

STATISTICA I e II

Renato Guseo

Seminario: Modelli Lineari

Padova, 15 maggio 2009

Seminario

Modelli Lineari

1 Regressione e correlazione parziale

1.1 I modelli interpretativi

Siano ${}_1X, {}_2X, {}_3X, \dots, {}_{K-2}X, {}_{K-1}X, {}_KX$, K variabili statistiche sostanzialmente esauritive nella descrizione di un fenomeno la cui matrice di varianze e covarianze, Σ_K , è di rango pieno. Come è noto, l'interdipendenza tra le variabili che insistono sulla medesima popolazione è un fatto piuttosto comune. Si ipotizzi, in particolare, che le variabili ${}_KX$ e ${}_{K-1}X$ possano subire simultaneamente l'influenza delle rimanenti ${}_1X, {}_2X, \dots, {}_{K-2}X$.

Lo studio della correlazione diretta tra ${}_KX$ e ${}_{K-1}X$ può fornire indicazioni scorrette o *spurie* proprio perché tale legame è indotto talvolta dalle variazioni dei livelli delle variabili concomitanti ${}_1X, {}_2X, \dots, {}_{K-2}X$.

Un modo per "eliminare" l'effetto *lineare* delle prime $K - 2$ variabili si basa sui residui di regressione. Precisamente, si definiscono due nuove variabili residuo

$$\begin{aligned} {}_K\mathcal{X} &= {}_KX - \alpha_0^* - \alpha_1^* {}_1X - \dots - \alpha_{K-2}^* {}_{K-2}X \\ {}_{K-1}\mathcal{X} &= {}_{K-1}X - \beta_0^* - \beta_1^* {}_1X - \dots - \beta_{K-2}^* {}_{K-2}X, \end{aligned} \quad (1)$$

ove i coefficienti α_j^*, β_j^* , $j = 1, 2, \dots, K - 2$ sono determinati secondo il criterio dei minimi quadrati e si riferiscono, è bene ribadirlo, a variabili esplicative espresse nella scala originaria. Si osservi, a margine, che una eventuale standardizzazione delle variabili esplicative ${}_1X, {}_2X, \dots, {}_{K-2}X$ non modificherebbe assolutamente le variabili depurate ${}_K\mathcal{X}$ e ${}_{K-1}\mathcal{X}$ in virtù dell'invarianza assicurata dalle trasformazioni lineari destre, $\mathbf{Z} = \mathbf{X}\mathbf{A}$.

Il coefficiente di correlazione al quadrato tra ${}_K\mathcal{X}$ e ${}_{K-1}\mathcal{X}$,

$$\rho_{{}_K\mathcal{X} {}_{K-1}\mathcal{X}}^2, \quad (2)$$

misura la cosiddetta *correlazione parziale al quadrato* tra ${}_KX$ e ${}_{K-1}X$ "al netto" dell'effetto *lineare* delle altre $K - 2$ variabili ${}_1X, {}_2X, \dots, {}_{K-2}X$.

Di solito tale correlazione si annota come segue,

$$\tilde{R} = {}_KX {}_{K-1}X \rho_{{}_1X, \dots, {}_{K-2}X} = \rho_{{}_K\mathcal{X} {}_{K-1}\mathcal{X}}. \quad (3)$$

Il segno del coefficiente di correlazione parziale ${}_KX {}_{K-1}X \rho_{{}_1X, \dots, {}_{K-2}X}$ coincide con quello di $\sigma_{{}_K\mathcal{X} {}_{K-1}\mathcal{X}}$.

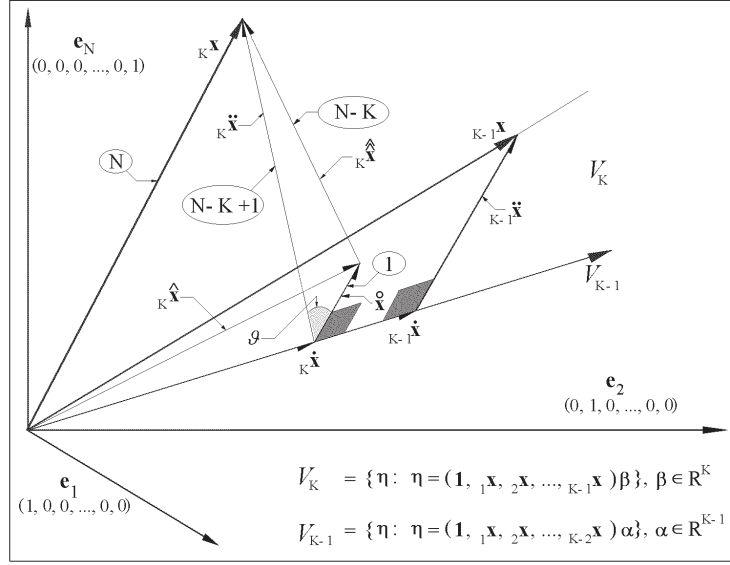


Figura 1: Regressione e correlazione parziale come miglioramento tra modelli nidificati.

