Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di laurea triennale in
STATISTICA PER L'ECONOMIA E L'IMPRESA

RELAZIONE FINALE

Selezione del modello tramite la statistica PRESS

Relatrice: Prof.ssa Luisa Bisaglia
Dipartimento di Scienze Statistiche

Laureanda: Sara De Tommaso
Matricola: 2015216

Anno Accademico 2022/2023

Indice

Introduzione

La selezione del modello è una fase cruciale dell'analisi dei dati statistici, che mira a identificare il modello migliore per descrivere i dati osservati.
Un obiettivo fondamentale della selezione del modello è trovare un equilibrio tra la sua complessità e la sua capacità di adattarsi ai dati disponibili.
Infatti, un modello troppo semplice potrebbe risultare sottodimensionato, non riuscendo a catturare adeguatamente la complessità dei dati. D'altra parte, un modello troppo complesso potrebbe essere sovradimensionato, formato quindi da un eccessivo numero di parametri, e potrebbe soffrire di "overfitting", ossia di un buon adattamento ai dati di addestramento ma una scarsa capacità di generalizzazione a nuovi dati.
Per selezionare un buon modello, esistono diverse metodologie disponibili. Esse si differenziano per gli algoritmi utilizzati per identificare i sottoinsiemi di variabili candidate e per i criteri utilizzati per valutare la bontà di adattamento dei modelli considerati.
I criteri di valutazione possono essere basati sulla penalizzazione della log-verosimiglianza, sull'indice di determinazione, sulla minimizzazione del rischio o su misure di bontà di adattamento. L'obiettivo di tali criteri è selezionare il modello che offre il miglior compromesso tra complessità e bontà di adattamento ai dati, consentendo una descrizione accurata e generalizzabile del fenomeno studiato.
Le misure di selezione del modello più comuni sono:

- coefficiente di determinazione R^2;
- R^2 aggiustato o corretto (R^2_{adj});
- criterio di informazione di Akaike (AIC);
- criterio Bayesiano di Schwarz (SBC);
- criterio di Mallow (CP).

Questi indicatori di valutazione possono essere confrontati tra loro per determinare il migliore tra diversi modelli candidati. Tuttavia, è importante notare che l'utilizzo di queste misure non garantisce che il modello selezionato non includa variabili non significative

o che sia in grado di fare previsioni accurate su valori futuri. Ciò è dovuto al fatto che tutti i dati disponibili sono stati utilizzati per la costruzione del modello, quindi ciò che il modello fornisce sono informazioni sulla capacità di prevedere le osservazioni presenti.
In statistica, il coefficiente di determinazione R^2 è comunemente utilizzato per la selezione di un modello di regressione. Esso misura la forza della relazione lineare tra la variabile esplicativa e la variabile risposta e fornisce una misura complessiva della bontà di adattamento del modello ai dati osservati.
Tuttavia, ci sono alcune limitazioni computazionali e interpretative associate a R^2.
Infatti, con l'aggiunta di variabili esplicative al modello, il suo valore non può diminuire e tende a selezionare modelli con un numero maggiore di predittori, anche se non tutti sono significativi per spiegare la variabile di risposta. Pertanto, un aumento del valore di R^2 non significa necessariamente che il nuovo predittore contribuisca in modo significativo alla spiegazione della variabile di risposta. R^2 fornisce così un indice di "eccesso" e da solo non è sufficiente per valutare la qualità del modello.
Una versione corretta di R^2 che tiene conto del numero di variabili nel modello rispetto alla dimensione del campione è il coefficiente di determinazione aggiustato, R^2_{adj}. Questo coefficiente corregge il potenziale problema di sovradimensionamento del modello considerando la sua complessità e la numerosità del campione.
La statistica PRESS (Predicted Residual Sum of Squares) affronta il problema di sovradimensionamento valutando l'errore di previsione del modello su dati di previsione indipendenti che non sono stati utilizzati per l'addestramento del modello. Questa statistica fornisce una stima dell'errore di previsione sui dati futuri e può essere utilizzata come criterio per la selezione del modello.
Nel presente lavoro, si concentra l'attenzione sull'analisi del comportamento di R^2_{PRESS}, un criterio di valutazione basato sulla statistica PRESS, confrontando le sue prestazioni rispetto ad altri criteri di selezione del modello quando sono presenti predittori non necessari. Utilizzando dati simulati, si osserva che R^2_{PRESS} generalmente fornisce le migliori prestazioni nella selezione del modello vero come migliore per la previsione tra le diverse misure considerate.

