Relazione Assegno di ricerca.
La selezione delle variabili nell'analisi multivariata.
01/09/2006 – 31/08/2007

LA SELEZIONE DI VARIABILI NELLA REGRESSIONE LINEARE. IL MODELLO BERDS

Massimo Guerriero[∴]

Summary: The problem of variable selection is one of the most pervasive model selection problems in statistical applications. Often referred to as the problem of subset selection, it arises when one wants to model the relationship between a variable of interest and a subset of potential explanatory variables or predictors, but there is uncertainty about which subset to use. This paper reviews some of the key developments which have led to the wide variety of approaches for this problem. In the section 3, for example, a new algorithm— backward elimination via repeated data splitting (BERDS)— is proposed for variable selection in regression. We also discuss the following problem: given a random sample from an unknown probability distribution, estimate the sampling distribution of some prespecified random variable, on the basis of the observed data. A general method, called the *bootstrap*, is introduced.

Keywords: Linear Regression, Variable Selection, Omission Bias, Data Splitting, Bootstrap.

1. Introduzione

La realtà operativa richiede spesso l'analisi di fenomeni che dipendono da più fattori. Analisi di questo tipo si possono affrontare grazie alla regressione multivariata che costruendo un modello

∴ Dipartimento di Economie Società e Istituzioni, Sezione Statistica, Università degli Studi di Verona, Via dell'Artigliere, 19 – 37129 Verona (e-mail: massimo.guerriero@univr.it).

adeguato basato su un'idonea funzione, appunto la *funzione di regressione*, descrive nel modo migliore come varia una variabile al variare delle altre, e quindi permette di analizzare e descrivere nei migliori dei modi un fenomeno specifico dipendente da una molteplicità di fattori. Nella regressione multivariata o multipla un fenomeno da analizzare si può descrivere nel seguente modo:

$$Y=f(X_1, X_2, \ldots, X_k),$$

dove la variabile dipendente Y (fenomeno da analizzare) dipende da k variabili indipendenti potenzialmente esplicative o predittive X_1, $X_2,\ldots, X_k$.

Naturalmente la fase più importante dell'analisi in questione è quella di trovare la funzione di regressione più idonea a rappresentare la relazione tra le varie variabili quantitative; un lavoro che trova la migliore funzione di regressione richiede l'espletamento di quattro fasi[1]: la rilevazione dei valori, la scelta del tipo di modello, il calcolo dei parametri incogniti e la verifica della bontà d'adattamento.

La rilevazione dei valori

In questa fase si dovranno rilevare su ognuna delle n unità statistiche elementari, n determinazioni di X_i. In questo modo le informazioni si collocheranno in una matrice di n righe e $k+1$ colonne, che rappresentano rispettivamente quanti sono i casi osservati e quante le variabili prese in considerazione.

La scelta del tipo di modello

Scegliere la funzione di regressione più idonea vuol dire scegliere il modello che descriva il fenomeno nei migliori dei modi. In questo senso si deve tener conto della specificità della struttura del fenomeno e delle conseguenti ipotesi costruite per rappresentarlo nel modo più preciso possibile. I modelli che si prenderanno in considerazione nel presente lavoro, sono di tipo ***parametrico***. Nello specifico il modello adottato sarà ***lineare*** nei parametri delle variabili esplicative, del tipo:

[1] Dario Olivieri: *"Fondamenti di statistica"*, Cedam, 1998 seconda edizione.

$$Y=\beta_0+\sum_{i=1}^{k}\beta_i X_i+\varepsilon, \qquad (1)$$

dove i residui ε hanno media zero e varianza σ^2 (cui valore non si conosce e quindi da stimare dai dati disponibili), e sono campionati indipendentemente dalla distribuzione. I coefficienti $\beta_0, \beta_1,\ldots, \beta_k$ sono sconosciuti e quindi anch'essi da stimare.

Il calcolo dei parametri incogniti

Il coefficiente di regressione incognito β_i, con $0\leq i\leq k$, rappresenta il legame di Y rispetto ad X_i, con $0\leq i\leq k$, tenendo costanti tutte le altre variabili. In altre parole l'inclinazione β_i spiega come varia Y in corrispondenza di una variazione unitaria della variabile X_i. A tal fine nel presente lavoro si userà il ***metodo dei minimi quadrati***[2]. Con esso si calcoleranno i $k+1$ parametri del modello che minimizzano la somma dei quadrati delle distanze tra i valori osservati e quelli interpolati:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_{i,1}-\beta_2 x_{i,2}-\ldots-\beta_k x_{i,k}\right)^2.$$

Rispetto al modello lineare di tipo parametrico *(1)*, le stime dei coefficienti di regressione, b, chiamate in letteratura *stime ordinarie a minimi quadrati*, risulteranno[3]:

$$b=(X'X)^{-1}X'Y$$

dove $b'=(b_0, b_1, \ldots, b_k)$. Introducendo nel modello di regressione i parametri stimati, calcoliamo il valore della variabile Y usando i valori conosciuti delle variabili esplicative. Infatti sapendo che n è il numero delle osservazioni e k il numero delle variabili esplicative, identifichiamo X come una matrice $n \times (k+1)$, dove nella riga i-esima si ha 1 seguito dai valori delle variabili $X_1, X_2, \ldots, X_k$ per l'i-esima

[2] G. A. F. Seber: *"Linear Regression Analysis"*, Wiley, New York, 1977.
[3] A.J. Miller: *"Subset selection in regression"*, Chapman and Hall, 1990.