1.2 La regressione parziale

Miglioramento tra modelli regressivi e regressione parziale.

Esiste una dimostrazione, piuttosto complessa (si veda, ad esempio, Landenna (1984, 308–17) che consente di ricavare, in forma alternativa, il coefficiente di correlazione parziale al quadrato sfruttando un indice di miglioramento relativo tra due modelli regressivi opportunamente nidificati. Una prova più agevole dell'equivalenza asserita poggia su semplici considerazioni geometriche.

Teorema. La correlazione parziale al quadrato

$$\tilde{R}^2 = {}_{KX}{}_{K-1X} \rho_{1X, \dots, K-2X}^2,$$

coincide con il miglioramento relativo che si consegue introducendo, in un modello di regressione multipla del tipo,

$${}_K X = \alpha_0 + \alpha_1 {}_1 X + \dots + \alpha_{K-2} {}_{K-2} X + \varepsilon, \quad (4)$$

la nuova variabile ${}_{K-1} X$. In altri termini, se si indica con ${}_{KX} \sigma_{1X, \dots, K-2X}^{*2}$ la varianza residua del modello con $K - 2$ variabili e con ${}_{KX} \sigma_{1X, \dots, K-1X}^{*2}$ la corrispondente del modello lineare completato con la variabile ${}_{K-1} X$, si ottiene

$$\tilde{R}^2 = {}_{KX}{}_{K-1X} \rho_{1X, \dots, K-2X}^2 = \frac{{}_{KX} \sigma_{1X, \dots, K-2X}^{*2} - {}_{KX} \sigma_{1X, \dots, K-1X}^{*2}}{{}_{KX} \sigma_{1X, \dots, K-2X}^{*2}}. \quad (5)$$

Prova. Sia \mathcal{V}_{K-1} un sottospazio lineare di R^N generato dalle prime $K - 1$ colonne \mathbf{X} di una matrice $\mathbf{Z} = (\mathbf{1}, {}_1 \mathbf{x}, {}_2 \mathbf{x}, \dots, {}_{K-2} \mathbf{x}, {}_{K-1} \mathbf{x}, {}_K \mathbf{x}) = (\mathbf{X}, {}_{K-1} \mathbf{x}, {}_K \mathbf{x})$ ove ${}_{K-1} \mathbf{x}$

e ${}_K\mathbf{x}$ sono vettori colonna corrispondenti alle variabili statistiche ${}_{K-1}X$ e ${}_KX$ con $N \geq K$. Si veda a questo proposito la Figura 1.

Siano ${}_K\hat{\mathbf{x}}$ e ${}_{K-1}\hat{\mathbf{x}}$ le proiezioni ortogonali di ${}_K\mathbf{x}$ e ${}_{K-1}\mathbf{x}$ su \mathcal{V}_{K-1} . I corrispondenti vettori residui, ${}_K\ddot{\mathbf{x}}$ e ${}_{K-1}\ddot{\mathbf{x}}$, sono ortogonali a \mathcal{V}_{K-1} . Il quadrato del prodotto scalare tra ${}_K\ddot{\mathbf{x}}$ e ${}_{K-1}\ddot{\mathbf{x}}$ dà luogo a

$$({}_K\ddot{\mathbf{x}}' {}_{K-1}\ddot{\mathbf{x}})^2 = \|{}_K\ddot{\mathbf{x}}\|^2 \|{}_{K-1}\ddot{\mathbf{x}}\|^2 \cos^2(\vartheta),$$

ove ϑ è l'angolo compreso tra ${}_K\ddot{\mathbf{x}}$ e ${}_{K-1}\ddot{\mathbf{x}}$.

