UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA ECONOMIA E FINANZA

Tesi di laurea

Selezione delle variabile per migliorare le Previsioni: il LASSO

Relatore:

Prof. Guido Masarotto

Laureanda:

Djeuteu Celimene

Anno Accademico 2009-2010

Dedico questa tesi ai miei parenti che mi hanno sopportato e aiutato in modo indiretto al compimento di questo importante lavoro.

Inoltre a tutti quelli che mi hanno permesso di portare avanti questi studi.

INDICE

INTRODUZIONE

In genere, l'obiettivo delle scienze statistiche consiste nel trovare quale modello probabilistico possa generare dei dati tratti da una popolazione, in modo tale da fornire una descrizione sintetica del fenomeno oggetto di studio, che ne permetta l'interpretazione e la previsione.

Dato un vettore casuale delle risposte Y, l'analisi di regressione permette di selezionare tra un insieme di variabili concomitanti, quello parsimonioso per una efficiente previsione della variabile risposta. Tuttavia essa presenta alcune problematiche in quanto a volte ci si trova davanti a un modello quasi spurio, ossia un modello con variabili aggiunte non necessarie per la stima del modello stesso, oppure ad un campo con numero di osservazioni abbastanza grande.

Gli algoritmi di selezione delle variabili, detti stepwise regression (backward e forward regression), ridge regression e all-subsets regression, non includono sempre accurate predizioni, ma presentano risultati parziali instabili nei test ipotesi durante o dopo la selezione delle variabili.

Per cercare di ovviare a ciò, nel 1996 Tibshirani sviluppò un metodo efficiente che migliorò i problemi di stabilità e di predizione, detto "LASSO", una versione

interessante di minimi quadrati ordinari che limita la somma dei coefficienti di regressione in assoluto. Dalla sua scoperta, però, tale metodo non è stato mai usato perché presenta comunque difetti nell'implementazione.

Nel 2004, EFron Bradley, basandosi sempre il suo lavoro sulla risoluzione di problemi reali, introdusse il problemapiù importante nella regressione lineare, la selezione dei regressori. Questa riflette la selezione dei variabili tra un insieme di candidati, stimando i parametri per quelli ultimi per fare inferenza e per calcolare intervalli di confidenza.

In più, egli dimostrò l'esistenza di un legame tra il LASSO e un altro chiamato LAR, un nuovo algoritmo di selezione nei forward regression, sviluppando una struttura algoritmica che mette li insieme e procura un'implementazione rapida, a cui viene dato perciò il nome di "LARS" (la "s"è sottointesa come la stagewise o il LASSO). LARS è un metodo potenzialmente rivoluzionario, perfezionando difetti degli altri metodi di selezione e offrendo in più grafici che mostrano i passi nella complessità dei modelli e un semplice database con regole per determinare il livello ottimo di complessità che evita al massimo pregiudizi nei test d'ipotesi.

Nel mio lavoro, quindi, parlerò innanzi tutto della regressione lineare e di tutti gli aspetti ad essa collegati, per soffermarmi poi sulla selezione delle variabili con i metodi stepwise, LASSO, e più brevemente LAR. L'ultima parte sarà focalizzata sul metodo LARS, che richiede solo lo stesso ordine di grandezza di sforzo

computazionale, come i minimi quadrati ordinari applicati all'insieme completo di variabili esplicative.

Capitolo 1

LA REGRESSIONE LINEARE

1.1 Definizione

In statistica, la regressione lineare si riferisce ad un approccio di modellizzazione del rapporto tra una variabile scalare Y e una o più variabili denominate X. Nella regressione lineare, i modelli dei parametri ignoti sono stimati sulla base dei dati utilizzando funzioni lineari. Tali modelli sono chiamati modelli lineari. Più comunemente, la regressione lineare si riferisce ad un modello in cui la media condizionata di Y, dato il valore di X, è una funzione affine di X. Meno comunemente, la regressione lineare potrebbe riferirsi ad un modello in cui la mediana o qualche altro quantile della distribuzione condizionata di Y dato X, è espresso in una funzione lineare di X. Come tutte le forme di regressione, quella lineare si concentra sulla distribuzione di probabilità condizionata di Y dato X,

piuttosto che sulla probabilità condizionata di Y e X, che è il dominio di analisi multivariata.

La regressione lineare è stata il primo tipo di analisi di regressione ad essere studiata con rigore e ad essere ampiamente utilizzata nelle applicazioni pratiche. Questo perché, da una parte, i modelli che dipendono linearmente dai loro parametri ignoti sono più facili da stimare rispetto ai modelli che sono legati in modo non lineare per i loro parametri, e dall'altra perché le proprietà statistiche degli stimatori risultanti sono più facili da determinare.

Due grandi passi sono presenti nella regressione lineare: se l'obiettivo è la previsione, la regressione lineare può essere utilizzata per adattare un modello predittivo di un insieme di dati osservati Y usando le variabile esplicative X. Dopo lo sviluppo di un tale modello, possiamo prevedere o calcolare le Y per i nuovi valori di X. Quindi per un X dato senza il relativo Y, il modello adattato può essere usato per fare una previsione del valore di Y. Data una variabile Y e una serie di variabili $X_1,...,X_P$ che possono essere correlate alla y, l'analisi di regressione lineare può essere applicata sia per quantificare la forza della relazione tra Y e le x_j, e valutare quali delle x_j possono non avere rapporto con la y, sia per identificare quali sottoinsiemi dei x_j contengono informazioni ridondanti su y. Una volta che uno di loro è noto, gli altri non sono più informativi.

I modelli di regressione lineare sono spesso stimati con il metodo dei minimi quadrati, ma possono essere modellati anche in altri modi, come ad esempio riducendo al minimo la "mancanza di adattamento" in qualche altra norma. Al contrario, l'approccio dei minimi quadrati può essere utilizzato per adattare i modelli che non sono lineari. Per cui, i termini "minimi quadrati" e "modello lineare", sono strettamente collegati, ma non sono sinonimi.

1.2 Formulazione del modello di regressione lineare

Dato un insieme di dati $Y\ \{y_{i,}\ x_{i1,}\ldots,x_{ip}\}_{i=1}^{n}$ unità statistiche, una regressione lineare assume che la relazione tra la variabile d'interesse y_i e il p-vettore dei regressori x_i, sia approssimativamente lineare. Questo rapporto approssimativo è modellato attraverso un cosiddetto "termine di disturbo" ε_i, una variabile casuale non osservata che aggiunge errori tra la variabile dipendente e il regressore. Per cui si può scrivere

$y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \quad , \quad i = 1,\ldots.n$ ε_i sono identicamente distribuiti come une normale di media 0 e di varianza di media σ^2, che uguale anche la varianza della variabile risposta, y_i.

