

UNIVERSITA' DEGLI STUDI DI PADOVA
FACOLTA' DI SCIENZE STATISTICHE
CORSO DI LAUREA IN SCIENZE
STATISTICHE ED ECONOMICHE

TESI DI LAUREA

CONTROLLO DELLA MOLTEPLICITA'
ASSOCIATA AI MODELLI DI
REGRESSIONE DI TIPO STEPWISE

Relatore: Dott. Livio Finos

Laureanda: Erica Tosato

ANNO ACCADEMICO 2008/2009

a mia madre

Indice

Introduzione	7
1. Capitolo: La regressione stepwise	9
1.1. La Regressione Stepwise.....	9
1.2. Tipi di regressione stepwise	11
1.3. Criteri di arresto	12
1.4. Organizzazione dei dati	15
1.5. Stepwise in SAS	16
2. Metodi non parametrici per il controllo della molteplicità	18
2.1. Problema della molteplicità.....	18
2.2. Procedura di “Optimal Subset”	21
2.3. La statistica non parametrica.....	26
2.4. Procedura non parametrica di correzione della molteplicità per la stepwise ..	29
2.5. Esempio d’applicazione	36
3. Programmazione in SAS.....	44
3.1. Il SAS	44
3.2. Macro SAS	49
3.3. Funzioni e Procedure della macro PvalueADJ.SAS	51
4. Macro SAS: Applicazione del TEST DI PERMUTAZIONE alla regressione STEPWISE.	62
4.1. Logica dell’algoritmo	62
4.2. Criticità della macro	63
4.3. Richiamo di una macro in ambiente SAS e caricamento dati: Richiamodati.SAS.....	64
4.4. Macro: PvalueADJ.SAS.....	66
4.5. Conclusioni.....	78
Conclusioni	79
NOTE.....	80
Appendice	84
Richiamodati.sas	84
PValueADJ.sas.....	86
Separa.sas.....	88
Riferimenti Bibliografici.....	90
Ringraziamenti.....	95

Introduzione

La statistica è una disciplina che permette alla scienza di prendere decisioni in condizioni di incertezza, offre quindi la possibilità di misurare il grado di affidabilità quando il quadro generale di un fenomeno non è perfettamente delineato e chiaro.

Il grado di conoscenza del fenomeno deve indirizzare lo statistico nella scelta della metodologia di studio più appropriata.

Là dove si ha una conoscenza del fenomeno tale da garantire che gli assunti del modello lineare (generalizzato) siano rispettati, il metodo può essere adottato per descrivere in modo analitico il fenomeno stesso. Questo tipo di modellazione permette di valutare la relazione tra un insieme di predittori ed una variabile risposta.

I modelli di regressione hanno quindi, il compito di quantificare le relazioni tra aspetti (le variabili) oggetto dell'indagine scientifica. La fase di modellizzazione comprende anche delle forme di "sintesi" e semplificazione nella descrizione del fenomeno.

Questo implica, solitamente la riduzione del totale di predittori inclusi nel modello. La selezione viene fatta mantenendo le variabili che risultano maggiormente associate alla variabile risposta, con lo scopo generico di mantenere una adeguata rappresentazione del fenomeno nel suo complesso.

In questa fase di selezione è pratica comune fare uso di metodi di selezione automatica che vengono generalmente definiti "metodi stepwise".

A fronte delle grandi potenzialità di tali metodi, emerge una lacuna legata all'affidabilità dei risultati ottenuti. Infatti, il test che verifica la presenza di un legame tra il complesso di variabili selezionate e la variabile dipendente (usualmente test F o test del rapporto di verosimiglianza) perde il controllo della probabilità di errore di primo tipo. I p-value ottenuti da questi test risulteranno quindi, troppo ottimistici (cioè troppo bassi).

In questa tesi si definisce un metodo di correzione di tale p-value. In particolare si definirà l'algoritmo e lo si implementerà in una macro SAS.

Nel primo capitolo, si fa una panoramica teorica sui metodi di selezione stepwise, descrivendone le caratteristiche ed illustrando come il SAS affronta l'argomento. Nel secondo capitolo, si definirà fondamentalmente il metodo non parametrico di correzione dei p-value, vedendo un esempio di applicazione. Nel terzo capitolo, si introdurrà l'altro importante argomento, che riguarda il linguaggio SAS.

1. Capitolo: La regressione stepwise

1.1. La Regressione Stepwise

La stepwise è un metodo di regressione automatico, consigliabile soprattutto in studi esplorativi in presenza di molteplici variabili, per la stima di modelli lineari multivariati e/o generalizzati (*glm*).

Nasce dalla necessità di selezionare un sottoinsieme “ottimo” tra un gran numero di variabili esplicative per la costruzione di un modello efficiente.

Qualora si debba selezionare un modello regressivo da un insieme limitato di variabili predittive, l'operazione è gestibile, in termini di tempo e risorse, con i consueti metodi regressivi “manuali”.¹ Quando, invece, il numero delle variabili esplicative è elevato, è ragionevole ricorrere a metodi più efficienti, come appunto la stepwise, che garantisce una soluzione più pratica e veloce al problema, evitando di considerare esplicitamente tutte le soluzioni.

Il presente lavoro è focalizzato sulla regressione stepwise per la stima di modelli lineari multivariati, ma il progetto è facilmente estendibile ad un concetto più generale quali i modelli lineari generalizzati.

Entrando nel dettaglio, quando si studia un fenomeno y su un insieme di k variabili predittive $\mathbf{x}'=(x_1, x_2, \dots, x_k)$, la stepwise, dopo una serie di passi, permette di raggiungere un modello ottimo:

$$\hat{y}_j = b_0 + \sum_{i=1}^k b_i x_{ji} \quad (j=1,2,\dots, n)$$

dove alcuni coefficienti b_i potrebbero essere risultati nulli durante il processo di stima.

¹ Con “manuale” si intende che saranno considerate tutte le 2^k possibili combinazioni di variabili esplicative, risulteranno quindi 2^k modelli confrontabili tra loro uno a uno in termini di efficienza. (k è il numero di variabili predittive a disposizione).

Per arrivare a definire questo modello si devono innanzi tutto:

- *decidere il criterio di selezione da adottare* che dipende dall'obiettivo di ricerca, dal tipo di dati su cui si sta lavorando e dai vari legami che sussistono di fondo.
- *definire l'insieme dei dati su cui procedere con l'analisi*: è importante capire se è possibile lavorare sui dati originali o si renda necessario apportare qualche trasformazione alle variabili, sia per gestire meglio eventuali relazioni non lineari, sia per un maggior controllo dei dati anomali.
- *stabilire i parametri dell'analisi o successive ri-analisi*, cioè il numero di variabili esplicative da considerare, oppure il criterio che fa candidare o meno un predittore all'ingresso nel modello (ad esempio il suo contributo alla diminuzione\aumento della varianza della variabile dipendente).

Vediamo di seguito i tre punti nel dettaglio.

1.2. Tipi di regressione stepwise

In merito alle scelte del criterio di analisi stepwise, nella teoria statistica abbiamo a disposizione principalmente tre possibilità:

- la “*forward selection*”: costruisce il modello ottimo partendo da zero, inserendo una alla volta le varie variabili esplicative da considerare, valutate secondo il loro “contributo” predittivo. Il processo di inclusione\esclusione si blocca soddisfatti i criteri di arresto. Regole poco restrittive rischiano di creare un modello troppo complesso e poco efficace\efficiente.
- la “*backward elimination*”: partendo da un modello “completo” elimina in successione, secondo i criteri di arresto impostati, tutte le variabili che apportano uno scarso contributo predittivo, cioè che non dimostrano una forte correlazione con la variabile dipendente.
- la “*stepwise regression analysis*”: si tratta della combinazione delle prime due opzioni. Una variabile predittiva inclusa in un primo momento, il cui apporto esplicativo rispettava i criteri, potrebbe essere successivamente eliminata se l’introduzione di altre variabili ne determinasse una “perdita” di efficacia.

Molti statistici preferiscono l’eliminazione a ritroso poiché permette di individuare situazioni che altrimenti si perderebbero, come ad esempio particolari iterazioni tra i predittori che li collegano alla risposta. Tuttavia il metodo combinato è il solo in grado di soddisfare ogni necessità generica di selezione.

Nella macro SAS implementata in questa tesi, si ha la possibilità di scegliere a priori il tipo di metodo con cui si intende lavorare, ma si è privilegiato il terzo perché più diffuso.

1.3.Criteri di arresto

Scelto il criterio di analisi, resta da capire come definire i criteri di arresto della procedura di selezione e come “organizzare” i dati.

La logica di fondo che guida l’inclusione o meno di una variabile è legata a vari fattori, più attenti a specificare correttamente il modello e a contenere gli errori di stima che non alla verifica statistica dei risultati. Va detto che è comunque difficile individuare criteri assolutamente oggettivi e validi.

Nel decidere i criteri d’arresto, solitamente si tengono presente tre principi (Singh e Tayal, 1980):

- Una variabile omessa erroneamente crea una distorsione di tutte le stime dei coefficienti, ma garantisce una varianza minima.
- Includere una variabile correlata con le altre non provoca la distorsione dei coefficienti del modello, ma fa aumentare la varianza.
- Si può procedere all’eliminazione di una variabile se il suo coefficiente è più piccolo del suo scarto quadratico medio, poiché le stime acquisiranno maggiore accuratezza.

Nella fattispecie, quando si lavora con l’analisi di regressione stepwise, i criteri di arresto potrebbero essere scelti tra i seguenti:

- *definire un numero massimo di predittori inseribili.* In genere si cerca di mantenere un’equa corrispondenza tra numero di variabili presenti nel modello e numero delle osservazioni. Questo perché la relazione tra predittore e risposta è misurata dal coefficiente di regressione al netto di tutti gli altri, i cui gradi di libertà² influenzano l’attendibilità della stima dei coefficienti del modello. Maggiore è il numero delle osservazioni rispetto al numero delle variabili, maggiori sono i gradi di libertà, e quindi maggiore la validità della stima.
Tuttavia, dato l’alto grado di soggettività di questo punto, sovente lo si accompagna con criteri più obiettivi.

² L’aumento dei gradi di libertà fa diminuire α . Di conseguenza l’area critica, cioè la probabilità di accentazione di H_0 , aumenta.

- *definire la porzione di devianza\varianza complessivamente spiegata.* Non saranno introdotte altre variabili esplicative se si è spiegata una data frazione della devianza\varianza con le variabili già incluse. Il principio è che è sufficiente spiegare una porzione significativa, così da mantenere più chiara l'interpretabilità del modello. La devianza viene misurata dal coefficiente di determinazione R^2 (relazione tra y e l'insieme dei predittori inclusi nel modello); per la varianza lo si aggiusta in rapporto ai gradi di libertà.
- *capire quanto l'ultimo predittore inserito nel modello spiega della devianza\varianza di y non ancora "coperta" dalle variabili già annesse al modello.*
- *si procede all'analisi la varianza.* Mentre con i due precedenti metodi si esegue un'analisi quantitativa, qui si compie una valutazione qualitativa. Si misura la significatività statistica del contributo dei predittori nell'interpretazione della variabilità di y , attraverso l'applicazione del test F di Snedecor.

Per la stepwise si possono valutare sia la significatività complessiva, cioè della totalità delle variabili introdotte nel modello individuato, sia di un singolo predittore. Nel primo caso si confronta il valore del test ottenuto³ dal modello con quello teorico⁴. Se il primo è maggiore del secondo l'equazione è significativa e quindi il modello è valido; diversamente, si accetta l'ipotesi nulla di non significatività del modello. Lo stesso confronto vale per il singolo predittore, cambiando solo le formule dei test F⁵. In tal modo si possono valutare sia l'effetto dell'inclusione che quello dell'esclusione di una determinata variabile, con logica identica. Se si sta considerando l'inclusione di una variabile e il modello ottenuto è significativo, la variabile verrà inclusa. Qualora invece se ne stia esaminando l'esclusione, se il modello stimato risulta significativo essa verrà esclusa.

³ $F = \frac{\text{Devianza di regressione} \setminus k}{\text{Devianza residua} \setminus (n - k - 1)} = \frac{R_{y,1,2,\dots,k}^2 \setminus k}{(1 - R_{y,1,2,\dots,k}^2) \setminus (n - k - 1)}$ dove k è il numero di variabili inserite nel

modello e R^2 è la frazione di devianza spiegata dalla combinazione lineare di k .

⁴ F_α è il valore teorico che ha k gradi di libertà al numeratore e $n-k-1$ al denominatore.

⁵ $F = \frac{\text{Devianza ridotta dal } k - \text{esimo predittore}}{\text{Devianza residua} \setminus (n - k - 1)} = \frac{r_{y(k,1,2,\dots,k-1)}^2}{(1 - R_{y,1,2,\dots,k}^2) \setminus (n - k - 1)}$

- *determinare l'originalità* del nuovo predittore, cioè quante nuove informazioni introduce nel modello. Per fare ciò si utilizza il coefficiente di correlazione tra i predittori già inclusi e quello appena inserito ($R^2_{k,1,2,\dots,k-1}$ è compreso tra 0 e 1). Nel caso l'indice fosse significativo, cioè vicino a 1, significherebbe che il nuovo predittore non spiega nulla in più, essendo correlato con quelli già inclusi. Il valore fissato come criterio di ammissione/esclusione è detto di *tolleranza*; più è alto, più c'è il rischio di includere variabili "inutili". D'altro canto, arrestare il processo alle prime battute con regole troppo restrittive non è sempre proficuo, specie in certi tipi di dati quali le serie storiche.

1.4. Organizzazione dei dati

Due sono i problemi che riguardano questa parte: la necessità di una trasformazione dei dati e i valori anomali.

La trasformazione dei dati è indispensabile per una comparazione paritaria tra le differenti variabili prese in esame. Molto spesso, infatti, nell'analisi di un campione si trovano a confronto variabili sia qualitative che quantitative. Le prime, per essere utilizzate correttamente nello studio con le seconde, devono essere trasformate in variabili *dummy*.

I dati anomali, detti anche *outliers*, possono essere presenti o perché ci sono dei risultati "eccezionali", oppure a causa di errori di misurazione. La loro esistenza provoca scompensi nella stima delle correlazioni tra variabili poiché il risultato dell'analisi della regressione può essere condizionato anche da un solo valore.

Il modo più facile per rilevarli è attraverso la proiezione grafica dei dati.

La stepwise può decidere di intervenire eliminandoli o trasformandoli per ridurre l'impatto.

Si deve però tenere presente che, anche se l'eliminazione sembrerebbe essere la via più "facile", c'è il rischio di perdere un'informazione importante qualora i dati fossero davvero un'espressione speciale del fenomeno.

1.5.Stepwise in SAS

L'applicazione della stepwise nel SAS avviene attraverso l'utilizzo della PROC REG per l'individuazione di modelli lineari multivariati e della PROC GLMSELECT nella stima di modelli *glm*. Nella macro le due funzionalità sono facilmente interscambiabili, dato che la PROC REG potrebbe essere un sotto caso della PROC GLM.

Nel presente lavoro ci si è concentrati sulla PROC REG. Tramite essa si possono:

- maneggiare un modello di regressione multipla;
- utilizzare nove metodi di selezione;
- fare dei cambiamenti interattivi sia sul modello che sui dati;
- introdurre restrizioni lineari sui parametri;
- testare ipotesi singole o multiple;
- ricorrere alla diagnostica per la collinearità;
- salvare stime, previsioni, residui, intervalli di confidenza e altre statistiche diagnostiche in appositi dataset SAS;
- generare diversi tipi di grafici e statistiche.

La nostra analisi si focalizza sul secondo punto: selezionare il metodo di regressione.

Tra i nove disponibili ci interessano ovviamente quelli legati alla regressione stepwise, cioè:

FORWARD

Questo metodo parte senza variabili nel modello e le aggiunge una a una. Ad ogni step la variabile aggiunta è quella che massimizza la stima del modello. Con tale approccio è possibile specificare i gruppi di variabili che si intendono trattare durante il processo. Un'opzione permette di precisare il criterio di inclusione. In pratica, per ogni variabile indipendente il metodo calcola la statistica F, che riflette il suo contributo nel modello qualora essa sia inclusa. I p-value per questa F sono comparati con SLENTY (di default 0.5 ma si può decidere diversamente). Ciò significa che, per essere inclusa nel modello, la variabile deve risultare significativa ad un livello dello 0.5. Fino al raggiungimento della soglia di arresto, si procede con le variabili che, tra quelle rimaste, hanno la statistica F maggiore nel modello. In tal modo le variabili sono

aggiunte una a una fino a che non rimangono solo quelle con una F trascurabile, ovvero si sia raggiunto il limite di arresto.

BACKWARD

Si parte con il modello completo e si eliminano una dopo l'altra le variabili. Ad ogni passo, la variabile che contribuisce meno al modello viene eliminata. Come nel metodo precedente è possibile specificare il gruppo di variabili da trattare durante la selezione. Un'opzione rende possibile indicare il criterio d'esclusione. Il metodo comincia con il calcolo della F per l'intero modello. Quindi le variabili sono cancellate dal modello una ad una, finché tutte le rimanenti producono una F significativa al SLSTAY, che di default è 0.10. Ad ogni passo, la variabile che mostra il più piccolo contributo al modello viene cancellata.

STEPWISE

E' la combinazione dei due metodi precedenti. Come nella selezione forward, le variabili sono aggiunte una a una nel modello. La statistica F per una variabile che può essere aggiunta deve essere significativa al livello SLENTY (0.15 di default). Ad ogni aggiunta, la stepwise controlla tutte le variabili già incluse e cancella quelle che non producono una F significativa al livello SLSTAY (0.15 di default). Solo dopo questo controllo, eventualmente con la dovuta eliminazione, può essere aggiunta un'altra variabile. Il processo termina quando nessuna delle variabili escluse ha una F significativa al livello SLENTY e nessuna variabile nel modello è significativa al livello SLSTAY, oppure quando la variabile da aggiungere è già stata eliminata in precedenza, o, infine, il modello rispetta il termine di arresto, si è cioè raggiunto il numero di effetti che si intendevano includere.

Come si nota, la teoria esposta in precedenza rispetta perfettamente quanto proposto dal SAS.

2. Metodi non parametrici per il controllo della molteplicità

2.1. Problema della molteplicità

Il problema della molteplicità sopraggiunge in tutti i casi in cui si deve testare più di un'ipotesi, cioè qualora si stia studiando un fenomeno che contempla più di una variabile. Ciò rappresenta lo scenario più comune nella realtà.