Poiché ${}_K\ddot{\mathbf{x}}'\mathbf{1} = {}_{K-1}\ddot{\mathbf{x}}'\mathbf{1} = 0$ per l'ortogonalità dei residui ${}_K\ddot{\mathbf{x}}$ e ${}_{K-1}\ddot{\mathbf{x}}$ rispetto al sottospazio lineare \mathcal{V}_{K-1} e quindi, in particolare, con riferimento al vettore $\mathbf{1}$, si può scrivere,

$$\begin{aligned} \frac{({}_K\ddot{\mathbf{x}}' {}_{K-1}\ddot{\mathbf{x}})^2}{\|{}_K\ddot{\mathbf{x}}\|^2 \|{}_{K-1}\ddot{\mathbf{x}}\|^2} &= \frac{(M({}_K\ddot{X} {}_{K-1}\ddot{X}))^2}{M({}_K\ddot{X})^2 M({}_{K-1}\ddot{X})^2} \\ &= \rho_{{}_K\ddot{X} {}_{K-1}\ddot{X}}^2 = {}_{KX}{}_{K-1X}\rho_{1X, \dots, K-2X}^2 = \cos^2(\vartheta), \end{aligned} \quad (6)$$

il quadrato del coefficiente di correlazione tra le variabili residuo ${}_K\ddot{X}$ e ${}_{K-1}\ddot{X}$, depurate degli effetti lineari indotti dalle variabili $1, {}_1X, {}_2X, \dots, {}_{K-2}X$, coincide con il quadrato del coseno dell'angolo ϑ compreso tra i vettori ${}_K\ddot{\mathbf{x}}$ e ${}_{K-1}\ddot{\mathbf{x}}$.

Si consideri ora la matrice $\mathbf{W} = (\mathbf{X}, {}_{K-1}\mathbf{x})$ ed il modello lineare

$${}_K\mathbf{x} = \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Applicando i minimi quadrati si consegue la proiezione ${}_K\hat{\mathbf{x}} \in \mathcal{V}_K$,

$${}_K\hat{\mathbf{x}} = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'{}_K\mathbf{x},$$

e il residuo ${}_K\hat{\mathbf{x}}$ il cui modulo $\|{}_K\hat{\mathbf{x}}\|$ è inferiore al modulo di ${}_K\mathbf{x}$ per il teorema di Pitagora.

Il vettore $\overset{\circ}{\mathbf{x}} = {}_K\hat{\mathbf{x}} - {}_K\mathbf{x}$ è ortogonale a \mathcal{V}_{K-1} per il teorema delle differenze di proiezioni o di residui in modelli lineari nidificati.

L'angolo compreso tra $\overset{\circ}{\mathbf{x}}$ e ${}_K\ddot{\mathbf{x}}$ coincide con l'angolo ϑ più sopra indicato. Il confronto tra i quadrati dei moduli dei vettori residuo porge, ancora una volta,

$$\cos^2(\vartheta) = \frac{\delta_{K-2}^{*2} - \delta_{K-1}^{*2}}{\delta_{K-2}^{*2}} = \frac{\overset{\circ}{\mathbf{x}}' \overset{\circ}{\mathbf{x}}}{{}_K\ddot{\mathbf{x}}' {}_K\ddot{\mathbf{x}}} = \rho_{{}_K\ddot{X} {}_{K-1}\ddot{X}}^2 = {}_{KX}{}_{K-1X}\rho_{1X, \dots, K-2X}^2, \quad (7)$$

un indice di miglioramento nell'adattamento relativo all'interpretazione delle modalità di ${}_KX$ passando da un modello ridotto \mathcal{V}_{K-1} al modello completo includente ${}_{K-1}X$. \square

1.3 Selezione stepwise

Relazione tra rapporto di correlazione parziale e rapporto F .

Sia, per semplicità di notazione, \tilde{R}^2 il rapporto di correlazione parziale al quadrato tra ${}_K X$ e ${}_{K-1} X$ al netto dell'effetto lineare dovuto alle variabili ${}_1 X, {}_2 X, \dots, {}_{K-2} X$,

$$\tilde{R}^2 = {}_{KX}{}_{K-1X} \rho_{1X, \dots, K-2X}^2 = \frac{\delta_{K-2}^{*2} - \delta_{K-1}^{*2}}{\delta_{K-2}^{*2}} = \cos^2(\vartheta).$$