In forma matriciale $Y = X\beta + \varepsilon$

Dove

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Detto ciò, ci sono alcuni note da fare:

- le variabili indipendenti sono assunti stocastici, e misurati senza errore da un punto di visto sperimentale; più in generale sono assunti fissati. A volte uno dei regressori può essere una funzione non lineare di un altro regressore o dei dati, come nella regressione polinomiale. Dato che il problema di regressione consiste nella determinazione della legge di probabilità di y_i rispetto a x_i, definiamo la parte sistematica della distribuzione delle y_i, il valore atteso $E(y_i) = \mu_i$, come una funzione di x_i tramite parametri di regressione. Questi ultimi verranno stimati, per poi fare inferenze. Si nota che $\mu_i = (x_i, \beta_i)$ e che i β_i appartengono a un sottoinsieme di dimensione p e non dipendono da $n.$;
- y_i sono variabili quantitative. La loro variabilità è costante e indipendente dalle condizioni sperimentali. Quindi $\sigma_i = v(y_i) = E(y_i, \mu_i)^2$, $i = 1...n$: si parla di omoschedasticità. Sono assunti correlati oppure indipendenti rispetto alla

distribuzione congiunta di Y, che sarebbe normale di media μ_i, vincolata e di varianza σ^2. La decisione su quale variabile modellare come variabile

indipendente può essere basata sulla presunzione che il valore di una delle variabili sia causato, o direttamente influenzato, da altre variabili. In alternativa, si può ricorrere ad un modello operativo per una delle variabili espressa in funzione degli altri. In quest'ultimo caso non c'è bisogno di presunzione di causalità;

- ε_i è il termine d' errore, di disturbo oppure rumore. Questa variabile cattura tutti gli altri fattori che influenzano la variabile dipendente y_i diverso da x_i. Si assume anche che il termine d'errore e i regresso risono correlati; è un passo fondamentale nella formulazione di un modello di regressione lineare, in quanto determinerà il metodo da utilizzare per la stima. Sono indipendenti e identicamente distribuiti come una normale di media 0 e di varianza σ^2.

ASSUNZIONI

Due ipotesi di base sono comuni a tutti i metodi di stima utilizzati in analisi di regressione lineare:

1) La matrice X deve essere di rango p, cioè avere colonne indipendenti. Così dobbiamo avere $p < n$, dove n è la dimensione del campione (questa è una condizione necessaria ma non sufficiente). Se questa condizione non è rispettata si ha una multicollinearità nei regressori. In questo caso il vettore β parametro non sarà identificabile, al massimo si potrà restringere il suo valore per alcuni sottospazi lineari di R_p. In altre parole, ci devono essere abbastanza dati a disposizione rispetto al numero di parametri da stimare. Nel caso contrario, si ricorre ad un sistema di equazioni senza soluzione unica;
2) le variabile esplicative si presumono prive di errori, ossia non concomittante da errori di misurazione. Anche se non è realistico in molte impostazioni, lasciando cadere questa ipotesi, si arriverebbe ad errori molto più elevati.

Prima di fare la stima del modello, e per valori abbastanza grandi di osservazioni, si assume che le distribuzioni della variabile osservata si avvicinino ad una distribuzione normale.

1.3 Stima del modello attraverso i minimi quadrati

Un modello stimato può essere usato per identificare il rapporto tra un unico predittore x_j e la risposta y ,a parità di altri predittori nel modello. In particolare,

l'interpretazione del β_j è il cambiamento atteso in Y per un cambiamento in una sola unità x_j quando le altre sono tenute fisse. L'effetto marginale di x_j di y può essere valutato sulla base di un coefficiente di correlazione.

Numerose procedure sono state sviluppate per stimare parametri e fare inferenza nella regressione lineare. Tuttavia esse differiscono nella semplicità computazionale degli algoritmi (presenza di anomalie nelle soluzioni).

La tecnica la più usata è il metodo dei minimi quadrati (OLS). La robustezza rispetto alle distribuzioni a coda pesante, e ipotesi teoriche necessarie per convalidare auspicabili proprietà statistiche, come la coerenza e l'efficienza asintotica. Il metodo dei minimi quadrati ordinari (OLS) è concettualmente e computazionalmente semplice. Le stime OLS sono comunemente utilizzate per analisi sia in campo sperimentale, sia per l'osservazione dei dati. Il metodo OLS minimizza la somma dei residui al quadrato e conduce ad una espressione in forma chiusa per il valore stimato di β, parametro sconosciuto.

Sia $YN_n(\mu,\sigma^2 I_n), \sigma^2 \geq 0$ dove n è il numero di osservazioni, il metodo dei minimi quadrati si basa sulla funzione di verosimiglianza, $\ell(\mu,\sigma^2;y) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|y-\mu\|^2$ che si ricava Log-differenzziando la funzione di densità,

$$P_Y(y;\mu,\sigma^2) = (1/\sigma\sqrt{2\pi})^n \exp\{-1/2\sigma^2 \sum_{i=1}^{n}(y_i-\mu_i)^2\} \quad i=1...n$$

$$(\mu,\sigma^2) \in V \times (0,+\infty)$$

La stima di massima verosimiglianza di β è ottenuta minimizzando la funzione di log-verosimiglianza

$$\hat{\beta} = (X^T X)^{-1} X^T Y = \left(\frac{1}{n} \sum x_i x_i^{'} \right)^{-1} \left(\frac{1}{n} \sum x_i y_i \right)$$

Perché questa relazione sia vera, la matrice $(X^T X)$ deve avere rango pieno, per poter essere invertibile.

La stima è corretta e consistente se gli errori hanno una varianza minima e sono incorrelati con i covariati, $E[x_i \varepsilon_i] = 0$

Essa è anche efficiente sotto l'ipotesi che gli errori siano omoschedastici, ossia $E[\varepsilon_I^2 / x_i]$ non dipende da i. Questa condizione è valida sia con i dati sperimentali che con quelli osservazionali. Se l'obiettivo è l'inferenza oppure la modellazione predittitiva, le prestazioni delle stime OLS possono essere povere se è presente la multicollinearità, a meno che la dimensione del campione sia grande. Nella regressione lineare semplice, in cui vi è un solo regressore (con una costante), le stime OLS hanno una forma semplice che è strettamente legata al coefficiente di correlazione tra il regressore e la risposta.