osservazione; *Y* sarà un vettore lungo *n* valori osservati della variabile stimata. Usando i minimi quadrati si ottiene

$$var(x'b)=\sigma^2 x'(X'X)^{-1}x.$$

Facendo la fattorizzazione di Cholesky di *X'X* si ottiene una matrice *R* triangolare superiore[4] *(k+1)* × *(k+1)*, quindi $(X'X)^{-1}=R^{-1}R^{-T}$ dove *-T* è l'inversa della trasposta.
Si può concludere che la varianza di $\hat{Y}$ è data da:

$$var(x'b)=\sigma^2(x'R^{-1})(x'R^{-1})'.$$

La verifica della bontà d'adattamento

L'ultima fase della costruzione del modello è rivolta a verificare la bontà d'adattamento del modello stimato adottato, in altre parole, *valutare il grado d'approssimazione tra il valore reale Y e quello teorico* $\hat{Y}$. La misura della bontà d'adattamento si esegue con opportuni indici che tengano conto degli errori tra valori osservati e teorici. Qui di seguito si cita **l'indice di determinazione R^2** dato dal rapporto tra la Devianza di regressione (somma dei quadrati degli scarti tra i valori interpolati e la media o somma dei quadrati della regressione) e la Devianza totale (somma dei quadrati degli scarti tra valori osservati e la media o somma totale dei quadrati)[5]:

$$R^2 = \frac{Dev.\text{Re}\,gress.}{Dev.Totale} =$$

$$= \frac{\sum_{i=1}^{n}\left(y_i^{\wedge}-m_y\right)^2}{\sum_{i=1}^{n}\left(y_i-m_y\right)^2} =$$

[4] Una matrice triangolare inferiore (superiore) è una matrice quadrata i cui elementi al di sopra (sotto) della diagonale principale sono tutti nulli.
[5] Essa misura la variabilità in *Y* senza considerare la variabilità della variabile regressore *X*.

$$= 1 - \frac{Dev.\mathrm{Re}sidua}{Dev.Totale}.$$

Esso misura la proporzione di variabilità del fenomeno reale *Y* spiegata dal modello teorico di regressione costruito con le variabili *X*. L'indice di determinazione è compreso tra zero e uno. Sarà uguale a zero quando la devianza di regressione è nulla, quindi il modello lineare non fornirà motivazioni sulle variazioni del fenomeno; sarà uguale ad uno quando la devianza residua è nulla, quindi il modello spiega completamente le variazioni del fenomeno. Nella regressione multipla, tuttavia, è opportuno fare uso di un indice che tenga conto anche del numero di variabili esplicative e dei parametri inclusi nel modello, oltre che dell'ampiezza del campione, **l'R^2 *corretto***. L'R^2 corretto è dato dalla seguente espressione:

$$R^2_{adj} = 1 - \left[\left(1 - R^2_{y,12\ldots p}\right) \left(\frac{n-1}{n-k-1} \right) \right],$$

dove *k* è il numero delle variabili esplicative.

In molti casi però, in cui si ha un insieme di variabili indipendenti numeroso, costruire un buon modello regressivo può essere difficoltoso. Per tale ragione nasce l'esigenza di valutare la variabile *Y* sulla base solo di una parte dell'insieme dei fattori che la influenzano. A tal fine esistono delle apposite metodologie che cercano di costruire modelli regressivi che includono solo le variabili indipendenti più influenti, ma che in ogni caso garantiscono una buona precisione nelle valutazioni del fenomeno.

2. La selezione di variabili

Come detto sopra, l'esigenza èi pervenire a modelli di regressione in cui è preso in considerazione solo un *sottoinsieme* di variabili esplicative.
A tal fine nella costruzione del modello di regressione multipla il criterio principale da seguire è la *parsimonia*. Tale criterio impone di inserire in un modello il numero minimo di variabili indipendenti che consentano di spiegare la variabile risposta, quindi si intende inserire

solo le variabili esplicative che possono essere utili per la valutazione della variabile dipendente. In altre parole si analizzerà il fenomeno prendendo in considerazione non tutto il *set* di variabili esplicative disponibili ma una parte di esso, senza in ogni modo compromettere la possibilità di ottenere un risultato abbastanza preciso. Tuttavia per questo ultimo fine è possibile che al sottoinsieme di variabili prese in considerazione si aggiungano altre variabili calcolate da combinazioni lineari di variabili originarie presenti nel sottoinsieme. Tali variabili possono essere costanti ma anche quadrati, prodotti, rapporti e differenze tra variabili, nonché logaritmi di variabili ecc.; la condizione necessaria per la loro inclusione è che devono essere presenti nel sottoinsieme le variabili originarie da cui derivano.

Le ragioni che ci portano ad un lavoro di selezione di variabili, sono diverse e tutte non trascurabili in un'analisi statistica:

- si riesce a trovare le variabili che effettivamente incidono sulla variabile risposta *Y*;
- può essere dispendioso per elevati costi di rilevazione, raccogliere i dati di un *set* completo di variabili potenzialmente predittive;
- con una numerosità minore delle variabili dell'indagine si possono avere misure e rilevazioni più accurate, con tempi e costi di misurazione più contenuti;
- in modelli di regressione con meno variabili e quindi meno parametri da calcolare si possono ottenere dati più "vicini" alla realtà;
- i modelli di regressione con poche variabili esplicative sono di più facile interpretazione, soprattutto perché sono meno esposti al rischio di multicollinearità tra le variabili esplicative.

Il modello con meno variabili

Si propone qui l'analisi della variabilità di *Y* usando solo le prime p delle k variabili esplicative, con $p<k$, ottenendo così la seguente matrice delle variabili indipendenti disponibili

$$X= (X_A, X_B),$$

dove X_A consiste in una matrice con le prime $(p+1)$ colonne di X, e X_B nelle rimanenti $(k-p)$ colonne di X. Usando la fattorizzazione di

Cholesky (purché la $X'X$ sia una matrice positiva e simmetrica) si ottiene

$$X'_A X_A = R'_A R_A$$

dove R_A consiste in una matrice con le prime $(p+1)$ righe e colonne della matrice triangolare superiore R. Il modello con solo p variabili presenterà la seguente varianza dei valori stimati di Y

$$var(x'_A b_A) = \sigma^2 (x'_A R^{-1}_A)(x'_A R^{-1}_A), \qquad (2)$$

dove b_A è il vettore dei coefficienti di regressione del modello con p variabili e R^{-1}_A è una matrice con le prime $(p+1)$ righe e colonne di R^{-1}.