La devianza residua del modello completo è una frazione propria della devianza del *modello ridotto* ottenuto escludendo la variabile ${}_{K-1} X$,

$$\delta_{K-1}^{*2} = \delta_{K-2}^{*2}(1 - \tilde{R}^2). \quad (8)$$

Storicamente si è analizzato un particolare rapporto, il rapporto F , che è collegato ad una speciale distribuzione probabilistica, la cosiddetta variabile casuale \mathcal{F} di Snedecor. Il rapporto F viene introdotto nel modello regressivo multiplo campionario sotto le assunzioni di normalità, indipendenza e omoschedasticità dell'errore residuo ϵ , per effettuare il test di significatività sulle componenti esplicative. Nell'approccio descrittivo non è essenziale assumere una speciale forma distributiva per ϵ e questo fatto rende molto più interessante e generale il risultato conseguito. Risultato che consente l'espressione di un giudizio di significatività sulla rilevanza delle singole componenti esplicative.

Qui ci si limita quindi alla sua semplice definizione operativa sul piano geometrico trascurando, perché non perfettamente rilevante, l'aspetto distributivo più sopra menzionato.

Per rapporto F si intende

$$\begin{aligned} F &= \frac{(\delta_{K-2}^{*2} - \delta_{K-1}^{*2})/1}{\delta_{K-1}^{*2}/(N-K)} = \frac{(\delta_{K-2}^{*2} - \delta_{K-2}^{*2}(1 - \tilde{R}^2))}{(1 - \tilde{R}^2)\delta_{K-2}^{*2}/(N-K)} = \\ &= \frac{\tilde{R}^2(N-K)}{(1 - \tilde{R}^2)}, \quad N > K, \end{aligned} \quad (9)$$

ove si è tenuto conto della (8) e, alternativamente, invertendo la (9) si ha

$$\tilde{R}^2 = \frac{F}{F + N - K}. \quad (10)$$

In definitiva, la (9) e la (10) evidenziano l'esplicita relazione biunivoca monotona crescente tra il rapporto di correlazione parziale al quadrato \tilde{R}^2 e il rapporto F . In particolare, se la correlazione parziale è nulla anche F è nullo e, viceversa, se la correlazione parziale è massima, F tende a più infinito, $F \rightarrow +\infty$.

Il rapporto di correlazione parziale al quadrato \tilde{R}^2 o il rapporto F sono alla base delle procedure *stepwise regression* che consentono di trattare – in modo assistito via *software* – la costruzione di un modello regressivo mediante due sottoprocedure

alternative: a) eliminando variabili non significative – *backward elimination* – sulla base di bassi livelli della correlazione parziale oppure, biunivocamente, sulla base di un rapporto F modesto; b) aggiungendo variabili significative – *forward model building* – con alti livelli di correlazione parziale o di rapporto F .

Un valore tipico di soglia per la selezione/esclusione secondo il rapporto F è il valore 4. Il rapporto F qui definito viene più spesso rappresentato mediante la *statistica* t il cui segno concorda con il segno del coefficiente stimato della variabile esplicativa di interesse, $K_{-1}X$. Il quadrato di t coincide con F ,

$$t^2 = F, \quad (11)$$

per cui, vista la convenzionalità nella scelta dell'ultima variabile esplicativa, una componente di un modello regressivo multiplo è *significativa* se la statistica t assume valori all'*esterno* dell'intervallo $-2 < t < 2$.

Di solito tra le due sottoprocedure estreme di selezione (*backward*, *forward*) si preferisce, sia da un punto di vista sostantivo, sia da un punto di vista statistico, la prima tecnica (*backward*). La ragione principale è il riferimento ad un modello più generale di cui si tenta una semplificazione/riduzione. Questa impostazione aiuta a percepire più facilmente il ruolo delle variabili esplicative che potrebbero risultare impropriamente significative nei primi passi di una procedura *forward*, per essere escluse successivamente nel modello più ampio in cui compaiono, più ragionevolmente, le variabili più importanti dal punto di vista della capacità interpretativa in termini di riduzione consistente della variabilità residuale.

Tra i vari software statistici disponibili, Statgraphics, SPSS, SAS e Statistica, mettono a disposizione semplici procedure di *model building* (*backward*, *forward*) pilotate nella gestione manuale o automatica dalla correlazione parziale o da un suo sostituto F oppure t . Naturalmente, il software predisposto per il trattamento del modello lineare multiplo sotto l'ipotesi di normalità, indipendenza ed omoschedasticità dell'errore residuo, ϵ , può essere usato immediatamente per il più generale caso descrittivo.