Come sintetizzato in Benjamini and Hochberg 1995, quando si studia un fenomeno per arrivare a delle risposte, ci si trova di fronte alla seguente situazione:

	H0 accetta	H0 rifiutata	Totale
H0 è vera	U	V	m_0
H0 è falsa	T	S	$m - m_0$
Totale	$m - R$	R	M

Tabella 2.1

- m_0 è il numero di H_0 vere
- $m - m_0$ è il numero di H_0 false
- U è il numero di veri “negativi”
- V è il numero di falsi “positivi”: errore I tipo
- T è il numero di falsi “negativi”: errore II tipo
- S è il numero di veri “positivi”.
- $H_1 \dots H_m$ ipotesi nulle testate
- In m ipotesi nulle testate, dove in m_0 risulta vera H_0 , R è una variabile random osservabile, mentre S , T , U , e V sono variabili random non osservabili.

Nello studio di un fenomeno è importante non incorrere in errori che potrebbero portare fuori strada. In particolare evitare di rifiutare H_0 quando questa è vera, dato che

significherebbe accettare un modello di stima privo di significatività in termini spiegazione del fenomeno stesso.

Se si dovesse infatti, simulare uno studio⁶, utilizzando una stepwise con l'eliminazione a ritroso, generando covarianze indipendenti (distribuite come una normale standardizzata), indipendenti e incorrelate dalla variabile dipendente, troveremo che la probabilità di avere modelli significativi è piuttosto alta.

Infatti, generando 10 covarianze con 20 osservazioni, in 1000 ricampionamenti di Montecarlo, la probabilità di trovare un modello significativo considerando un $\alpha=0,05$ sotto H_0 è più del 50%,. Passando a 20 covarianze e 30 osservazioni, la percentuale si alza fino all'80%!

Con m covarianze si otterrebbero $M=2^m-1$ modelli, nessuno realmente significativo o sensato.

Molte sono le metodologie per controllare il problema, tra le più conosciute:

- Il Familywise Error Rate (FWER) rappresenta un controllo piuttosto forte, intendendosi la probabilità di fare almeno un errore di primo tipo.

Secondo Bauer, Hommel, ed Holm, nell'applicazione di test multipli con regioni critiche C_1, \dots, C_S per testare le ipotesi nulle $H_{0(1)}, \dots, H_{0(S)}$, si controlla FWE (l'errore di primo tipo), se la probabilità di rigettare erroneamente una qualsiasi ipotesi nulla è minore o uguale a α sotto H_0 (tutte le $H_{0(i)}$ sono vere), cioè

$$\Pr(\bigcup_{i=1}^S C_i) \leq \alpha \quad \text{per } I \in \{1, \dots, S\}$$

quando $H_{0(1)}, \dots, H_{0(S)}$ sono tutte vere.

Il metodo prevede la probabilità di rigetto sotto una configurazione completa, cioè controlla ogni singola ipotesi nulla univariata. La procedura è così severa da essere consigliabile più negli studi confermativi che in quelli esplorativi. Non è perciò sempre la soluzione

⁶ Finos, Brombin, Salmaso in "Adjusting Stepwise p-value in generalized linear model".

più utile ed appropriata, sia per la sua rigidità che per le assunzioni sulle distribuzioni che sottendono ai test,.

- Il False Discovery Rate (FDR) risulta più utile negli studi esplorativi, poichè molto meno conservativo del precedente. Infatti, secondo Benjamini e Hochberg, e considerando la tabella 2.1, in una lista di ipotesi rigettate rappresenta la porzione di erronei rigetti dell'ipotesi nulla ($E(V/R)$), e si nota che:
 - Con $m=m_0$, cioè se tutte le ipotesi nulle sono vere: $FDR=FWER$.
 - Con $m>m_0$, $FDR\leq FWER$, per cui ogni procedura che controlla FWER controlla anche FDR.

Va comunque sottolineato che in presenza di un numero elevato di ipotesi nulle da testare, i due metodi perdono di efficacia.

L'approccio che si analizza nel presente studio, "*Optimal Subset*", risulta essere un valido strumento alternativo, in particolare all'FDR. Si basa sia sulla selezione del migliore sotto insieme di ipotesi, sia sull'aggiustamento della molteplicità.

2.2.Procedura di “Optimal Subset”

Supponiamo di dover fare inferenza su m ipotesi, per cui avremo un insieme $H=\{H_1,\dots,H_m\}$ di ipotesi nulle elementari, delle quali m_0 sono vere e m_1 sono false. Consideriamo H_Ω insieme di cardinalità M di tutte possibili intersezioni delle m ipotesi in H , non tutte necessarie. H_Ω è quindi un insieme di ipotesi multivariate, dove un generico elemento $C=\{H_1\cap\dots\cap H_m\}$ con $C\in H_\Omega$ identifica un sottoinsieme che risulta sotto ipotesi nulla se ciascuna delle sue elementari ipotesi è sotto ipotesi nulla. Ω indica il criterio di costruzione di H_Ω .

Tre elementi sono necessari ad identificare una procedura di “*Optimal Subset*” :

1. Il criterio Ω , che dovrà generare H_Ω , l'insieme dei possibili sottoinsiemi di H .
2. Il test $\Phi(H_i; H_i\in C)$, $C\in H_\Omega$ per le ipotesi multivariate in H_Ω , che ci fornirà le H_i con i p-value minori. Il test è non distorto per le ipotesi elementari in H_Ω , e può essere basato su una combinazione di statistiche parziali o p-value. Un approccio non parametrico faciliterebbe le cose dato che aiuterebbe a controllare una possibile dipendenza tra variabili.
3. Il test $\Psi(C; C\in H_\Omega)$ aggiusta i p-value ed identifica il sotto insieme ottimale per le ipotesi di H_Ω . Anche in questo caso si consiglierebbe un test non parametrico dato che risulta piuttosto comune incorrere in una dipendenza tra gli elementi di H_Ω , in quanto le ipotesi univariate possono essere contenute contemporaneamente in più di un elemento di H_Ω .

Il punto 2 sostanzialmente screma le ipotesi elementari, individuando quelle con il minor p-value e riducendo la cardinalità di H_Ω . Tuttavia i p-value così ottenuti, soffrono, del problema della molteplicità, essendo stati scelti tra un insieme di altri. Quindi, per ottenere un test globale non distorto devono essere corretti, per cui la scelta di un appropriato test Ψ , potrebbe risolvere il problema.

Se Ψ conduce al rigetto di H_0 , si rigetteranno tutte le ipotesi univariate, che sono in $C_{\min}\in H_\Omega$ e dato che esso è non distorto per C_{\min} , Ψ rigetta C_{\min} con probabilità minore/uguale ad α quando tutte le sue ipotesi univariate sono sotto ipotesi nulla; al

contrario C_{\min} è sotto l'ipotesi alternativa se almeno una ipotesi elementare è sotto l'alternativa.

Questo spiega come la procedura non sia un controllo forte del FWE, che non testa la singola ipotesi ma solo le ipotesi multivariate generate dalla loro intersezione ottimale, ovvero le ipotesi multivariate associate al minimo p-value, ma un controllo debole del FWE su H_{Ω} , cioè non sull'insieme di ipotesi univariate, ma sull'insieme di ipotesi elementari prodotte da Ω .

Per il punto 2, nello sviluppo di una macro in SAS, si utilizzerà in particolare la stepwise, la quale soggetta ai classici problemi di distorsione degli stimatori, rende necessario l'aggiustamento dei p-value, utilizzando strumenti non parametrici, come le permutazioni.

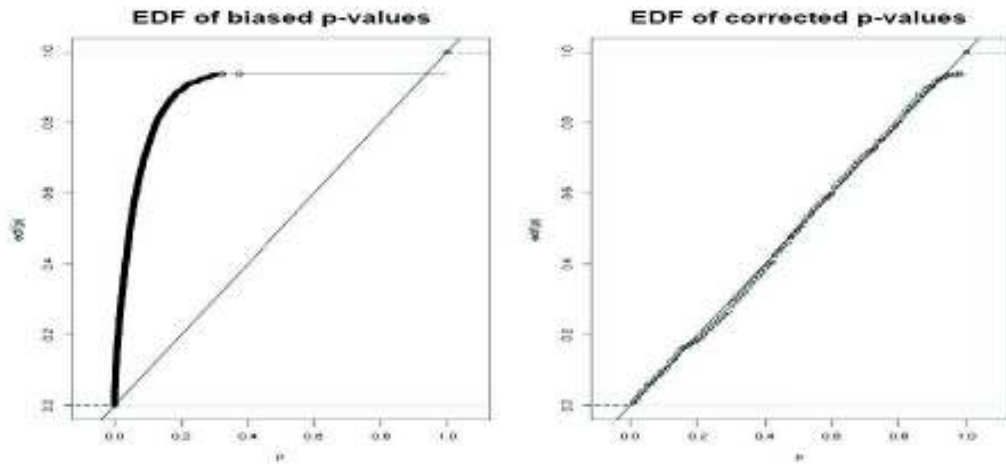
La stepwise, al di là dei limiti ovviabili, è un valido e comodo strumento per identificare un sotto insieme ottimo di ipotesi da testare. I suoi criteri di selezione, backward, forward, e etc rappresentano Ω per generare H_{Ω} , e le varie regressioni forniscono i p-value che aggiusteremo considerando il problema della molteplicità.

Infine con i p-value aggiustati misureremo Ψ .

Il metodo non controlla la molteplicità nei test univariati, ma seleziona ipotesi multivariate che producono risultati più significativi nei test combinati. I p-value ottenuti saranno corretti dalla molteplicità presente. Questo permette di avere un metodo più flessibile e meno restrittivo rispetto ai precedenti, in special modo in quegli studi dove si è interessati soprattutto all'esplorazione del fenomeno.

Simulando, un modello che dovrebbe essere completamente sotto ipotesi nulla, invece si ipotizza, la presenza di dipendenza con la variabile dipendente in parte delle covarianze. Il confronto dei risultati prima e dopo la procedura suggerita, evidenzia l'efficacia della proposta a livello teorico.

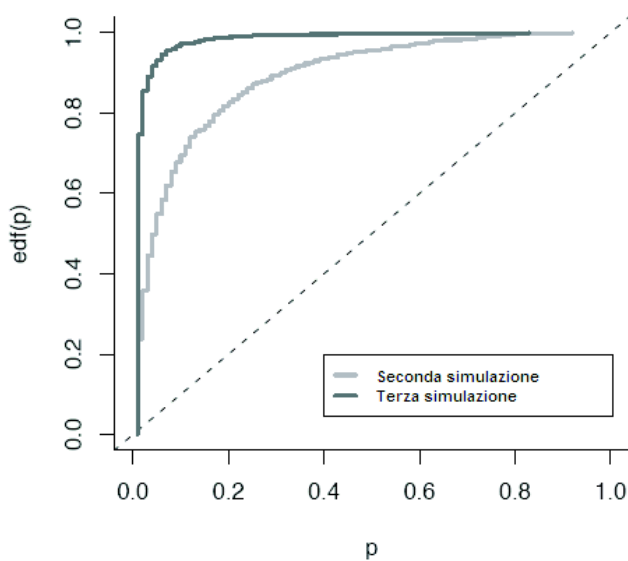
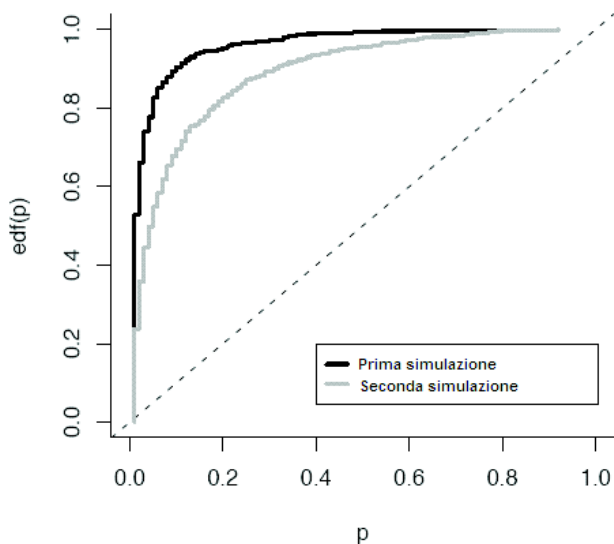
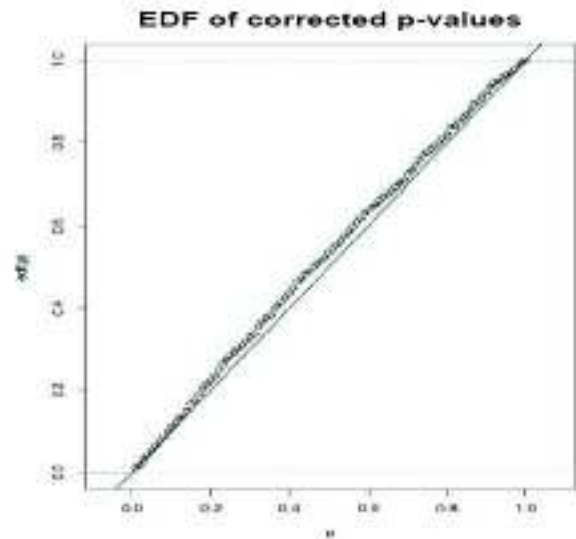
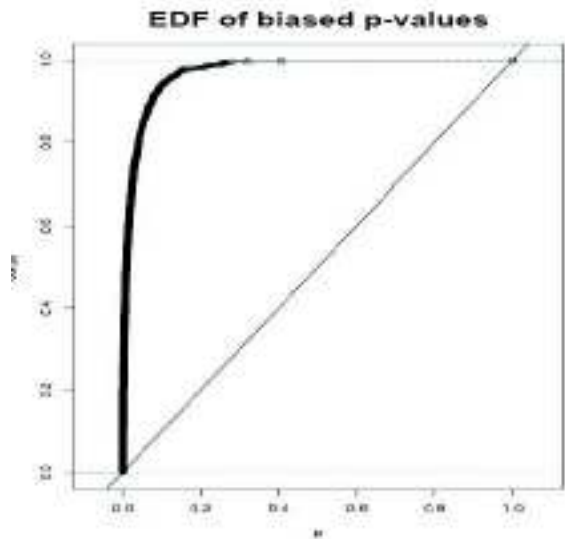
Prima simulazione: $m=10$ covarianze distribuite come una normale, indipendenti e incorrelate con la variabile dipendente, 5 covarianze dipendenti da essa, con numero di osservazioni $n=30$, indice di correlazione $\rho=0,4$, con $B=100$ permutazioni e $MC=1000$ ricampionamenti.



I due grafici rappresentano chiaramente l'effetto prima e dopo l'applicazione della procedura. A destra si ha la cumulata empirica dei p-value del test F per il modello selezionato. Si nota che la probabilità di ottenere un p-value tra 0 e 0,2 è piuttosto alta, nonostante si sappia che l'accentazione del modello non ha senso. Il secondo grafico, rappresenta l'andamento della distribuzione per i p-value del test proposto. Come si vede, questi si distribuiscono sulla retta, dove il test è a livello α per ogni p-value.

Seconda simulazione: $m=20$ covarianze distribuite come una normale, indipendenti e incorrelate con la variabile dipendente, e 5 covarianze dipendenti da essa, con numero di osservazioni $n=30$, indice di correlazione $\rho=0,4$, con $B=100$ permutazioni e $MC=1000$ ricampionamenti.

Abbiamo incrementato il numero delle covarianze impattando sulla molteplicità, ma diminuito la percentuale di variabili dipendenti, lavorando sulla porzione di variabili attive.



Mettendo a confronto le due simulazioni, la curva di potenza decresce all'aumentare delle variabili indipendenti.

Riapplicando l'algoritmo, aggiungendo una terza simulazione con $m=20$ covarianze distribuite come una normale (indipendenti e incorrelate con la variabile dipendente), e 5 correlate con essa, con numero di osservazioni $n=30$, alzando l'indice di correlazione $\rho=0,6$. Dal confronto con la seconda simulazione, si nota che la curva di potenza aumenta, all'aumentare dell'indice di correlazione.

Si affronterà ora l'argomento della statistica non parametrica e di come, grazie ai principi su cui si fonda, risulti un valido supporto alla proposta fatta.

2.3. La statistica non parametrica

Sovente nella ricerca sperimentale è possibile disporre solo di un limitato numero di osservazioni. Ciò mette nella difficile condizione di non poter fare assunzioni precise sul fenomeno studiato, in particolare sulla sua distribuzione (in special modo se si tratta di nuove applicazioni).

Infatti, negli approcci non parametrici, al contrario degli altri, non è indispensabile fare assunzioni sulla conoscenza della distribuzione di probabilità di un fenomeno. Per questo sono spesso detti “distribution free”.

Ideali in campioni non eccessivamente ampi, si possono utilizzare con qualsiasi tipo di dati (eterogenei, omogenei, continui, discreti, ordinati), e con valori che possono essere distribuiti come normali, quasi normali, e non normali.

Questa branca della statistica riguarda sostanzialmente:

- Modelli statistici non parametrici
- L’inferenza non parametrica
- I test statistici non parametrici

Inoltre, la statistica non parametrica, non richiedendo grandi assunzioni, è considerata più robusta⁷ rispetto a quella parametrica.

L’idea che sorregge l’intero processo è di ampliare il campione in studio, attraverso varie metodologie di ricampionamento, così da aggiungere una variabilità “artificiale” all’unico campione di cui si dispone. In tal si riescono a stimare la distorsione, la varianza o addirittura, l’intera distribuzione campionaria dello stimatore. Ciò richiede un buono sforzo computazionale.

Per tale ragione, approcci non parametrici hanno potuto avere spazio e godere di nuovo interesse, solo negli ultimi anni grazie all’introduzione di strumenti computazionali più evoluti.

⁷ Robusta è quella procedura inferenziale che permane stabile al variare delle condizioni sotto le quali è stata derivata. (Statistica. Il Mulino, 1998)

Non molto tempo fa, si ricorreva a questo tipo di metodo, solo qualora si fosse stati assolutamente certi che il fenomeno non avesse una distribuzione affine alla normale.

Ora invece, si suggerisce di ricorrere alla statistica non parametrica ogni volta sussista un qualche dubbio sugli assunti teorici della distribuzione del fenomeno.

Analizziamo di seguito i pro e contro dell'utilizzo dell'approccio sopra citato⁸.

