di trovare procedure che mantengano buone proprietà anche quando il modello ipotizzato è solo un'approssimazione del vero modello (Pelegatti, 2000). La teoria della robustezza si basa su procedure di inferenza necessarie a prevenire perdite di efficienza e consistenza rispetto a eventuali deviazioni dal modello e cerca di prevenire eventuali effetti causati da valori anomali o da osservazioni influenti.
Questa tesi costituisce un primo tentativo per valutare come la procedura di selezione del modello influisca sulla previsione. In particolare ci si è concentrati sul problema della previsione dopo la selezione di variabili nel modello lineare classico. La tesi è così strutturata:

- il primo capitolo analizza alcuni metodi per la selezione delle variabili per evidenziarne la molteplicità di scelta;
- il secondo capitolo racchiude una panoramica sulla previsione partendo dal motivo per cui è argomento di interesse di molti studiosi fino ad evidenziare le varie problematiche connesse. Il capitolo conclude con gli intervalli di previsione soffermandosi sul caso particolare della distribuzione normale;

- il terzo capitolo racchiude uno studio di simulazione dove si calcolano gli intervalli di previsione dal lasso, dai minimi quadrati e dopo la selezione stepwise, con l'obiettivo di confrontare fra loro i diversi metodi delle selezioni delle variabili.

Indice

Capitolo 1

La selezione delle variabili nel modello lineare

1.1 La regressione lineare

Negli ultimi 30 anni il modello lineare è stato ampiamente usato nelle statistiche e resta uno degli strumenti più importanti.
L'obiettivo fondamentale della regressione è rappresentato dalla previsione della variabile dipendente. Per farlo, la regressione deve rendere massima la capacità previsiva delle variabili indipendenti.
Data una o più variabili esplicative[1] diciamo $X_1, \ldots, X_p$ è possibile ottenere la risposta Y che risulta essere la somma di due componenti ed è costruita in modo tale da essere il miglior predittore della variabile dipendente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon.$$

[1] utilizziamo il termine variabile esplicativa e non indipendente in quanto potrebbe essere un po' fuorviante: le variabili indipendenti potrebbero essere fortemente correlate tra di loro

Il primo termine $f(X_1, \ldots, X_p)$ rappresenta la componente sistematica ed individua la relazione che vi è tra le variabili esplicative e la risposta; il secondo termine ε rappresenta la componente accidentale o di errore: individua gli scostamenti accidentali tra Y e $f(X_1, \ldots, X_p)$, naturalmente quest'ultima componente, ε, è priva di connessione con le variabili esplicative.
La nostra attenzione si sofferma nel caso in cui $f(\cdot)$ sia una funzione lineare nei parametri ovvero:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

I parametri non noti del modello sono: β_0 (l'intercetta), e $\beta_1, \beta_2, \ldots, \beta_p$ (i coefficienti di regressione). $X_1, \ldots, X_p$ sono variabili deterministiche, ovvero misurate senza errore anche se nell'analisi di alcuni fenomeni reali tale ipotesi non è realistica.
Supponiamo di avere n osservazioni $y_1, \ldots, y_n$. Tali osservazioni sono determinazioni di variabili casuali $(Y_1, \ldots, Y_n)$ per cui vale:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i \qquad i = 1, \ldots, n$$

Per maggior semplicità si può utilizzare la notazione matriciale:

$$Y = X\beta + \varepsilon$$

Dove indichiamo con:

- Y il vettore $n \times 1$ dei valori della variabile dipendente per le n unità del campione;
- X la matrice $n \times (p + 1)$ dei valori dei p regressori per le n unità del campione. La matrice contiene, oltre ai valori dei regressori, una prima colonna composta da n valori tutti pari a 1 in corrispondenza dell'intercetta del modello;

- β il vettore $(p+1)$ dei parametri del modello;
- ε il vettore $n \times 1$ dei termini d'errore.

Per quanto riguarda le esplicative X_i sono variabili che possono risultare da diverse fonti:

- input quantitativi;
- trasformazioni di input quantitativi, come il logaritmo, la radice quadrata o il quadrato;
- espansioni di una base, come $X_2 = X_1^2$, $X_3 = X_1^3$, portando ad una rappresentazione polinomiale;
- binarie o "dummy" per la codifica dei livelli di input qualitativi;
- interazioni tra variabili, per esempio, $X_3 = X_1 \cdot X_2$.

Ciascuna osservazione del campione può essere interpretata come una realizzazione empirica delle corrispondenti variabili nella popolazione. Si possono quindi descrivere le ipotesi degli ε:

- $E(\varepsilon) = 0$ vettore nullo di n elementi da cui $E(Y) = X\beta$;
- $Var(\varepsilon) = E(\varepsilon\varepsilon') = \Sigma = \sigma^2 I_n$,
  dove I_n rappresenta la matrice identità di ordine n;
- $\varepsilon \sim NMV(0, \sigma^2 I_n)$ da cui $Y \sim NMV(X\beta, \sigma^2 I_n)$.

Nell'equazione del modello troviamo $p+1$ parametri ignoti più σ^2 che dovranno essere stimati. Il metodo dei minimi quadrati può essere utilizzato per stimare il vettore di parametri incogniti β. Prendendo un campione di n unità si può calcolare il vettore delle stime $\hat{\beta}$ da cui è possibile determinare

il vettore $\hat{y}$ dei valori teorici della variabile dipendente per le n unità del campione nell'ipotesi di perfetta dipendenza lineare tra Y ed i p regressori:

$$\hat{y} = X\hat{\beta}.$$

I residui vengono calcolati attraverso la differenza tra gli n valori empirici e i corrispondenti valori teorici:

$$e = y - \hat{y} = y - X\hat{\beta}.$$

Gli n valori di e_i sono n determinazioni campionarie del termine d'errore ε del modello.
Attraverso il metodo dei minimi quadrati, viene calcolato il vettore di coefficienti β in modo da rendere minima la somma dei quadrati degli scarti tra ordinate empiriche e ordinate teoriche, o equivalentemente, la somma dei residui al quadrato. Si può interpretare il problema anche geometricamente: la funzione che si vuole minimizzare si può leggere come la distanza del vettore y dal vettore $X\beta$ che, al variare di β descrive lo spazio colonna della matrice X. La distanza di un vettore da un sottospazio è minima se si valuta la distanza del vettore dalla sua proiezione ortogonale sul sottospazio, si tratta di trovare quel vettore $\hat{\beta}$ che definisce la proiezione ortogonale di y sullo spazio colonna di X. Il vettore $y - X\hat{\beta}$ è ortogonale a ogni vettore dello spazio colonna di X:

$$\langle X\beta, y - X\hat{\beta}\rangle = 0 \qquad \forall b$$

che corrisponde:

$$(X\beta)'(y - X\hat{\beta}) = 0$$

Essendo β' diverso dal vettore nullo dovrà essere:

$$X'X\hat{\beta} = X'y$$

Se $X'X$ è invertibile allora il sistema ammette una unica soluzione:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Se le colonne di X sono linearmente dipendenti la matrice $X'X$ è singolare e il sistema normale ha infinite soluzioni.
I coefficienti di regressione, β_j, in un modello di regressione multiplo si dicono coefficienti di regressione parziale ed esprimono la variazione media della variabile dipendente, per ogni variazione unitaria della corrispondente variabile indipendente, a parità di valori assunti rispetto agli altri regressori nel modello. I coefficienti di regressione del modello multiplo coincidono infine con quelli di altrettanti modelli semplici qualora i regressori siano fra loro incorrelati o, in termini geometrici, qualora le colonne della matrice X siano tra loro ortogonali.