1 Metodi di selezione del modello

La selezione del modello di regressione lineare multipla rappresenta un passaggio cruciale nell'analisi statistica e nell'interpretazione dei risultati. Esistono diversi criteri disponibili per selezionare il modello finale, e la scelta del criterio migliore rimane oggetto di discussione tra i ricercatori. Non esiste una misura che sia considerata superiore a tutte le altre in modo assoluto, e ogni criterio può raccomandare un modello diverso come il migliore. Nella presente sezione, saranno brevemente esaminati i vantaggi e gli svantaggi di alcuni criteri comunemente utilizzati per tale selezione.

L'obiettivo è fornire una panoramica delle diverse opzioni disponibili e offrire un'analisi oggettiva delle loro caratteristiche. E' importante considerare attentamente le peculiarità del proprio studio e le implicazioni di ogni criterio prima di prendere una decisione sulla selezione del modello.

In definitiva, la scelta del criterio di selezione richiede una valutazione attenta dei vantaggi e degli svantaggi di ciascuna misura disponibile, tenendo conto del contesto specifico dello studio e degli obiettivi di ricerca. E' quindi fondamentale adottare un approccio ponderato e informato nella selezione del modello finale.

1.1 Coefficienti di determinazione R^2 ed R^2 corretto

La misura più comune per valutare la bontà di adattamento di un modello è il coefficiente di determinazione R^2: $R^2 = 1 - \frac{SSE}{SST}$ con SSE (Sum of Squares Error) $= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ e SST (Sum of Squares Total) $= \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

Esso fornisce una stima della proporzione di variazione della variabile dipendente che può essere spiegata dal modello. Il suo valore varia da 0 a 1 e viene espresso in percentuale, indicando quanto bene il modello riesce a spiegare la variazione dei dati. Un valore di R^2 più vicino a 1 (o al 100%) indica un modello migliore.

Tuttavia, un problema dell'utilizzo di R^2 è che il suo valore tende ad aumentare quando vengono aggiunti più parametri al modello. Ciò significa che il modello con il più alto valore di R^2 sarà sempre quello composto da più parametri, indipendentemente dal rea-

le contributo delle variabili alla previsione. Ciò può portare a un fenomeno noto come "overfitting" (sovradimensionamento), in cui il modello diventa troppo complesso rispetto alla quantità di dati disponibili per l'addestramento. Di conseguenza, il modello può avere una capacità di generalizzazione ridotta, una minore abilità predittiva e un aumento dell'errore nelle previsioni su dati non osservati.

Per mitigare questo problema, è importante utilizzare tecniche di regolarizzazione e validazione. Questo può includere l'uso di metodi di selezione delle variabili per ridurre la complessità del modello.

Un'alternativa al coefficiente di determinazione R^2 è il coefficiente di determinazione corretto (R^2_{adj}). Questo coefficiente tiene conto del numero di predittori nel modello e penalizza l'inclusione di variabili che non contribuiscono significativamente alla spiegazione della variazione della variabile dipendente. Il coefficiente R^2_{adj} è calcolato utilizzando la seguente formula: $R^2_{adj} = 1 - \frac{SSE/n-p}{SST/n-1} = 1 - \frac{MSE}{MST}$, con:

- n : la numerosità campionaria;
- p : la numerosità dei predittori compresa l'intercetta;
- MSE (Mean Squared Error) : l'errore quadratico medio, una misura della dispersione dei dati intorno alla linea di regressione, dove valori più bassi indicano una migliore aderenza del modello ai dati;
- MST (Mean Squared Total) : la varianza media totale, una misura della variabilità totale dei dati in un modello di regressione. Rappresenta la somma della varianza spiegata (MSE) e della varianza residua (errore non spiegato dal modello).

A differenza del coefficiente di determinazione R^2, il coefficiente di determinazione corretto presenta un comportamento diverso. Esso aumenta solo se l'aggiunta di nuovi predittori al modello contribuisce effettivamente a migliorarne la qualità (Dunlop, Tamhane, 2000). L'utilizzo del coefficiente di determinazione corretto permette di ottenere una stima più accurata della bontà di adattamento del modello, tenendo conto del numero di predittori utilizzati.