La stima della varianza si ottiene considerando nota la media $\mu_{\sigma^2} = \mu = Py$, e risulta essere scritta $\hat{\sigma}^2 = \frac{e^T e}{n} = \frac{(y - \hat{y})^T (y - \hat{y})}{n}$, dove e è il vettore degli errori stimati, ovvero i residui. Sono ortogonali allo spazio formato dalle covariate e alla variabile risposta.

1.4) Verifica d'ipotesi

Se un coefficiente di regressione è pari a zero viene tolto dal modello. Più in generale, se i coefficienti di regressione sono tutti pari a zero allora i regressori sono irrilevanti per l'analisi dei dati y e si ritorna al modello di campionamento casuale semplice.

Sia y una realizzazione di $Y\left(\mu,\sigma^2 I_n\right)$, dove $\sigma \geq 0$, $\mu \in V$, sottospazio p-dimensionale di R^n, generato dalle colonne della matrice X. Le ipotesi fatte su un modello di regressione sono l'uguaglianza a zero di un coefficiente $\beta_r, r = 2...n$ oppure di un insieme di coefficienti (si esclude l'intercetta). L'ipotesi lineare generale è data da

$H_0: \ \mu \in V_o$

in cui V_0 è un sotto spazio di V con dimensione $p_0 < p$. L' ipotesi nulla corrisponde ad

$H_0: \ \beta_r = 0, \ r = 2...p$

generato dalle $p_0 = p - 1$ colonne di X esclusa l'r-esima. L'ipotesi lineare generale permette di trattare unitariamente molti importanti problemi di verifica di ipotesi per i modelli lineari normali.

Si considera l'ipotesi nulla

$H_0: \ \mu \in V_o$

contro l'ipotesi alternativa

$H_1: \ \mu \in V \setminus V_o$

La funzione di verosimiglianza $L(\mu,\sigma^2;y) = (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\|y-\mu\|^2\right\}$ è massima in $(\hat{\mu},\hat{\sigma}^2) = (Py, \|y-Py\|^2/n)$ sotto il modello completo mentre è $(\hat{\mu},\hat{\sigma}^2) = (P_0 y, \|y-P_0 y\|^2/n)$ sotto il modello ridotto.

Il test è equivalente a rifiutare l'ipotesi nulla per valori grandi di

$$F = \frac{\|(P-P_0)y\|^2/(p-p_0)}{\|(I_n-P)y\|^2/(n-p)} = \frac{(\tilde{\sigma}^2-\hat{\sigma}^2)/(p-p_0)}{\hat{\sigma}^2/(n-p)}$$

F esprime dunque la diminuzione relativa nella varianza stimata dell'errore che si ha passando dal modello ridotto al modello completo, corretto con i gradi di libertà.

Una grande diminuzione di F è sintomo di inadeguatezza del modello ridotto.

Il test sulla significatività di un parametro, il test t, equivale al test per verificare l'ipotesi nulla che valga il modello ridotto, avendo solo una sola variabile esplicativa contro l'alternativa che valga il modello completo, composto dal resto delle variabili esplicative. Si tratta di modelli annidati e il test del rapporto di verosimiglianza è equivalente al test che rifiuta l'ipotesi nulla per valori grandi di F.

Capitolo 2

SELEZIONE DELLE VARIABILI

2.1)Introduzione

Scegliere un modello lineare con l'idea di impostarlo ai dati a cui il modello verrà applicato, rileva di un' importanza statistica nel senso che, si sceglie tra le covariate date nella partenza, l'insieme parsimonioso per l'efficiente previsione della variabile risposta.

Nella regressione lineare, i classici metodi di selezione delle variabile sono la backward elemination oppure la forward stepwise selection. In particolare, dato un insieme di p variabili esplicative, la forward stepwise regression sceglie quella che produce una correlazione più alta con la variabile risposta y, che chiameremo x_{j1}, $j = 1,...,p$, e esegue una semplice regressione lineare di y su x_{j1}. Questo produce il più piccolo valore R^2 e lascia un vettore residuo ortogonale al x_{j1}, considerato come il vettore risposta. Si inserisce quindi nel modello uno fra gli altri predittori, quello

che risulta produrre una miglior estimazione, e si ripete il processo di selezione fino ad avere un insieme di k predittori $x_{j1},...,x_{jk}$. La backward elemination, d'altro canto, comincia l'estimazione del modello con tutti i predittori e sequenzialmente toglie le variabili che contribuiscono ad avere un valore piccolo della R^2.

Questi metodi non sono stabili perché possono produrre modelli sbagliati, in quanto un cambiamento nei dati può portare a scegliere una variabile invece di un'altra. Da ciò deriva che le inferenze possano risultare non corrette, visto il numero grande di modelli proposti. Inoltre, se è presente un numero elevato di osservazioni oppure il numero delle variabili esplicative è superiore al numero dei dati, la scelta delle variabili diventa più difficile e perciò si hanno modelli spuri.

Efron promette di produrre modelli interpretabili, predizioni accurate e inferenze approssimativamente non errate: introduce il "LASSO" (Tibshirani 1996).

2.2) LASSO

È una versione limitata del metodo dei minimi quadrati ordinari (OLS), nel senso che minimizza il quadrato dei residui più un termine, $l1$ penalty, sui coefficienti di regressione: limita la somma del valore assoluto dei coefficienti di regressione. Sia il vettore di p regressori, Y vettore risposta e $\beta = (\beta_1,...,\beta_p)$, abbiamo la formulazione seguente

$$\|Y - X\beta\|_2^2 + \theta\sum_{j=1}^{p}|\beta_j|$$

Dato $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)'$, un opportuno vettore di coefficienti stimati per dare il vettore di previsione $\hat{\mu}$

$$\hat{\mu} = \sum_{j=1}^{p} x_j \hat{\beta}_j = X\hat{\beta}$$

Con l'errore quadratico totale $S(\hat{\beta}) = \|y - \hat{\mu}\|^2$.

Sia $T(\hat{\beta}) = \sum_{j=1}^{p}|\hat{\beta}_j|$, la norma assoluta di $\hat{\beta}$, LASSO sceglie $\hat{\beta}$ minimizzando $S(\hat{\beta})$ soggetto di un t vincolato su $T(\hat{\beta})$

LASSO tende a ridurre a 0 i coefficienti OLS uno alla volta. La riduzione spesso migliora la previsione. Quindi per le variabili che sono più valide, che chiaramente devono essere incluse nel modello e dove il restringimento verso lo zero è meno auspicabile, $l1$ penalty stringe meno. Questo è importante per fornire previsioni migliori dei valori futuri.