1.4 Il modello lineare multiplo con errore gaussiano

In questo paragrafo si farà vedere che l'approccio campionario al modello lineare (regressione multipla e sue varianti), sotto specifiche assunzioni distributive normali dell'errore ϵ , conferma l'approccio "distribution free" che fa capo alla parallela analisi descrittiva. Tale risultato mette in chiara luce le difficoltà della tradizione statistica, basata su assunzioni di verosimiglianza, in relazione al rischio comune di errata specificazione della forma distributiva dell'errore.

Per stabilire il collegamento formale che intercorre tra gli approcci descrittivo e campionario nella modellazione parsimoniosa di relazioni lineari multiple è conve-

niente utilizzare il medesimo supporto geometrico adottato in precedenza con alcuni aggiustamenti associati all'ipotesi di gaussianità dell'errore ϵ .

Si consideri il modello lineare completo con matrice \mathbf{W} di rango pieno, $r(\mathbf{W}) = K$,

$${}_K\mathbf{x} = \mathbf{W}\boldsymbol{\beta} + \epsilon, \quad (12)$$

ove ${}_K\mathbf{x}, \epsilon \in R^N$; $\mathbf{X} = (\mathbf{1}, {}_1\mathbf{x}, {}_2\mathbf{x}, \dots, {}_{K-2}\mathbf{x})$ e $\mathbf{W} = (\mathbf{X}, {}_{K-1}\mathbf{x})$.

Si ipotizzi la sua applicazione in un contesto campionario per cui la risposta ${}_K\mathbf{x} = \boldsymbol{\eta} + \epsilon$ può essere interpretata come somma di due componenti, una deterministica, $\boldsymbol{\eta} = \mathbf{W}\boldsymbol{\beta}$, ed una stocastica, l'errore ϵ .

Nel modello lineare multiplo campionario si considerano alcune ulteriori ipotesi sulla struttura dell'errore ϵ per conseguire una chiara definizione dei test statistici rivolti alla ricognizione della significatività delle componenti. Le assunzioni più comuni sono le seguenti:

- a) errore ϵ dotato di media nulla, $M(\epsilon) = \mathbf{0}$;
- b) incorrelazione ed omoschedasticità degli errori, $M(\epsilon\epsilon') = \sigma^2\mathbf{I}$;
- c) normalità degli errori per cui si ha $\epsilon \sim \mathcal{N}_N(\mathbf{0}, \sigma^2\mathbf{I})$. Quest'ultima assunzione facilita la trattazione dell'inferenza sulla significatività delle componenti mediante un vincolo distributivo speciale.

Sulla "naturalità" di queste assunzioni si ritornerà successivamente in sede di discussione dei risultati ottenuti. Come si è visto nel caso descrittivo si distinguono due modelli nidificati, il modello completo sotto l'ipotesi generale, $H = H_0 \cup H_1$,

$${}_K\mathbf{x} = \mathbf{W}\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim \mathcal{N}_N(\mathbf{0}, \sigma^2\mathbf{I}), \quad H, \quad (13)$$

ed il modello ridotto, sotto l'ipotesi nulla H_0 , in cui si esclude convenzionalmente la variabile ${}_{K-1}\mathbf{x}$,

$${}_K\mathbf{x} = \mathbf{X}\boldsymbol{\alpha} + \epsilon, \quad \epsilon \sim \mathcal{N}_N(\mathbf{0}, \sigma^2\mathbf{I}), \quad H_0. \quad (14)$$

Si veda a questo proposito la Figura 2. Come è noto, sotto le precedenti assunzioni, le stime ottenute secondo il criterio della massima verosimiglianza coincidono con quelle dei minimi quadrati. Non solo, per le assunzioni fatte la norma al quadrato del vettore ϵ è proporzionale ad una distribuzione chi quadrato, precisamente,

$$\epsilon'\epsilon \sim \sigma^2\chi_N^2, \quad H = H_0 \cup H_1. \quad (15)$$

È in questo caso palese il ruolo semplificatore esercitato dall'assunzione di omoschedasticità poiché la riproducibilità della distribuzione χ^2 è limitata solo alla somma di componenti indipendenti della medesima famiglia.