Consideriamo il caso in cui $p = 2$ e la prima colonna di X sia 1_n:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \tag{1.1}$$

Il modello di regressione è detto lineare semplice:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

esso interpola i valori di y attraverso una retta di regressione.

Il termine β_0 è l'intercetta, noto anche come polarizzazione nell'apprendimento automatico e assumiamo sia inclusa in β.
Visto come funzione oltre lo spazio di input *p-dimensionale*, $f(X) = X'\beta$ è lineare, e l'inclinazione $f'(X) = \beta$ è un vettore nello spazio di input che

punta verso la direzione con salita più ripida [2] .

Vi sono molti metodi lineari per stimare i parametri del modello, ma quello maggiormente famoso è il metodo dei minimi quadrati. In questo approccio, si considerano i coefficienti β per minimizzare la somma dei quadrati dei residui

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

$RSS(\beta)$ è una funzione quadratica dei parametri, e quindi il suo minimo esiste sempre, ma non è l'unico. Una migliore comprensione si ha con la forma matriciale, ossia

$$RSS(\beta) = (y - X\beta)'(y - X\beta)$$

in cui X è una matrice $n \times p$ con ogni riga come vettore di input, e y è un *n-vettore* degli output dell'insieme dei dati. Semplificando la formula, si avrà:

$$X'(y - X\beta) = 0$$

Se $X'X$ non è singolare, l'unica soluzione viene fornita da:

$$\hat{\beta} = (X'X)^{-1}X'y$$

e il valore stimato per l'input x_i è $y_i = y(xi) = x_i'\hat{\beta}$.

1.1.1 Teorema di Gauss-Markov

Uno dei risultati più importanti riguardo i modelli lineari afferma che gli stimatori ottenuti attraverso i minimi quadrati dei parametri β hanno la più piccola varianza tra tutti gli stimatori lineari non distorti.

[2] HASTIE T., TIBSHIRANI R., Independent components analysis through product density estimation, in S. T. S. Becker and K. Obermayer (eds), Advances in Neural Information Processing Systems 15, MIT Press, Cambridge, MA, 2003.

Concentriamoci sulla stima di una combinazione lineare dei parametri $\theta = a'\beta$. La stima dei minimi quadrati di $a'\beta$ è

$$\hat{\theta}_a = a'\hat{\beta} = a'(X'X)^{-1}X'y$$

Se si assume la correttezza del modello lineare lo stimatore $a'\hat{\beta}$ è non distorto perché

$$\begin{aligned} E(a'\hat{\beta}) &= E(a'(X'X)^{-1}X'y) \\ &= a'(X'X)^{-1}X'X\beta \\ &= a'\beta. \end{aligned} \tag{1.2}$$

Il teorema di Gauss-Markov afferma che se abbiamo un qualsiasi altro stimatore lineare $\hat{\theta}_c = c'y$ che è non distorto per $\theta = a'\beta$, cioè, $E(c'y) = a'\beta$, allora

$$Var(a'\hat{\beta}) \leq Var(c'y).$$

Consideriamo l'errore quadratico medio di uno stimatore $\tilde{\theta}$:

$$\begin{aligned} MSE(\tilde{\theta}) &= E(\tilde{\theta} - \theta)^2 \\ &= Var(\tilde{\theta}) + [E(\tilde{\theta}) - \theta]^2. \end{aligned} \tag{1.3}$$

Il teorema di Gauss-Markov implica che lo stimatore dei minimi quadrati abbia il minimo errore quadratico medio tra tutti gli stimatori lineari senza distorsione. Tuttavia, vi può esistere uno stimatore distorto con minore errore quadratico medio. Tale stimatore può scambiare una piccola distorsione per una maggiore riduzione della varianza. Stime distorte sono comunemente usate. Qualsiasi metodo che riduce o pone a zero alcuni coefficienti dei minimi quadrati comporta una stima distorta.

L'errore quadratico medio è intimamente legato all' accuratezza della stima. Consideriamo la predizione della nuova risposta al input x_0,

$$Y_0 = f(x_0) + \varepsilon_0.$$

L'errore di previsione atteso di $\tilde{f}(x_0) = x_0^T\tilde{\beta}$ è:

$$\begin{aligned} E(Y_0 - \tilde{f}(x_0))^2 &= \sigma^2 + E(x_0'\tilde{\beta} - f(x_0))^2 \\ &= \sigma^2 + MSE(\tilde{f}(x_0)). \end{aligned} \tag{1.4}$$

Pertanto, l'errore di previsione atteso e l'errore quadratico medio differiscono solo per la costante σ^2, che rappresenta la varianza della nuova osservazione y_0.

1.2 La scelta dei regressori

Ci sono due motivi per cui spesso le stime dei minimi quadrati non sono soddisfacenti:

- La prima è la precisione: le stime ottenute attraverso i minimi quadrati hanno spesso distorsione bassa ma grande varianza. L'accuratezza della stima può essere talvolta migliorata riducendo o impostando alcuni coefficienti a zero.
- La seconda ragione riguarda l'interpretazione. Spesso viene analizzato un sottoinsieme dei predittori per ottenere un quadro generale sacrificando i piccoli dettagli.

Cercheremo quindi di descrivere alcuni modi per selezionare un sottoinsieme di variabili con la regressione lineare.

1.2.1 Forward & Backward Stepwise Selection

Forward-stepwise è chiamato *"algoritmo greedy"* in quanto produce una sequenza di modelli annidati. Attraverso la forward-selection, i singoli regressori vengono aggiunti in maniera sequenziale al modello: inizia con l'intercetta ed aggiunge passo a passo tutti i regressori a partire da quello che

descrive meglio i dati. Si parte con una sola covariata, quella con la maggiore correlazione significativa[3] con la variabile risposta. Dopo aver fissato un livello di significatività, si inserisce la seconda variabile: quella che presenta il coefficiente di correlazione parziale più elevato e significativo. Si prosegue fino a quando il coefficiente di correlazione parziale dell'ultima variabile inserita non è più significativo rispetto al livello prefissato; il modello definitivo è quello ottenuto al penultimo passo.
Backward-stepwise fa il procedimento opposto della forward-stepwise: parte considerando il modello che include tutte le variabili a disposizione, e dopo aver fissato un livello di significatività, elimina la variabile con il coefficiente di regressione meno significativo in base al test t. Si calcolano di nuovo le stime dei coefficienti delle variabili rimaste e si ripete il procedimento sino a quando non vi sono più covariate che risultano non significative al livello prefissato.
Gli algoritmi portano spesso a risultati molto simili. Alcuni pacchetti software implementano delle strategie ibride di selezione che sfruttano entrambi i metodi: valutano ad ogni passo le due strategie e selezionano la migliore delle due.