LASSO ha una proprietà tale per cui si stipula che, per qualsiasi valore vincolato di t, solo un sottoinsieme di repressori hanno un valore diverso dallo zero di $\hat{\beta}_j$.

In conclusione, se la bontà del modello scelto risulta soddisfacente, il problema che si pone è di sapere se i regressori sono quelli più importanti.

2.3) L'algoritmo di LARS

Least Angle Regression è una versione stilizzata della stagewise regression per migliorare il LASSO. È una procedura che utilizza una semplice formula matematica per accelerare i calcoli. Solo k passaggi sono necessari per l'insieme completo di regressori adatti , dove k rappresenta il numero di covariate.

Il funzionamento di LARS è, brevemente, il seguente. Come nella classica selezione forward, esso prevede che si inizi con tutti i coefficienti pari a 0, per poi trovare il predittore, x_{j1}, con la correlazione più alta con la risposta. Prendiamo il più grande passo possibile nella direzione di questa fino al predittore o qualche altro fattore predittivo, x_{j2}, che ha correlazione più elevata con il vettore residuo. Invece di continuare in questo modo, LARS procede in una direzione equiangolo tra i due predittori fino a una terza variabile, x_{j3}, che abbia una correlazione alta con i residui. Si procede poi in maniera equiangolo tra x_{j1}, x_{j2}, x_{j3}, cioè lungo "la direzione di minimo angolo", fino all'entrata della quarta variabile e così via.

La figura sotto illustra la situazione nel caso di due variabile esplicative $X = (x_1, x_2)$

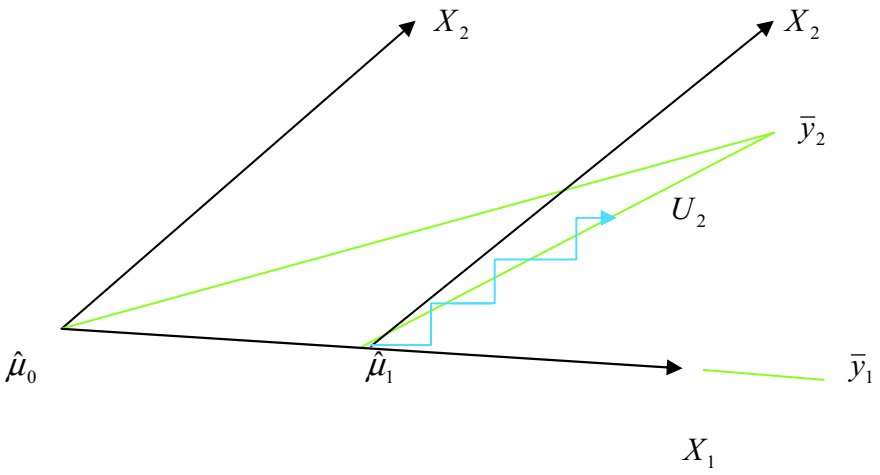


Figura 1: $\bar{y}_2$ è la proiezione di y nello spazio L (x1, x2). A partire da $\hat{\mu}_0 = 0$, il vettore residuo $\bar{y}_2 - \hat{\mu}_0$ ha una correlazione maggiore con X_1 che X_2; la prossima stima di LARS sarà $\hat{\mu}_1 = \hat{\mu}_0 + \hat{\gamma}_1 X_1$, dove γ_1 scelto è tale che $\bar{y}_2 - \hat{\mu}_1$ sia la bissezione dell'angolo formato da X_1 e X_2; allora $\hat{\mu}_2 = \hat{\mu}_1 + \hat{\gamma}_2 u_2$, in cui u_2 è la bisettrice unità; $\hat{\mu}_2 = \bar{y}_2$ nel caso $p = 2$, ma non in quello $p > 2$. Qui LARS tende a LASSO, ma una modifica è necessaria per garantire accordi in più dimensioni.

Tre sono le caratteristiche principali derivate:

1) Una semplice modifica del LARS implementa il LASSO e calcola tutte le stime possibili di LASSO, usando un ordine di grandezza, meno computazionale, minore rispetto ai metodi precedenti. Una diversa modificazione efficiente del LARS implementa la FORWARD STAGEWISE REGRESSION o più comunemente Stagewise, un altro modello di selezione.

2) LARS ha un vantaggio sulla velocità. Infatti raggiunge l'ampia soluzione con le soluzione dei minimi quadrati, usando tutte le variabili in p tappi, vale a dire che in LAR le variabili non vengono mai tolte dal modello una volta entrate.

3)C'è la disponibilità di una statistica C_p per scegliere il numero di passi da usare fino alla soluzione,

$$C_p = \left(\frac{1}{\hat{\sigma}^2}\right)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 - n + 2k$$

Dove k è il numero di tappi e $\hat{\sigma}^2$ rappresenta la varianza stimata dal modello saturo con $n > p$. È basato sul teorema 3 in Bradley Efron (2004), il quale stipula che dopo k tappi di LAR il numero di gradi di libertà $\sum_{i=1}^{n}\text{cov}(\hat{\mu}_i, Y_i)/\sigma^2$ sia approssimativamente k. Questo fornisce una semplice regola per smettere dopo k tappi che minimizza la statistica C_p.

Notiamo che questo k grado libertà è valido solo per la statistica C_p e si rivela falso dal momento in cui i gradi di libertà sono p per il modello completo, ma il numero totale di passi compiuti può superare p. Tuttavia si dimostra empiricamente che, per un risultato intuitivamente plausibile, bisogna che i gradi di libertà siano ben approssimati dal numero di predittori non nulli nel modello. In particolare, né la diminuzione dello scarto quadratico medio, né l'errore di previsione atteso sono

lineari in k (sotto Ipotesi nulla che $\beta_j = 0$ per ogni j). Anche questi due effetti rendono meno accettabili la statistica C_p. Questo criterio è stato introdotto da Mallows (1973) per essere usato in OLS, ma ci si è rivelato un metodo di scelta inconsistente (Shao (1993)), dato che sovrastima il modello.

Loubes and Massart (2004), Madigan and Ridgeway (2004) e Stine (2004) propongono allora un altro criterio di selezione, la "cross-validazione". La cross-validazione è una tecnica statistica di ricampionamento che consiste nel suddividere il campione di partenza in k sottoinsiemi e, ad ogni passo, sulla parte (1/k)-esima del dataset, chiamato training data set, si costruisce il modello, ossia si stimano i parametri di un modello, che verranno misurati dalle loro capacità predittive nella restante parte che costituisce il training dataset.