Quindi

$$var(x'b) \geq var(x'_A b_A),$$

cioè la varianza dei valori stimati di Y è maggiore con un numero più alto di variabili indipendenti usate nel modello (Miller 1990).
Usando il modello (1) ed utilizzando solo un sottoinsieme di variabili predittive, le stime dei coefficienti di regressione risulteranno

$$b_A = (X'_A X_A)^{-1} X'_A Y$$

quindi

$$\begin{aligned} E(b_A) &= (X'_A X_A)^{-1} X'_A X\beta \\ &= (X'_A X_A)^{-1} X'_A (X_A, X_B)\beta \\ &= (X'_A X_A)^{-1} (X'_A X_A, X'_A X_B)\beta \\ &= \beta_A + (X'_A X_A)^{-1} X'_A X_B \beta_B, \end{aligned}$$

dove β_A e β_B consistono rispettivamente nei primi $(p+1)$ e negli ultimi $(k-p)$ elementi di β. Osservando l'ultima equazione esposta si nota che il secondo termine della stessa rappresenta il ***bias*** che assumono i $(p+1)$ coefficienti di regressione, omettendo dalla valutazione le $(k-p)$ variabili indipendenti. In questo caso specifico si ha la presenza di ***bias* da omissione di variabili**. Nella stima di Y per un dato x, il *bias* da omissione sarà:

$$x'\beta - E(x'_A b_A) =$$
$$= x'_A\beta_A + x'_B\beta_B - x'_A\beta_A - x'_A(X'_A X_A)^{-1} X'_A X_B \beta_B =$$
$$= [x'_A x'_A (X'_A X_A)^{-1} X'_A X_B]\beta_B.$$

Si rileva la presenza di un *trade-off* tra *bias* e varianza delle stime. In generale un aumento della presenza di variabili indipendenti nel modello, causerà una diminuzione di *bias* ma accrescerà la varianza descritta nell'equazione *(2).*

I metodi classici per la formazione del sottoinsieme di variabili

L'obiettivo è di selezionare da un *set* di k variabili esplicative disponibili un sottoinsieme di $p<k$ variabili $X_{(1)}, X_{(2)}\ldots,X_{(p)}$ che minimizzi la seguente equazione:

$$S = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} b_{(j)} x_{i,(j)}\right)^2,$$

dove $b_{(j)}$ è il coefficiente di regressione stimato con il metodo dei minimi quadrati, mentre $x_{i,(j)}$ *e* y_i sono rispettivamente le i-esime osservazioni delle variabili $X_{(j)}$ e della variabile Y, per $i=1,2,\ldots,n$. I metodi disponibili in letteratura sono tanti, ma in via di classificazione si possono individuare i tre metodi più usati per l'individuazione del modello "adeguato", senza dover tuttavia considerare tutti i 2^k possibili modelli risultanti da un *set* di k variabili esplicative. Si tratta del ***Forward Selection*** o Selezione in Avanti, ***Backward Elimination*** o Eliminazione all'Indietro, e ***Stepwise Regression*** detta Regressione a "Passi Saggi". In ogni caso i modelli di regressione che si possono ottenere con uno dei metodi disponibili, devono essere poi valutati e confrontati facendo ricorso a diversi criteri disponibili (come ad esempio l'R^2_{adj}*).*

Forward Selection

L'algoritmo della Selezione in Avanti inizia con l'assunzione che nel modello non ci sia nessuna variabile indipendente se non l'intercetta β_0. La prima variabile regressore selezionata nel modello è quella con più elevata correlazione con la variabile dipendente; inoltre

le ulteriori variabili scelte, saranno anche quelle che avranno la più alta correlazione parziale in valore assoluto con *Y*, dopo aver scelto $X_{(j)}$. Quindi la prima variabile selezionata, X_j, è quella che minimizza:

$$S= \sum_{i=1}^{n} (y_i - b_j x_{ij})^2,$$

dove b_j risulta:

$$b_j = \frac{\left(\sum_{i=1}^{n} x_{ij}\, y_i \right)}{\left(\sum_{i=1}^{n} x_{ij}^2 \right)};$$

sostituendo questa ultima espressione nella *S* si ottiene:

$$S = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} x_{ij}\, y_i \right)^2}{\sum_{i=1}^{n} x_{ij}^2}.$$

Si può dire allora che la variabile scelta è quella che massimizza

$$\frac{\left(\sum_{i=1}^{n} x_{ij}\, y_i \right)^2}{\sum_{i=1}^{n} x_{ij}^2}. \qquad (3)$$

La prima variabile scelta, che si chiamerà $X_{(1)}$, si introdurrà poi in ogni sottoinsieme ulteriore. A questo punto si trova quella variabile $X_{j.(1)}$ che massimizza l'espressione *(3)*, posto che la *Y* è sostituita con $Y - X_{(1)} b_{(1)}$ *e* X_j è sostituita con $X_{j.(1)}$. Il procedimento prosegue selezionando quelle variabili che minimizzano la **Devianza Residua** (in letteratura abbreviata **con *RSS*, *Residual Sum of Squares***), date le variabili già scelte. Nel metodo della Selezione in Avanti è molto usato il test della *statistica F*. Una variabile predittiva entra nel

modello se la statistica F sarà maggiore di un valore prescelto $F_{to\text{-}enter}$, con F dato da:

$$F_{to-enter} = \frac{RSS\,(b_j \mid b_1)}{MSEP\,(X_j, X_2)},$$

dove *MSEP* è la media degli errori quadratici di previsione.