1.2.2 Forward-Stagewise Regression

Forward-Stagewise Regression (FS) inizia con l'intercetta uguale a $\bar{y}$, e i predittori con coefficienti inizialmente tutti a 0. Ad ogni passo l'algoritmo identifica la variabile più correlata con i residui. Si calcola quindi il coefficiente di regressione lineare semplice del residuo su questa variabile scelta, e quindi si aggiunge il coefficiente stesso per quella variabile. Questo si continua fino a quando nessuna delle variabili ha correlazione con il residui. Come

[3] si utilizza il test t

conseguenza, forward-Stagewise Regression può necessitare di molti passaggi per raggiungere i minimi quadrati, e storicamente è stata respinta in quanto inefficiente.

1.3 Shrinkage Methods

Mantenendo un sottoinsieme dei predittori e scartando il resto, il modello da studiare che ne esce ha la possibilità di avere degli errori di previsione minori rispetto al modello completo. Tuttavia, essendo un processo discreto (le variabili sono mantenute o scartate) spesso presenta alta varianza, e quindi non riduce l'errore di predizione del modello completo. I metodi Shrinkage non scartando le variabili, non soffrono tanto l'alta variabilità.

1.3.1 Ridge Regression

La ridge regression è un metodo di stima che ottiene uno stimatore distorto simile nella forma allo stimatore dei minimi quadrati ma con varianza più piccola.
I coefficienti Ridge minimizzano la somma dei quadrati dei residui:

$$\hat{\beta}^{ridge} = \underset{\beta}{argmin}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\}.$$

Qui $\lambda \geq 0$ è un parametro di complessità che controlla la quantità di contrazione: maggiore è il valore di λ, maggiore è la quantità di restringimento. I coefficienti non sono mai ridotti a zero, ma possono assumere valori molto vicini allo zero. Un modo equivalente di scrivere il problema ridge è:

$$\hat{\beta}^{ridge} = \underset{\beta}{argmin} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$$vincolato \quad \sum_{j=1}^{p} \beta_j^2 \leq t.$$

Un coefficiente estremamente positivo su una variabile può essere annullato da un altrettanto grande coefficiente negativo sulla sua correlata. Imponendo un vincolo sulla dimensione dei coefficienti, questo problema è alleviato. La scomposizione della matrice in valori singolari (SVD) della matrice di input X ci da una maggiore conoscenza della natura della Ridge Regression. La SVD della $p \times n$ matrice X ha la forma:

$$X = UDV'.$$

U e V sono matrici ortogonali rispettivamente di dimensione $n \times p$ e $p \times p$, le colonne di U sono generate dalle colonne di X e le colonne di V sono generate dalle righe di D. D è una matrice diagonale che ha come valori $d_1 \geq d_2 \geq, \ldots, \geq d_p \geq 0$ chiamati i valori singolari di X. Se uno o più $d_j = 0$ allora è una matrice singolare. Usando gli SVD, si può scrivere il vettore dei minimi quadrati:

$$\begin{aligned} X\hat{\beta}^{ls} &= X(X'X)^{-1}X'y \\ &= UU'y. \end{aligned} \tag{1.5}$$

Le soluzioni Ridge sono:

$$\begin{aligned} X\hat{\beta}^{ls} &= X(X'X + \lambda I)^{-1}X'y \\ &= UD(D^2 + \lambda I)^{-1}DU'y \\ &= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j' y. \end{aligned} \tag{1.6}$$

dove gli u_j sono le colonne di U. La SVD della matrice X sono un altro modo di rappresentare le componenti principali della variabile X. La matrice di covarianza è data da $S = X'X/n$ da cui abbiamo

$$X'X = VD^2V',$$

gli autovettori v_j (colonne di V) sono anche chiamate componenti principali di X. La prima componente principale ha come proprietà che $z_1 = Xv_1$ ha la più grande varianza campionaria di tutte le combinazioni lineari normalizzate delle colonne di X.

$$Var(z_1) = Var(Xv_1) = \frac{d_1^2}{n}$$

Le successive componenti principali z_j hanno varianza massima d_j^2/n e sono soggette ad essere ortogonali alle precedenti. I piccoli valori d_j corrispondono alla direzione nello spazio delle colonne di X ed hanno una piccola varianza quindi la Ridge Regression preferisce queste direzioni.

1.3.2 Il Lasso

Il metodo Lasso è un metodo di restringimento, la cui stima viene definita nel seguente modo:

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$$vincolato \quad \sum_{j=1}^{p} |\beta_j^2| \leq t.$$

In tal caso, è possibile riparametrizzare la costante β_0 mediante la standardizzazione dei predittori; la soluzione per β_0 è $\bar{y}$ e successivamente si ha un modello senza intercetta[4]. Il metodo Lasso viene anche conosciuto come base di perseguimento e il problema lasso può anche essere scritto nell'equivalente forma Langragiana, ossia:

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{2} |\beta_j^2| \}$$

[4]TIBSHIRANI R., Regression shrinkage and selection via the lasso, in Journal of the Royal Statistical Society, Series B58, 1996.

Si può calcolare la soluzione Lasso come un problema di programmazione quadratica, in quanto gli algoritmi efficienti sono disponibili per calcolare l'intero percorso di soluzioni di come λ viene variata. A causa della natura del vincolo, da una variabile t molto piccola ne scaturirà un coefficiente con una variabile pari a zero. In pratica, il lasso compie una continua selezione del sottoinsieme. Se t viene scelto maggiore rispetto a $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j|$, le stime di lasso saranno di $\hat{\beta}_j$. D'altro canto, si potrebbe anche indicare con $t = t_{0/2}$ i coefficienti minimi quadrati che vengono ridotti mediamente del 50%[5]. Ad ogni modo, la natura di tale restringimento non è ovvia e la variabile t può essere scelta per minimizzare una stima dell'errore di predizione previsto. Lasso traduce ogni coefficiente con un fattore costante λ, troncato allo zero[6].

Ridge Regression e il Lasso

Cercheremo ora di confrontare la Ridge Regression e il Lasso per restringere il modello di regressione lineare. Nel caso in cui la matrice degli input X sia ortonormale, la Ridge Regression e il Lasso hanno delle soluzioni chiare. Entrambi i metodi applicano una semplice trasformazione ai minimi quadrati per stimare i β_j. La Ridge Regression fa una contrazione proporzionale. Il Lasso converte ogni coefficiente per un fattore costante λ, troncandolo a zero. Si puó generalizzare la Ridge Regression e il Lasso, e visualizzarli come stime di Bayes. Si consideri il criterio

$$\tilde{\beta} = \underset{\beta}{argmin}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{2} |\beta_j^2|^q\}$$

$q = 1$ corrisponde al Lasso, mentre $q = 2$ alla Ridge Regression.

[5]WU T., LANGE K., Coordinate descent procedures for lasso penalized regression, in Annals of Applied Statistics2(1), 2008.

[6]ZHAO P., YU B., On model selection consistency of lasso, in Journal of Machine Learning Research7, 2006.