Vantaggi	Svantaggi
1. Non si richiedono particolari assunzioni sulla distribuzione, in particolare sulla normalità della popolazione da cui verrà estratto il campione e su cui verranno effettuate le analisi	Non ci si avvale di tutta l'informazione disponibile e a volte si tende a perderne durante le rielaborazioni
2. Dato che si poggia su un numero limitato di assunzioni di partenza, le sue statistiche sono robuste	
3. Quando i campioni sono piccoli, anche con calcoli elementari, si può stimare un valore esatto di probabilità per i test e gli intervalli di confidenza, senza richiedere la normalità della distribuzione.	Quando i campioni sono grandi, le analisi diventano impegnative da affrontare e gestibili solo attraverso l'informatica.
4. Meno sensibile ai dati anomali, può avere più ampio utilizzo e le non risposte sono in numero minori	Per molti test non parametrici è complesso valutare la significatività delle ipotesi, poiché è difficile disporre delle tavole dei valori critici pensati per essi.
5. Facile da gestire con le nuove tecnologie computazionali	
6. I test n.p. ⁹ sono facili da capire	
7. Sono innumerevoli, a causa della diversità dei problemi che si sono susseguiti nella storia della loro evoluzione; c'è quindi ampia scelta sulle possibili analisi da svolgere	

⁸ Hollander M. e Wolfe D.A.. *Nonparametric Statistical Methods*, (1999, 2nd ed. John Wiley & Sons, New York)

⁹ N.p.: non parametrici.

Va sottolineato che, al di là dei paragoni, è importante che per ogni metodo che si intende utilizzare, ci siano sempre le condizioni e i presupposti perchè per ottenere risultati inconfutabili. Infatti, per quanto un test sia robusto e potente, perde tutta la sua efficacia se applicato a situazioni che non rispettano le condizioni della sua validità.

La statistica parametrica e quella non parametrica possono, convivere e collaborare. Questo lavoro ne è un esempio. Qui, infatti, si associa una regressione parametrica (il metodo stepwise), a metodi non parametrici al fine di correggerne i p-value, quale antidoto al problema della multicollinearità. Ciò tramite l'utilizzo di un'appropriata strategia di permutazioni, come le seguenti:

Bootstrap

Forse è in assoluto il più diffuso. Dal campione osservato (x_1, \dots, x_n) si estrae un campione casuale, quindi con ripetizione, di numerosità pari ad n , definito *campione bootstrap*. Non è altro che il campione originale in cui certi dati saranno ripetuti ed altri di conseguenza, saranno persi.

Nel caso è proprio l'estrazione con ripetizione a generare la variabilità "artificiale" dell'unico campione di cui si dispone.

Permutazione

Questo metodo è quello che viene usato nel progetto esposto. E' preferito al bootstrap perchè più soddisfacente per piccoli campioni. Al contrario del precedente non prevede la ripetizione, quindi la variabilità artificiale che si produce è connessa a tutte le osservazioni campionate, non c'è quindi perdita di informazione.

2.4.Procedura non parametrica di correzione della molteplicità per la stepwise

Come accennato in precedenza, grazie agli strumenti computazionali potenti e sofisticati realizzati negli ultimi anni, si è potuto riprendere e riscoprire la statistica non parametrica.

L'idea dell'utilizzo di metodi non parametrici a supporto di problematiche degli approcci parametrici è davvero recente. In particolare, per quanto riguarda il problema della molteplicità nella stepwise attraverso l'utilizzo di metodi di ricampionamento (in Salmaso e Finos, 2009), da cui questa tesi prende spunto.

L'uso della regressione stepwise è diffuso perché permette di selezionare un sottoinsieme "ottimo" di variabili tra quelli possibili, allo scopo di descrivere le informazioni disponibili e di generare ipotesi per modelli interpretativi fini, attraverso una soluzione efficiente, senza dover valutare esplicitamente tutte le possibili soluzioni, nonostante si sappia:

- che il test sulla devianza spiegata dal modello selezionato è distorto, perché l'ordinaria statistica test, sulla quale questi metodi si fondano, sono utilizzati per testare ipotesi pre-specificate, quando invece, il modello è selezionato attraverso una procedura guidata dai dati.
- che trattandosi di metodo di regressione multipla, è affetto da multicollinearità. Un'esplicativa può quindi mostrare un legame con la variabile risposta solo perché legata ad un'altra, che presenta una reale connessione con la dipendente.

Il SAS è uno dei tanti software statistici, che utilizza questa metodologia pur presentando le suddette problematiche. Nonostante sia ampiamente in uso, non si è ancora stati in grado d'implementare il sistema di regole statistiche che guidano e soggiacciono alla procedura, al fine di garantire una stima del modello empiricamente più significativa.

Infatti, le regole di inclusione o permanenza, di un regressore nel modello, sono ancora connesse all'applicazione di test fondati su assunti rigidi che forniscono p-value basati su distribuzioni minate nella loro validità dallo sgretolarsi dei loro assunti teorici.

Si rende quindi, necessario migliorare l'affidabilità dell'approccio, magari proprio utilizzando soluzioni non parametriche. In tal modo, attraverso la correzione dei p-value dei modelli ottenuti, si andrebbe a controllare il problema della molteplicità, cioè la possibilità di commettere un errore di I° tipo.

L'errore di I° tipo (FWE:family wise error) riguarda l'accettazione dell'ipotesi alternativa, quando questa non è vera, si identifica cioè un modello significativo, quando in realtà questo dovrebbe essere sotto H_0 , espressione di incorrelazione tra le esplicative e la dipendente.

Il SAS per decidere l'inclusione/esclusione di un regressore, utilizza di default l'indice di determinazione lineare R^2 e la statistica F . Questo vale per tutte le tre tipologie di stepwise, a condizione che i loro presupposti teorici permangano; ciò difficilmente accade in uno scenario reale di studio.

Infatti, molti autori affermano (ad esempio nell'ipotesi di una stepwise forward) che, se anche la distribuzione F è corretta (perchè tutti i regressori, precedentemente inclusi nel modello, non sono stati guidati, nello svolgersi del processo, dai dati per la loro accettazione, cercando il miglior regressore), si possono comunque riscontrare delle discontinuità ad ogni passo, che invalideranno sicuramente la distribuzione F .

Questo insinua il dubbio, che variabili accettate nel modello potrebbero essere state erroneamente incluse.

La molteplicità che segue la stepwise, quindi rende il suo uso controverso:

- R^2 piuttosto alto
- La stima degli errori standard è bassa
- I risultati dipendono dalle correlazioni tra i predittori
- I test ordinari sul quale la stepwise è basata, sono definiti per testare ipotesi prestabilite

I problemi che possono sorgere:

- validità dell'inferenza
- sotto o sovra stima del problema
- convergenza del modello: questo aspetto può far sorgere il paradosso di Freedman che vede il numero delle variabili esplicative uguali al numero di osservazioni.
- incerta selezione del modello: non siamo sicuri di avere davvero il miglior modello. Due stime fatte su due campioni provenienti dallo stesso esperimento potrebbero dare soluzioni differenti.

Molte sono le soluzioni studiate negli ultimi anni¹⁰ per ovviare al problema. La maggior parte si concentra su ciascun p-value fornito ad ogni passo della stepwise. Questa tesi, invece, pone l'attenzione sul p-value fornito dall'intero modello selezionato, poiché sembra essere il solo in grado di valutare, se almeno una variabile, tra quelle selezionate, è associato con la variabile dipendente.

Questo approccio sebbene non particolarmente "aggressivo", è l'unico applicabile a qualsiasi modello lineare generalizzato e metodo stepwise, e garantisce un minimo di validità inferenziale ai risultati.

Entriamo nella questione.

I modelli potenzialmente stimabili, con m covarianze disponibili, possono essere $M=2^m-1$.

Definiamo Ω l'insieme di tutti questi modelli, insieme di cardinalità $C(\Omega)=M$, dove ω sono i suoi elementi e p_ω il p-value associato al modello ω .

Inoltre, $\Omega_0 \subset \Omega$ è il sotto insieme di tutti i modelli sotto l'ipotesi nulla H_0 , che prevede la non correlazione di \mathbf{X} con \mathbf{Y} .

¹⁰ Finos L., Brombin C., Salmaso L.(2007). Adjusting stepwise p-value in generalized linear model.

Il nostro problema riguarda quei modelli, contenuti in Ω_0 , che verranno erroneamente attribuiti a $\Omega - \Omega_0$, cioè all'ipotesi H_1 ; questo errore è definito di *I° tipo*, e la sua probabilità è data da:

$$P^{II}(p_\omega \leq \alpha \mid \omega \in \Omega_0) = P(\min_{\omega \in \Omega_0} p \leq \alpha) \leq \alpha$$

e non è considerata banale, in quanto è al di fuori del nostro controllo.

Questo, solleva problemi pratici per i p-value che tendono ad essere davvero piccoli, anche quando \mathbf{Y} non è associato con alcuni dei predittori.

In più, una ricerca esaustiva di tutti gli elementi di Ω non è sempre possibile. Inoltre, tutti i metodi stepwise mirano a contenere i costi di queste ricerche approfondite, limitando l'esplorazione, ad un sottoinsieme di tutte le possibili soluzioni (Ω). Il prezzo pagato è che non sempre la soluzione trovata è quella ottima per l'intero dataset, e i p-value non sono sempre il minimo (infatti, ogni metodo può produrre un differente modello).

Sotto l'ipotesi nulla globale (tutte le covarianze tra la risposta e i predittori sono incorrelate), si propone di correggere il p-value del modello selezionato in modo:

- da controllare l'errore di *I° tipo* a livello α
- garantire la non distorsione
- garantire la consistenza del p-value del modello selezionato

Basandosi su \mathbf{Y} preso da uno spazio campionario \mathcal{Y} , si testerà H_0 , che si fonda su una legge di probabilità f generata da \mathcal{Y} , appartenente ad una certa famiglia di distribuzione F_0 .

Si trova $t(\mathbf{Y}, \mathbf{X}) = t(\mathbf{X})$, il p-value del modello selezionato dalla procedura stepwise per il modello $\mathbf{Y} = h(\mathbf{X})$, dove con t si intende un test statistico adatto.

Si crea G , l'insieme finito di g trasformate di \mathbf{Y} in G , poiché definisce l'orbita $\mathcal{Y}_{|\mathbf{Y}}$ di \mathbf{Y} , e dove G è l'insieme di tutte le possibili permutazioni di \mathbf{Y} .

¹¹ la probabilità che si attribuisca un modello all'ipotesi alternativa quando, invece, è valida l'ipotesi nulla

L'ipotesi nulla implica che la distribuzione di $t(\mathbf{Y})$ sia invariante sotto le trasformazioni di G , che sono $g(\mathbf{Y})$, per ogni g in G , e che \mathbf{Y} abbia la stessa distribuzione, in qualsiasi momento \mathbf{Y} abbia distribuzione f in F_0 . Questo è direttamente applicabile alle distribuzioni di $t(g(\mathbf{Y}))$ e $t(\mathbf{Y})$.

In altre parole, ciò significa che l'ipotesi nulla, di non associazione tra \mathbf{Y} e \mathbf{X} , garantisce che ogni permutazione casuale delle osservazioni di \mathbf{Y} abbia la stessa verosimiglianza, e che la legge del processo causale che seleziona il miglior modello e calcola i p-value (il valore della statistica t) sia la stessa per la variabile \mathbf{Y} e ciascuna delle sue permutazioni $g(\mathbf{Y})$.

Gli assunti precedentemente esposti, serviranno ora a definire l'algoritmo per l'aggiustamento dei p-value, ripreso nella macro SAS che analizzeremo successivamente:

1. si avvia una stepwise standard per un modello lineare generalizzato, per $\mathbf{Y}=\mathbf{h}(\mathbf{X})$
2. Si estrae il p-value associato alla statistica F (test sui residui della devianza per *glm*), definito p-value osservato, p_{raw} .
3. Si procede con le B permutazioni della variabile osservata e si ripetono per esse i punti 1 e 2.
4. Si otterranno B nuovi modelli, se tutte le stepwise ne hanno generato uno.
5. L'aggiustamento del p-value (p_{adj}) è esattamente la frazione dei p-value permutati, p^* , più piccoli o uguali a quello osservato:

$$\frac{\#(p^* \leq p_{raw})}{p - value\ totali}$$

Le proprietà del test sono:

- Poiché il test è invariante¹² rispetto la sua misura, esso è di misura α (cioè controlla debolmente FWE¹³)
- Il test è non distorto¹⁴ perché:

$$F(t(\mathbf{Y})) > F(t(g(\mathbf{Y}))), g(\mathbf{Y}) \in \mathcal{Y}_{|\mathbf{Y}}$$

¹² Si dice invariante quando la regione critica (cioè di rifiuto dell' H_0), non si modifica rispetto ad una prefissata trasformazione dei valori campionari.

¹³ FWE: Family wise errors

¹⁴ Pesarin 2001.

Quando Y dipende almeno da una variabile in X .

Sotto l'ipotesi alternativa H_1 , la distribuzione dei p-value del modello selezionato è stocasticamente più grande per Y che per una generica permutazione casuale.

- Definiamo come consistente quella procedura di selezione che individua almeno una covarianza associata, con probabilità 1 quando $n \rightarrow \infty$ (per esempio, la selezione forward considera tutti modelli univariati, mentre la backward considera i modelli completi). Quindi, se la procedura stepwise è consistente, anche il test è consistente perché:

$$t(\mathbf{Y}) \rightarrow 0 \quad \text{con} \quad n \rightarrow \infty$$

In particolare, la distribuzione dei p-value del modello selezionato converge a 0, al contrario della distribuzione dei p-value del modello selezionato, per ogni permutazione di $g(\mathbf{Y})$.

Ricapitolando:

- La procedura proposta fornisce l'aggiustamento del p-value del modello selezionato, piuttosto che l'aggiustamento del p-value per ciascun coefficiente. Questo perché controlla la FWE debolmente: il p-value globale aggiustato, indica solo se esiste almeno una covarianza non nulla nel modello selezionato.
- Il controllo globale dell'errore di I° tipo è sempre garantito dal fatto che il processo di selezione è lo stesso sia per i dati osservati che per quelli permutati. Il controllo individuale dipende fortemente dal tipo di selezione usata., poiché alla stepwise è connessa al selezione delle variabili piuttosto che alla correzione del p-value.
- Oggi con la stepwise il modello è selezionato tra un subset di possibili modelli, così che il risultato ottenuto non è altro che una soluzione sub-ottima, e il p-value non è sempre il minimo. La situazione gioca un ruolo cruciale nella forza della correlazione e della potenza globale, anche se la procedura darà il modello "giusto" non sempre produrrà il p-value più basso.

D'altro canto, se la procedura desse il miglior modello, quello con il più piccolo p-value, lo otterrebbe solo con un'esplorazione quasi esaustiva, cosa che farebbe accrescere il problema della molteplicità ($C(\Omega)$)

All'estremo, un basso numero di iterazioni, quindi un numero limitato di modelli vagliati, con una bassa abilità di selezione, ridurrà la molteplicità a scapito dei p-value, che incrementeranno di valore.

Ecco perché, questa proposta d'approccio alla correzione della multicollinearità, applicata ad un'esplorazione parsimoniosa dei modelli candidabili, e connessa allo scopo di trovare un buon modello (cioè con p-value bassi), sembra essere la miglior soluzione proponibile.

- Il condizionamento su a set di statistiche sufficienti in H_0 (i dati osservati lo sono sempre) e l'assunzione di scambiabilità dei dati osservati, definisce un test di permutazione indipendente dal modello di verosimiglianza dei p-value generati dalla stepwise. Quindi siamo in grado di avere un buon test pur non conoscendo la distribuzione dei p-value generati dalle osservazioni e il metodo stepwise usato.
- E' importante sottolineare che non si richiede che $p_{\omega} \in \Omega$ sia di misura α sotto H_0 , in quanto il controllo del livello di α per la correlazione è ottenuto anche quando i test individuali dei modelli (cioè dei p_{ω}) non sono di misura α . Ogni statistica che abbia un comportamento differente sotto H_0 e H_1 è una valida scelta. Pertanto la statistica F, il rapporto di massima verosimiglianza, AIC, il R^2 o la devianza del modello sono tutti buoni metodi di scelta, ma nel caso in esame si preferisce considerare il p-value perché la sua distribuzione non dipende dal numero di variabili indipendenti X .

Da questo si evince che il metodo descritto, tra quelli proposti per controllare l'errore di primo tipo, è il più flessibile perché legato a meno assunti di partenza e di più ampia applicabilità; sostanzialmente ha lo scopo di validare il modello trovato.

Nel nostro caso, lavoreremo quindi con i p-value forniti dalla stepwise generati dalla PROC REG.

2.5. Esempio d'applicazione

Di seguito, vediamo un esempio applicativo dell'algoritmo descritto in precedenza nella procedura di "Optimal subset", utilizzando una macro SAS, che spiegheremo più avanti.

Consideriamo un campione¹⁵ di dati relativi al crimine ed a statistiche demografiche per 47 stati degli USA nel 1960, 13 con variabili esplicative, ed una dipendente. I dati sono stati raccolti dall' FBI's *Uniform Crime Report* e da altre agenzie governative per determinare se l'indice del crimine dipendesse dalle variabili misurate nello studio.

R: Tasso di criminalità: # di reati riportati alla polizia per milione di popolazione.

Age: Il numero di maschi tra 14-24 anni d'età per 1000 di popolazione

S: Variabile indicatore per gli stati del sud (0 = No, 1 = SI')

Ed: media del numero di anni di scolarizzazione x10 per persone dai 25 anni in su

Ex0: spesa pro capite nel 1960 per la polizia per stato e governo

Ex1: spesa pro capite nel 1959 per la polizia per stato e governo

LF: Tasso di partecipazione alle forze lavoro per 1000 cittadini maschi dai 14-24 anni d'età

M: Numero di maschi per 1000 femmine

N: Misura della popolazione per Stato in centinaia di migliaia

NW: Numero di non bianchi per un milione di popolazione

U1: Tasso di disoccupazione di cittadini maschi per 1000 tra i 14-24 anni d'età

¹⁵ Vandaele, W. (1978) Participation in illegitimate activities: Erlich revisited. In *Deterrence and incapacitation*, Blumstein, A., Cohen, J. and Nagin, D., eds., Washington, D.C.: National Academy of Sciences, 270-335. Methods: A Primer, New York: Chapman & Hall, 11. Anche in: Hand, D.J., et al. (1994) *A Handbook of Small Data Sets*, London: Chapman & Hall, 101-103.

U2: Tasso di disoccupazione di cittadini maschi per 1000 tra i 35-39 anni d'età

W: Valore medio di beni trasferibili e capitali o redditi familiari in decine di \$

X: Numero dei guadagni di famiglia per 1000, sotto 1/2 del reddito medio.

Si potrebbe incorrere in molti problemi nell'analisi di questi dati con la regressione perché alcune esplicative sono altamente correlate. Per esempio EX0 e EX1, che misurano la spesa in polizia in anni consecutivi, hanno una correlazione dello 0.99. Anche ricchezza (W) e disparità dei redditi (X) sono altamente correlati, come pure U1 e U2, che misurano la disoccupazione di gruppi d'età differenti. Quando le esplicative sono altamente correlate, i coefficienti stimati sono instabili. Rimuovendo dal modello una variabile i risultati potrebbero cambiare drasticamente.

Inoltre, la relazione causale tra EX0 (spesa nel 1960) e tasso di criminalità non è chiara. E' l'aumento della spesa ad avere effetto sul crimine o è il crimine ad avere effetto sulla spesa?