La procedura iterativa termina quando una variabile regressore porta ad una statistica F più bassa del valore $F_{to\text{-}enter}$, oppure quando tutte le variabili disponibili sono state inserite. Il criterio del test F parziale comporta il calcolo del contributo che ciascuna variabile esplicativa dà alla somma dei quadrati dopo che altre variabili esplicative sono state incluse nel modello. La nuova variabile, quindi, è inclusa nel modello solo se lo stesso ne risulta significativamente migliorato.

In un modello con due variabili, supponendo di includere X_2, per stabilire se X_1 migliora in maniera significativa il modello, si possono impostare le seguenti ipotesi:

$\boldsymbol{H_0}$: La variabile X_1 non migliora in maniera significativa il modello in cui la variabile X_2 sia stata inclusa.

$\boldsymbol{H_1}$: La variabile X_1 migliora in maniera significativa il modello in cui la variabile X_2 sia stata inclusa.

Il test F è dato dalla seguente espressione:

$$F = \frac{RSS\,(b_1 \mid b_2)}{MSEP\,(X_1, X_2)},$$

dove F indica la statistica con distribuzione *F di Snedecor* con *1* e *n-p-1* gradi di libertà. Se il valore osservato della statistica F è maggiore del valore critico, si decide di rifiutare H_0 e si conclude che l'inserimento della variabile X_1 migliora in maniera significativa un modello di regressione che già contenga la variabile X_2. Se il valore osservato della statistica F è minore del valore critico allora si accetta H_0. Un ulteriore modo per decidere se includere o meno una variabile potrebbe anche essere quello del *p-value* per il test F, in cui si decide di accettare l'entrata di una nuova variabile se il suo *p-value* è minore di un determinato *cutoff* (usualmente tra 0,05 e 0,20).

Stepwise Regression

La regressione a "Passi Saggi" può essere considerata come una modifica della Selezione in Avanti. Questo algoritmo proposto da Efroymson(1960)[6] si differenzia dal precedente poiché in questo metodo una variabile introdotta precedentemente può risultare sovrabbondante in virtù della entrata di nuove variabili regressore, infatti ad ogni *step* tutte le variabili esplicative considerate precedentemente nel modello sono testate di nuovo attraverso la valutazione della relativa F per vedere se la loro presenza accresce *RSS* notevolmente. E' ovvio che con questa differenza di fondo è probabile che il sottoinsieme ottenuto con la regressione a "Passi Saggi" sia differente dal sottoinsieme ottenuto con il metodo della Selezione in Avanti.

Backward Elimination

Il metodo dell'Eliminazione all'Indietro (conosciuto anche come il metodo "*Pruning*" di variabili) è in un certo modo speculare al metodo della Selezione in Avanti. La procedura parte con un modello che prevede la presenza di tutte le k variabili esplicative candidate più una costante, per rimuovere poi passo dopo passo, tutte le variabili "non necessarie", senza la possibilità dopo la loro cancellazione di poterle reinserire. Ricordando che RSS_k è l'*RSS* con tutte le k variabili nel *subset*, il metodo esclude quella variabile che dopo la sua cancellazione fa ottenere il minor RSS_{k-1}, o che con la sua presenza fa presentare il più alto *RSS*. Si può dimostrare (Miller 1990) che l'aumento di *RSS* dopo la cancellazione di una variabile predittiva X_i è

$$\frac{b_i^2}{c^{ii}},$$

dove b_i è il coefficiente di regressione a minimi quadrati di X_i con tutte le variabili nel modello, e c^{ii} è l'i-esimo elemento della diagonale di $(X'X)^{-1}$. Con l'uso della statistica F parziale si calcola questa ultima per

[6] M. A. Efroymson: "*Multiple regression analysis. Mathematical Methods for Digital Computers"*, Wiley, New York, pag. 191-203, 1960.

ogni variabile indipendente, come se essa fosse l'ultima variabile regressore ad entrare nel modello. Il valore più piccolo di queste statistiche parziali è confrontato con un valore predefinito $F_{to\text{-}delete}$. Se $F<F_{to\text{-}delete}$, la variabile predittiva è rimossa e la procedura è ripetuta di nuovo con il modello superstite, tenendo conto che dopo aver rimosso una variabile la statistica F delle variabili rimanenti cambia. Tale procedura è iterativa sino a quando il più piccolo valore F non è inferiore al valore $F_{to\text{-}delete}$.

3. L'algoritmo *BERDS*

I metodi classici della selezione di variabili, nonostante siano stati usati moltissimo e seppur siano da ritenere le fondamenta di tutti i metodi nuovi rivolti a risolvere problemi di selezione di sottoinsiemi di variabili, sono stati oggetto di critiche da parte della letteratura. Non si può fare a meno, infatti, di sottolineare che i metodi classici spesso identificano e *selezionano erroneamente variabili rumore* come variabili reali predittive, *facendo ottenere modelli instabili*[7]. Alcuni studi[8] hanno dimostrato che il numero di variabili rumore selezionate dai metodi classici aumentano all'aumentare delle variabili indipendenti disponibili, cosicché la probabilità di identificare correttamente variabili reali è inversamente proporzionale al numero di variabili predittive prese in considerazione. L'algoritmo *BERDS*[9] (*Backward Elimination via Repeated Data Splitting*, -Eliminazione all'Indietro con Ripetute Suddivisioni dei Dati-) qui di seguito esplicitato ha l'obiettivo di eliminare l'instabilità dei modelli selezionati; esso usa il già visto metodo dell'Eliminazione all'Indietro con ripetute ripartizioni dei dati. In generale, per una individuazione della logica di funzionamento dell'algoritmo che possa offrire valutazioni precise, di fronte ad una *variabile rumore* X_j con relativo

---

[7] S. Derksen e H.J. Keselman: "*Backward, Forward and Stepwise Automated Subset Selection Algorithms: Frequency of Obtaining Authentic and Noise Variables"*, volume 45, pag. 265-282, British Journal of Mathematical and Statistical Psychology, 1992.