1.3.3 Least Angle Regression

Least Angle Regression (LAR) è relativamente nuovo (Efron et al., 2004), e può essere visto come una sorta di versione democratica della regressione forward stepwise; è intimamente connesso con il Lasso, e in effetti fornisce un algoritmo estremamente efficiente i cui passi si possono così sintetizzare:

- Standardizza i predittori per avere media zero e norma unitaria partendo dai residui: $r = y - \bar{y}, \beta_1, \beta_2, \ldots, \beta_p = 0$;
- Trova il predittore x_j più correlato con $\mathbf{r}$;
- Muove i β_j da 0 verso i coefficienti dei minimi quadrati $\langle x_j, r \rangle$ finché non si incontri un altro x_k con una correlazione così alta con i residui come x_j;
- Muove β_j e β_k nella direzione definita dai coefficienti dei minimi quadrati dei residui su (x_j, x_k) finché qualche altro regressore x_l abbia una correlazione alta con i residui;
- Continua in questo modo fino a quando tutti i predittori p sono stati inseriti. Dopo $min(n-1, p)$ passi, si arriva alla completa soluzione dei minimi quadrati.

Questo processo va avanti finché tutte le variabili sono inserite nel modello.

1.4 Combinazioni lineare di regressori

In molte situazioni abbiamo un gran numero di input, spesso correlati. Si possono comporre delle combinazioni lineari $Z_m, m = 1, \ldots, M$ delle X_j originali, e utilizzarle al posto delle X_j come input nella regressione. I metodi differiscono nel modo in cui le combinazioni lineari sono costruite.

1.4.1 Componenti Principali

Lo scopo principale di questa tecnica è la riduzione di un numero più o meno elevato di variabili tramite una trasformazione lineare delle variabili che proietta quelle originarie in un nuovo sistema cartesiano nel quale le variabili vengono ordinate in ordine decrescente di varianza. La variabile con maggiore varianza viene proiettata sul primo asse, la seconda sul secondo asse e così via. La riduzione della complessità avviene limitandosi ad analizzare le principali (per varianza) tra le nuove variabili ottenute. In questa tecnica sono i dati stessi che determinano i vettori di trasformazione. Questo approccio utilizza come combinazione lineare Z_m le componenti principali esposte con la Ridge Regression. Poiché gli z_m sono ortogonali, questa regressione è la somma di regressioni univariate:

$$\hat{y}^{pcr}_{(M)} = \bar{y}1 + \sum_{m=1}^{M} \hat{\theta}_m z_m,$$

dove $\hat{\theta}_m = \langle z_m, y\rangle / \langle z_m, z_m\rangle$. Poiché gli z_m sono combinazioni lineari degli x_j originali, possiamo esprimere così la soluzione:

$$\hat{\beta}^{pcr}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m.$$

La Ridge Regression è simile alla regressione delle componenti principali: entrambe operano con le componenti principali della matrice degli input. La prima restringe i coefficienti delle componenti principali; la seconda elimina i $p - M$ componenti più piccoli.

1.4.2 Minimi Quadrati Parziali (PLS)

Quando si calcola una regressione lineare non è detto che le componenti principali più importanti risultino quelle con autovalore più alto. La tecnica dei minimi quadrati parziali è simile alla regressione delle componenti

principali, ma le variabili scelte si correlano sia alla varianza delle variabili indipendenti (x) che a quella delle variabile dipendente (y). L'algoritmo si può così definire:

- Standardizzare ogni x_j per avere media 0 e varianza 1;
- Per $m = 1, 2, \ldots, p$:

  a) $z_m = \sum_{j=1}^{p} \hat{\varphi}_{mj} x_j^{m-1}$, dove $\hat{\varphi}_{mj} = \langle x_j^{(m-1)}, y \rangle$.

  b) $\hat{\theta}_m = \langle z_m, y \rangle / \langle z_m, z_m \rangle$.

  c) $\hat{y}^{(m)} = \hat{y}^{(m-1)} + \hat{\theta}_m z_m$.

  d) rendere ortogonale ogni $x_j^{(m-1)}$ con i rispettivi z_m: $x_j^{(m)} = x_j^{(m-1)} - [\langle z_m, x_j^{(m-1)} \rangle / \langle z_m, z_m \rangle] z_m, j = 1, 2, \ldots, p$.

- $\{z_l\}_1^m$ sono originariamente lineari (x_j), così lo è anche $\hat{y}^{(m)} = X\hat{\beta}^{pls}(m)$. Questi coefficienti lineari possono essere recuperati dalla sequenza di trasformazioni di PLS.

Capitolo 2

La previsione

Prevedere è un'attività che fa parte del nostro quotidiano. Ogni giorno frasi del tipo "che tempo farà oggi?", "che tipo di giornata mi spetterà al lavoro?", "quanto sarà lunga la coda alle poste?"..., occupano molti dei nostri pensieri.
Molti lavoratori basano la loro attività sulle previsioni come ad esempio i metereologi per ovvi motivi o i ristoratori, che devono fare i conti con la previsione di afflusso nei loro locali per poter organizzare al meglio il loro lavoro ottimizzando le risorse. Pensiamo infine agli azionisti che basano costantemente le loro scelte in base a previsioni fondate su determinate considerazioni, ma gli esempi potrebbero essere migliaia. La previsione è probabilmente il pilastro principale su cui poggia la gestione dell'impresa.

Le nostre decisioni sono spesso dettate dalla convinzione che le nostre previsioni si realizzeranno. Una cosa è certa: non esistono previsioni senza errori, solo raramente e casualmente l'errore è nullo. Chi lavora infatti con le previsioni sa bene che l'obiettivo principale è ridurre l'errore di previsione. La maggior parte dei modelli che vengono sviluppati riescono a studiare abbastanza bene il passato, ma il futuro è spesso imprevedibile.

La formulazione della previsione è basata sulla conoscenza e sulla rielaborazione di informazioni derivanti dal passato del fenomeno oggetto di studio: "sono un collegamento tra passato e futuro"(Cipolletta 1992).
Ogni previsione deriva da un insieme preciso di informazioni che possono portare a previsioni condizionali o condizionate da tale insieme. L'esplicitazione delle informazioni che condizionano la previsioni sono oggetto chiave dello studio infatti cambiando condizioni iniziali, lo studioso può ritrovarsi ad osservare scenari diversi. Questa visione ha ampliato il numero di alternative che la previsione deve individuare e che si possono verificare in dipendenza dell'andamento di variabili esogene[1] non controllabili.
La previsione statistica si applica a fenomeni definiti in modo da essere passibili da misurazioni oggettive.
Un argomento connesso alla previsione, ma di cui non ci dilunghiamo perché ovvio, è indubbiamente la qualità dei dati a disposizione.
Per comodità si possono suddividere i metodi statistici di previsione in base al tipo di analisi che viene effettuata e dagli obiettivi che ci si prefigge di raggiungere. La prima suddivisione può essere fatta rispetto all'asse temporale: possiamo distinguere previsioni a breve, medio e lungo termine. Questa prima distinzione risulta essere importante perché nel passare dal breve al lungo periodo, le tecniche previsive devono necessariamente semplificarsi e diminuire in qualità: diventa sempre meno verosimile la presunzione di poter descrivere il futuro a partire da relazioni osservate nel passato.
Si possono inoltre distinguere due grandi gruppi di tecniche previsive: i metodi esogeni e i metodi endogeni. Si parla di metodi endogeni quando la variabile di interesse è prevista ricorrendo a modelli che considerano le sole realizzazioni passate, con l'obiettivo di prolungarle nel tempo. L'ipotesi

[1]Una variabile è esogena se il suo valore è determinato da processi esterni al modello

è che nel passato sia contenuta tutta l'informazione necessaria per definire le dinamiche future. Tra i metodi di previsione endogeni rientrano quelli come il livellamento esponenziale, l'estrapolazione grafica, ed, in generale, i metodi per l'analisi stocastica delle serie temporali.
I metodi esogeni invece sono tali perché la variabile di interesse è prevista facendo ricorso a modelli che considerano alcuni fattori esogeni, o cause, che si ritengono importanti. Il presupposto su cui si basa l'impiego di questi modelli per fare le previsioni è che le medesime cause producono i medesimi effetti. Nell'impostazione di tipo esogeno rientrano i modelli di regressione uni o pluri equazionali.