Sotto l'ipotesi generale H e quindi anche sotto l'ipotesi nulla, H_0 , la proiezione secondo i minimi quadrati di ${}_K\mathbf{x}$ o di ϵ sullo spazio complementare rispetto alla

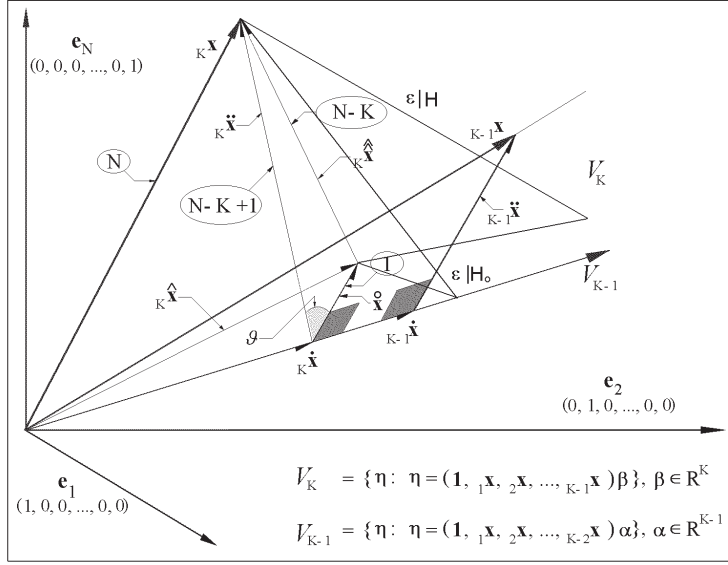


Figura 2: Regressione e correlazione parziale come miglioramento tra modelli nidificati: residui come proiezione dell'errore sugli spazi complementari.

varietà lineare \mathcal{V}_K generata dalle colonne di \mathbf{W} porge comunque il residuo ${}_K\hat{\mathbf{x}}$. Precisamente, posto ${}_W\mathbf{P} = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$ il proiettore di rango $r({}_W\mathbf{P}) = K$, si ha

$${}_K\hat{\mathbf{x}} = (\mathbf{I} - {}_W\mathbf{P})_K\mathbf{x} = (\mathbf{I} - {}_W\mathbf{P})\boldsymbol{\epsilon}, \quad (16)$$

vettore residuo appartenente ad uno spazio lineare di dimensione $N - K$.

Sotto l'ipotesi restrittiva, H_0 , la proiezione secondo i minimi quadrati di ${}_K\mathbf{x}$ o di $\boldsymbol{\epsilon}$ sullo spazio complementare rispetto alla varietà lineare \mathcal{V}_{K-1} generata dalle colonne di \mathbf{X} porge comunque il residuo ${}_K\tilde{\mathbf{x}}$. Precisamente, posto ${}_X\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ il proiettore di rango $r({}_X\mathbf{P}) = K - 1$, si ha

$${}_K\tilde{\mathbf{x}} = (\mathbf{I} - {}_X\mathbf{P})_K\mathbf{x} = (\mathbf{I} - {}_X\mathbf{P})\boldsymbol{\epsilon}, \quad (17)$$

vettore residuo appartenente ad uno spazio lineare di dimensione $N - K + 1$.

Le Equazioni (16) e (17) evidenziano il fatto che i due vettori residuo ${}_K\hat{\mathbf{x}}$ e ${}_K\tilde{\mathbf{x}}$ possono essere pensati come vettori casuali normali ottenuti mediante l'applicazione di due trasformazioni lineari rette da due matrici idempotenti, $(\mathbf{I} - {}_W\mathbf{P})$ e $(\mathbf{I} - {}_X\mathbf{P})$.

Il lemma ed il teorema di Cochran che seguono consentono di risolvere il problema della forma distributiva delle norme al quadrato dei due vettori residuo, ${}_K\hat{\mathbf{x}}$ e ${}_K\tilde{\mathbf{x}}$.

Lemma. Sia $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_m$ una successione di matrici simmetriche di dimensione N tali che $\sum_{j=1}^m \mathbf{A}_j = \mathbf{I}$, allora le condizioni che seguono sono equivalenti:

$$(i) \sum_j r(\mathbf{A}_j) = N, \quad (ii) \mathbf{A}_i\mathbf{A}_j = \mathbf{0} \text{ per } i \neq j, \quad (iii) \mathbf{A}_i^2 = \mathbf{A}_i, \quad i = 1, 2, \dots, m. \quad (18)$$