[8] P. A. Murtaugh: "*Methods of Variable Selection in Regression Modelling"*, Communications in Statistics-Simulation and Computation, volume 27, pag. 711-734, 1998.

[9] P.F. Thall, K.E. Russel, R.M. Simon: "*Variable selection in regression via repeated data splitting*", volume 6 n. 4 pag 416-434, Journal of Computational and graphical statistics, The American statistician, 1997.

$\beta_j=0$, non altamente correlata con una variabile predittiva reale, si preferisce escludere X_j poichè la sua inclusione aumenta la varianza predittiva. Al contrario, se $\beta_j\neq0$ con X_j che è una *variabile reale* non altamente correlata con altre predittive reali, si preferisce includere X_j. In riferimento a questi aspetti è dimostrato[10] che con *set* che includono un gran numero di variabili rumore, *BERDS* presenta, rispetto a *BECV (Backward Elimination via Cross-Validation)* e al classico algoritmo dell'Eliminazione all'Indietro, maggiori probabilità di includere nel modello le predittive reali e di escludere le variabili rumore, oltre a selezionare meno variabili altamente correlate.

Nello specifico in *BERDS* i dati vengono divisi in due sottoinsiemi *{E,V}*. In *E* si eseguono Eliminazioni all'Indietro rispetto un certo *cutoff* α, e per ogni α usato, il modello ottenuto in *E* viene poi validato in *V* calcolando la somma delle deviazioni quadrate dei valori previsti rispetto a quelli osservati. Il processo viene ripetuto *m* volte in modo che l'α che minimizza la somma delle *m* somme di quadrati sarà scelto come *cutoff* da applicare in una finale Eliminazione all'Indietro sull'intero *set* di dati. In via generale la procedura dell'algoritmo si sviluppa in cinque fasi:

1. i dati si ripartono casualmente in due sottoinsiemi complementari, il *set* di stima *E* e il *set* di validazione *V*;
2. si esegue una Eliminazione all'Indietro sui dati che compongono *E*, registrando il *p-value* α_j del *test F* parziale, (ma si usano anche *test* $\beta_j=0$ contro $\beta_j\neq0$ quando X_j è cancellato) con $1\leq j \leq p$, e con $\alpha_0=1$ per il modello completo composto da tutte le *p* predittive. Si denota con $\Lambda(E)$ e con $A(E)$ rispettivamente il massimo e il minimo di $\alpha_1,\ldots, \alpha_p$. Come visto prima, per ogni $\alpha\in[0,1]$ la somma dei quadrati di validazione corrispondente alla scissione *{E,V}* risulta

$$SS_{E,V}(\alpha)= \sum_{i\in V} \left[Y_i-X_i\beta(\alpha,E)\right]^2$$

   dove $\beta(\alpha,E)= \beta(\alpha_j,E)$ per ogni $\alpha_j=min\{ \alpha_r: \alpha_r \geq \alpha, r=1,\ldots, p\}$;
3. al fine di eliminare la sensitività rispetto ad una particolare partizione usata, si ripetono le fasi *(1)* e *(2)* *m* volte[11]. Per ogni

[10] *Ibidem*, nota 10.

[11] R. Picard e K. N. Berk: *"Data Splitting"*, The american statistician, volume 44, pag. 140-147, 1990.

$\alpha \in [0,1]$, si denota l'r-esimo *set* di stima, il *set* di validazione, e la somma dei quadrati di validazione $SS_r(\alpha)=SS_{Er,Vr}(\alpha)$, con $r=1,\ldots, m$. Si definisce[12] la somma dei quadrati di validazione completa, $SS(\alpha)$, il 20% della media ordinata di $\{SS_1(\alpha),\ldots, SS_m(\alpha)\}$.

Come si può notare, per un dato *cutoff* α, i modelli $\beta(\alpha,E_1),\ldots,\beta(\alpha,E_m)$ ottenuti dalle m ripetizioni, non saranno mai gli stessi. Grazie a questa modifica si può dire che $SS(\alpha)$ stima realmente l'abilità predittiva del modello che si seleziona;

4. per determinati percentili q, si denota con L_q e U_{100-q} rispettivamente il q-esimo e il $(100-q)$-esimo quantile di $\{A(E_1),\ldots, A(E_m)\}$ e di $\{\Lambda(E_1),\ldots, \Lambda(E_m)\}$. Si definisce con α^* il *cutoff* che minimizza $SS(\alpha)$ su $[L_q,U_{100-q}]$. In pratica $q=0$ vuol dire non troncazione; $q=90$ vuol dire troncare sotto il *90*-esimo percentile di $\{A(E_1),\ldots, A(E_m)\}$ e sopra il *10*-ecimo percentile di $\{\Lambda(E_1),\ldots, \Lambda(E_m)\}$; $q=100$ vuol dire troncare al massimo e al minimo di questi sottoinsiemi. In conclusione questo *step* blocca il dominio α su cui la funzione obiettivo è minimizzata, per evitare l'alto grado di variazione locale di $SS(\alpha)$ prossima a *0*, che persiste con m grande;
5. si ottiene il modello finale eseguendo sull'intero *set* dei dati, una ultima Eliminazione all'Indietro con *cutoff* α^*.

Semplificando si può dire che l'algoritmo *BERDS* si basa principalmente su quattro *parametri*:

- la percentuale dei dati assegnato e distribuito nel *set* di stima *E*. Facendo l'esempio di assegnare ad *E* il *90%* dei dati e a *V* il rimanente *10%*, questo si indica con *E:V=90:10*;
- il numero delle ripetizioni *m* che si decide di eseguire;
- il metodo che si usa per fare la media nella fase *3*;
- l'ammontare del dominio α con cui nella fase *4* si impone il percentile *q*.