2.1 Intervalli di previsione

Poiché non è possibile avere delle previsioni perfette, calcolare delle stime puntuali non rappresenta sicuramente il metodo migliore per fare previsione; risulta più efficace e di facile lettura, stimare l'intervallo dei valori che potremmo aspettarci di osservare.
Sia $Y = (Y_1, Y_2, \ldots, Y_n)$ un vettore casuale continuo osservabile; il problema della previsione, concernente un ulteriore variabile casuale continua Z, consiste nel dare una stima della osservazione futura z sulla base di un campione y derivante da Y. Supponiamo quindi di avere un vettore (Y, Z) casuale con distribuzione congiunta in funzione di un parametro ignoto, dove Y rappresenta la variabile osservabile mentre Z indica la variabile futura. La distribuzione congiunta di Z e Y si presume essere conosciuta, a meno di un parametro k-dimensionale.
Nel caso in cui Z sia una variabile casuale unidimensionale, una possibile soluzione può essere data in termini di limiti di previsione, cioè funzioni di

$h_\alpha(y)$ tali che:

$$P_{Y,Z}\{Z \leq h_\alpha(Y)\} = \alpha$$

per ogni $\alpha \in (0,1)$. La probabilità così descritta, rappresenta la probabilità di copertura di $h_\alpha(y)$ e si intende calcolata rispetto alla distribuzione congiunta di Y e Z.

Una soluzione semplice consiste nel considerare come limiti di previsione i quantili della distribuzione condizionata di Z dato $Y = y$ dopo aver sostituito il parametro non noto con una sua stima[2] opportuna. Sfortunatamente questa soluzione, detta estimativa, ha una probabilità di copertura diversa da α.

In realtà, in alcuni casi particolari, è possibile trovare una soluzione esatta attraverso la quantità pivotale, che è una funzione di Y e Z ma la sua distribuzione non dipende dal parametro ignoto.

Per quanto riguarda tutti i casi in cui non si possa calcolare la quantità pivotale, si cerca di correggere la funzione di distribuzione estimativa per ridurre al minimo gli errori dei limiti di previsione. Questo è un argomento ampiamente discusso che vede tutt'ora impegnati diversi studiosi[3].

2.1.1 Il caso Normale

Consideriamo ora il problema della previsione nel modello di regressione lineare multipla: $Y_1, \ldots, Y_n$, $Z = Y_{n+i}$ con $n \geq 1$ sono normali indipendenti con σ^2 ignoto e media dipendente dalle variabili esplicative X_j, $j = 1, \ldots, p$. Vogliamo prevedere Z in corrispondenza di p valori fissati di $X_j = x_{n+1,j}$

[2]La stima dei parametri è il procedimento tramite il quale, sulla base delle osservazioni campionarie, si assegna al parametro incognito della popolazione un valore o un insieme di valori. Nel primo caso si parla di stima puntuale e nel secondo di stima per intervalli.

[3]Barndorff-Nielsen e Cox (1996), Vidoni (1998), Ueki e Fueda (2007)

con $j = 1, \ldots, p$ misurati su una unità statistica non inclusa nel campione. Costruiamo quindi un intervallo di previsione, di livello $1 - \alpha$ per

$$\begin{aligned} Z &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{n+1,j} + \varepsilon_{n+1} \\ &= x'_{n+1}\beta + \varepsilon_{n+1}, \end{aligned} \tag{2.1}$$

dove $x_{n+1} = (1, x_{n+1,1}, \ldots, x_{n+1,p})'$.
Essendo il modello gaussiano, possiamo scrivere:

$$Z \sim N(x'_{n+1}\beta, \sigma^2)$$

Prevediamo Z con $\hat{Z} = x'_{n+1}\hat{\beta}_j$, l'errore di previsione si può scrivere come:

$$Z - x'_{n+1}\hat{\beta} \sim N(0, \frac{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}{X\sigma^2})$$

La quantità pivotale sarà data da:

$$\frac{Z - x'_{n+1}\hat{\beta}}{\sqrt{\hat{\sigma}^2(1 + x'_{n+1}(X'X)^{-1}x_{n+1})}} \sim t_{n-p-1}.$$

Ponendo $\hat{\sigma}_z = \hat{\sigma}\sqrt{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}$, possiamo così scrivere l'intervallo di previsione:

$$[x'_{n+1}\hat{\beta} - t_{n-p-1,1-\alpha}\hat{\sigma}_z, x'_{n+1}\hat{\beta} + t_{n-p-1,1-\alpha}\hat{\sigma}_z].$$

2.2 Importanza del modello

La maggior parte delle applicazioni statistiche che partono dai dati osservati, iniziano lo studio con la selezione del modello per poi procedere alle stime dei parametri e alla fase di inferenza. Il modello utilizzato per analizzare i dati viene determinato solo dopo l'esame dei dati stessi.
Le proprietà degli stimatori e le procedure di inferenza (come predittori, test,

o intervalli di confidenza) sono ovviamente influenzati dall'incertezza di tale modello.
La teoria tradizionale della statistica parametrica ha come presupposto cardine che il modello sia noto al ricercatore prima dell'analisi statistica, eccetto il vero valore del vettore dei parametri. Di conseguenza, le effettive proprietà statistiche degli stimatori e le procedure di inferenza in presenza di incertezza modello, non sono descritte dalla teoria tradizionale che assume un modello a priori dato.
La scelta dei regressori da includere nel modello rappresenta una fase cruciale nell'analisi della regressione multipla. In fase di costruzione del modello, si possono commettere inoltre due tipi di errori derivanti dalla scelta dello stesso:

1. omissione di regressori rilevanti;

2. inclusione di regressori irrilevanti.

Non bisogna dimenticare che un modello è tanto più utile, in quanto più facilmente interpretabile, quanto è più parsimonioso (poche variabili esplicative). L'inserimento di variabili irrilevanti può ridurre l'affidabilità delle stime, e tale effetto si riflette sulle procedure di costruzione degli intervalli di confidenza e verifica d'ipotesi.