Con un modello formato da Age, Ed, U2, X, and Ex0, si dimostra che è importante dare un occhio alla direzione dei coefficienti. Da questi coefficienti si evince che più educazione e spesa fanno aumentare il crimine, poiché c'è un'altra variabile (variabile appoggio non connessa con questi dati) che causa l'incremento contemporaneo del tasso di educazione e criminalità.

Per ciò questi dati rappresentano un buon esempio di come possa essere facile sbagliare una analisi.

Applicando una stepwise in cui vengono incluse tutte e 13 le variabili, con il campione dei dati osservati otteniamo sei possibili modelli

Passo 1

Modello stimato: $R=14,44+0,89Ex_0+\varepsilon$

Analisi della varianza					
Origine	DF	Somma dei quadrati	Media dei quadrati	Valore F	Pr > F
Modello	1	32533	32533	40.36	<.0001
Errore	45	36276	806.13907		
Totale corretto	46	68809			

Variabile	Stima dei parametri	Errore standard	SS Tipo II	Valore F	Pr > F
Interc	14.44640	12.66926	1048.15843	1.30	0.2602
Ex0	0.89485	0.14086	32533	40.36	<.0001

Passo 2

Modello stimato: $R=-94,46+1,24Ex_0+0,40X+\varepsilon$

Analisi della varianza					
Origine	DF	Somma dei quadrati	Media dei quadrati	Valore F	Pr > F
Modello	2	39931	19966	30.42	<.0001
Errore	44	28878	656.31982		
Totale corretto	46	68809			

Variabile	Stima dei parametri	Errore standard	SS Tipo II	Valore F	Pr > F
Interc	-94.46616	34.39470	4950.90882	7.54	0.0087
Ex0	1.24148	0.16375	37726	57.48	<.0001
X	0.40953	0.12198	7398.18643	11.27	0.0016

Passo 3

Modello stimato: $R = -327,54 + 1,57Ed + 1,24Ex0 + 0,75X + \varepsilon$

Analisi della varianza					
Origine	DF	Somma dei quadrati	Media dei quadrati	Valore F	Pr > F
Modello	3	45802	15267	28.53	<.0001
Errore	43	23008	535.05987		
Totale corretto	46	68809			

Variabile	Stima dei parametri	Errore standard	SS Tipo II	Valore F	Pr > F
Interc	-327.54088	76.91367	9703.45462	18.14	0.0001
Ed	1.57869	0.47661	5870.49757	10.97	0.0019
Ex0	1.24314	0.14785	37827	70.70	<.0001
X	0.75058	0.15077	13261	24.78	<.0001

Passo 4

Modello stimato: $R = -424,92 + 0,76Age + 1,66Ed + 1,29Ex0 + 0,64X + \varepsilon$

Analisi della varianza					
Origine	DF	Somma dei quadrati	Media dei quadrati	Valore F	Pr > F
Modello	4	48196	12049	24.55	<.0001
Errore	42	20614	490.79828		
Totale corretto	46	68809			

Variabile	Stima dei parametri	Errore standard	SS Tipo II	Valore F	Pr > F
Interc	-424.92222	85.85140	12023	24.50	<.0001
Age	0.76022	0.34421	2394.04639	4.88	0.0327
Ed	1.66050	0.45797	6452.18670	13.15	0.0008
Ex0	1.29804	0.14377	40008	81.52	<.0001
X	0.64091	0.15270	8646.71205	17.62	0.0001

Passo 5

Modello stimato: $R = -524,37 + 1,01Age + 2,03Ed + 1,23Ex0 + 0,91U2 + 0,63X + \varepsilon$

Analisi della varianza					
Origine	DF	Somma dei quadrati	Media dei quadrati	Valore F	Pr > F
Modello	5	50206	10041	22.13	<.0001
Errore	41	18604	453.74745		
Totale corretto	46	68809			

Variabile	Stima dei parametri	Errore standard	SS Tipo II	Valore F	Pr > F
Interc	-524.37433	95.11557	13791	30.39	<.0001
Age	1.01982	0.35320	3782.81167	8.34	0.0062
Ed	2.03077	0.47419	8322.11780	18.34	0.0001
Ex0	1.23312	0.14163	34394	75.80	<.0001
U2	0.91361	0.43409	2009.88258	4.43	0.0415
X	0.63493	0.14685	8482.73176	18.69	<.0001

Passo 6

Modello stimato: $R = -618,5 + 1,12Age + 1,81Ed + 1,05Ex0 + 0,82U2 + 0,15W + 0,82X + \varepsilon$

Analisi della varianza					
Origine	DF	Somma dei quadrati	Media dei quadrati	Valore F	Pr > F
Modello	6	51458	8576.36928	19.77	<.0001
Errore	40	17351	433.77652		
Totale corretto	46	68809			

Variabile	Stima dei parametri	Errore standard	SS Tipo II	Valore F	Pr > F
Interc	-618.50284	108.24560	14162	32.65	<.0001
Age	1.12518	0.35086	4461.00468	10.28	0.0026
Ed	1.81786	0.48027	6214.72948	14.33	0.0005
Ex0	1.05069	0.17522	15597	35.96	<.0001
U2	0.82817	0.42740	1628.68477	3.75	0.0597
W	0.15956	0.09390	1252.58447	2.89	0.0970
X	0.82357	0.18149	8932.27708	20.59	<.0001

Sintesi di Selezione stepwise

Step	Variabile immessa	Variabile rimossa	Etichetta	Numero var in	R-quadro parziale	R-quadro del modello
1	Ex0			1	0.4728	0.4728
2	X			2	0.1075	0.5803
3	Ed			3	0.0853	0.6656
4	Age			4	0.0348	0.7004
5	U2			5	0.0292	0.7296
6	W			6	0.0182	0.7478

Step	C(p)	Valore	
		F	Pr > F
1	32.3913	40.36	<.0001
2	19.0160	11.27	0.0016
3	8.8156	10.97	0.0019
4	5.8402	4.88	0.0327
5	3.6631	4.43	0.0415
6	3.0599	2.89	0.0970

Sul campione dei dati originale si accetterebbe il modello con un p-value di <0,0001. Applicando la correzione, con 1000 ricampionamenti si ottiene un PVALUE aggiustato di 0,001, che conferma la bontà del modello scelto.

Se ora, dai precedenti dati togliamo le variabili selezionate dalla precedente analisi e stimiamo un nuovo modello regredito su S LF M N NW U, da una stepwise a tre passi otteniamo:

1 Passo

Modello stimato: $R=77,9+0,34N+\epsilon$

Analisi della varianza					
Origine	DF	Somma dei quadrati	Media dei quadrati	Valore F	Pr > F
Modello	1	7836.60195	7836.60195	5.78	0.0204
Errore	45	60973	1354.94833		
Totale corretto	46	68809			

Variabile	Stima dei parametri	Errore standard	SS Tipo II	Valore	
				F	Pr > F
Interc	77.95482	7.48845	146833	108.37	<.0001
N	0.34284	0.14256	7836.60195	5.78	0.0204

2 Passo

Modello stimato: $R = -475,5 + 0,55M + 0,51N + \varepsilon$

Analisi della varianza					
Origine	DF	Somma dei quadrati	Media dei quadrati	Valore	
				F	Pr > F
Modello	2	18120	9060.03194	7.86	0.0012
Errore	44	50689	1152.02756		
Totale corretto	46	68809			

Variabile	Stima dei parametri	Errore standard	SS Tipo II	Valore	
				F	Pr > F
Interc	-475.55159	185.38981	7580.30059	6.58	0.0138
M	0.55648	0.18626	10283	8.93	0.0046
N	0.51970	0.14416	14971	13.00	0.0008

3 Passo

Modello stimato: $R = -542,76 + 0,66M + 0,54N - 0,44U1 + \varepsilon$

Analisi della varianza					
Origine	DF	Somma dei quadrati	Media dei quadrati	Valore	
				F	Pr > F
Modello	3	20710	6903.21796	6.17	0.0014
Errore	43	48100	1118.59588		
Totale corretto	46	68809			

Variabile	Stima dei parametri	Errore standard	SS Tipo II	Valore	
				F	Pr > F
Interc	-542.76208	187.94481	9328.92353	8.34	0.0060
M	0.66736	0.19747	12776	11.42	0.0016
N	0.54685	0.14317	16319	14.59	0.0004
U1	-0.44810	0.29451	2589.58999	2.32	0.1354

Il p-value della regressione sui dati originali farebbe accettare il modello con un p-value=0,0014 con un livello $\alpha=0,01$. Applicando l'algoritmo descritto su 1000 permutazioni, il modello non risulta più significativo a parità di livello con il p-value_{adj}=0,02 aggiustato.

3. Programmazione in SAS

3.1. Il SAS

In questo capitolo, si introducono le nozioni necessarie a chi si avvicina per la prima volta al SAS per poter comprendere il lavoro svolto. Dopo una panoramica generale sul software e la sua struttura, si entrerà nello specifico delle Macro, con una carrellata sulle principali funzioni usate.

Il SAS è uno dei più noti e utilizzati sistemi per la gestione e l'elaborazione statistica di dati. La sua diffusione è legata alla capacità nel gestire enormi quantità di dati senza problemi e alla possibilità di compiere diversi tipi di elaborazioni.

Nasce prevalentemente come strumento statistico, ma si sta evolvendo in molte altre direzioni, legate soprattutto alle necessità delle realtà aziendali. Questo attraverso lo sviluppo di soluzioni che permettono la creazione e l'amministrazione di data warehousing e di data mining adatti, ad esempio alla gestione delle risorse umane, al supporto alle decisioni, alla gestione finanziaria, ecc.

Il SAS è sviluppato dalla SAS Institute Inc., ed è utilizzabile nei principali sistemi operativi e piattaforme hardware. E' costituito da un linguaggio di programmazione e da un insieme di moduli. In specifiche librerie, ha inoltre, una serie di procedure (PROC), che equivalgono a dei veri e propri programmi. La versione che utilizziamo in questa sede è la 9.1 versione italiana.

Il "modulo base" (SAS/BASE) permette ad un generico utente le seguenti operazioni:

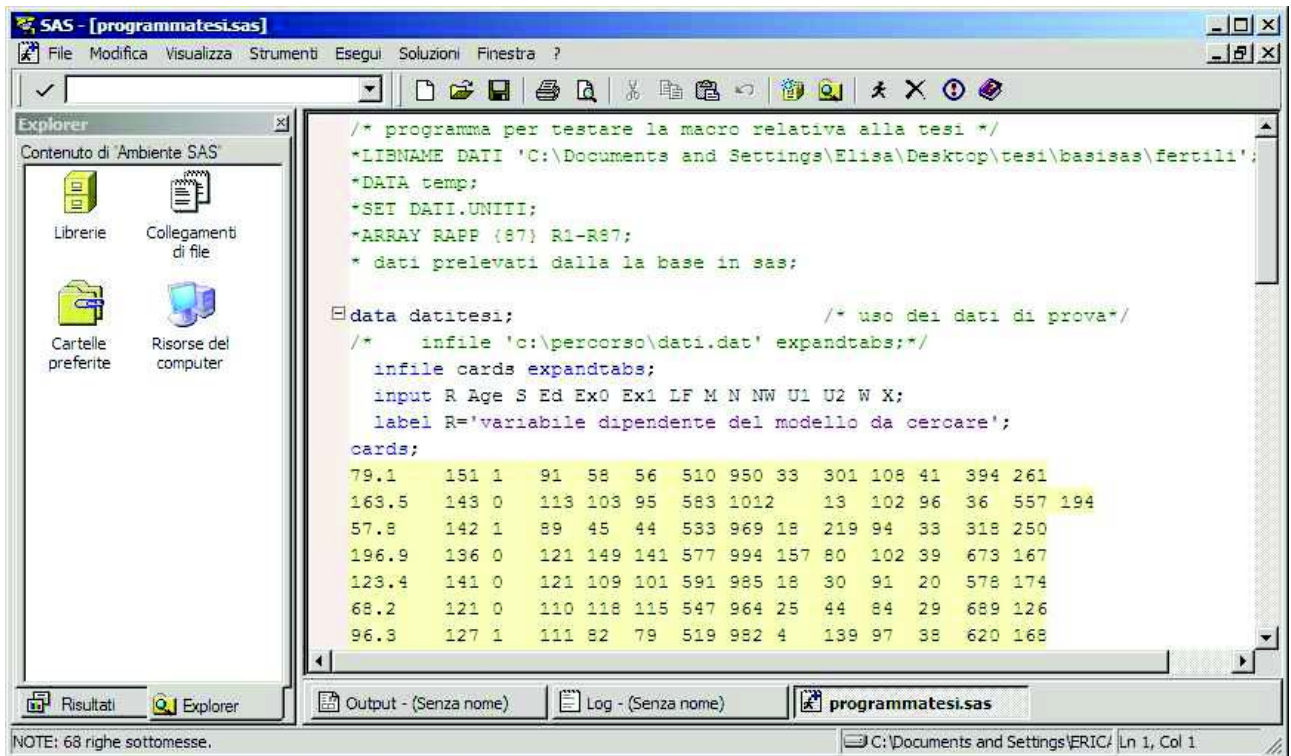
- inserimento, ricerca e gestione di dati;
- generazione di report e grafici;
- analisi statistica e matematica;
- pianificazione, previsione e supporto alle decisioni;
- ricerca operativa e project management;
- gestione della qualità;
- sviluppo di applicazioni;

Questo attraverso l'utilizzo di:

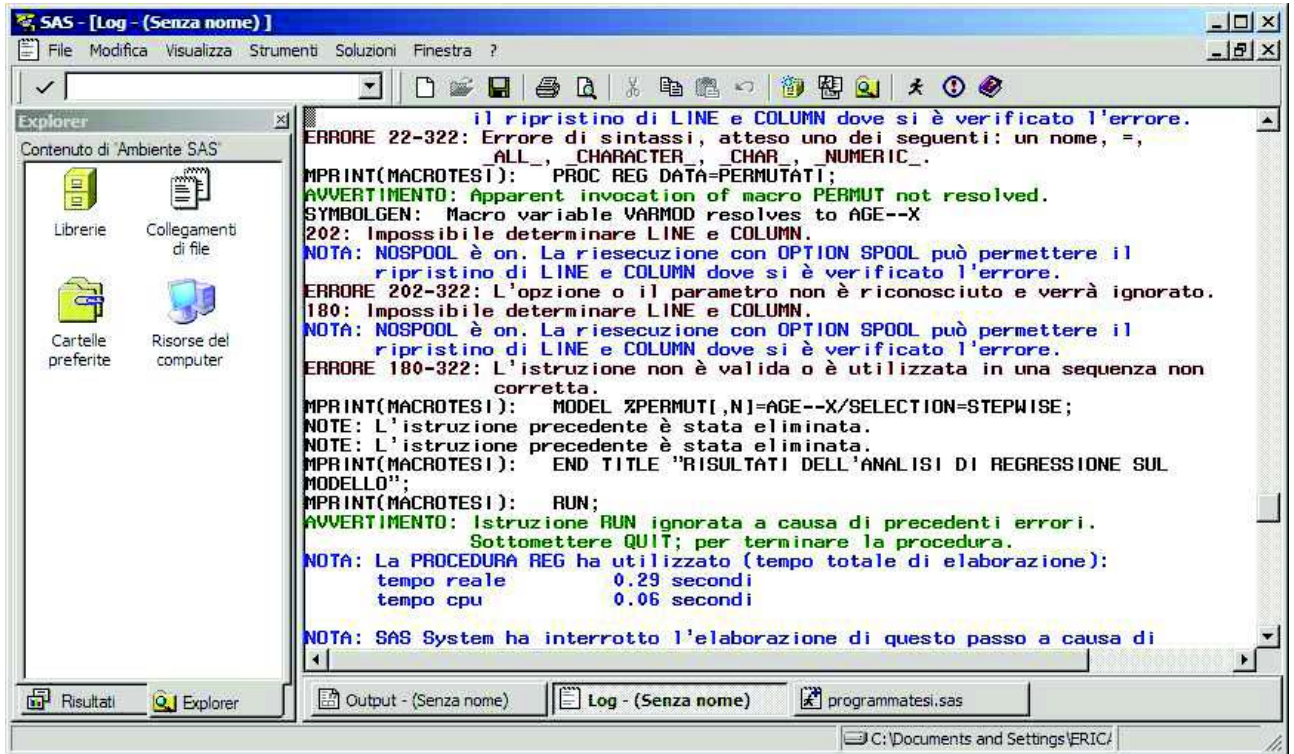
- **Base SAS Software** necessario alla gestione dei dati;
- **SAS procedures software** per l'analisi ed il reporting;
- **Macro facility** uno strumento per estendere e personalizzare le applicazioni;
- **DATA step debugger** per individuare gli eventuali problemi nelle applicazioni sviluppate;
- **Output Delivery System (ODS)**, un modulo che elabora i risultati, e li restituisce in formati standard facilmente gestibili, quali SAS data sets, listing files o Hypertext Markup Language;
- **SAS windowing environment** un'interfaccia grafica ed interattiva per eseguire e testare le applicazioni sviluppate nell'ambiente SAS.