Rispetto a R^2, l'R^2_{adj} considera l'effetto delle variabili indipendenti che possono influenzare in modo distorto i risultati del coefficiente di determinazione.
Il valore di R^2_{adj} è sempre minore o uguale al valore di R^2. La sua scala va da meno infinito a 1, infatti può assumere anche valori negativi se il valore di R^2 è prossimo allo zero. Tuttavia, un valore negativo di R^2_{adj} indica che il modello non offre alcuna capacità di spiegazione rispetto alla variabilità della variabile dipendente.
In conclusione, l'utilizzo del coefficiente di determinazione corretto (R^2_{adj}) aggiunge precisione e affidabilità nella valutazione del modello, considerando l'effetto dei predittori aggiuntivi e fornendo una misura più accurata della sua capacità di spiegare la variazione dei dati.

1.2 Criteri di informazione AIC e SBC

L'R^2_{adj} rappresenta una misura corretta di bontà di adattamento che offre un buon compromesso tra la precisione dei risultati e la parsimonia del modello, con il minor numero di aggiustamenti per le variabili esplicative aggiuntive.
Tuttavia, esistono altre misure di selezione dei modelli che possono essere prese in considerazione. Tra queste misure, molti ricercatori hanno sostenuto l'utilizzo di criteri informativi come l'Akaike Information Criterion (AIC) e lo Schwarz Bayesian Information Criterion (SBC). Entrambi i criteri riassumono l'adattamento del modello ai dati combinando la verosimiglianza nel punto di ottimo con un termine di penalizzazione che dipende dal numero dei parametri del modello (Burnham, Anderson, 2004).
In particolare, il criterio SBC applica una penalità maggiore rispetto all'AIC quando si considerano modelli più complessi. Di seguito sono riportati alcuni possibili vantaggi e svantaggi che caratterizzano i due criteri di informazione.

- AIC = n ln(SSE) - n ln(n) + 2p, con:
  - n : numerosità campionaria;
  - p : la numerosità dei predittori compresa l'intercetta;

- SSE (Sum of Squared Errors) è un termine che indica la somma dei quadrati degli errori residui nella regressione o nell'analisi dei dati.

Vantaggi:

- è utile per modelli di previsione, in quanto tiene conto della complessità del modello e della qualità dell'adattamento ai dati;
- è semplice da calcolare e può essere facilmente utilizzato per confrontare modelli diversi;
- fornisce una buona adattabilità ai dati.

Svantaggi:

- per grandi campioni tende a selezionare modelli complessi;
- non tiene conto della dimensione del campione, il che significa che può essere influenzato da campioni di dimensioni diverse;
- può essere soggetto al problema di "overfitting".

L'AIC penalizza la complessità del modello aggiungendo un termine di penalità proporzionale al numero di parametri stimati nel modello. L'obiettivo è selezionare il modello con il valore minimo di AIC, poiché un valore più basso indica un migliore bilanciamento tra la bontà di adattamento e la complessità del modello.

- SBC = n ln(SSE) - n ln(n) + p ln(n)

  Il criterio di Schwarz Bayesian Information Criterion (SBC) è una misura di selezione del modello sviluppata dallo statistico americano Gideon Schwarz, basata sulla teoria bayesiana. È simile all'AIC, ma si differenzia per l'applicazione di una penalità più forte per la complessità del modello.

  Un vantaggio dell'utilizzo del criterio SBC è che favorisce la scelta di modelli più parsimoniosi, cioè con un numero ridotto di parametri, il che può portare a una migliore generalizzazione e predizione. Tuttavia, questa parsimonia potrebbe comportare una perdita di alcune informazioni contenute nei dati.

Il criterio SBC si basa sull'utilizzo di una distribuzione di probabilità sulla complessità del modello e introduce un termine di penalità basato sul logaritmo del numero di osservazioni nel campione. L'obiettivo è selezionare il modello che presenta il valore minimo di SBC e il miglior trade-off tra la bontà di adattamento del modello e la sua complessità, secondo l'approccio bayesiano.