Capitolo 3

Simulazione

Nei capitoli precedenti abbiamo fatto una panoramica teorica di varie tecniche di selezione delle variabili nello studio di un modello lineare.
In questa sezione la nostra attenzione si concentrerà nella parte principe di questo lavoro: lo studio di una simulazione per gli intervalli di previsione ottenuti con due diversi metodi di selezione delle variabili: il Lasso e la selezione stepwise.
In generale le stime ottenute attraverso il metodo dei minimi quadrati (OLS), forniscono una buona accuratezza, ma peccano nella generalizzazione in quanto tendono semplicemente ad interpolare i punti dell'insieme di osservazione a disposizione. Risulta quindi molto interessante mettere a confronto il metodo dei minimi quadrati con un metodo riconducibile ad un problema dei minimi quadrati ma vincolato in termini di regolarizzazione sui coefficienti. Il lasso, infatti, riduce i coefficienti di regressione vincolandone la norma L_1 e permette di selezionare le variabili. Tibshirani (1996) mostra infatti che:

$$\hat{\beta}_j^{lasso} = sign(\hat{\beta}_j^{OLS})(|\hat{\beta}_j^{OLS}| - \lambda_1)^+$$

Abbiamo quindi ritenuto opportuno mettere a confronto il metodo dei minimi quadrati, il Lasso ed infine un metodo consigliato quando si è in presenza si

numerose variabili, per la stima di modelli lineari mutivariati e generalizzati: lo stepwise. Per potere analizzare e discutere tali metodi, abbiamo scelto lo studio attraverso la simulazione dei dati.

3.1 Costruzione del modello

Per prima cosa abbiamo costruito un data set di dati proveniente da 22 regressori realizzando così la nostra matrice X:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,22} \\ x_{2,1} & x_{2,2} & \dots & x_{2,22} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,22} \end{bmatrix} \tag{3.1}$$

La composizione della variabile risposta Y è stata effettuata attraverso una composizione lineare di soli 5 regressori dei 22:

$$Y = \beta_1 X_1 + \beta_3 X_3 + \beta_8 X_8 + \beta_{10} X_{10} + \beta_{12} X_{12} + \beta_{18} X_{18} + \varepsilon$$

I valori dei regressori x_{ij} solitamente sono dei valori noti in quanto raccolti in fase di sperimentazione, ma in questo studio verranno simulati attraverso diverse distribuzioni o combinazioni degli stessi come indicato in tabella 3.1.

$y = (y_1, \dots, y_n)'$ è il vettore contenente le n osservazioni della variabile risposta ed è stato calcolato combinando la matrice di regressione con un vettore di parametri di regressione anch'esso dato:

$$\beta = (4, 0, 3, 0, 0, 0, 0, 0.4, 0, 7, 0, 1, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0)$$

A questo punto, avendo costruito il modello, conosciamo tutto. Risulta quindi interessante dimenticare il vettore β e studiare i dati attraverso i tre modelli di selezione delle variabili per confrontarne la bontà. In realtà, il vettore

Regressore	Distribuzione	Regressore	Distribuzione
X_1	cost.1	X_{12}	Po(45)
X_2	N(80,30)	X_{13}	$x_{10} * x_{11}$
X_3	N(70,10)	X_{14}	U(40,70)
X_4	$x_2 * x_3$	X_{15}	N(90,10)
X_5	Bin(50,0.5)	X_{16}	$x_4 * x_{11}$
X_6	W(28,73)	X_{17}	Po(52)
X_7	B(4,5)	X_{18}	U(1,100)
X_8	N(60,40)	X_{19}	F(53, 68)
X_9	$x_3 + x_7 * x_8$	X_{20}	$x_{12} * x_{18}$
X_{10}	G(0.04)	X_{21}	T(81)
X_{11}	$x_{10} * x_3 + x_2$	X_{22}	$x_{17} * x_{21}$

Tabella 3.1: distribuzione variabili esplicative

Y non è una semplice combinazione lineare tra le esplicative e i relativi parametri, ma dipende anche dalla variabile errore ε. Si assume che:

$$E(\varepsilon_i) = 0, \qquad Var(\varepsilon_i) = \sigma^2, \qquad Cov(\varepsilon_i, \varepsilon_j) = 0,$$

Infine, per poter sottoporre a test le ipotesi sui parametri del modello assumeremo anche che:

$$\varepsilon_i \sim N(0, \sigma^2) \qquad i = 1, 2, \ldots, n$$

Se consideriamo congiuntamente le quattro assunzioni, possiamo riassumerle come segue:

$$\varepsilon_i \sim IN(0, \sigma^2) \qquad i = 1, 2, \ldots, n;$$

dove I sta ad indicare che gli errori sono indipendentemente distribuiti.
Stiamo assumendo le X non aleatorie.

3.2 Gli intervalli di previsione

La costruzione di intervalli di previsione risulta essere il nostro obiettivo principale e come abbiamo già annunciato, per completezza, vogliamo confrontare gli intervalli ottenuti dal lasso, dai minimi quadrati e dallo stepwise. Per studiare gli intervalli di previsione abbiamo implementato in R una funzione che abbiamo denominato "simulazione ". Questa funzione prende come input la numerosità del campione (n), il numero di replicazione Monte Carlo ($repl$) e il livello di significatività ($alpha$) e restituisce un vettore con di dimensione 4 con le probabilità di copertura degli intervalli di previsione ad un livello α dei 4 modelli.
I passaggi salienti della funzione descritta in appendice sono:

1. costruzione della matrice X $(n \times 22)$. A tal fine abbiamo creato una funzione "osservazione"che genera n numeri causali provenienti dalle singole distribuzioni dei regressori come descritti nella sezione precedente. Le distribuzioni sono molto variegate sia nella forma che nei parametri, per questo motivo abbiamo fatto in modo che la funzione "osservazione"restituisca non solo la matrice delle osservazioni, ma che queste siano standardizzate;

2. implementazione del modello con i veri parametri;

3. utilizzo di un ciclo *for* per replicare *repl* volte i seguenti passaggi:

   - studio dei dati attraverso la stima dei minimi quadrati del modello lineare completo ossia il modello che prende in considerazione tutti i regressori senza scartarne nemmeno uno. Previsione su una nuova unità x_{n+1}. Calcolo del limite superiore dell'intervallo

di previsione della nuova unità a disposizione col modello appena ottenuto;

- studio dei dati attraverso il lasso. La scelta del parametro di *shrinkage* è stata fatta attraverso la *cross-validation*. La *cross-validation* è una tecnica che consiste nel suddividere la matrice dei dati in k sottoinsiemi; per ogni sottionsieme viene calcolato il complementare da cui si stimano i parametri (*training set*), mentre sul sottoinsieme accantonato (*validation set*) vengono testati i risultati ottenuti. Previsione su una nuova unità x_{n+1}. Calcolo del limite superiore dell'intervallo di previsione della nuova unità a disposizione col modello appena ottenuto;
- studio dei dati attraverso la regressione stepweise. Ottenuta la stima dei parametri, abbiamo costruito il limite superiore dell'intervallo di previsione su una nuova unità x_{n+1}.

4. calcolo quante volte, in percentuale, y_{n+1} risulta inferiore al limite superiore di previsione.

A questo punto abbiamo lanciato la funzione con diverse combinazioni di input per vedere come varia la probabilità di copertura dei limiti di previsione al variare della numerosità del campione e del livello di copertura α. I risultati ottenuti sono nella tabella 3.2.

n	repl	*alpha*	OLS m.Completo	lasso	stepwise
50	10000	0.90	0.9011	0.899	0.9343
50	10000	0.95	0.9458	0.955	0.9878
50	10000	0.99	0.9911	0.9898	0.9985
100	5000	0.90	0.9038	0.9030	0.9366
100	5000	0.95	0.9548	0.949	0.971
100	5000	0.99	0.988	0.9872	0.9968
100	10000	0.90	0.9013	0.9021	0.9276
100	10000	0.95	0.9490	0.9519	0.9621
100	10000	0.99	0.9898	0.9895	0.9948

Tabella 3.2: Risultati Simulazione

3.3 Conclusioni

In questo lavoro abbiamo focalizzato la nostra attenzione sulla previsione. Ci siamo particolarmente concentrati sugli intervalli di previsione dal lasso, dai minimi quadrati e dallo stepwise. Tutti i modelli, guardando la tabella 3.2, soddisfano le aspettative: all'aumentare delle repliche l'*alpha* stimato si avvicina sempre più a quello vero e lo *standard error* relativo è sempre inferiore a 0.0029.