L'utente si relaziona con il software attraverso le sue interfacce (DMS: display manager system):

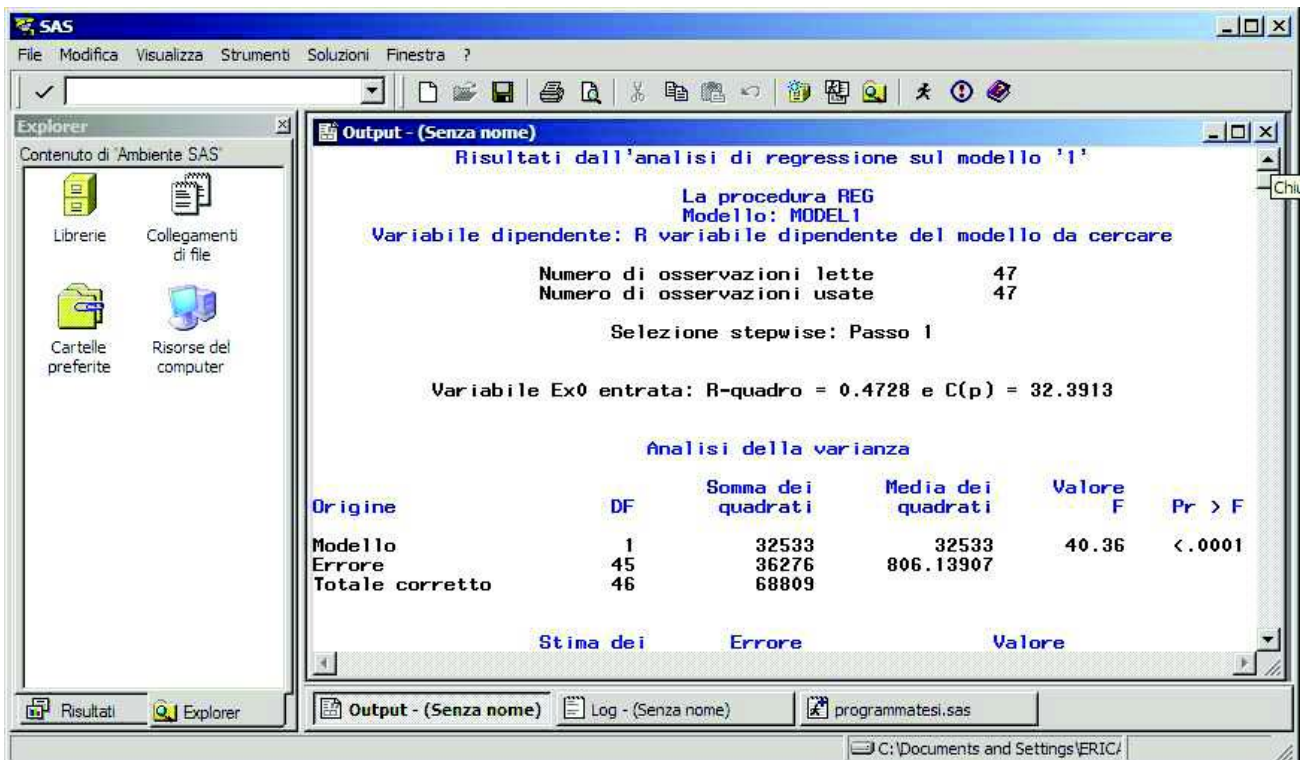
- **Program editor**: qui si scrivono le istruzioni che si vogliono far eseguire al SAS



- **Log:** una volta svolte le istruzioni dell'utente, in questa pagina compaiono le informazioni sulle elaborazioni e gli eventuali errori intercorsi.



- **Output:** si visualizzano i risultati ottenuti dall'elaborazione delle istruzioni eseguite.



E' inoltre possibile integrare il pacchetto base con i seguenti moduli, indipendenti uno dall'altro:

- SAS/STAT: analisi statistica dei dati;
- SAS/GRAPH: rappresentazione grafica dei dati;
- SAS/IML: algebra matriciale;
- SAS/FSP: gestione dati interattiva (input, edit...);
- SAS/AF: ambiente di sviluppo di applicazioni visuali in linguaggio SCL;
- SAS/ASSIST: interfaccia per l'uso guidato del sistema;
- SAS/CALC: foglio elettronico;
- SAS/ACCESS: importazione dati da altri formati e database
- SAS/ETS: analisi di serie storiche;
- SAS/QC: controllo qualità;
- SAS/OR: ricerca operativa.

Per poter eseguire delle procedure in SAS si deve partire dall'acquisizione di dati,. Questi possono essere rappresentati in molteplici formati (ASCII, text, xls, ecc.); al momento dell'importazione verranno tradotti in formato SAS. I dati codificati costituiscono i DATASET SAS, in cui le righe rappresentano le unità osservate e le colonne le variabili.

I DATASET possono essere temporanei o permanenti. Nel primo caso, vengono archiviati temporaneamente per la sola durata della sessione di lavoro. Nel secondo caso vengono memorizzati in file formato SAS (formati *.SSD, *.SD2, ecc.) e potranno essere riutilizzati in per successive sessioni SAS. Tutti i dataset creati e utilizzati nella macro oggetto del lavoro saranno archiviati temporaneamente, per cui ad ogni nuova sessione verranno rigenerati.

Per soddisfare le richieste dell'utente, il SAS impiega una sorta di programmi composti da istruzioni, che si articolano in STEP:

- DATA STEP creano e modificano i database;
- PROC STEP riguardano l'utilizzo delle procedure presenti nelle librerie dei vari moduli disponibili.

Per la realizzazione di un programma in SAS è essenziale una corretta sintassi e l'utilizzo appropriato delle varie procedure. Molto spesso per facilitare certe operazioni di routine vengono creati e salvati dei programmi ad hoc, utilizzando le MACRO SAS, opzione disponibile nella SAS Macro Facility. La macro, formata da una successione di procedure sarà quindi richiamata ogni qual volta ne avremo necessità.

3.2. Macro SAS

Spesso si ricorre alle macro per creare programmi su misura, che migliorano o “personalizzano” le procedure già disponibili. Nel caso specifico si è lavorato sulla stepwise definendo un test di livello α , non distorto e consistente, utile a testare la significatività dell’insieme dei predittori selezionati nel modello.

Le macro utilizzano una sintassi propria, perciò le opzioni e compilatori sono in qualche caso differenti dalla normale programmazione SAS. Ci sono comunque molte similitudini con il linguaggio dei DATA e PROC steps. Inoltre, anche le macro non distinguono tra minuscolo e maiuscolo.

Il linguaggio delle macro consente il passaggio di informazioni tra i SAS steps, permettendo la creazione di codici dinamici.

Per far eseguire una macro è necessario importarla nell’EDITOR SAS attraverso la seguente istruzione:

%include ‘directory contenente la macro’;

%nome della macro (parametri necessari al funzionamento della macro);

Per creare una macro si deve sempre iniziare con

%MACRO nome-macro;

e terminare con

%MEND;

in *MEND*, per ragioni di chiarezza si può riprendere il nome della macro, ma è opzionale.

Le macro possono includere:

- *Testo Costante*: il linguaggio macro lo tratta come ogni altra stringa viene trattata in SAS. Non viene valutato, risolto oppure esaminato, serve solo a rendere più leggibile il programma.

- *Macro variabili*: una volta definite possono assumere qualsiasi valore. Il loro nome può avere una lunghezza massima di 32 caratteri, può cominciare con una lettera o con “_”, ed ogni combinazione di lettere, trattini e numeri può seguire la prima lettera. Per essere facilmente distinguibile, il nome delle variabili è preceduto nel testo da “&”.
- *Comandi di programma*: sono valutati ed eseguiti quando la macro è richiamata in ambiente SAS. Molti comandi possono essere usati nelle macro ma non in codice aperto. Solitamente sono riconoscibili perché preceduti da “%”. (%LET, %DO)
- *Macro espressioni*: funzionano come nell’ordinario ambiente SAS, salvo che lavorano con variabili macro. Possono essere usate per condizionare il flusso del processo o per creare nuove assegnazioni di valore. (%IF, %THEN)
- *Macro funzioni*: operano su le variabili macro e sul testo costante. Oltre quelle usate anche in DATA step ci sono un certo numero di funzioni uniche del linguaggio macro. Lo stesso tipo di funzione può essere eseguita con nomenclature differenti; ad esempio, se si vuole scrivere una parola in maiuscolo si possono utilizzare sia %UPCASE che %SYSFUNC.

Di seguito si presentano le principali funzioni usate per elaborare la macro sviluppata nel lavoro.

3.3. Funzioni e Procedure della macro PvalueADJ.SAS

Le possibilità e le funzioni messe a disposizione dal linguaggio delle macro sono molte. Un programmatore dotato di esperienza e fantasia ha sicuramente opportunità.

Nel caso in esame non si ha la pretesa di essere esaustivi sul SAS (non si sta preparando un manuale) l'intento è almeno di essere chiari sui punti trattati.

Creazione di una macro

La macro può iniziare e terminare con:

```
% PvalueADJ(dati da analizzare, definizione variabili, ed altri input che serviranno a far girare il programma);
```

```
.....codice della macro
```

```
%MEND PvalueADJ;
```

PvalueADJ è il nome della macro. I commenti nel programma, necessari per renderlo leggibile sono compresi tra /*...*/. In tal modo non interagiscono con le funzioni e i comandi del programma stesso.

Definizione e uso di una variabile Macro

Il modo più facile per definire una variabile è usare **%LET**.

%LET è seguito dal nome della variabile, dal segno uguale "=" e dal valore da assegnare a questa variabile.

```
%LET nome variabile = testo o valore;           %LET pippo = 5;
```

Il linguaggio macro non supporta i valori mancanti, si parla solo di valore nullo:

```
%LET pippo =;                               %LET pippo = ;
```

La variabile "pippo" verrà richiamata nel programma dove è necessario, apponendo

davanti al nome “&”, &pippo.

%PUT permette di scrivere del testo in LOG. Per essere eseguito, diversamente dell’ambiente SAS base, non richiede di essere inserito in un DATA step.

```
%LET pippo = ROSA;
```

```
%PUT          Chi vince il giro d’Italia, vince la maglia &pippo;
```

Nel LOG apparirà “Chi vince il giro d’Italia, vince la maglia **ROSA**”

Quest’ultimo comando è piuttosto comodo perché semplifica la leggibilità delle routine. Infatti, il SAS solitamente non dà indicazioni su come certe parti del programma girino in LOG. Inserendo %PUT si riesce a capire e osservare meglio come una variabile si comporta nei vari step della routine, rendendosi conto a che punto si è arrivati. Un altro comando utile per rendere più leggibile il LOG è **SYMBOLGEN**, che permette di vedere con un breve commento quando una variabile è risolta.

```
OPTION SYMBOLGEN;
```

Non si è ritenuto di dover aggiungere altro al LOG per evitare diventasse troppo prolisso e dispersivo.

```
ODS OUTPUT SELECTION ANOVA=SS;
```

permette di immagazzinare i risultati di una determinata procedura in un certo dataset, per poterli richiamare ed utilizzare in successive elaborazioni. Nello specifico, ANOVA è il dataset SAS creato da specifiche statistiche ottenute nel processo della stepwise. Si è deciso di salvare nella tabella SS. La tabella ANOVA è stata individuata tramite l’utilizzo di:

```
ODS TRACE ON
```

```
ODS TRACE OFF
```

Che debitamente posizionate in “PROC”, mappano in LOG i vari dataset dei risultati, creati nell’applicazione del programma.

Esecuzioni iterative

%DO e %WHILE questi comandi sono simili a quelli usati nei corrispondenti DATA step nel SAS base. Queste due funzioni sono solo alcune delle tante iterazioni logiche che si possono usare. Permettono di eseguire un blocco di operazioni ripetutamente finchè non si verifica una determinata condizione:

```
%MACRO ALLYR(start, stop);

    %LET YEAR = &START;

    %DO %WHILE (&YEAR<= &STOP);

        DATA TEMP;

            SET YR&YEAR;

            YEAR = 1900 + &YEAR;

        RUN;

        PROC APPEND BASE=ALLYEAR DATA=TEMP

    %LET YEAR= %EVAL(&YEAR + 1);

    %END;

%MEND ALLYR;
```

In questo esempio, l'utente assegna l'inizio (START) e la fine (STOP) come input, che saranno richiamati nella macro.

Alla variabile YEAR si dà START come valore di partenza e si inizia il ciclo. Si fanno una serie di operazioni finché la variabile YEAR raggiunge il valore di STOP, utilizzando la variabile contatore:

```
%LET YEAR=%EVAL(&YEAR+1)
```

che incrementa la variabile &YEAR, per numeri interi, fino al limite superiore "STOP".

Si sono utilizzate queste routine, ad esempio, per creare sub-dataset da altri. La procedura è stata introdotta all'interno di più cicli, ed è stato inoltre possibile specificare che il nuovo dataset contenesse solo una parte del precedente, dove ad esempio, una certa variabile aveva un certo valore (WHERE):

```
%DO %WHILE (&j<=&nsample);
```

```
DATA SOTTOMATRICE&j (WHERE=( _SAMPLE_ =&j));
```

```
SET SASUSER.PERMUTA;
```

```
RUN;
```

```
%LET j=%EVAL(&j+1);
```

```
%END;
```

Per stampare un dataset in OUTPUT, si potrà usare PROC PRINT

```
PROC PRINT DATA=SASUSER.PVECT;
```

```
TITLE 'VETTORE P-VALUE';
```

```
RUN;
```

Con TITLE si può far apparire del testo a spiegazione di ciò che abbiamo stampato in OUTPUT.

L'esempio seguente, non crea nessun dataset, ma ha consentito di lavorare con essi, infatti, permette di assegnare il valore contenuto nella colonna riferita alla variabile “_temg001” del dataset (WORK.COUNT) j-esimo, ad una variabile di nome “COUNT”, che si potrà richiamare nei passi successivi.

```
DATA _NULL_;
```

```
SET WORK.COUNT&i.;
```

```
CALL SYMPUT("count",_temg001);
```

```
RUN;
```

Analisi dai dati

PROC REG è una delle molte procedure di regressione disponibili nel SAS. Questa è una procedura generale per la regressione, mentre altre hanno applicazioni più specifiche.

```
PROC REG;
```

```
MODEL variabile dipendente = variabile esplicativa;
```

```
RUN;
```

```
QUIT;
```

“RUN” sottomette la procedura, e “QUIT” dichiara la sua conclusione.

PROC REG consente molte possibilità d’analisi:

- Fornisce sette metodi di selezione;
- Cambia modelli e dati interattivamente;
- Lavora con molteplici MODEL comandi;
- Fornisce previsioni, residui, residui standardizzati, bande di confidenza, che possono essere salvati in un dataset;
- Può fornire statistiche speciali;
- Testa ipotesi lineari;
- Testa ipotesi multiple;
- e molte altre...

La procedura usa la stima dei quadrati per l’analisi di modelli lineari regressivi. Importante è la possibilità di servirsi del metodo di regressione stepwise¹⁶, utilissima per le analisi esplorative.

¹⁶ Un dettaglio dei vari metodi di selezione lo si può trovare in NOTE.

Specificazione della procedura:

PROC REG *opzioni*;

MODEL *variabile dipendente*=*regressori* / *opzioni*;

VAR *variabili*;

FREQ *variabile*;

WEIGHT *variabile*;

ID *variabile*;

PRINT *opzioni*;

ADD *variabili*;

DELETE *variabili*;

DELOBS *numero di osservazioni*;

OUTPUT OUT=*SASdataset* "*nome SAS del dataset X*"=*nome*;

TEST *equazione₁...equazione_k*/*opzioni*;

MTEST *equazione₁...equazione_k*/*opzioni*;

BY *variabili*;

....

In particolare alcune opzioni riferite nello specifico PROC REG possono essere:

DATA=: indica il nome del dataset che deve essere usato dalla procedura per ottenere le analisi e possibili modelli.

OUTEST= si richiede che i parametri stimati e le statistiche opzionali siano salvati in questo dataset.

OUTSSCP=richiede che la somma dei quadrati e la matrice dei prodotti incrociati siano salvati in questo dataset TYPE=SSCP.

NOPRINT permette di non stampare il normale output in OUTPUT.

MODEL *variabile dipendente*=*regressori* / *opzioni*;

dopo aver specificato il modello con la variabile dipendente e i vari regressori. Si possono specificare delle opzioni, tra cui:

NOPRINT: non stampa i normali risultati della regressione.

NOINT: esclude l'intercetta dal modello.

ALL: permette di avere le caratteristiche dei dettagli di stima, somma dei quadrati, coefficienti standardizzati, matrice di correlazione, la collinearità.

SELECTION permette di specificare il metodo che si vuole adottare, stepwise, forward (f), none per il modello completo, ecc.

SLENTRY o SLE\SLSTAY o SLS rispettivamente criteri di entrata e permanenza per il metodo stepwise.

Molte altre sono le opzioni disponibili per la parte MODEL sulle regressioni: i dettagli di stima, le previsioni e i residui sui risultati di ogni specifico metodo di selezione scelto. Per maggiori dettagli si rimanda ad una guida o documentazione SAS.

VAR *variabili*;

Con questa opzione è possibile aggiungere delle variabili, non specificate nel precedente modello, nella matrice dei prodotti incrociati.

FREQ *variabile*;

E' utile quando una variabile nel dataset rappresenta la frequenza delle osservazioni. Questo è tenuto in considerazione nelle analisi come una sorta di "peso" dell'osservazione. Molto simile all'opzione WEIGHT, ma calcola diversamente i gradi di libertà.

WEIGHT variabile;

E' utilizzabile quando nel gruppo delle variabili ce n'è una che rappresenta il peso delle osservazioni, che non necessariamente corrispondono alle frequenze con cui esse si presentano. Sotto determinati presupposti è utile per il calcolo del miglior stimatore non distorto (BLUE).

ID variabile;

Associa e quindi individua le osservazioni tramite un'altra variabile

PRINT opzioni;

Raccoglie le varie opzioni di MODEL in modo interattivo, scegliendole.

ADD variabili;

Permette di aggiungere variabili indipendenti, precedentemente listate in MODEL o VAR, è utilizzabile interattivamente per allargare il modello.

DELETE variabili;

Permette al contrario della precedente opzione di cancellare variabili indipendenti dal modello. E perciò utilizzabile interattivamente per ridurre il modello.

DELOBS numero di osservazioni;

Elimina un dato numero di osservazioni dai calcoli.

TEST equazione₁...equazione_k/opzioni;

Permette di verificare delle ipotesi riguardo i parametri di un modello precedentemente dato. Per esempio, particolari relazioni che devono esistere o meno.

MTEST equazione₁...equazione_k/opzioni;

Come nel precedente caso multivariato, ma con più variabili dipendenti.

BY variabili;

Può essere utile per fare analisi separate su gruppi di osservazioni. Quando riceve questo input organizza i dati, ordinandoli secondo questa variabile.

PROC REG è piuttosto versatile ed in base alle varie opzioni scelte si hanno differenti opportunità, come la diagnostica sulla collinearità, problema che rende le stime instabili e gli standard errors alti.

Creazione, manutenzione, e gestione dei dataset

PROC SQL, dove SQL è l'acronimo di Structured Query Language, è una procedura utilizzata in molti altri programmi. E' piuttosto flessibile e utile nella creazione, manutenzione, e modifica delle tabelle contenenti dati utilizzati nei programmi o dei risultati delle loro elaborazioni.

PROC SQL;

.....

QUIT;

“QUIT” ne dichiara al conclusione

Tra le sue molteplici funzioni:

- Crea estrazioni di dati da dataset SAS (è utilizzata anche nella macro).
- Genera report da dataset.
- Permette di combinare i dati SAS in molti modi, nel lavoro si sono divise e unite tabelle, eliminate o incluse colonne.
- Crea, aggiorna e cancella dati

Molte funzioni dell'SQL sono eseguibili anche con i soliti passi DATA ma, come detto essendo un linguaggio piuttosto comune, il suo utilizzo facilita l'interpretazione del programma, anche a chi non è esperto di SAS.