Questo criterio è utilizzabile in presenza di una buona numerosità del training dataset. Così, per ognuna delle k parti (di solito k = 10) si allena il modello, evitando quindi problemi di overfitting, ma anche di campionamento asimmetrico (e quindi affetto da bias) del training dataset, tipico della suddivisione del dataset in due sole parti (ovvero training e validation dataset). Variando la composizione dei due campioni si valuta la migliore stima e la sua variabilità.

Capitolo 3

Simulazione: studio di confronto tra gli algoritmi LARS, LASSO e stepwise regression

Finora abbiamo sviluppato la teoria sul funzionamento dei metodi di selezione di variabili tenendo conto dell'accuratezza delle stime e della bontà del modello. In questo Capitolo la nostra attenzione si baserà sull'applicazione della procedura LARS con altri metodi di selezione, ovvero LAR, Lasso e la Stepwise. Per potere analizzare e discutere in merito all'efficienza di tali metodi, ricorreremo all'applicazione della procedura ad un esempio pratico: lo studio dei dati relativi alla progressione del diabete su 442 pazienti. Questa progressione dipende, tra le altre cose, dall'età, dal sesso, dal peso, dalla media della pressione del sangue e altre sei diverse misure relative alle caratteristiche del sangue.

L'obiettivo è quello di creare un modello che preveda i risultati da queste variabili esplicative, tramite la ripetizione di questa procedura per i tre algoritmi, al

fine di identificare quello più efficiente, tenendo conto che tra i buoni modelli il migliore è quello con il più piccolo sottoinsieme dei predittori.

Il nostro dataset sarà composto da una variabile risposta y, da una matrice X con dieci regressori e da una matrice X2 che comprende i dieci repressori più alcuni interazioni tra i repressori. I dati sono stati standardizzati per avere una media 0 e di lunghezza pari all'unità su ogni colonna. Al fine di raggiungere i nostri scopi, considereremo la prima parte del modello con i soli regressori, ossia senza fattori d'interazione e la seconda con tutti i fattori d'interazione.

3.1 Modello senza iterazione

Denotiamo con x.age, x.sex, x.bmi, e x.map le variabili rispettivamente dell'età, del sesso, del peso e della media della pressione sanguigna, mentre le altre sei variabili (x.tc, x.ldl, x.hdl, x.tch, x.ltg, e x.glu) sono le caratteristiche del sangue. Dobbiamo assumere che i regressori siano stati standardizzati per avere media pari a 0 e lunghezza, l'unità, e la risposta ha media 0

$$\sum_{i=1}^{n} y_i = 0 \qquad \sum_{i=1}^{n} x_{ij} = 0 \qquad \sum_{i=1}^{n} x_{ij}^2 = 1 \quad , \qquad j = 1 \ldots p \quad \text{e} \quad i = 1 \ldots n$$

Riorganizziamo i dati in un nuovo dataset, che chiamiamo ,dati, in cui abbiamo considerato solo i regressori senza interazione tra loro e con la variabile risposta.

>library(lars)

>attach(diabetes)

>dati<-data.frame(x,y)

>dati

	x.age	x.sex	x.bmi	x.map	x.tc	x.ldl	x.hdl	x.tch	x.ltg	x.glu	y
P											
1	0.03807	0.05068	0.061696	0.021872	-0.04422	-0.03482	-0.04340	-0.00259	0.01990	-0.0176	151
2	-0.0018	-0.04464	-0.05147	-0.02632	-0008. 44	-0.01916	0.074411	-0.03949	-0.06832	-0.0922	75
3	0.08529	0.050680	0.044451	-0.00567	-0.04559	-0.03419	-0.03235	-0.00259	0.002863	-0.0259	141
4	-0.0890	-0.04464	-0.01159	-0.03665	0.012190	0.024990	-0.03603	0.03430	0.022692	-0.0093	206
5	0.00538	-0.04464	-0.03638	0.021872	0.003934	0.015596	0.008142	-0.00259	-0.03199	-0.0466	135
6	-0.0926	-0.04464	-0.04069	-0.01944	-0.06899	-0.07928	0.041276	-0.07639	-0.04118	-0.0963	97
	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.
441	-0.09634	-0.04464	0.039062	0.000121	0.016318	0.001.52	-0.02867	0.02655	0.044528	-0.0259	220
442	-0.04547	-0.04464	-0.07303	-0.0081	0.083740	0.00278	0.173815	-0.03949	-0.00421	0.00306	57

Analisi primilinare dei dati

> cor(dati)

	x.age	x.sex	x.bmi	x.map	x.tc	x.ldl
x.age	1.00000000	0.17373710	0.1850847	0.3354267	0.26006082	0.2192431
x.sex	0.17373710	1.00000000	0.0881614	0.2410132	0.03527682	0.1426373
x.bmi	0.18508467	0.08816140	1.0000000	0.3954153	0.24977742	0.2611699
x.map	0.33542671	0.24101317	0.3954153	1.0000000	0.24246971	0.1855578
x.tc	0.26006082	0.03527682	0.2497774	0.2424697	1.00000000	0.8966630
x.ldl	0.21924314	0.14263726	0.2611699	0.1855578	0.89666296	1.000000
x.hdl	-0.07518097	-0.37908963	-0.3668110	-0.1787612	0.05151936	-0.1964551
x.tch	0.20384090	0.33211509	0.4138066	0.2576534	0.54220728	0.6598169
x.ltg	0.27077678	0.14991756	0.4461586	0.3934781	0.51550076	0.3183534
x.glu	0.30173101	0.20813322	0.3886800	0.3904294	0.32571675	0.2906004
y	0.18788875	0.04306200	0.5864501	0.4414838	0.21202248	0.1740536

	x.hdl	x.tch	x.ltg	x.glu	y
x.age	-0.07518097	0.2038409	0.2707768	0.3017310	0.1878888
x.sex	-0.37908963	0.3321151	0.1499176	0.2081332	0.0430620
x.bmi	-0.36681098	0.4138066	0.4461586	0.3886800	0.5864501
x.map	-0.17876121	0.2576534	0.3934781	0.3904294	0.4414838
x.tc	0.05151936	0.5422073	0.5155008	0.3257168	0.2120225
x.ldl	-0.19645512	0.6598169	0.3183534	0.2906004	0.1740536
x.hdl	1.00000000	-0.7384927	-0.3985770	-0.2736973	-0.3947893
x.tch	-0.73849273	1.0000000	0.6178574	0.4172121	0.4304529
x.ltg	-0.39857700	0.6178574	1.0000000	0.4646705	0.5658834
x.glu	-0.27369730	0.4172121	0.4646705	1.0000000	0.3824835
y	-0.39478925	0.4304529	0.5658834	0.3824835	1.0000000

I dati sopra riportati rappresentano la correlazione esistente tra le variabile del dataset diabete. Tra le variabile repressori la più forte correlazione pari a -0.73849273 è tra la variabile x.tch e la variabile x.hdl. Notiamo che la terza variabile, x.bmi, seguita direttamente dalla 9, x.ltg, hanno coefficienti di correlazione (rispettivamente 0.5864501 e 0.5658834) più elevati con la variabile risposta. Da cui sappiamo che per tutte le procedure, il peso(x.bmi) è più rilevante e quindi introdotto per primo nel modello.