Teorema di Cochran. Siano \mathbf{x} un vettore normale, $\mathbf{x} \sim \mathcal{N}_N(\mathbf{0}, \mathbf{I})$, e $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_m$ una successione di matrici simmetriche di rango r_1, r_2, \dots, r_m tali che $\sum_{j=1}^m \mathbf{A}_j = \mathbf{I}$. Se vale una (e quindi tutte) delle condizioni del lemma precedente, allora le forme quadratiche $\mathbf{x}'\mathbf{A}_j\mathbf{x}$ sono distribuite secondo chi quadrati indipendenti,

$$\mathbf{x}'\mathbf{A}_j\mathbf{x} \sim \chi_{r_j}^2. \quad (19)$$

□

Nel caso in esame le coppie di proiettori ${}_w\mathbf{P}$, $(\mathbf{I} - {}_w\mathbf{P})$ e ${}_x\mathbf{P}$, $(\mathbf{I} - {}_x\mathbf{P})$ soddisfano i requisiti richiesti dal teorema di Cochran per cui si ha subito,

$${}_K\hat{\mathbf{x}}'{}_K\hat{\mathbf{x}} \sim \sigma^2 \chi_{N-K}^2 | H; \quad {}_K\ddot{\mathbf{x}}'{}_K\ddot{\mathbf{x}} \sim \sigma^2 \chi_{N-K+1}^2 | H_0. \quad (20)$$

Anche in questo caso vale l'osservazione circa il ruolo dell'assunzione di omoschedasticità dell'errore in relazione alla riproducibilità del χ^2 .

Resta da studiare il comportamento distributivo del vettore $\overset{\circ}{\mathbf{x}} = ({}_K\ddot{\mathbf{x}} - {}_K\hat{\mathbf{x}})$. Poiché $(\mathbf{I} - {}_w\mathbf{P}) + ({}_w\mathbf{P} - {}_x\mathbf{P}) + {}_x\mathbf{P} = \mathbf{I}$, si ha $[(\mathbf{I} - {}_w\mathbf{P}) + ({}_w\mathbf{P} - {}_x\mathbf{P})]\boldsymbol{\epsilon} = (\mathbf{I} - {}_x\mathbf{P})\boldsymbol{\epsilon}$, ovvero, $[(\mathbf{I} - {}_x\mathbf{P}) - (\mathbf{I} - {}_w\mathbf{P})]\boldsymbol{\epsilon} = ({}_w\mathbf{P} - {}_x\mathbf{P})\boldsymbol{\epsilon} = ({}_K\ddot{\mathbf{x}} - {}_K\hat{\mathbf{x}}) = \overset{\circ}{\mathbf{x}}$ e quindi, sotto H_0 , si ottiene che la differenza delle devianze dei due modelli nidificati (guadagno) è distribuita proporzionalmente ad un chi quadrato con un grado di libertà

$$\overset{\circ}{\mathbf{x}}'\overset{\circ}{\mathbf{x}} = ({}_K\ddot{\mathbf{x}} - {}_K\hat{\mathbf{x}})'({}_K\ddot{\mathbf{x}} - {}_K\hat{\mathbf{x}}) \sim \sigma^2 \chi_1^2 | H_0. \quad (21)$$

ed è stocasticamente indipendente dalla norma al quadrato del residuo del modello completo, ${}_K\hat{\mathbf{x}}'{}_K\hat{\mathbf{x}}$.

Sotto l'alternativa H_1 , la statistica $\overset{\circ}{\mathbf{x}}'\overset{\circ}{\mathbf{x}}$ è un chi quadrato non centrale con un grado di libertà ed è stocasticamente superiore rispetto a $\sigma^2 \chi^2$.

Si può ora saggiare la rilevanza della variabile inserita come ultima tra le esplicative nel modello di regressione multipla mediante il rapporto F , che, a meno del fattore $N - K$, è definito come la cotangente al quadrato dell'angolo ϑ ,

$$F = \cot^2(\vartheta)(N - K) = \frac{\overset{\circ}{\mathbf{x}}'\overset{\circ}{\mathbf{x}}(N - K)}{{}_K\hat{\mathbf{x}}'{}_K\hat{\mathbf{x}}} = \frac{\boldsymbol{\epsilon}'({}_w\mathbf{P} - {}_x\mathbf{P})\boldsymbol{\epsilon}/1}{\boldsymbol{\epsilon}'(\mathbf{I} - {}_w\mathbf{P})\boldsymbol{\epsilon}/(N - K)}. \quad (22)$$

Tale rapporto risulta avere, sotto l'ipotesi nulla, H_0 , ed in presenza delle condizioni di gaussianità dell'errore, $\boldsymbol{\epsilon} \sim \mathcal{N}_N(\mathbf{0}, \sigma^2 \mathbf{I})$, una distribuzione \mathcal{F} di Snedecor centrale, non dipendente dall'incognita varianza condizionale σ^2 , precisamente,

$$\mathcal{F}_{(1, N-K)} = \frac{\chi_1^2/1}{\chi_{N-K}^2/(N - K)} = t_{N-K}^2. \quad (23)$$

Sotto l'alternativa, H_1 , il rapporto F genera una corrispondente distribuzione \mathcal{F} di Snedecor non centrale che è stocasticamente superiore rispetto alla (23).