PROC MULTTEST è nata per controllare la verifica multipla delle ipotesi su un campione di dati. Questo test dà la possibilità di aggiustare i p-value in modo da controllare l'errore di primo tipo (FWE). Si può scegliere tra differenti possibilità:

- Bonferroni
- Sidak
- Stepdown methods
- Hochberg
- Hommel
- Fisher Combination
- False Discovery Rate
- Bootstrap
- Permutation

L'aggiustamento dei p-value è definito come il più piccolo livello di significatività per il quale le ipotesi date potrebbero essere rigettate, quando l'intera famiglia di test è considerato, si rigetta l'ipotesi nulla quando i p-value aggiustato è più piccolo di α ; in molti casi, questa procedura controlla il FWE al livello o sotto.

Quello che dovrebbe interessare nello studio di un fenomeno non è solo controllare l'FWE, ma soprattutto riuscire a individuare un dataset ottimo di ipotesi. Questa funzione si limita alla sola verifica di quelle che le vengono sottoposte. Ci si è perciò serviti della funzione non nel suo utilizzo standard, ma in modo improprio, per sfruttarne la funzionalità legata al ricampionamento, come strumento nella costruzione del nostro algoritmo. Con tale funzione si generano, tramite le permutazioni nuovi campioni dall'originale. In tal modo si possono applicare altre stepwise, i cui p-value saranno utilizzati, per aggiustare il p-value della stepwise sui dati originali.

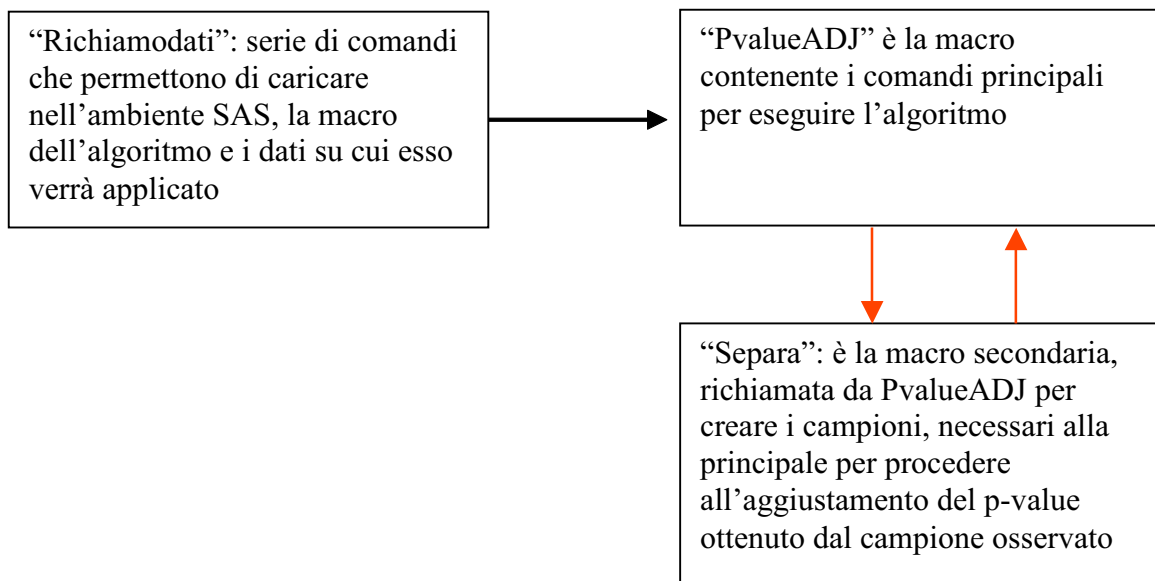
PROC IML (Iterative Matrix Language), procedura che come altre presenta una molteplicità di opzioni ed eccezioni. In particolare è stata utilizzata per elaborare i dataset, facendoli interpretare al SAS come matrici, su cui si sono applicate determinate operazioni, al fine di ottenere i risultati necessari.

Richiamare una macro secondaria

Per necessità operative, a volte, annidate in una macro ce ne possono essere delle altre, in quanto la macro principale per ottenere un determinato risultato necessita di sub-routine, per cui ad un certo punto nel flusso di processo, sono richiamate macro secondarie. Il processo prevede e permette lo scambio di parametri o le variabili o le risultati. Le sub-routine possono essere veri e propri programmi, che si può decidere di tenere separati dalle macro principale, per agevolarne la manutenzione e la leggibilità. Nel nostro caso, all'interno della macro principale:

```
%include 'C:\pericorso\separa.sas';
```

```
%separa(&nsample.);
```

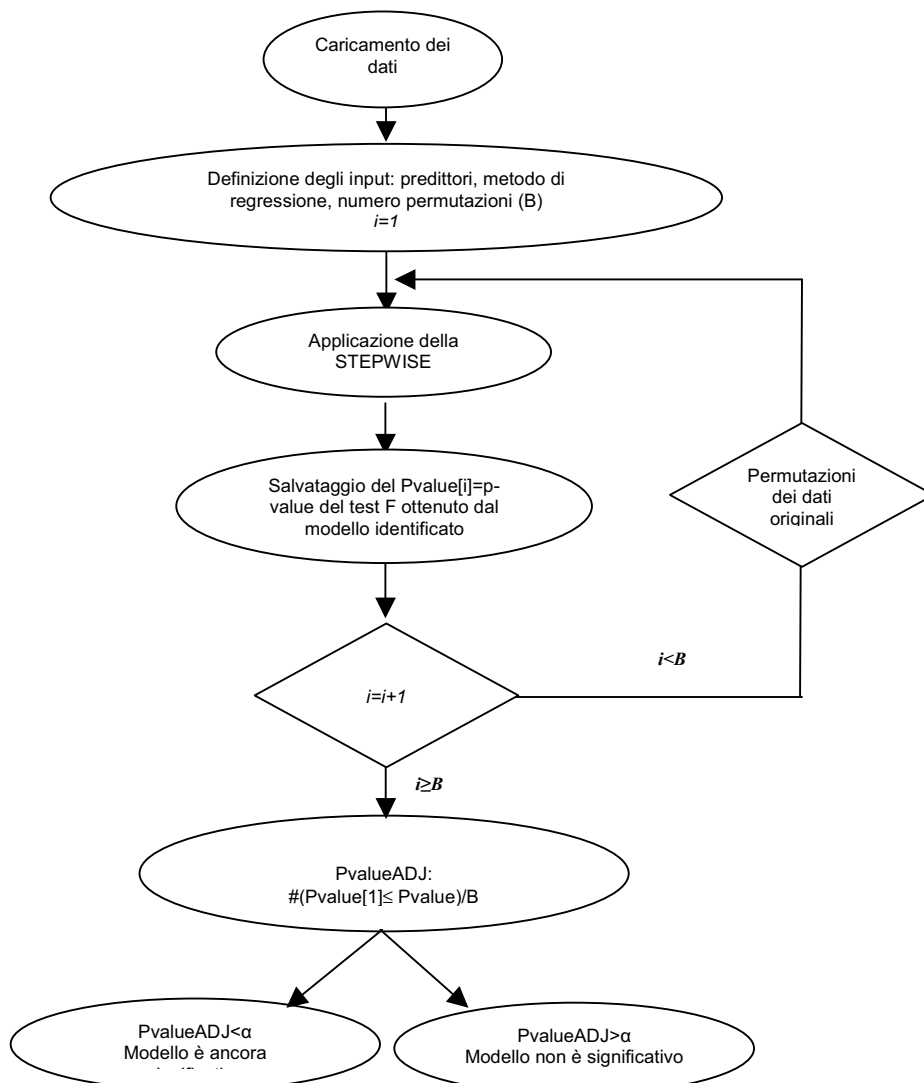


4. Macro SAS: Applicazione del TEST DI PERMUTAZIONE alla regressione STEPWISE.

4.1. Logica dell'algoritmo

Per lo sviluppo della macro si è seguita la proposta presentata nel paragrafo 2.2, ottenendo l'algoritmo sottostante.

Attraverso un approccio non parametrico si è cercato di ovviare ai problemi di affidabilità di stima. Applicando delle permutazioni sulla variabile dipendente, sono creati nuovi campioni su cui eseguire delle stepwise, i cui p-value sono serviti per correggere il p-value globale del modello ottenuto dai dati originali, che ci informa sulla reale validità della stima ottenuta



4.2.Criticità della macro

Nella costruzione della macro le criticità principali incontrate sono state:

- capire come ottenere\formare gli B+1 campioni ottenuti tramite B permutazioni della variabile dipendente Y:

$(Y_{Oss} X_1 \dots X_K) (Y_1 X_1 \dots X_K) \dots (Y_B X_1 \dots X_K)$

- “conservare” i campioni e applicarvi un ciclo di B+1 stepwise
- capire come raccogliere B+1 p-value della stepwise, stoccandoli in un vettore
- conservare i p-value e raggrupparli tutti assieme per poter applicare la correzione

Si è deciso quindi dividere le questioni, risolvendole indipendentemente una dall'altra, poiché la cosa risultava più snella in ambiente SAS. Una volta ottenute le varie soluzioni, dalla loro unione, si sarebbe ottenuto il diagramma esposto sopra.

Si otterranno tre file, tramite uno si caricheranno anche gli altri nell'ambiente SAS:

- Richiamodati: carica dati e la macro principale
- pvalueadj: esegue l'algoritmo e carica una macro secondaria
- separa: genera B campioni creati dalle permutazioni dalla variabile dipendente.

A seguire, il dettaglio di quanto appena illustrato. I commenti sono evidenziati in blue e i risultati in verde.

4.3. Richiamo di una macro in ambiente SAS e caricamento dati: Richiamodati.SAS

1. In questa prova, per comodità, si caricherà direttamente un esempio di SAS, che sarà utilizzato per formare il primo database, nominando le variabili come segue.

```
data datitesi;          /* si utilizzano dati di prova*/

infile cards expandtabs;

input R Age S Ed Ex0 Ex1 LF M N NW U1 U2 W X;

label R='variabile dipendente del modello da cercare';

cards;

79.1      151      1      91      58      56      510      950      33      301      108
      41      394      261

....;
```

Il database così formato potrà essere anche utilizzato successivamente per analisi o ulteriori manipolazioni dalle varie macro.

2. Con i seguenti comandi:

```
%include 'C:\PERCORSO\ PvalueADJ.SAS '; (1)

%PvalueADJ(datitesi(2),R(3),age—
x(4),13(4),13(5),3(6),47(7),stepwise(8),67890(9));

run;
```

si procede a:

1. richiamare la macro principale PvalueADJ.SAS
2. richiamare i dati su cui si intendo fare le analisi

3. definire la variabile dipendente, la numerosità campionaria
4. definire il range e numero dei predittori candidati.
5. fornire il numero di variabili che si intendono includere nel modello
6. dichiarare il B di permutazioni che si intendono applicare
7. precisare il numero di osservazioni nel campione
8. scegliere il tipo di stepwise da introdurre: stepwise, forward, backward
9. fornire il valore di inizio per il generatore random del ricampionamento.

Gli input che si daranno in questo step, potranno essere ripresi dalle varie macro che si utilizzeranno e si esamineranno in seguito.

Alla fine si otterranno i seguenti risultati:

- La stepwise e p value del test F del modello selezionato dai dati del campione originale
 - La stepwise e p value del test F dei modelli selezionati dai dati dei B nuovi campioni
 - Il p-value aggiustato
3. Nell'esempio si hanno 47 osservazioni, 14 variabili, la variabile dipendente è stata designata come R per la stepwise si decide di considerare tutte le variabili e 3 permutazioni.

4.4.Macro: PvalueADJ.SAS

Come detto in precedenza per agire in modo snello in ambiente SAS si è pensato che l'ideale sarebbe stato dividere i problemi in sub-routine, lavorando indipendentemente una dall'altra per poi unirle, in modo da poter muoversi in direzioni separate e indipendenti nello sviluppo delle stesse. Infatti, in questa macro, ne troviamo altre due annidate che abbiamo chiamato SEPARA e EX_REG.

1. Carico gli input dichiarati in Richiamodati.sas:

%macro Richiamodati (INGRESSO, vardip, varmod, nvar, nvar_tot, nsample, noss, selection, nseed);

INGRESSO: dati del campione (datitesi)

vardip: variabile dipendente (R)

varmod: l'insieme delle variabili indipendenti (age—x)

nvar: numero di variabili da considerare nel modello (13)

nvar_tot: numero totale delle variabili indipendenti (13)

nsample: numero permutazioni (3)

NOSS: numero osservazioni (47)

Selection¹⁷: è possibile specificare il metodo che si preferisce tra quelli disponibili FORWARD (or F), BACKWARD (or B), o STEPWISE, e volendo anche MAXR, MINR, RSQUARE, ADJRSQ, CP (Stepwise)

NSEED: restituisce un valore d'inizio per il numero casuale generatore del ricampionamento (67890)

¹⁷ Vedi nota (1) alla fine del word.

2. Si fa una regressione dei dati originali e raccolgo i p-value

```
%LET N='0';
```

```
ODS OUTPUT ANOVA=SS; si specifica che i risultati della stepwise andranno salvati in SS
```

```
ODS TRACE ON; si tracciano i risultati
```

```
PROC REG DATA=&INGRESSO.; si richiamano i dati originali
```

```
MODEL &VARDIP.=&VARMOD. su questi si specifica il modello
```

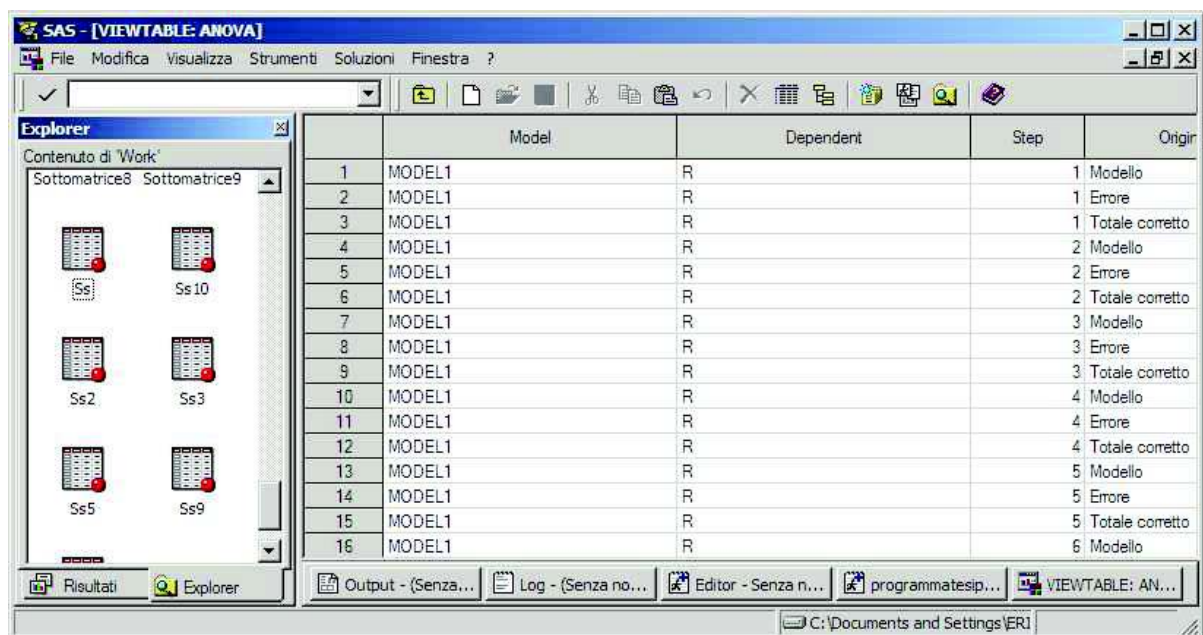
```
/SELECTION=&selection.; si indica il tipo di regressione che si vuole applicare
```

```
TITLE "Risultati dall'analisi di regressione sul modello &N.";
```

```
run;
```

```
ODS TRACE OFF;
```

In SAS sarà generata una tabella contenente tutte le statistiche riferite alla suddetta regressione, che sarà salvata nell'area temporanea e cancellata al termine della sessione di lavoro.



The screenshot shows the SAS software interface with the ANOVA table results. The table has five columns: Step, Model, Dependent, Step, and Origin. The data is as follows:

	Model	Dependent	Step	Origin
1	MODEL1	R	1	Modello
2	MODEL1	R	1	Errore
3	MODEL1	R	1	Totale corretto
4	MODEL1	R	2	Modello
5	MODEL1	R	2	Errore
6	MODEL1	R	2	Totale corretto
7	MODEL1	R	3	Modello
8	MODEL1	R	3	Errore
9	MODEL1	R	3	Totale corretto
10	MODEL1	R	4	Modello
11	MODEL1	R	4	Errore
12	MODEL1	R	4	Totale corretto
13	MODEL1	R	5	Modello
14	MODEL1	R	5	Errore
15	MODEL1	R	5	Totale corretto
16	MODEL1	R	6	Modello

3. Si trasferiscono i dati che serviranno per le elaborazioni successive da SS ad una vettore PVALTAB, che conterrà i p-value ad ogni passo della stepwise.

```
PROC SQL;
```

```
CREATE TABLE WORK.PVALTAB AS (SELECT STEP, SOURCE, ProbF AS P FROM  
WORK.SS WHERE SOURCE IN ("Modello"));
```

```
QUIT;
```

4. La seguente procedura servirà al punto successivo, in quanto si devono individuare e raccogliere solo i p-value calcolati all'ultimo step della regressione. Quindi si calcola il numero totale di righe di WORK.PVALTAB, che rappresenta anche il numero totale di step eseguiti dalla stepwise, si assegna quel valore ad una variabile, che si richiamerà e che rappresenta il p-value che serve

```
PROC SQL;
```

```
CREATE TABLE WORK.COUNT AS (SELECT COUNT(*) FROM WORK.PVALTAB);
```

```
QUIT;
```

```
DATA _NULL_;
```

```
SET WORK.COUNT;
```

```
CALL SYMPUT("COUNT",_temg001);
```

```
RUN;
```

5. Si crea quindi il PVECT in cui in seguito si immagazzineranno anche i p-value all'ultimo passo delle altre stepwise sui campioni generati dalle permutazioni sulla variabile dipendente.

```
PROC SQL;
```

```
CREATE TABLE WORK.PVECT AS (SELECT PVALTAB.P AS PV FROM WORK.PVALTAB  
WHERE STEP=&count);
```

```
QUIT;
```

Di seguito le due matrici create e il contenuto di PVECT a questo punto.