La selezione di variabili che effettua l'algoritmo si valuta rispetto a tre fattori: la probabilità media di includere variabili predittive reali non correlate tra loro, o alternativamente la probabilità di selezionare esattamente *j* di *k* predittive reali correlate, con

[12] *Ibidem*, nota 10.

j=0, 1,…, k; il numero di variabili rumore incluse nel modello finale; la media degli errori quadratici di previsione, *MSEP*,

$$MSEP= \frac{1}{n}\sum_{i=1}^{n} \left[\hat{Y}_i - E(Y_i|X_i)\right]^2$$

Per capire e conoscere l'algoritmo in tutti i suoi aspetti è interessante studiare la sua sensitività sia rispetto ai parametri su cui è stato creato, sia rispetto alla struttura dei dati presa in esame.

Per quanto riguarda gli ***effetti dei parametri*** dell'algoritmo si può dire che nel caso del modello con solo una variabile reale correlata nei dati, *BERDS* si dimostra sensitivo solo a *q*. Infatti con l'uso di un *q=90* piuttosto che un *q=100* si ottengono meno variabili rumore selezionate e un minor *MSEP*. Per dimostrare l'esistenza della sensitività rispetto a *m*, traccie di *MSEP* rispetto le ripetizioni dimostrano[13] che generalmente all'aumentare di *m*, gli errori di previsione diminuiscono (questo non accade per il solo caso di *q=100* con *E:V=50:50*). Riguardo a ciò si può dire che un uso efficiente di *BERDS* si può ottenere con un dominio $\alpha \in [L_{90}, U_{10}]$, con partizione *E:V=50:50* e *m=20*. Con un *set* di dati modesto invece si consiglia l'uso di una partizione *E:V=90:10*.

In ultimo si analizza la *sensitività di BERDS*[14] *alle differenti* ***strutture dei dati*** che si possono investigare. Per fare ciò è stata costruita una *matrice di grafici* dove nelle colonne si trovano quattro elementi della struttura dei dati: dimensione del campione (dato dal numero delle osservazioni *n*), numero delle variabili rumore nei dati, numero delle predittive reali correlate nei dati (con correlazione *0,90*), e il parametro β di una predittiva reale; nelle righe, invece, si trovano i risultati ottenuti nel modello finale, in termini di: *MSEP*, numero delle variabili rumore selezionate, potenza del modello, α empirico finale scelto. Partendo dalle relazioni che coinvolgono la prima colonna, la dimensione del campione, si nota subito che l'esecuzione di *BERDS* migliora nettamente con un aumento del campione stesso. Infatti all'aumentare di *n* oltre ad un'ottimizzazione della potenza del modello, si registra una netta diminuzione sia degli errori quadratici di

[13] *Ibidem*, nota 10.

[14] S. Russo, *La selezione di variabili nella regressione lineare*, Tesi di laurea, Università degli Studi di Verona, Facoltà di Economia, 2005, Verona.

previsione, sia delle variabili rumore selezionate, che del *cutoff* del *p-value* finale. La seconda colonna, il numero di variabili rumore nei dati, riassume una delle più importanti qualità di *BERDS*. Osservando i grafici interessati, si nota che *BERDS* è insensibile alla loro presenza nei dati; infatti all'aumentare delle variabili rumore l' α empirico del modello finale decresce, e la potenza del modello seppur registra un lieve e trascurabile declino, rimane pressoché stabile. Anche nella terza colonna, il numero di predittive reali correlate con correlazione *0,90*, *BERDS* fa notare un'altra sua qualità non trascurabile. Infatti la proprietà che si registra è che l'algoritmo seleziona esattamente la metà delle predittive reali correlate presenti nei dati. Questo accade con un numero di almeno 4 predittive correlate presente nei dati, viceversa *BERDS* ne seleziona comunque una. Quindi, vista la proprietà di selezionare metà delle predittive reali correlate presenti nei dati, è normale selezionarne *1* quando nei dati ce ne sono *2*, ma sorprende vedere che con *3* predittive reali correlate nei dati *BERDS* ne seleziona comunque *1* (per la precisione *1,08*). Infine rispetto alla potenza del modello finale, si nota che all'aumentare di collinearità nei dati essa rimane comunque stabile su quota *0,5*. Per concludere resta l'analisi della quarta colonna, il coefficiente di regressione β di una predittiva reale, che in un certo senso misura la forza di una determinata predittiva. Se questo è vero non deve sorprendere che all'aumentare di β, *BERDS* registra un aumento considerevole della potenza del modello. Inoltre all'aumentare di β, nonostante gli errori di previsione diminuiscano di poco, la traccia rispetto alle variabili rumore selezionate registra un netto declino, questo vuol dire che l'algoritmo ha una accentuata abilità ad eliminare automaticamente le variabili rumore per trattenere quelle reali. Anche il *cutoff* α diminuisce marcatamente all'aumentare di β. A questo punto si analizza la matrice dei grafici rispetto le righe. In questo senso, osservando la prima riga si riesce a vedere la marcata sensitività degli errori di previsione sia rispetto a *n* che rispetto al numero delle variabili predittive correlate presenti nei dati. Nello specifico, mentre all'aumentare di *n* l'indice *MSEP* migliora, all'aumento delle seconde peggiora. Al contrario, rispetto alle variabili rumore nei dati e al coefficiente β di una variabile reale, si può dire che *MSEP* appare non sensibile, rimanendo pressoché stabile. Nella seconda riga si osserva la sensibilità ai dati delle variabili rumore selezionate. Qui si mostra come in *BERDS* le variabili rumore vengono escluse dal modello,

tranne nello scenario che presenta una struttura dei dati con più di *3* variabili reali correlate, dove appunto il modello è instabile e si iniziano a selezionare variabili rumore. Per quanto riguarda la potenza di *BERDS* bisogna osservare le tracce della terza riga. In questi grafici si nota che *BERDS* offre una buona potenza in ogni scenario che possa caratterizzare il *set* dei dati, infatti persino con dati che includono molte variabili correlate la potenza non decresce ma rimane stabile, seppur su una quota di *0,5*. In ultimo la quarta riga, l' α empirico. I grafici mostrano come questo parametro del modello migliora, e quindi diminuisce in ogni scenario (ovviamente tranne nel caso in cui siano presenti tante variabili correlate), confermando la capacità di *BERDS* di aggiustarsi automaticamente ad ogni struttura dei dati[15].