L'utilizzo del criterio SBC fornisce una strategia per selezionare modelli che siano in grado di bilanciare la bontà di adattamento ai dati con la complessità del modello, secondo i principi della teoria bayesiana. Tuttavia, come per qualsiasi criterio di selezione del modello, è importante valutare attentamente i vantaggi e gli svantaggi specifici di SBC nel contesto del proprio studio e adattarli alle esigenze dell'analisi statistica.

1.3 Criterio CP di Mallow

Un'importante misura utilizzata per valutare l'adattamento dei modelli di regressione è la statistica CP di Mallow, sviluppata dallo statistico e matematico irlandese John Mallow. Essa fornisce una valutazione complessiva della bontà di adattamento del modello, tenendo conto sia dell'errore di previsione sia del numero di variabili indipendenti incluse nel modello.

La statistica CP è definita per un modello di regressione lineare con p coefficienti di regressione e n osservazioni. La formula è espressa come:

$$CP = \frac{SSE_p}{\hat{\sigma^2}} + 2(p+1) - n$$

nella quale SSE_p rappresenta la somma dei quadrati degli errori del modello ridotto e $\hat{\sigma}$ è l'errore quadratico medio del modello completo. La procedura per utilizzare la statistica CP di Mallow coinvolge un approccio graduale in cui si aggiungono o eliminano predittori nel modello. L'obiettivo è trovare il valore minimo di CP (Murtaugh, 1998). È importante notare che la relazione tra CP e il numero di coefficienti di regressione nel modello (p) è significativa. Modelli con CP > p possono produrre previsioni distorte.

La statistica CP di Mallow permette il confronto tra modelli che hanno diversi sottoinsiemi di parametri rispetto al modello completo (Gilmour, 1996).
Nel processo di confronto, il modello con il valore minore viene considerato un modello valido e raccomandato come modello finale.
L'utilizzo di questa statistica fornisce un metodo oggettivo per valutare e selezionare il modello che offre il miglior compromesso tra adattamento e complessità.
Tuttavia, è importante considerare anche altri criteri di valutazione e adattarli alle specifiche esigenze dell'analisi statistica.

1.4 Statistica PRESS

Il modello selezionato utilizzando uno qualiasi dei criteri precedentemente rappresentati rischia di essere meno parsimonioso, non incrementando allo stesso tempo la capacità predittiva.
La statistica PRESS (Predicted Residual Error Sum of Squares) è definita come una misura di validazione del modello che può essere utilizzata anche per confrontare i modelli di regressione e serve a misurare la capacità predittiva di un modello (Chen, S. , Hong, X., Harris, Sharkley P.M. 2004).
La PRESS viene prodotta per un insieme di dati di dimensione n escludendo ogni osservazione separatamente, e poi le $n-1$ osservazioni rimanenti vengono utilizzate per creare un'equazione di regressione che predice il valore della risposta omessa (etichettata come $\hat{y_{i(i)}}$). Una buona regressione avrà una somma dei quadrati predittiva minima.
È stato dimostrato che la statistica PRESS è una funzione ponderata dei residui dei minimi quadrati e nel seguito è riportata la rispettiva dimostrazione.
Il residuo e_i è calcolato prendendo la differenza tra la variabile di risposta Y_i ed il suo valore previsto quando l'i-esima osservazione è esclusa dalla costruzione del modello, $\hat{Y_i}$: $\epsilon_i = Y_i - \hat{Y_i}$ e la statistica PRESS è calcolata come somma dei quadrati di ϵ_i (Allen, 1971)

$$PRESS = \sum_{i=1}^{n}(e_i^2) = \sum_{i=1}^{n}(Y_i - \hat{y_{i(i)}})^2$$

Il metodo di calcolo richiede molto tempo, quindi i ricercatori hanno ideato un metodo alternativo per calcolare questa statistica, consentendo di determinare la PRESS con un unico adattamento del modello ai dati.