The left screenshot shows the SAS Output window with the following text:

Risultati dall'analisi di regressione sul modello '1'

La procedura REG
Modello: MODEL1
Variabile dipendente: B variabile dipendente del modello da cercare

Numero di osservazioni lette 47
Numero di osservazioni usate 47

Selezione stepwise: Passo 1

Variabile Ex0 entrata: R-quadro = 0.4728 e C(p) = 32.3913

Analisi della varianza

Origine	DF	Somma dei quadrati	Media dei quadrati	Valore F
Modello	1	32533	32533	40.36
Errore	45	36276	806.13947	
Totale corretto	46	68809		

Tabella dei parametri

Variabile	Stima dei parametri	Errore standard	BS Tipo II	Valore F	Pr > F
Interc	14.44840	12.66926	1048.15843	1.30	0.260
Ex0	0.89485	0.14086	32533	40.36	<.0001

Limiti sul numero di condizioni: 1, 1

Selezione stepwise: Passo 2

Variabile X entrata: R-quadro = 0.5803 e C(p) = 19.0160

The right screenshot shows the SAS Explorer window with the following table:

	Pr > F
1	<.0001

6. Con PROC MULTTEST si è riusciti a permutare direttamente i dati. Il comando è stato utilizzato in modo improprio, ma piuttosto efficace:

```
PROC MULTTEST DATA=&INGRESSO. PERMUTATION NSAMPLE=&NSAMPLE.  
NOPRINT SEED=&NSEED. OUTSAMP=PERM;
```

```
TEST MEAN (&VARDIP.);
```

```
CLASS &VARDIP.;
```

```
RUN;
```

Dopo aver specificato il database d'origine, il numero di permutazioni il tipo di permutazione, e la variabile da permutare, otteniamo una tabella chiamata PERM che contiene i valori delle Y.

	Sample	Class	Observation	R
40	1	122.5	19	75
41	1	123.4	15	79.8
42	1	127.2	8	155.5
43	1	155.5	24	96.8
44	1	163.5	25	52.3
45	1	167.4	43	82.3
46	1	196.9	39	82.6
47	1	199.3	12	84.9
48	2	34.2	46	50.8
49	2	37.3	32	75.4
50	2	43.9	6	68.2
51	2	45.5	20	122.5
52	2	50.8	21	74.2
53	2	51.1	3	57.8
54	2	52.3	7	96.8
55	2	53.9	10	70.5
56	2	54.2	44	103
57	2	56.6	33	107.2
58	2	57.8	26	199.3
59	2	65.3	36	127.2
60	2	66.4	16	94.6
61	2	68.2	14	66.4

7. I risultati ottenuti nel punto precedente, cioè i B sub-campioni di Y (nel particolare le colonne SAMPLE ed R di PERM), saranno salvati in una tabella chiamata PERMUTA che SEPARA.SAS userà per formare nuovi B campioni.

PROC SQL;

CREATE TABLE WORK.PERMUTA AS (SELECT _SAMPLE_,&VARDIP. FROM WORK.PERM);

QUIT;

Di seguito la tabella WORK.PERMUTA, che contiene uno sotto l'altro il riferimento e i valori di questi Y sub-campioni.

	Sample	R	
38	1	122.5	
39	1	121.6	
40	1	75	
41	1	70.0	Y1
42	1	155.5	
43	1	96.8	
44	1	52.3	
45	1	82.3	
46	1	82.6	
47	1	84.9	
48	2	50.8	
49	2	75.4	
50	2	68.2	
51	2	122.5	Y2
52	2	74.2	
53	2	57.8	
54	2	96.3	
55	2	70.5	
56	2	103	
57	2	107.2	
58	2	199.3	
59	2	127.2	

Dove "b" è il numero di permutazioni richieste

8. A questo punto si richiama SEPARA.SAS:

```
%INCLUDE 'C:\percorso\separa.sas';
```

```
%SEPARA(&nsample.);
```

```
RUN;
```

9. Come input, la macro secondaria richiede solo il numero di campioni (*nsample*) che si dovranno creare, cioè quante sono le permutazioni fatte. Di seguito il codice di SEPARA.SAS in rosso.

La macro, quindi permetterà di costruire B nuovi campioni prendendo le varie Y permutate, ottenute al punto 7, unendole una alla volta al “blocco” delle sole esplicative .

%MACRO SEPARA (nsample);

10. Si inizia il lavoro con un ciclo che va da 1 a B, numero di permutazioni\subcampioni

%LET j=1;

%DO %WHILE (&j<=&nsample);

%PUT THE VALUE OF j IS &j;

11. Nel ciclo si divide PERMUTA in submatrici(j) con j=1...B, B numero totale di permutazioni.

(j) è individuato nella prima colonna di PERMUTA (SAMPLE).

Le nuove submatrici includeranno tutti i valori della variabili dipendente (R) associati a j=i, con i=1...B

Row	Sample	R
32	1	57.3
33	1	100
34	1	189.9
35	1	58.2
36	1	83.1
37	1	186.9
38	1	122.9
39	1	121.6
40	1	75
41	1	79.8
42	1	155.5
43	1	96.8
44	1	52.9
45	1	82.3
46	1	82.8
47	1	74.9
48	2	50.8
49	2	75.4
50	2	88.2
51	2	122.9
52	2	74.2
53	2	87.8

```

DATA SOTTOMATRICE&j (WHERE=( _SAMPLE_ =&j));
SET WORK.PERMUTA;
RUN;
%LET j=%EVAL(&j+1);
%END;

```

Nel caso in esempio si sono specificate 3 permutazioni e quindi si otterranno 3 submatrici:

	Sample	R
38	1	122.5
39	1	121.6
40	1	75
41	1	79.8
42	1	155.5
43	1	96.8
44	1	52.3
45	1	82.3
46	1	82.6
47	1	84.9
48	2	50.8
49	2	75.4
50	2	68.2
51	2	122.5
52	2	74.2
53	2	57.8
54	2	96.3
55	2	70.5
56	2	103
57	2	107.2
58	2	199.3
59	2	127.2

12. Con le B submatrici, ora si costruiscono B nuovi campioni, associando le nuove B variabili dipendenti al blocco delle esplicative originali, lasciate inalterate.

Per cui,, si prende il campione originale, togliendo la variabile dipendente, e nominando questo nuovo set di dati “VARMOD”, lo si unisce con un “MERGE” alle B nuove sottomatrici create, e che saranno chiamate INGRESSOj, con j=1...B.

```

DATA VARMOD; SET &INGRESSO; DROP &VARDIP.;
RUN;
%LET j=1;
%DO %WHILE (&j<=&NSAMPLE);

```

```

DATA INGRESSO&J;
MERGE SOTTOMATRICE&j VARMOD;
DROP _SAMPLE_;
RUN;
%LET j=%eval(&j+1);
%end;
%MEND SEPARA;

```

Con questo ultimo passo, la macro ha espletato alla sua funzione. Infatti, esplorando una delle tabelle (ex: INGRESSO1), si capisce come la variabile dipendente permutata ($Y^*=R^*$) sia stata associata al set delle esplicative, dando origine a uno dei B nuovi campioni.

The screenshot shows the SAS Explorer interface. On the left, a directory tree under 'Contenuto di Work' contains several datasets: Datitesi, Ingresso1 (highlighted with a red box), Ingresso2, Ingresso3, Perm, Sasmacr, Sottomatrice1, Sottomatrice2, Sottomatrice3, and Varmod. On the right, the 'VIEWTABLE: Work.Ingresso1' window displays a data table with 20 rows and 8 columns: R, Age, S, Ed, Ex0, and Ex1. The first few rows are as follows:

	R	Age	S	Ed	Ex0	Ex1
1	51.1	151	1	91	58	
2	70.5	143	0	113	103	
3	34.2	142	1	89	45	
4	92.3	136	0	121	149	1
5	92.9	141	0	121	109	1
6	37.3	121	0	110	118	1
7	127.2	127	1	111	82	
8	50.8	131	1	109	115	1
9	45.5	157	1	90	65	
10	74.2	140	0	118	71	
11	53.9	124	0	105	121	1
12	96.3	134	0	108	75	
13	167.4	128	0	113	67	
14	75.4	135	0	117	62	
15	84.9	152	1	87	57	
16	94.6	142	1	88	81	
17	107.2	143	0	110	66	
18	85.6	135	1	104	123	1
19	104.3	130	0	116	128	1
20	56.6	125	0	108	113	1

13. Di seguito si utilizza la macro EX_REG, annidata nella principale, per l'applicazione della procedura stepwise su ciascuno dei B campioni ottenuti:

```

%macro ex_reg(ds=);           Richiama il/i campioni su cui deve lavorare

ODS OUTPUT SELECTIONSUMMARY=SS&i.;   indirizza i risultati verso SS; dove i=1..B

ODS TRACE ON;                traccia i risultati

```



```

proc reg data=&ds;                                indica su che dati fare la regressione

model &vardip.=&varmod.                          definisce il modello

/selection=&selection;                            definisce il tipo di regressione

run;

ODS TRACE OFF;

```

Le procedure che seguono sono le stesse eseguite sulla stepwise dei dati originali, iterate per i nuovi B campioni:

14. Sii raccolgono i risultati in PVALUTAB_i pescandoli dagli SS_i, con $i=1 \dots B$

```

PROC SQL;

CREATE TABLE WORK.PVALTAB&i. AS (SELECT SS&i..STEP, SS&i..Fvalue, SS&i..ProbF as P
FROM WORK.SS&i. WHERE SOURCE IN ("Modello"));

QUIT;

```

15. Da PVALUTAB_i li si salva in PVECT_i, prendendo solo quelli che sono riferiti all'ultimo step della procedura di regressione.

```

DATA _NULL_;

SET WORK.COUNT&i.;

CALL SYMPUT("COUNT",_temg001);

RUN;

PROC SQL;

CREATE TABLE WORK.PVECT&i. AS (SELECT PVALTAB&i..P AS PV FROM
WORK.PVALTAB&i. where step=&count);

QUIT;

```

16. Si unisce il p-value del campione originale con gli altri B, e si chiude la macro:

```
PROC APPEND BASE=WORK.PVECT DATA=WORK.PVECT&i.;
```

```
RUN;
```

```
QUIT;
```

```
%MEND;
```

17. Per sottomettere la macro EX_REG nel momento che interessa:

```
%MACRO MAIN;
```

```
%LET N=&nsample.;
```

qui si capisce quante stepwise si eseguiranno

```
%LET i=0;
```

si inizializza la variabile "i" che non dovrà essere maggiore di n (n. permutazioni)

```
%PUT &i. &n. ;
```

```
%DO %WHILE (&i.<&n.);
```

```
%LET i=%EVAL(&i.+1);
```

```
%EX_REG(DS=INGRESSO&i.);
```

```
%PUT MODELLO &i.;
```

si rendono più leggibili i risultati, specificando su che campione si sta eseguendo la procedura

```
%END;
```

```
%MEND;
```

```
OPTION MACROGEN SYMBOLGEN;
```

```
%MAIN;
```

si termina l'operazione

18. In seguito si termina utilizzando i pvalue immagazzinati in precedenza per correggere il pvalue del modello osservato. Per fare ciò si utilizza la PROC IML, che permette di compiere operazioni matriciali:

```
PROC IML;
```

```
USE WORK.PVECT;
```

```
Read ALL into X;
```

```
TITLE "Riepilogo";
```

```
NRAW=NRAW(X);
```

```
E=&NSAMPLE.-NRAW;
```

```
A=J(E,1,1);
```

```
PV=X//A;
```

```
NRAWPV=NRAW(PV);
```

```
PRINT PV;
```

```
COUNT=0;
```

```
DO i=1 to NRAWPV;
```

```
IF PV[1]>=PV[i] THEN COUNT=COUNT+1;
```

si conta il numero di pvalue da permutazioni minori o uguali del p-value dei dati osservati

```
END;
```

```
PVALUE_ADJ= COUNT/&NSAMPLE.;
```

si calcola il p-value aggiustato

```
PRINT PVALUE_ADJ;
```

si stampano nell'output i risultati attenuti

```
QUIT;
```

```
%MEND PVALUEADJ;
```

fine macro.

4.5. Conclusioni

La macro al momento presenta due limiti:

- Non si riesce a nascondere l'output delle stepwise che non interessa. Durante la regressione su 1000 campioni, chiede di pulire la pagina dell'output manualmente.
- Non è possibile definire a priori e contemporaneamente i criteri di accesso\permanenza delle variabili per i vari metodi stepwise. Anche se le modifiche sono semplici, si deve comunque creare una macro specifica per ciascuno metodo.

La macro è sicuramente migliorabile, e raffinabile, ma risponde allo scopo per cui è stata ideata. Essa utilizza la PROC REG, ma è possibile adattarla facilmente anche ad altri metodi di stima piuttosto.

Conclusioni

In questo lavoro si è visto come garantire la validità dell'inferenza fatta a partire da modelli selezionati tramite metodi stepwise e si è realizzato una macro SAS che ne favorisce una agevole applicazione.

Si è scelto di soffermarsi sulla stepwise perchè è una metodologia piuttosto diffusa, dato che permette di selezionare efficientemente ed efficacemente di modelli esplicativi senza dover valutare esplicitamente tutte le possibili soluzioni.

La fase di modellizzazione comprende delle forme di "sintesi" e semplificazione nella descrizione del fenomeno.

Questo implica, solitamente la riduzione del totale di predittori inclusi nel modello. La selezione viene fatta mantenendo le variabili che risultano maggiormente associate alla variabile risposta, con lo scopo generico di mantenere una adeguata rappresentazione del fenomeno nel suo complesso.

A fronte delle grandi potenzialità di tale metodo, emerge una lacuna legata all'affidabilità dei risultati ottenuti. Infatti, i comuni test impiegati per verificare la presenza di un legame tra il complesso di variabili selezionate e la variabile dipendente perdono il controllo della probabilità di errore di primo tipo. I p-value ottenuti da questi test risulteranno quindi troppo ottimistici (cioè troppo bassi).

Con gli attuali strumenti computazionali si è potuto introdurre metodologie prima difficilmente utilizzabili, questo senza stravolgere, o suggerendo metodi alternativi di stima, ma semplicemente aggiungendo all'inferenza fatta una regola extra, per rafforzare la validità dei risultati ottenuti.

Per esprimere tutto questo, si è scelto il SAS perchè come la stepwise è uno strumento piuttosto conosciuto.

La macro è facilmente estendibile a modelli più generali, come *glm* e così pure, ai differenti metodi stepwise.

NOTE

Metodi di selezione del modello (dal SAS manual users on line)

I nove metodi di selezione del modello implementati dalla PROC REF sono specificati con SELECTION= opzione nella definizione del modello. Ogni metodo è discusso in questa sezione.

Modello completo (NONE)

Questo metodo è di default, e non fornisce la possibilità di selezione di un modello. Il modello completo specificato in MODEL è usato per stimare il modello. Per molte analisi regressive, questo può essere il sono metodi di cui si ha bisogno.

Selezione Forward (FORWARD)

La tecnica di selezione forward inizia con nessuna variabile nel modello. Per ogni variabile indipendente, la forward calcola la statistica F, che riflette il contributo della variabile nel modello, se inclusa. Il p-value per queste statistiche F sono comparate a SLENTY=valore che può essere specificato da noi, o di 0.5 se omesso. Se la statistica F non ha un livello di significatività maggiore di SLENTY (criterio d'accesso), la forward si ferma, altrimenti il metodo aggiunge la variabile che ha la più grande statistica F al modello. Il metodo allora calcola la F per le variabili che rimangono ancora fuori, e si riavvia il processo di valutazione. Così le variabili sono aggiunte ad una ad una fino a che non restano variabili che producano una statistica F. Una volta che una variabile è nel modello, lì resta.

Backward Elimination (BACKWARD)

L'eliminazione backward inizia calcolando la statistica F per il modello complete, con tutte le variabili indipendenti. Allora le variabili sono cancellate dal modello una a una fino a che le variabili rimanti nel modello producono un statistica F significativa al valore SLSTAY (criterio di permanenza), che di default è 0.1. Ad ogni passo la variabile che mostra il contributo più piccolo viene cancellata.

Stepwise (STEPWISE)

Il metodo stepwise è una modifica della selezione forward e differisce nel fatto che le variabili nel modello, una volta incluse possono anche essere tolte. Come la selezione forward, le variabili sono aggiunte una ad una al modello, e la statistica F per una variabile che può essere aggiunta deve essere significativa al livello SLENTY- dopo che una variabile è stata aggiunta comunque la stepwise guarda a tutte le variabili già incluse nel modello e cancella tutte quelle che non producono un F significativa al livello SLSTAY. Solo dopo che questo controllo è fatto e fatte le eliminazioni necessarie, si procede ad aggiungere un'altra variabile nel modello. Il processo finisce quando nessuna delle variabili fuori del modello ha un F significativa a SLENTY e le variabili introdotte sono significative al livello SLSTAY, o quando le variabili che devono essere aggiunte nel modello sono già state vagliate e eliminate.

Altri tipi di metodi che potrebbero essere considerati con qualche modifica:

Maximum R² Improvement (MAXR)

La tecnica di miglioramento massimo R² non si assesta su un singolo modello. Infatti, esso prova a trovare il migliore modello ad una variabile, il miglior modello a due variabile e così via, benché con esso non si ha la garanzia di trovare il modello con il più grande R² per ogni taglia.

Il MAXR inizia a trovare il modello univariate con il più alto R². Allora un'altra variabile, quella che produce il più grande incremento di R² viene aggiunta. Una volta ottenuto un modello bivariato, ciascuna delle variabili nel modello è comparata alle altre variabili non incluse. Per ciascuna comparazione, MAXR decide se rimuovere o meno una delle variabili e rimpiazzarla con un'altra variabile che incrementa R². Dopo la comparazione con tutte le possibili cambiamenti, MAXR fa le modifiche che produce il più alto R². Le comparazioni indiziano di nuovo e il processo continua fino a che MAXR trova che non ci sono cambiamenti che incrementano R². Così il modello bivariato è considerato il migliore che la tecnica possa trovare. Allora un'altra variabile è aggiunta e si riparte con il processo di comparazione e cambiamento fino a trovare il miglior modello trivariato, e così via.

La differenza tra stepwise e questo metodo è che tutti i cambiamenti sono valutati prima che siano fatti. Nella stepwise, la peggiore variabile può essere rimossa senza

considerare che si può concludere il processo aggiungendo la migliore variabile rimanente. MAXR può richiedere molto più tempo che la stepwise

Minimum R² (MINR) Improvement

Il MINR assomiglia MAXR, ma il cambiamento scelto è quello che produce il più piccolo incremento in R². Per un dato numero di variabili nel modello, MAXR e MINR solitamente producono lo stesso miglior modello, ma MINR considera più modelli per ogni misura

R² Selection (RSQUARE)

RSQUARE trova un sottoinsieme di variabili indipendenti che meglio rappresenta la dipendente con una regressione lineare nel campione dato. Si può specificare il numero massimo e minimo di variabili indipendenti da includere in un subset e il numero di subsets di diversa misura da selezionare. RSQUARE può fare efficientemente tutti i possibili subset di regressioni e mostrare i modelli che decrementano R² in ciascuna misura del subset. Altre statistiche sono disponibili per comparare i subset di differenti misure. Queste statistiche con i coefficienti della regressione stimata, possono essere mostrate o fornite da un SAS data set.