In conclusione si può dire che il successo di *BERDS*, ovviamente oltre a trovare un sottoinsieme di variabili idoneo a prevedere attendibili valori futuri nella regressione lineare, è stato, in primo luogo, proprio quello di riuscire a calcolare empiricamente dai dati il *cutoff* da usare per selezionare le variabili predittive. Questo primo obiettivo è da ritenere importante se si considera che il metodo classico dell'Eliminazione all'Indietro sceglie un arbitrario *cutoff* α del *p-value*. Oltre a ciò *BERDS* presenta qualità non indifferenti nell'escludere dal sottoinsieme sia variabili rumore che metà delle variabili reali correlate tra loro. Queste principali caratteristiche fanno di *BERDS* un algoritmo che offre modelli di variabili stabili e capaci ad aggiustarsi ad ogni struttura dei dati che si presenti all'analisi.

4. Il metodo *bootstrap*

Lo scopo di questo paragrafo è quello di esaminare un nuovo metodo di selezione di variabili che la letteratura sta proponendo come alternativo ai metodi classici, proprio per migliorarne i punti deboli. Questo metodo chiamato *bootstrap*, seleziona modelli predittivi usando combinazioni di ricampionamenti *bootstrap* e metodi classici di selezione di variabili automatizzati. Nella letteratura la prima grande introduzione al metodo *bootstrap* è stata presentata da Efron nel 1979[16]. Dato un campione casuale di dimensione n, $X=(X_1,$

[15] *Ibidem*, nota 10.

[16] B. Efron: *"Bootstrap methods: another look at the jackknife"*, Annals of Statistics, volume 7, 1-26, 1979.

$X_2,\ldots, X_n)$, ottenuto da una distribuzione di probabilità sconosciuta F, e il vettore delle sue realizzazioni osservate, $x=(x_1, x_2,\ldots, x_n)$,

$$X_i=x_i, \qquad X_i \sim F \qquad i=1,2,\ldots,n$$

l'intento è quello di stimare la distribuzione di campionamento di una prespecificata variabile casuale R sulla base dei dati osservati x, con R che dipende sia da X che da F, $R(X, F)$. A questo punto il metodo *bootstrap* si può riassumere nei tre punti seguenti:

1. si sostituisce la sconosciuta funzione di densità F con una stima f, inserendo $1/n$ ad ogni punto $x_1, x_2,\ldots, x_n$;
2. con f stimato, si traccia un campione casuale di dimensione n da f, il campione *bootstrap*

$$X_i^* = x_i^*, X_i^* \sim f \qquad i=1,2,\ldots,n \qquad (4)$$

   dove $X^*_i=(X^*_1, X^*_2,\ldots, X^*_n)$ e $x^*=(x^*_1, x^*_2,\ldots, x^*_n)$. A differenza del tradizionale *jackknife*, i valori di X^* si ottengono con ricampionamento di tipo casuale semplice e con reimmissione di X; si ricorda che quest'ultima deriva da F[17];
3. si approssima la distribuzione di campionamento di $R(X,F)$ dalla distribuzione *bootstrap* di

$$R^* = R(X^*, f),$$

   dove la distribuzione di R^* è stata prodotta grazie al meccanismo casuale *(4)*, con f tenuto fissato al suo valore stimato.

L'approccio *bootstrap* come una strategia di selezione

Una recente proposta di implementazione della tecnica *bootstrap* nella selezione di variabili è stata fatta da Austin e Tu[18]. L'approccio da loro proposto basa la selezione di modelli predittivi

[17] M. Guerriero e G. De Luca: *"La tecnica bootstrap per gli intervalli di previsione nei modelli AR(p)-ARCH(q)"*, Università degli Studi di Verona -Facoltà di Economia-, Quaderni di Statistica, n. 2, 2001.

[18] P.C. Austin e J.V. Tu: *"Bootstrap methods for developing predictive models"*, volume 58 n. 2, The American Statistician, 2004.

sull'estrazione, dall'insieme dei dati disponibili, di campioni *bootstrap* ripetuti, per poi eseguire in ogni campione *bootstrap* una Eliminazione all'indietro per selezionare un modello parsimonioso. Più in generale l'idea di base dell'esecuzione del metodo *bootstrap* nella selezione di variabili prevede l'uso di un metodo classico di selezione di variabili ad ogni replicazione *bootstrap*, al fine di identificare le variabili significative. Le variabili rilevanti molto probabilmente verranno incluse nella maggior parte dei modelli selezionati nelle replicazioni *bootstrap*, dove si assume che ad ogni replicazione, essendo un campione casuale dei dati in analisi, si rifletta la struttura dei dati. Premesso ciò si può dire che la probabilità a posteriori di ogni variabile nei modelli sarà un criterio per valutare l'importanza predittiva di una variabile.