Nonostante ciò, si possono comunque avanzare delle critiche:

- Il modello completo è di difficile interpretazione a causa della ridondanza di regressori, presenta inoltre un limite importante dato dall'instabilità delle stime;
- Il lasso oltre a dare risultati in linea con le aspettative presenta gli intervalli di previsione più stretti rispetto ai minimi quadrati[1]. Questo

[1]Risultato dovuto alla varianza: il lasso ha varianza inferiore.

risultato è particolarmente confortante visto che il lasso come abbiamo già esposto è riconducibile ad una rivisitazione migliorativa del metodo dei minimi quadrati.

- Lo stepwise, che rappresenta il modello più conoscito in quanto efficace ed efficiente, ha ottenuto i risultati meno soddisfacenti, dimostrando così i suoi limiti.

Il risultato più importante della simulazione è la conferma che la scelta del modello influenza la previsione.
La numerosità dei metodi per la selezione del modello presenti in letteratura, è argomento attuale in quanto come abbiamo appena evidenziato anche attraverso questa prima analisi, bisogna riporre la massima attenzione quando si fa inferenza.

Appendice A

Comandi di R

```
osservazioni<-function(num){
 x1<-rep(1,num)
 x2<-rnorm(num,80,30)
x2<-(x2-mean(x2))/sqrt(var(x2))
 x3<-rnorm(num,70,10)
x3<-(x3-mean(x3))/sqrt(var(x3))
 x4<-x2*x3
x4<-(x4-mean(x4))/sqrt(var(x4))
 x5<-rbinom(num,50,0.5)
x5<-(x5-mean(x5))/sqrt(var(x5))
 x6<-rweibull(num,28,73)
x6<-(x6-mean(x6))/sqrt(var(x6))
 x7<-rnorm(num,4,5)
x7<-(x7-mean(x7))/sqrt(var(x7))
 x8<-rnorm(num,60,40)
x8<-(x8-mean(x8))/sqrt(var(x8))
 x9<-x3+x7*x8
```

```
x9<-(x9-mean(x9))/sqrt(var(x9))
 x10<-rgamma(num,0.04)
x10<-(x10-mean(x10))/sqrt(var(x10))
 x11<-x10*x3+x2
x11<-(x11-mean(x11))/sqrt(var(x11))
 x12<-x2*x7
x12<-(x12-mean(x12))/sqrt(var(x12))
 x13<-x10*x11
x13<-(x13-mean(x13))/sqrt(var(x13))
 x14<-runif(num,40,70)
x14<-(x14-mean(x14))/sqrt(var(x14))
 x15<-rnorm(num,90,10)
x15<-(x15-mean(x15))/sqrt(var(x15))
 x16<-x4*x11
x16<-(x16-mean(x16))/sqrt(var(x16))
 x17<-rpois(num,52)
x17<-(x17-mean(x17))/sqrt(var(x17))
 x18<-rnorm(num,18,56)
x18<-(x18-mean(x18))/sqrt(var(x18))
 x19<-rf(num,53,68)
x19<-(x19-mean(x19))/sqrt(var(x19))
 x20<-x12*x18
x20<-(x20-mean(x20))/sqrt(var(x20))
 x21<-rt(num,81)
x21<-(x21-mean(x21))/sqrt(var(x21))
 x22<-x17*x21
x22<-(x22-mean(x22))/sqrt(var(x22))
```

```
X<-data.frame(x1=x1,x2=x2,x3=x3,x4=x4,x5=x5,x6=x6,
x7=x7,x8=x8,x9=x9,x10=x10,x11=x11,x12=x12,x13=x13,
x14=x14,x15=x15,x16=x16,x17=x17,x18=x18,x19=x19,
x20=x20,x21=x21,x22=x22)

return(X)}



simulazione<-function(num, repl, alpha){
contatore_A<-rep(0,repl)
contatore_LASSO<-rep(0,repl)
contatore_STEP<-rep(0,repl)
beta<-c(4,0,3,0,0,0,0,0.4,0,7,0,1,0,0,0,0,0,2,0,0,0,0)

for(i in 1:repl)
 {
## aggiornamento dei dati ad ogni replicazione
DATI<-osservazioni(num+1)
x_n1<-DATI[num+1,]
X<-DATI[1:num,]
y_1<-beta[1]*x_n1$x1+x_n1$x3*beta[3]+beta[8]*x_n1$x8+
   beta[10]*x_n1$x10+beta[12]*x_n1$x12+beta[18]*x_n1$x18+
   rnorm(1)
Xin<-svd(t(as.matrix(X))%*%as.matrix(X))
```

```
Xin<- Xin$v%*%diag(1/Xin$d)%*%t(Xin$u)
radice<-sqrt(1+(as.matrix(x_n1)%*%Xin%*%t(as.matrix(x_n1))))
den<-NULL
prova<-NULL
er<-rnorm(num)

## creazione modello simulato

y<-beta[1]*X$x1+X$x3*beta[3]+beta[8]*X$x8+beta[10]*X$x10+
   beta[12]*X$x12+beta[18]*X$x18+er

## studio A: modello completo

A<-lm(y~X$x2+X$x3+X$x4+X$x5+X$x6+X$x7+X$x8+X$x9+X$x10+
   X$x11+X$x12+X$x13+X$x14+X$x15+X$x16+X$x17+X$x18+X$x19+
   X$x20+X$x21+X$x22, x=T)
beta_A<-A$coefficient
Y_A<-as.matrix(X)%*%beta_A
Y_A1<-as.matrix(x_n1)%*%beta_A
s_A<-sqrt(sum(A$res^2)/(length(Y_A)-length(beta_A)))
## calcolo il quantile del modello completo
qp_A<-Y_A1+qt(alpha,num-length(beta_A))*s_A*radice

## studio lasso

matrice<- as.matrix(X[-1])
LASSO<-cv.glmnet(matrice,y,,family="gaussian", alpha=1)
```

```
Y_LASSO<-predict(LASSO, matrice, s=LASSO$lambda.min,
  type="response")
s_LASSO<-sqrt(sum((Y_LASSO-y)^2)/(length(Y_LASSO)-
  sum(coef(LASSO)[,1]>0)))
Y_LASSO1<-as.matrix(x_n1)%*%as.matrix(coef(LASSO))
## calcolo il quantile del lasso
qp_LASSO<-Y_LASSO1+qt(alpha,num-sum(coef(LASSO)[,1]>0))*
  s_LASSO*radice

## studio stepwise

mod<-lm(ys~Xs$x2+Xs$x3+Xs$x4+Xs$x5+Xs$x6+Xs$x7+Xs$x8+Xs$x9
  +Xs$x10+Xs$x11+Xs$x12+Xs$x13+Xs$x14+Xs$x15+Xs$x16+Xs$x17
  +Xs$x18+Xs$x19+Xs$x20+Xs$x21+Xs$x22, x=T)
STEP<-step(mod, direction="both", trace=F)
beta_STEP<-coef(STEP)
new=as.data.frame(x_n1)
Y_STEP1<-predict(STEP, new, interval="prediction")
s_STEP<-sqrt(sum(STEP$res^2)/(length(Y_STEP)-length(beta_STEP)))
## calcolo il quantile del modello completo
qp_STEP<-Y_STEP1+qt(alpha,num-length(beta_STEP))*s_STEP*radice



## aggiorno i contatori
if(y_1<=qp_A){
contatore_A[i]<-1}
```

```
if(y_1<=qp_LASSO){
contatore_LASSO[i]<-1}

if(y_1s<=qp_STEP){
contatore_STEP[i]<-1}
##fine ciclo for
}
se_A<-sqrt(var(contatore_A)/repl)
se_LASSO<-sqrt(var(contatore_LASSO)/repl)
se_STEP<-sqrt(var(contatore_STEP)/repl)
Ris_A<-(sum(contatore_A))/repl
Ris_LASSO<-(sum(contatore_LASSO))/repl
Ris_STEP<-(sum(contatore_STEP))/repl

 return(c(Ris_A,se_A, Ris_LASSO, se_LASSO, Ris_STEP, se_STEP))
}
```