Il subset di modelli selezionato da questo metodo è ottimale in termini di R² per un dato campione, ma essi non sono necessariamente ottimali per la popolazione dal quale il campione è disegnato o per ogni altro campione per il quale si voglia fare predizione. Se un subset è selezionato per un modello sulla base di un ampio R² o su ogni altro criterio comunemente usato per la selezione di un modello, allora tutte le statistiche calcolate per quel modello sotto l'assunzione che il modello sia dato a priori, incluse tutte le statistiche calcolate da PROC REG, queste sono distorte. Il metodo è utile per costruire un modello esplicativo, ma per trovare un "vero" modello si dovrebbe implementare la teoria sottostante.

Ma questo metodo al contrario degli altri garantisce di trovare sempre un modello con un ampio R², esso necessita, comunque, di molto tempo nei calcoli, per cui quando si hanno molte variabili da vagliare viene preferito il metodo stepwise

Adjusted R² Selection (ADJRSQ)

Simile al precedente, ma per la selezione fine usato R^2 aggiustato

Mallows' C_p Selection (CP)

Simile ai precedenti solo che si usa la statistica Mallows' C_p , i modelli sono listati in ordine ascendente rispetto Mallows' C_p

Appendice

Richiamodati.sas

Data datitesi

```
infile cards expandtabs;
```

```
input R Age S Ed Ex0 Ex1 LF M N NW U1 U2 W X;
```

```
label R='Variabile dipendente del modello da cercare';
```

```
cards;
```

79.1	151	1	91	58	56	510	950	33	301	108	41
	394	261									
163.5	143	0	113	103	95	583	1012	13	102	96	36
	557	194									
57.8	142	1	89	45	44	533	969	18	219	94	33
	318	250									
196.9	136	0	121	149	141	577	994	157	80	102	39
	673	167									
123.4	141	0	121	109	101	591	985	18	30	91	20
	578	174									
68.2	121	0	110	118	115	547	964	25	44	84	29
	689	126									
96.3	127	1	111	82	79	519	982	4	139	97	38
	620	168									
155.5	131	1	109	115	109	542	969	50	179	79	35
	472	206									
85.6	157	1	90	65	62	553	955	39	286	81	28
	421	239									
70.5	140	0	118	71	68	632	1029	7	15	100	24
	526	174									
167.4	124	0	105	121	116	580	966	101	106	77	35
	657	170									
84.9	134	0	108	75	71	595	972	47	59	83	31
	580	172									
51.1	128	0	113	67	60	624	972	28	10	77	25
	507	206									
66.4	135	0	117	62	61	595	986	22	46	77	27
	529	190									
79.8	152	1	87	57	53	530	986	30	72	92	43
	405	264									
94.6	142	1	88	81	77	497	956	33	321	116	47
	427	247									
53.9	143	0	110	66	63	537	977	10	6	114	35
	487	166									
92.9	135	1	104	123	115	537	978	31	170	89	34
	631	165									
75.0	130	0	116	128	128	536	934	51	24	78	34
	627	135									

122.5	125	0	108	113	105	567	985	78	94	130	58
	626	166									
74.2	126	0	108	74	67	602	984	34	12	102	33
	557	195									
43.9	157	1	89	47	44	512	962	22	423	97	34
	288	276									
121.6	132	0	96	87	83	564	953	43	92	83	32
	513	227									
96.8	131	0	116	78	73	574	1038	7	36	142	42
	540	176									
52.3	130	0	116	63	57	641	984	14	26	70	21
	486	196									
199.3	131	0	121	160	143	631	1071	3	77	102	41
	674	152									
34.2	135	0	109	69	71	540	965	6	4	80	22
	564	139									
121.6	152	0	112	82	76	571	1018	10	79	103	28
	537	215									
104.3	119	0	107	166	157	521	938	168	89	92	36
	637	154									
69.6	166	1	89	58	54	521	973	46	254	72	26
	396	237									
37.3	140	0	93	55	54	535	1045	6	20	135	40
	453	200									
75.4	125	0	109	90	81	586	964	97	82	105	43
	617	163									
107.2	147	1	104	63	64	560	972	23	95	76	24
	462	233									
92.3	126	0	118	97	97	542	990	18	21	102	35
	589	166									
65.3	123	0	102	97	87	526	948	113	76	124	50
	572	158									
127.2	150	0	100	109	98	531	964	9	24	87	38
	559	153									
83.1	177	1	87	58	56	638	974	24	349	76	28
	382	254									
56.6	133	0	104	51	47	599	1024	7	40	99	27
	425	225									
82.6	149	1	88	61	54	515	953	36	165	86	35
	395	251									
115.1	145	1	104	82	74	560	981	96	126	88	31
	488	228									
88.0	148	0	122	72	66	601	998	9	19	84	20
	590	144									
54.2	141	0	109	56	54	523	968	4	2	107	37
	489	170									
82.3	162	1	99	75	70	522	996	40	208	73	27
	496	224									
103.0	136	0	121	95	96	574	1012	29	36	111	37
	622	162									
45.5	139	1	88	46	41	480	968	19	49	135	53
	457	249									
50.8	126	0	104	106	97	599	989	40	24	78	25
	593	171									
84.9	130	0	121	90	91	623	1049	3	22	113	40
	588	160									

```

;
%include 'C:\... \PValueADJp.sas';
%PValueADJ(datitesi,R,age--x,13,13,1000,47,STEPWISE,67890);

run;

```

PValueADJ.sas

```

%macro PValueADJ(ingresso,vardip,varmod,nvar,nvartot,nsample,noss,selection,nseed);

/*Regressione dei dati originali e raccolgo i p-vaue*/

%LET N='0';
  ODS OUTPUT ANOVA=SS;
  PROC REG DATA=&ingresso.;
    MODEL &vardip.=&varmod.
          /SELECTION=&selection.;

TITLE "Risultati dall'analisi di regressione sul campione &N.";

RUN;

PROC SQL;
  CREATE TABLE WORK.PVALTAB AS (SELECT STEP, SOURCE, ProbF AS P
  FROM WORK.SS WHERE SOURCE IN ("Modello"));
QUIT;

/*Creo il PVECT*/

PROC SQL;

  CREATE TABLE WORK.COUNT AS (SELECT COUNT(*) FROM
  WORK.PVALTAB);

QUIT;

DATA _NULL_;
  SET WORK.COUNT;
  CALL SYMPUT("COUNT",_temg001);
RUN;

PROC SQL;

CREATE TABLE WORK.PVECT AS (SELECT PVALTAB.P AS PV FROM
work.PVALTAB WHERE STEP=&count);
QUIT;

```

```

PROC MULTTEST DATA=&ingresso. PERMUTATION NSAMPLE=&nsample. NOPRINT
    SEED=&nseed. OUTSAMP=PERM;
    TEST MEAN (&vardip.);
    CLASS &VARDIP.;
RUN;

PROC SQL;
    CREATE TABLE WORK.PERMUTA AS (SELECT _SAMPLE_,&vardip. FROM
    WORK.PERM);
QUIT;

%INCLUDE 'C:\...\SEPARAP.SAS;

    %SEPARA(&nsample.);

RUN;

%MACRO EX_REG(ds=);
    ODS OUTPUT ANOVA=SS&i.;
    PROC REG DATA=&ds;
    MODEL &vardip.=&varmod.
        /SELECTION=&selection;
TITLE "Risultati dall'analisi di regressione sul campione &i";
RUN;

PROC SQL;
    CREATE TABLE WORK.PVALTAB&i. AS (SELECT SS&i..STEP, SS&i..Fvalue,
    SS&i..ProbF as P FROM WORK.SS&i. WHERE SOURCE IN ("Modello"));
QUIT;

PROC SQL;
    CREATE TABLE WORK.COUNT&i. AS (SELECT COUNT(*) FROM
    WORK.PVALTAB&i.);
QUIT;

DATA _NULL_;
    SET WORK.COUNT&i.;
    CALL SYMPUT("count",_temg001);
RUN;

PROC SQL;
    CREATE TABLE WORK.PVECT&i. AS (SELECT PVALTAB&i..P AS PV FROM
    WORK.PVALTAB&i. WHERE STEP=&count);
QUIT;

PROC APPEND BASE=WORK.PVECT DATA=WORK.PVECT&i.;
RUN;

QUIT;

%MEND MACRO EX_REG;

%MACRO MAIN;
    %LET n=&nsample.;
    %LET i=0;

```

```

        %PUT &i. &n. ;
        %DO %WHILE (&i.<&n.);
        %LET i=%EVAL(&i.+1);
        %EX_REG(ds=ingresso&i.);
        %PUT MODELLO &i.;
        %END;
%MEND MAIN;

OPTION MACROGEN SYMBOLGEN;
%MAIN;

PROC IML;
    USE WORK.PVECT;
    READ ALL INTO X;
        TITLE "Riepilogo";
    nrow=NROW(X);
    E=&nsample.-nrow;
    A=j(E,1,1);
    PV=X//A;
    nrowpv=NROW(PV);
    PRINT pv;
    count=0;
    DO i=1 TO nrowpv;
        IF PV[1]>=PV[i] THEN count=count+1;
    END;
    pvalue_adj= count/&nsample.;
    PRINT count;
    PRINT nrowpv;
    PRIND pvalue_adj;

QUIT;

%MEND PValueADJ;

```

Separa.sas

```

%MACRO SEPARA (nsample);
    %LET j=1;
    %DO %WHILE (&j<=&nsample);
        %PUT THE VALUE OF j IS &j;

DATA SOTTOMATRICE&j (WHERE=( _SAMPLE_ =&j));

SET WORK.PERMUTA;

RUN;

%LET j=%EVAL(&j+1);
%END;

DATA VARMOD; SET &ingresso; DROP &vardip.;
RUN;

%LET j=1;

```

```
%DO %WHILE (&j<=&NSAMPLE);  
DATA INGRESSO&J;  
MERGE SOTTOMATRICE&j VARMOD;  
DROP _SAMPLE_ ;  
RUN;  
%LET j=%eval(&j+1);  
%END;
```

```
%MEND SEPARA;
```

Riferimenti Bibliografici

Austin P.C., Tu J.V.. (2004) Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology* 57, 1138-1146.

Fabbris L. (1997). *Statistica Multivariata – Analisi esplorativa dei dati*. McGraw Hill.

Basso D., Finos L. and Salmaso L. (2005), A new association test in linear models with application to experimental designs, *Journal of Applied Statistical Science*, vol. 15.

Finos L. and Salmaso L.. *Nuove Proposte per il controllo della molteplicità negli studi esplorativi con applicazioni ai microarrays (Presentazione)*

Finos L., Brombin C., Salmaso L. (2007). Adjusting stepwise p-values in generalized linear models.

Finos L., Brombin C., Salmaso L. (Vienna, 2007). Adjusting stepwise p-values in generalized linear models. (Presentazione)

Finos L., Pesarin F., Salmaso L., Solari A. (2004). Nonparametric Iterated Procedure for Testing Genetic Differentiation. *Atti dello XLD Riunione Scientifica della Società Italiana di Statistica, Bari, Sessioni spontanee*, PP. 95-98.

Finos, L. and Salmaso, L. A new nonparametric approach for multiplicity control: Optimal Subset procedures.

Finos, L., Pesarin, F., Salmaso, L. (2003). Confronti multipli tramite metodi di permutazione. *Statistica Applicata*, vol. 15, 2, PP.275-300.

Friendly M. (2006). *Contour. SAS® System for Statistical Graphics*, First Edition by SAS® Institute Inc da internet <http://www.math.york.ca/SCS/sssg/contour.html>. Ultimo accesso 21/04/2007.

Ge Y., Dudoit S., Terence P. Speed. (2003) Resampling-based Multiple Testing for Microarray Data Analysis. *Sociedad de Estadística e Investigación Operativa Test* Vol. 12, No. 1, pp. 1-77.

Grechanovsky E., Pinsker I. (1994) Conditional p-values for F-statistic in a forward selection procedure. *Computational Statistics & Data Analysis* 20 (1995), pp. 239-263.

Good P. (1994). *Permutation Test*. Springer-Verlag

Pesarin F. (2001). *Multivariate Permutation Tests: with applications in biostatistics*. John Wiley & Sons, LTD

Piccolo, D. (1998). *Statistica*. Il Mulino.

Publicazioni inerenti al SAS:

Agostinelli C. and Sartorelli S. (2002). *Quaderni ASID – Introduzione al linguaggio SAS*. Facoltà di Scienze Statistiche dell'Università degli Studi di Padova.

Arboretti R., Pesarin F., Salmaso L. (1999). SAS macro NPC for nonparametric combination of dependent permutation tests by Conditional Resampling Techniques. Working papers 17-1999

Carpenter A. (2004). *Carpenter's Complete Guide to the SAS[®] Macro Language* (Second Edition). SAS Institute Inc., Cary, NC 27513, USA.

Deqing Pei, Wei Liu and Cheng Cheng (2004). %ArrayPerm: A SAS Macro for Permutation Analysis of Microarray Data. Da internet:
<http://www.lexjansen.com/pharmasug/2004/coderscorner/cc06.pdf>. Ultimo accesso 21/04/2007.

Everitt B.S & Der G. (2000). *A Handbook of Statistical Analysis Using SAS[®]*. Chapman&Hall/CRC.

Gerber K., Hall E., Holcomb K., Tolson T.F.J. (23/03/2005). SAS/IML – Interactive Matrix Language (Presentazione). ITC Research Computing Support Group

Hood K. and Miller S.. Resampling Using the SAS System. Price Waterhouse, LLP. SUGI 20, Paper 272

Joyner S.P. (1990). SAS/STAT™ Guide for Personal Computer. SAS Institute Inc., Cary, NC, USA.

Leighton R.W. (2002). Some uses (and handy abuses) of PROC TRANSPOSE. SUGI27. Paper16-27. Da internet <http://www2.sas.com/proceedings/sugi27/p016-27.pdf> . Ultimo accesso 21/04/2007.

Mikkelsen J.D. and Peters A. (2001). SQL Processing with the SAS® System Course Notes. SAS Institute Inc., Cary, NC 27513, USA.

Pesarin F. and Salmaso L. Il Sistema SAS. Da internet http://homes.stat.unipd.it/pesarin/Documentation_NPC.rtf ultimo accesso 21\04\2007

Plymouth Meeting, PA. Da internet <http://www.ats.ucla.edu/STAT/sas/library/nesug99/bt150.pdf>, ultimo accesso 21/04/2007.

Regression Analysis Using SAS, autore sconosciuto. Da internet <http://statistica.unimib.it/utenti/lovaglio/rego.doc>. Ultimo accesso 21/04/2007.

Robin High R. An Introduction to Appending Two or More SAS Datasets or Merging Them Together. Statistical Programmer and Consultant. Da internet <http://uoregon.edu> ultimo accesso 21/04/2007.

Rossi S. (2006-2007). Introduzione alla programmazione e alla analisi della statistica mediante

SAS documentation (Autore sconosciuto), PROC MULTTEST Statement. Da internet <http://v8doc.sas.com/sashtml/stat/chap43/sect5.htm>, ultimo accesso 19/06/2007.

SAS SAMPLE LIBRARY (26/06/2007), REGRESSION OF SUBSETS OF VARIABLES. Da internet http://ftp.sas.com/techsup/download/sample/samp_lib/ . Ultimo accesso 26/06/2007

SAS SAMPLE LIBRARY (26/06/2007), TESTS FOR MULTICOLLINEARITY. Da internet http://ftp.sas.com/techsup/download/sample/samp_lib/ . Ultimo accesso 26/06/2007

SAS SAMPLE LIBRARY. RIDGE REGRESSION USING IML. Da internet http://ftp.sas.com/techsup/download/sample/samp_lib/ . Ultimo accesso 26/06/2007

SAS\IML – Interactive Matrix Language. Da internet www.rhoworld.com/pdf/ch1099.pdf , ultimo accesso 21/04/2007

SAS[®]. Università degli studi di Padova, Scuola di dottorato in territorio, ambiente, risorse e salute (Presentazione).

Schuster K., Sipe L. 50 ways to merge your data – instalment 1. Coders' Corner, Paper 103-26. Da internet www.lexjansen.com/mwsug/2001/Posters/POS-034-waystomerg.pdf . Ultimo accesso 21/04/2007.

Shtatland E.S. , Kleinman K., Cain E.M.. Stepwise Methods in using SAS[®] PROC LOGISTIC and SAS[®] ENTERPRISE MINER[™] for prediction, Statistics and Data Analysis, SUGI¹⁸28, Paper 258-28.

Spector P. Array from A to Z. Statistical Computing Facility. Department of Statistics. University of California, Berkeley. Da internet <http://www.stat.berkeley.edu/~spector>. Ultimo accesso 21/04/2007.

Stanford University (2005). SAS Macro Language. Social Science Data and Software.

¹⁸ SAS Users Group International conferences (SUGI).

Technical Support SAS (19/06/2007). Generating Combinations and Permutations. Da internet <http://support.sas.com/techsup/technote/ts498.html>. Ultimo accesso 21/04/2007.

University of Massachusetts, (autore sconosciuto), March, 2006. SAS Data Management. Biostatistics Consulting Center. School of Public Health.

Vedrashko A.. Haas MFE SAS Workshop (Presentazione). Da internet <http://faculty.haas.berkeley.edu/peliu/computing>. Ultimo accesso 21/04/2007

Virgile B. (1999). Changing the Shape of your data: PROC TRANSPOSTE vs. ARRAYS. Beginning tutorials. Paper60. Da internet <http://ssc.utexas.edu/docs/sashelp/sugi/24/Begtutor/p60-24.pdf>. Ultimo accesso 21/04/2007.

Wieczkowski M.J. (1999). Alternatives to merging SAS data sets...but be careful. IMS HEALTH.

Ringraziamenti

La persona che più devo ringraziare è il Dott. Livio Finos, che mi ha assicurata, seguita ed è sempre stato disponibile. A seguire, il Professor Pesarin, che mi ha dato l'opportunità di svolgere questa tesi.

Ringrazio il Professor Colombo che è stato maestro e guida preziosa, non solo nelle discipline statistiche, ma, e soprattutto nella vita.

Ringrazio mia madre, e miei fratelli perchè nonostante tutto hanno sempre creduto che ce la potessi fare.

Ringrazio Pierangelo che mi ha dato l'opportunità di entrare in Alfa Laval, a Sergio che ha mi ha voluto nel suo team, a Massimiliano che mi sopporta ogni giorno ed infine ma non ultimo, a Jo che continua a fidarsi.

Ringrazio Lorenzo per avermi corretto la sintassi, e ancor di più, Silvia Sartorelli per il suo fondamentale supporto in SAS.

Ringrazio tutti gli amici e colleghi che non mi hanno permesso di dimenticare di chiudere questo capitolo della mia vita.

Ringrazio Silvia e Lucia che mi sono state vicine sempre: prima, durante ed ora.