Usiamo il comando seguente per stimare il modello sia con il metodo LASSO che con il metodo LAR.

```
da.lars1<-lars(x,y,type=c("lasso"),trace=T)
LASSO sequence
Computing X'X .....
LARS Step 1 :   Variable 3    added
LARS Step 2 :   Variable 9    added
LARS Step 3 :   Variable 4    added
LARS Step 4 :   Variable 7    added
LARS Step 5 :   Variable 2    added
LARS Step 6 :   Variable 10   added
LARS Step 7 :   Variable 5    added
LARS Step 8 :   Variable 8    added
```

LARS Step 9 : Variable 6 added

LARS Step 10 : Variable 1 added

LASSO Step 11 : Variable 7 dropped

LARS Step 12 : Variable 7 added

Computing residuals, RSS etc .....

Call:

lars(x = x, y = y, type = "lasso", trace = T)

R-squared: 0.518

Sequence of LASSO moves:

	bmi	ltg	map	hdl	sex	glu	tc	tch	ldl	age	hdl	hdl
Var	3	9	4	7	2	10	5	8	6	1	-7	7
Step	1	2	3	4	5	6	7	8	9	10	11	12

<-lars(x,y,type=c("lar"),trace=T)

LAR sequence

Computing X'X .....

LARS Step 1 : Variable 3 added

LARS Step 2 : Variable 9 added

LARS Step 3 : Variable 4 added

LARS Step 4 : Variable 7 added

LARS Step 5 : Variable 2 added

LARS Step 6 : Variable 10 added

```
LARS Step 7 :    Variable 5     added
LARS Step 8 :    Variable 8     added
LARS Step 9 :    Variable 6     added
LARS Step 10 :   Variable 1     added
Computing residuals, RSS etc .....
Call:
lars(x = x, y = y, type = "lar", trace = T)
R-squared: 0.518
Sequence of LAR moves:
      bmi  ltg  map  hdl  sex  glu  tc  tch  ldl  age
Var   3    9    4    7    2    10   5   8    6    1
Step  1    2    3    4    5    6    7   8    9    10
```

Guardando i due risultati del LASSO e dell LAR, il primo elemento rilevante che notiamo è per tutti i due metodi, la cronologia (la priorità) in cui le variabili entrano nel modello è la stessa; la prima variabile che integra il modello è la 3, la seconda è la 9, la terza è la 4 …

Poi, l'applicazione dell'algoritmo Lars al Lasso è modificata al passo 11,ciò per dare un risultato Lasso: il numero di passi usati è 11,superiore al numero di variabili esplicative.Infatti, fino al passo 10 tutti i regressori sono già nel modello, ma Lars toglie la variabile 7 perché cambia di segno mentre la correlazione attiva non cambia

e poi la reinserisce. Allora, il numero d'interazione risulta più grande(12),da cui viene spiegata l'implementazione più lunga di lasso.

Nel frattempo, con LAR, il numero di passi necessario per implementare il modello è uguale al numero dei regressori presenti nel modello: i repressori non sono mai tolti dal modello