Bibliografia

[1] ANDERSON T., *An Introduction to Multivariate Statistical Analysis*, 3rd ed., Wiley, New York, 2003.

[2] BAIR E., HASTIE T., PAUL D., TIBSHIRANI R., *Prediction by supervised principal components*, in Journal of the American Statistical Association, 2006.

[3] BATTAGLIA FRANCESCO, *Metodi di previsione statistica*, Springer, 2007.

[4] CHRIS CHATFIELD, *Model uncertainty and forecast accuracy*, in journal of Forecasting,1996.

[5] EFRON B., HASTIE T., JOHNSTONE I., TIBSHIRANI R., *Least angle regression (with discussion)*, in Annals of Statistics 2004.

[6] FARCHIONE D., KABAILA P., *Confidence intervals for the normal mean utilizing prior information*. in Stat. Probab. Lett. 2008.

[7] GIOVANNI FONSECA, FEDERICA GIUMMOLÈ, PAOLO VIDONI *A note about calibrated prediction regions and distributions* in Journal of Statistical Planning and Inference, 2012.

[8] GIOVANNI FONSECA, FEDERICA GIUMMOLÈ, PAOLO VIDONI *Calibrating predictive distributions* in Journal of Statistical Computation and Simulation, 2012

[9] FRIEDMAN J., HASTIE T., HOEFLING H., TIBSHIRANI R., *Pathwise coordinate optimization*, in Annals of Applied Statistics, 2007.

[10] FRIEDMAN J., HASTIE T., ROSSET S., TIBSHIRANI R., ZHU J., *Discussion of three boosting papers* by Jiang, Lugosi and Vayatis, and Zhang, in Annals of Statistics, 2004.

[11] FRIEDMAN J., HASTIE T., TIBSHIRANI R., *Regularization paths for generalized linear models via coordinate descent*, in Journal of Statistical Software, 2010.

[12] F*RIEDMAN J., HASTIE T., TIBSHIRANI R., Response to Mease and Wyner, Evidence contrary to the statistical view of boosting*, in Journal of Machine Learning Research, 2008a.

[13] FRIEDMAN J., HASTIE T., TIBSHIRANI R., *Sparse inverse covariance estimate on with the graphical lasso*, in Biostatistics, 2008.

[14] GIRI, K., KABAILA, P., *The coverage probability of confidence intervals in 2 factorial experiments after preliminary hypothesis testing*. in Austral. N. Z. J. Stat. 2008.

[15] HASTIE T., TAYLOR J., TIBSHIRANI R., WALTHER G., *Forward stagewise regression and the monotone lasso*, in Electronic Journal of Statistics1, 2007.

[16] HASTIE T., TIBSHIRANI R., *Efficient quadratic regularization for expression arrays*, in Biostatistics5(3), 2004.

[17] HASTIE T., TIBSHIRANI R., FRIEDMAN J., *A note on Comparison of model selection for regression by Cherkassky and Ma*, in Neur al computation15, 2003.

[18] HASTIE T., TIBSHIRANI R., *Independent components analysis through product density estimation*, in S. T. S. Becker and K. Obermayer (eds), *Advances in Neural Information Processing Systems* 15, MIT Press, Cambridge, MA, 2003.

[19] HASTIE T., ZHU J., *Discussion of Support vector machines with applications by Javier Moguerza and Alberto Munoz*, in Statistical Science, 2006.

[20] HOEFLING H., TIBSHIRANI R., *Estimation of sparse Markov networks using modified logistic regression and the lasso*, submitted. Hoerl, A. E. and Kennard, R. (1970). Ridge regression, biased estimation for non orthogonal problems, Technometrics, 2008.

[21] KABAILA P., GIRI K., *Confidence intervals in regression utilizing prior information*. in J. Stat. Planning Infer, 2009.

[22] KABAILA P., GIRI K., *Large sample confidence intervals for the treatment difference in a two period crossover trial, utilizing prior information*. Statist. in Probab. Lett, 2009.

[23] KABAILA P., GIRI K., *Upper bounds on the minimum coverage probability of confidence intervals in regression after variable selection*. in Austral. N. Z. J. Stat.,2009.

[24] KABAILA P., LEEB H., *On the large sample minimal coverage probability of confidence intervals after model selection*. in J. Amer. Statist. Assoc., 101, 2006.

[25] KABAILA P., *On the coverage probability of confidence intervals in regression after variable selection*. in Austral. N. Z. J. Stat., 2005.

[26] KABAILA P., SYUHADA K., *The relative efficiency of prediction intervals*. in Comm. Statist. Theory Methods, 2007.

[27] KABAILA P., TUCK J., *Confidence intervals utilizing prior information in the BehrensFisher problem*. Austral. in N. Z. J. Stat.,2008.

[28] LEEB H., *Conditional predictive inference post model selection*. in Ann. Statist., 2009.

[29] LEEB H., *The distribution of a linear predictor after model selection: conditional finite-sample distributions and asymptotic approximations*. in Journal of Statistical Planning and Inference (2005)

[30] PAUL D., BAIR E., HASTIE T., TIBSHIRANI R., *Preconditioning for feature selection and regression in high dimensional problems*, in Annals of Statistics36(4), 2008.

[31] TIBSHIRANI R., *Regression shrinkage and selection via the lasso*, in Journal of the Royal Statistical Society, 1996.

[32] TIBSHIRANI R., SAUNDERS M., ROSSET S., ZHU J., KNIGHT K., *Sparsity and smoothness via the fused lasso*, in Journal of the Royal Statistical Society, 2005.

[33] ZOU H., HASTIE T., TIBSHIRANI R., *On the degrees of freedom of the lasso*, in Annals of Statistics, 2007.

[34] ZOU H., *The adaptive lasso and its oracle properties*, in Journal of the American Statistical Association, 2006.