Sia β_i il vettore dei coefficienti di regressione calcolati rimuovendo l'osservazione i-esima:

$$\beta_i = (X_i' X_i)^{-1} X_i' y_i$$

dove X_i è la matrice delle X esclusa l'osservazione i-esima e y_i è il vettore delle risposte esclusa l'osservazione i-esima. Sostituendo questa equazione nella formula per il calcolo del residuo della PRESS, si ottiene la seguente equazione:

$$\begin{aligned} \epsilon_i &= y_i - \hat{y}_i \\ &= y_i - x_i \hat{\beta}_i \\ &= y_i - x_i (X_i' X_i)^{-1} X_i' y_i \end{aligned}$$

Si consideri x come il vettore della i-esima riga, $X'X - x'x$ rappresenta la matrice $X'X$ con l'i-esima riga rimossa dalla costruzione del modello, che può essere espressa come (Shalabh):

$$[X'X - x'x]^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1} x' x (X'X)^{-1}}{1 - x(X'X)^{-1} x'}$$

di conseguenza,

$$\begin{aligned} [X_i' X_i]^{-1} &= (X'X)^{-1} + \frac{(X'X)^{-1} x_i' x_i (X'X)^{-1}}{1 - x(X'X)^{-1} x'} \\ &= (X'X)^{-1} + \frac{(X'X)^{-1} x_i' x_i (X'X)^{-1}}{1 - h_{ii}} \end{aligned}$$

dove h_{ii} rappresenta il valore leva che quantifica l'influenza che la risposta osservata Y_i ha sul suo valore pevisto $\hat{Y}_i$. Inserendo questa espressione nella formula dell'i-esimo residuo PRESS di cui sopra, si stabilisce la seguente relazione:

$$\begin{aligned} \epsilon_i &= y_i - x_i (X_i' X_i)^{-1} X_i' y_i \\ &= y_i - x_i [(X'X)^{-1} + \frac{(X'X)^{-1} x_i' x_i (X'X)^{-1}}{1 - h_{ii}}] X_i' y_i \\ &= y_i - x_i (X'X)^{-1} X_i y_i - \frac{x_i (X'X)^{-1} x_i' x_i (X'X)^{-1} X_i' y_i}{1 - h_{ii}} \\ &= y_i - x_i (X'X)^{-1} X_i y_i - \frac{h_{ii} x_i (X'X)^{-1} X_i' y_i}{1 - h_{ii}} \end{aligned}$$

$$= \frac{(1-h_{ii})y_i - (1-h_{ii})x_i(X'X)^{-1}X_i'y_i - h_{ii}x_i(X'X)^{-1}X_i'y_i}{1-h_{ii}}$$
$$= \frac{(1-h_{ii})y_i - x_i(X'X)^{-1}X_i'y_i}{1-h_{ii}}$$

Poiché $X'y = X_i'y_i + x_i'y_i$, si può sostituire $X'y - x_i'y_i$ con $X_i'y_i$ nella formula precedente, quindi l'espressione per ϵ_i sarebbe:

$$\epsilon_i = \frac{(1-h_{ii})y_i - x_i(X'X)^{-1}(X'y - x_i'y_i)}{1-h_{ii}}$$
$$= \frac{(1-h_{ii})y_i - x_i(X'X)^{-1}X'y + x_i(X'X)^{-1}x_i'y_i)}{1-h_{ii}}$$
$$= \frac{(1-h_{ii})y_i - x_i b + h_{ii}y_i}{1-h_{ii}}$$
$$= \frac{y_i - x_i b}{1-h_{ii}} = \frac{\epsilon_i}{1-h_{ii}}$$

Sostituendo la formula di ϵ_i appena ottenuta nella formula della statistica PRESS definita in precedenza, si ha:

$$\text{PRESS} = \sum_{i=1}^{n}(\epsilon_i^2) = \sum_{i=1}^{n}\left[\frac{\epsilon_i}{(1-h_{ii})}\right]^2$$

L'utilizzo di questa seconda espressione per la statistica PRESS porta a un calcolo più rapido, in quanto non è più necessario adattare il modello n volte.

I residui legati alle osservazioni con un'elevata variabilità di previsione sono ponderati meno pesantemente. L'indipendenza dei residui esterni ϵ_i consente alla statistica PRESS di essere una misura reale della validità o delle capacità di previsione del modello di regressione. In generale, più piccolo è il valore PRESS, migliore è la capacità predittiva del modello.