Figura 1

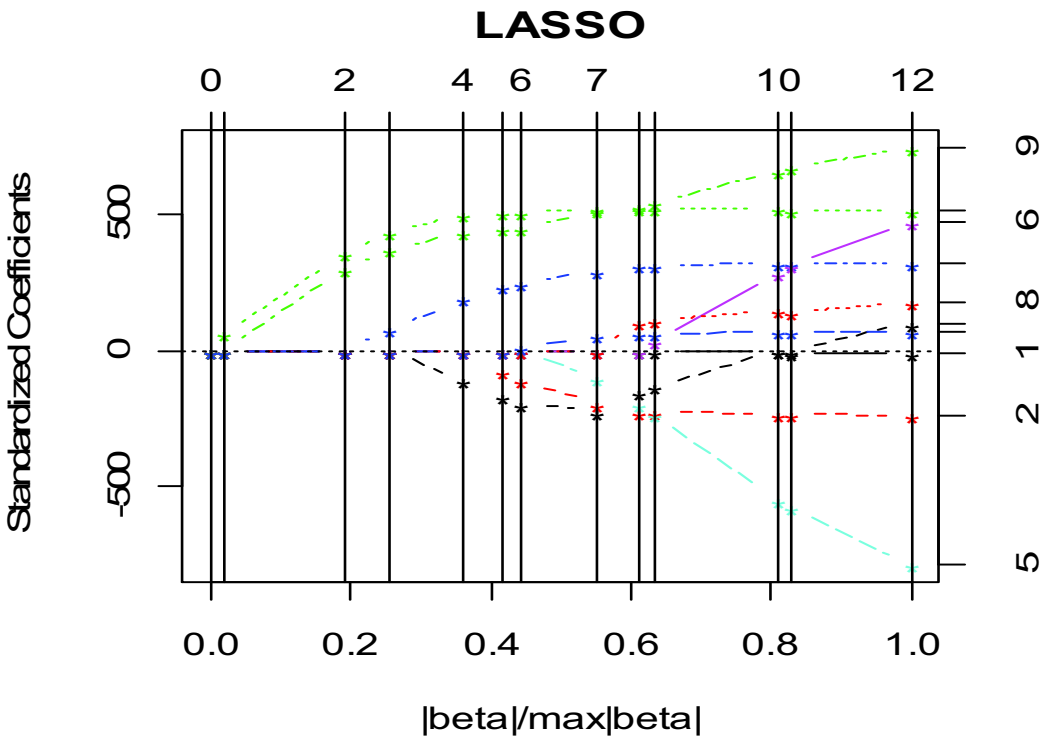
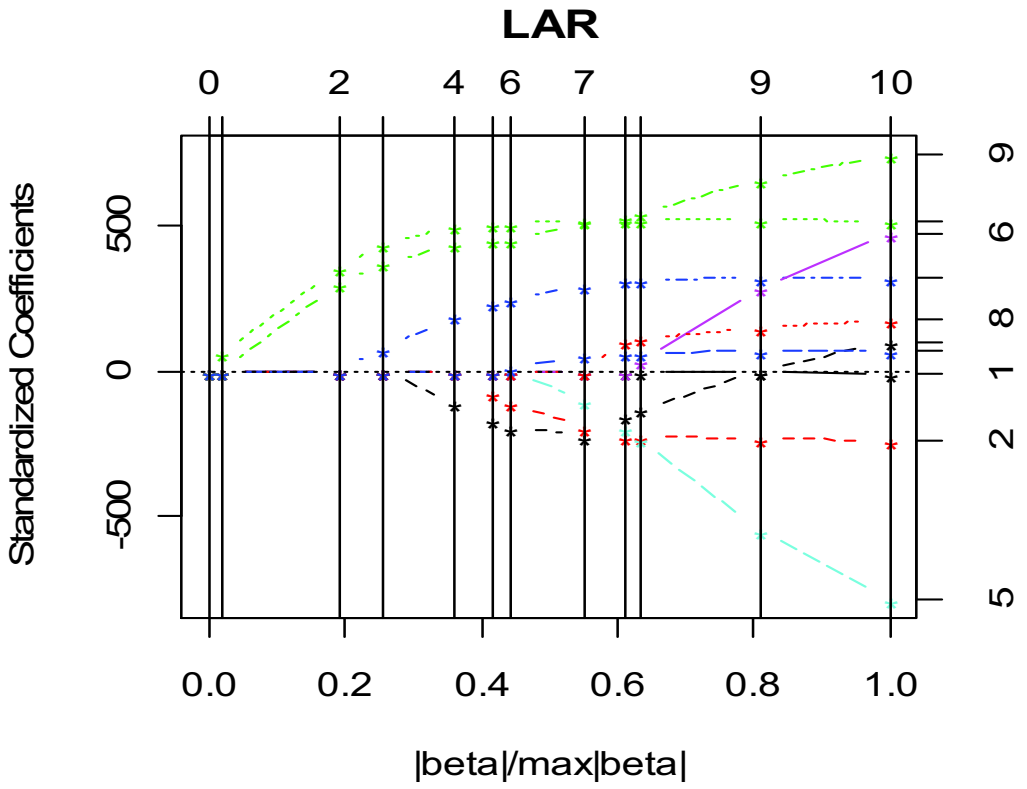


Figura 2



La **figura 1** e la **figura 2** illustrano rispettivamente i coefficienti stimati per LASSO e LAR. $|beta|/\max|beta|$ è sotto inteso come la norma assoluta $t = \sum_{j=1}^{p} |\hat{\beta}_j|$.

Al crescere di t, le variabile esplicative integrano il modello in maniere sequenziale.

Il summary dei due modelli stimati ci dà la statistica C_p, che decresce all'aumentare dei passi, ossia, man mano che le variabili entrano nel modello, fino a essere minimizzata per lasso, al passo 12 e per Lar al passo 10. Quindi, da un punto di vista computazionale Lar mette meno tempo rispetto a Lasso per trovare il risultato finale.

```
>summary(da.lars)
LARS/LASSO
Call: lars(x = x, y = y, type = c("lasso"), trace = T)
   Df      Rss          Cp
0   1  2621009    453.7263
1   2  2510465    418.0322
2   3  1700369    143.8012
3   4  1527165     86.7411
4   5  1365734     33.6957
5   6  1324118     21.5052
6   7  1308932     18.3270
7   8  1275355      8.8775
8   9  1270233      9.1311
9  10  1269390     10.8435
10 11  1264977     11.3390
11 10  1264765      9.2668
12 11  1263983     11.0000
```

```
> summary(da.lars1)
LARS/LAR
Call: lars(x = x, y = y, type = c("lar"), trace = T)
   Df      Rss          Cp
0   1  2621009   453.7263
1   2  2510465   418.0322
2   3  1700369   143.8012
3   4  1527165    86.7411
```

4	5	1365734	33.6957
5	6	1324118	21.5052
6	7	1308932	18.3270
7	8	1275355	8.8775
8	9	1270233	9.1311
9	10	1269390	10.8435
10	11	1263983	11.0000

```
plot.lars(fit1,xvar=c("norm"),plottype=c("Cp"))
```