Per quanto si è provato in precedenza nel contesto semplicemente descrittivo ove si evitano assunzioni distributive sulla forma dell'errore, il rapporto F è collegato direttamente al quadrato della correlazione parziale tra ${}_K\mathbf{x}$ e ${}_{K-1}\mathbf{x}$ al netto dei contributi lineari dovuti alle variabili ${}_1\mathbf{x}, {}_2\mathbf{x}, \dots, {}_{K-2}\mathbf{x}$, ovvero, posto $\tilde{R}^2 = \cos^2(\vartheta)$,

$$F = \cot^2(\vartheta)(N - K) = \frac{\cos^2(\vartheta)(N - K)}{1 - \cos^2(\vartheta)} = \frac{\tilde{R}^2(N - K)}{1 - \tilde{R}^2}. \quad (24)$$

Si tratta ora di esaminare il comportamento distributivo di F sotto l'ipotesi nulla, ovvero quando la variabile ${}_{K-1}\mathbf{x}$ risulta non significativa nell'ambito del modello regressivo completo. Se si fissa una soglia critica al livello di probabilità 95% si nota che le corrispondenti tavole della $\mathcal{F}_{(1, N-K)}$ presentano per $N - K$ superiore a 9 una soglia critica sostanzialmente costante per gli usi pratici e pari a 4.

Da un punto di vista interpretativo può essere istruttivo studiare il valore di soglia di \tilde{R}^2 in funzione di $(N - K)$ in corrispondenza al valore critico $F = 4$. Per valori di $(N - K)$ maggiori di 9 il livello di \tilde{R}^2 risulta monotono decrescente molto rapidamente a partire dal livello 0,28. Ad esempio, per $(N - K) = 96$ il corrispondente valore di soglia per \tilde{R}^2 è 0,04. In altri termini, la regola di selezione “comune” basata su $F = 4$ definisce una correlazione parziale al quadrato di soglia molto modesta e ampiamente superata dai criteri di esclusione/inclusione adottati direttamente sulla base di \tilde{R}^2 . Di solito i valori di riferimento sono superiori a 0,20 o 0,30.

In definitiva, l'assunzione piuttosto forte di nullità del valore medio, di omoschedasticità, indipendenza e gaussianità dell'errore, $\epsilon \sim \mathcal{N}_N(\mathbf{0}, \sigma^2 \mathbf{I})$, consente la definizione di un rapporto F con distribuzione nota, e non dipendente da σ^2 , sotto l'ipotesi nulla di non significatività di una componente regressiva di interesse. Questa assunzione complessa serve a giustificare “formalmente” la scelta della soglia $F = 4$ o sue varianti soprattutto per valori piccoli di $N - K$ (inferiori a 9). Se vengono tuttavia a cadere componenti specifiche dell'assunto $\epsilon \sim \mathcal{N}_N(\mathbf{0}, \sigma^2 \mathbf{I})$, viene meno la caratterizzazione dimensionale esatta in termini probabilistici della soglia $F = 4$.

I percorsi possibili possono essere molteplici ma due appaiono essere le direzioni principali.

- a) seguire un approccio basato su nuove specificazioni distributive della forma dell'errore e sul conseguente uso dei criteri della massima verosimiglianza e del rapporto di verosimiglianza. La forza e la simultanea fragilità della metodologia resta comunque legata alla specificazione distributiva. Un'errata specificazione inficia i livelli probabilistici nominali delle soglie utilizzate per il test d'ipotesi;
- b) rafforzare un approccio “distribution free” come prolungamento naturale dell'impianto descrittivo. Come si è visto, in questo contesto di base si sono definiti in modo normato – mediante indicatori non parametrici fondati sul quadrato della correlazione parziale – i test descrittivi concernenti la significatività di componenti esplicative in un modello di regressione multipla.