Nello specifico, per ogni variabile candidata, viene identificata la proporzione di campioni *bootstrap* in cui la variabile stessa è selezionata come predittiva indipendente rilevante, al fine di tracciare una sorte di classificazione delle variabili esplicative fatta in base alla proporzione di campioni in cui le stesse sono presenti. A questo punto si forma un modello predittivo preliminare composto da quelle variabili che risultano presenti in tutti i campioni *bootstrap*. La procedura va avanti aggiungendo al modello preliminare altre variabili predittive reali in base alla proporzione di campioni in cui, appunto, queste variabili "entranti" sono presenti. Ogni modello selezionato, poi, può essere valutato in base alla precisione predittiva, in modo da scegliere il modello finale da usare nella investigazione della variabile risposta. Il metodo *bootstrap* usa la probabilità a posteriori che ogni variabile sia stata inclusa nel modello, per valutarne la forza che la variabile stessa presenta nell'essere una predittiva indipendente. Infatti quello che accade è che una predittiva reale sia presente nella maggior parte dei modelli selezionati, mentre una variabile rumore risulta presente solo in una minoranza di modelli.

5. Conclusioni

L'alta probabilità di includere nei modelli predittivi variabili rumore al posto di variabili reali, fa si che i metodi classici automatizzati di selezione di variabili risultino instabili e rumorosi. Negli ultimi anni, sia grazie agli sviluppi delle tecnologie che hanno fornito calcolatori elettronici che riuscissero a far girare algoritmi

complessi, che grazie all'avanzare degli studi di ricerca, molti lavori offerti dalla letteratura hanno cercato di creare algoritmi che riuscissero a selezionare modelli predittivi poco rumorosi e allo stesso tempo precisi nelle valutazioni della variabile risposta (tra tutti si veda Austin-Tu 2004). I metodi *bootstrap*, seppur recenti e ancora da consolidare, in combinazione con i metodi classici, presentano una buona capacità di formare modelli predittivi stabili. Questi usando un ricampionamento casuale semplice con reimmissione delle variabili indipendenti e stimando la probabilità di entrare nei campioni che ogni variabile presenta, oltre ad eliminare dai modelli la rumorosità riescono ad eliminare ogni sorte di arbitrarietà nella scelta delle variabili influenti.

L'algoritmo *BERDS* ha lo scopo, appunto, di riuscire a selezionare da un *set* di variabili indipendenti, solo le variabili reali più influenti. Nonostante la sua versione precedente, l'algoritmo *BECV*, riuscisse ad escludere dal modello predittivo le variabili rumore, presentava comunque un problema rilevante: escludeva anche un buon numero di variabili reali. Questo era dovuto al fatto che *BECV* era sensitivo alla sua unica partizione dei dati scelta per eseguire la *cross-validation*. Nella modifica apportata e caratterizzante *BERDS*, si è riusciti ad ottenere un algoritmo che oltre ad escludere variabili rumore riuscisse anche ad includere le variabili reali influenti. Nello specifico *BERDS* esegue *m* ripartizioni dei dati per eseguirne al loro interno una Eliminazione all'Indietro. Registrando in ogni esecuzione sia l'α minimo che massimo rispetto ad ogni variabile, il *cutoff* che minimizza la somma dei quadrati di validazione sotto un determinato quantile di questi α viene usato per eseguire una Eliminazione all'Indietro finale sull'intero *set* dei dati. Questa modifica che riesce quindi ad adottare una sorte di α empirico nella selezione delle variabili è stato un passo rilevante negli studi regressivi.

Nonostante ciò, secondo un parere personale degli autori, il modello *BERDS* può essere ancora migliorato. Innanzitutto si sottolinea che nelle prime *m* Eliminazioni all'Indietro, il *cutoff* applicato è arbitrario (in genere *0,10*) seppur frutto di valutazioni dell'analista riguardo la struttura dei dati disponibili. Due aspetti rilevanti meriterebbero di essere approfonditi: le *m* ripetizioni degli *step* 1 e 2 da eseguire, e il metodo usato per trovare la somma dei quadrati di validazione completa $SS(\alpha)$. L'intento potrebbe essere

quello di trovare dei valori di ottimo sia per *m* che per *SS(α)*, al fine di sollevare il metodo dalla attuale arbitrarietà su cui poggia.

Riferimenti bibliografici

[2] D. Olivieri: "*Fondamenti di statistica*", Cedam, 1998, seconda edizione.

[3] G.A.F. Seber: "*Linear regression analysis*", Wilely, New York, 1977.

[4] A.J. Miller: "*Subset selection in regression*", Chapman and Hall, 1990.

[7] M.A. Efroymson: *"Multiple regression analysis. Mathematical Methods for Digital Computers",* Wiley, New York, pag. 191-203, 1960.

[8] S. Derksen e H.J. Keselman*: "Backward, Forward and Stepwise Automated Subset Selection Algorithms: Frequency of Obtaining Authentic and Noise Variables"*, volume 45, pag. 265-282, British Journal of Mathematical and Statistical Psychology, 1992.

[9] P. A. Murtaugh: "*Methods of Variable Selection in Regression Modelling"*, Communications in Statistics-Simulation and Computation, volume 27, pag. 711-734, 1998.

[10] P.F. Thall, K.E. Russel, R.M. Simon: "*Variable selection in regression via repeated data splitting*", volume 6 n. 4 pag 416- 434, Journal of Computational and graphical statistics, The American statistician, 1997.

[12] R. Picard e K. N. Berk: *"Data Splitting"*, The american statistician, volume 44, pag. 140-147, 1990

[15] S. Russo: *"La selezione di variabili nella regressione lineare"*, Tesi di laurea, Università degli Studi di Verona, Facoltà di Economia, 2005, Verona.

[17] B. Efron: *"Bootstrap methods: another look at the jackknife"*, Annals of Statistics, volume 7, 1-26, 1979.

[18] M. Guerriero e G. De Luca: *"La tecnica bootstrap per gli intervalli di previsione nei modelli AR(p)-ARCH(q)"*, Università degli Studi di Verona -Facoltà di Economia-, Quaderni di Statistica, n. 2, 2001.

[19] P.C. Austin e J.V. Tu: *"Bootstrap methods for developing predictive models"*, volume 58 n. 2, The american statistician, 2004.