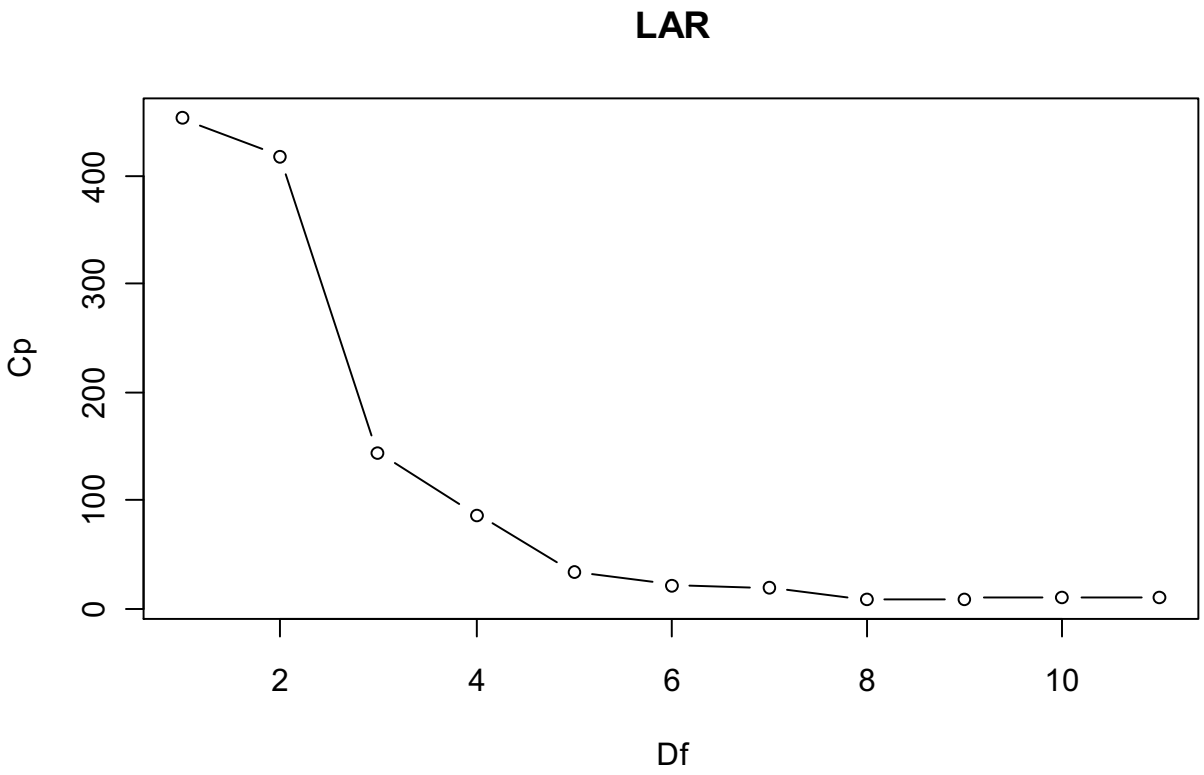


Diagramma rappresentando la statistica cp in funzione di gradi di libertà. Man mano che crescono i gradi di libertà, ossia le variabili entrano nel modello , la staistica Cp decresce fino a approssimare lo zero.

3.2) Modello generale

Consideriamoo la regressione usuale: abbiamo (x^i, y_i), $i = 1,...,n$, dove $x^i = (x_{i1},...,x_{ip})$ sono i 64 regressori chespiegano caratterizzando adesso la progressione del diabete e y_i è la variabile risposta. Il modello adesso è caratterizzato da un'insieme di 10 regressori,45 interazioni e il quadrato di 9 regressori tranne la seconda.

Importiamo i dati nel data frame int e costruiamo il nostro modello.

```
>int<-data.frame(x2,y)
> fit<-lars(x2,y,type="lasso",trace=T)
LASSO sequence
Computing X'X .....
LARS Step 1 :    Variable 3     added
LARS Step 2 :    Variable 9     added
LARS Step 3 :    Variable 4     added
.
.
.
Computing residuals, RSS etc .....
Call:
lars(x = x2, y = y, type = "lasso", trace = T)
R-squared: 0.592
```

Si nota che la prima variabile che viene scelta per spiegare la variabile risposta è sempre la terza. Durante l'implementazione con Lasso le variabili vengono tolti dal modello e poi rimessi ulteriormente e si raggiunge il modello finale dopo 104 passi. Però,il coefficiente di determinazione(0.592) è cambiata;la varianza della variabile risposta è spiegata di più rispetto al primo modello senza interazioni(0.518).

Lar ha un costo più basso perché il risultato è ottenuto dopo 64 passi pari al numero di repressori e ha lo stesso valore del coefficiente di determinazione(0.592) .

```
> fit1<-lars(x2,y,type="lar",trace=T)
LAR sequence
Computing X'X .....
LARS Step 1 :   Variable 3    added
LARS Step 2 :   Variable 9    added
LARS Step 3 :   Variable 4    added
.
.
.
Computing residuals, RSS etc .....
Call:
lars(x = x2, y = y, type = "lar", trace = T)
R-squared: 0.592
```

L’analisi della statistica cp dei due metodi confermano I risultati; il suo valore si abbassa mentre cresce I gradi di libertà.

```
> summary(fit)
LARS/LASSO
Call: lars(x = x2, y = y, type = "lasso", trace = T)
  Df    Rss    Cp
0   1 2621009 485.016
2 2510465 448.002
.
.
.
103 64 1068227  63.002
104 65 1068220  65.000
> summary.lars(fit1)
LARS/LAR
Call: lars(x = x2, y = y, type = "lar", trace = T)
  Df    Rss    Cp
0   1 2621009 485.016
1   2 2510465 448.002
3 1700369 164.100
.
.
.
63 64 1068229  63.003
64 65 1068220  65.000
```

CONCLUSIONE

Abbiamo visto che il metodo più comunemente usato per stimare un modello di regressione lineare è il metodo dei minimi quadrati. Però le stime ols così trovati, non soddisfano l'analisi: il primo problema è l'accuratezza delle previsioni,le stime ols sono consistenti ma hanno varianza larga, secondo è l'interpretazione soprattutto quando siamo in presenza di un numero grande di variabile esplicative.

Abbiamo proposto un nuovo metodo per stimare modelli lineari, il lasso minimizza la somma dei residui vincolata dalla somma del valore assoluto dei coefficienti sia inferiore a una constante. Dato la natura del vincolo, esso tende a mettere a zero alcuni coefficienti e gli altri sono ridotti al massimo. Così facendo sacrifica una tendenza ben poco per ridurre la varianza dei valori previsti e, quindi, può migliorare l'accuratezza complessiva della previsione ossia dà modelli interpretabili. Lo studio tramite un esempio ci permette di proporre algoritmi efficienti per gli estensioni di Lasso e Lar, e si dimostra che questi estensioni dà una performmenza superiore rispetto alla stepwise regression da una parte, e dall'altra ci premette di trovare similitudini tra lasso e lar. Anche se, modificando il Lars, (idea di Osborne,Presnelle

e Turlach)si ottiene facilmente e più velocemente i risultati di LASSO, si ritiene che il Lar, è il migliore.

Bibliografia

Bradley E., Trevor H., Iain J. e Robert T.(2004) . “least angle regression”. *The Annals of Statistics* 2004, Vol. 32, No. 2, P.407–499

Bradley Efron et.al. (2004). “Least angle regression”.
Stanford University.

George, E. I. (2000), The variable selection problem, J. Amer. Statist. Assoc., 95, 1304-1308.

Ming, Y. e Yi. L. (2004) “Model Selection and Estimation in Regression with Grouped Variables”. TECHNICAL REPORT NO. 1095

Osborne, M., Presnell, B. and Turlach, B. (2000a). A new approach to variable selection in least squares problems. IMA J. Numer. Anal. 20

P.389--403.

Pace L. e Salvan S.(2001). “introduzione alla statistica-II. Inferenza, verosimiglianza, modelli”. Cedam, padova.

Robert T.(1996). regression shrinkage and selection via the lasso

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, J. Royal. Statist. Soc. B., 58, 267-288.

Tim H., Nam Hee C., Lukas M., e Chris F. (2008). “Least angle and ℓ1 penalized regression:A review”. Statistics Surveys vol. 2, P. 61-93