Sostituendo questa espressione al posto di SSE nella formula dell'R^2 si può derivare una misura simile a R^2 che è generalmente più intuitiva da interpretare rispetto alla stessa PRESS. È definito come:

$$R^2_{PRESS} = 1 - \frac{PRESS}{SST}.$$

Si tratta di una tecnica utile per verificare la capacità predittiva del modello senza dover prelevare un altro campione o dividere i dati in set di addestramento e di validazione (Mediavilla F., Landram, F., Shan, V. 2008).

Poiché sia PRESS che R^2_{PRESS} sono derivate da osservazioni non incluse nel calcolo del

modello, possono aiutare a prevenire il sovradimensionamento riferito a modelli che sembrano fornire un forte adattamento all'insieme di dati in questione, ma che non riescono ad anticipare le nuove osservazioni. Pertanto, è opportuno esaminare più da vicino come la statistica PRESS possa essere utilizzata nella selezione del modello e nella stima dei parametri.

Si può notare che R^2 ed R^2_{PRESS} hanno una formula simile. Sebbene i valori non risultino uguali, è possibile che in presenza di predittori non necessari al modello, l'R^2 sia molto più elevato rispetto all'R^2_{PRESS}.

2 Il criterio R^2_{PRESS}

Al fine di esaminare il comportamento di R^2_{PRESS}, si è utilizzato un dataset simulato per condurre un'analisi esplorativa focalizzata sull'aggiunta di variabili non necessarie ai modelli di regressione lineare multipla.
L'obiettivo è valutare come viene influenzato il valore di R^2_{PRESS} con l'aggiunta di tali variabili e comprendere meglio il loro impatto sui risultati del modello.

2.1 Dati simulati

I dati simulati includono una variabile di risposta Y e tre predittori, i quali sono stati creati in base a un valore specifico di R^2. Questo approccio permette di studiare in dettaglio come i predittori influenzano la variabile di risposta, tenendo conto del livello di relazione tra di loro espresso dal valore di R^2. Il modello lineare per questi dati è :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon.$$

I valori dei coefficienti β_i sono fissati e valgono: β_0=0.2, β_1=0.3, β_2=0.5 e β_3=0.6.
Il vero valore di R^2 è espresso come:

$$R^2 = 1 - \frac{\sigma^2}{VAR(Y)}$$

e da questa formula si ottiene:

$$\text{VAR(Y)} = \frac{\sigma^2}{1-R^2}$$

con VAR(Y) ricavato dal modello di regressione assumendo che i predittori siano distribuiti secondo una normale standard:

$$\text{VAR(Y)} = \beta_1^2 + \beta_2^2 + \beta_3^2 + \sigma^2$$

e sostituendo la formula derivata in precedenza di VAR(Y) in questa equazione si ottiene:

$$\frac{\sigma^2}{1-R^2} = \beta_1^2 + \beta_2^2 + \beta_3^2 + \sigma^2$$

.

Risolvento l'equazione per σ^2

$$\sigma^2 = (1 - R^2)(\beta_1^2 + \beta_2^2 + \beta_3^2) + \sigma^2(1 - R^2)$$

$$\sigma^2 \text{ - } \sigma^2(1 - R^2) = (1 - R^2)(\beta_1^2 + \beta_2^2 + \beta_3^2)$$

$$\sigma^2(1 - (1 - R^2)) = (1 - R^2)(\beta_1^2 + \beta_2^2 + \beta_3^2)$$

$$\sigma^2(R^2) = (1 - R^2)(\beta_1^2 + \beta_2^2 + \beta_3^2)$$

$$\sigma^2 = \tfrac{1-R^2}{R^2}(\beta_1^2 + \beta_2^2 + \beta_3^2)$$

Per analizzare gli effetti di R^2 vengono fissati due valori di queste misure: 0.5 e 0.8. Utilizzando questi valori e la formula per σ^2 derivata in precedenza, si ottiene per $R^2 = 0.5$,

$$\sigma^2 = (\beta_1^2 + \beta_2^2 + \beta_3^2)\tfrac{1-R^2}{R^2}$$

$$= (0.3^2 + 0.5^2 + 0.6^2)\tfrac{1-0.5}{0.5}$$

$$= 0.7$$

Quindi ϵ_1 è generato da una N(0, 0.8367^2).

Invece per $R^2 = 0.8$:

$$\sigma^2 = (\beta_1^2 + \beta_2^2 + \beta_3^2)\tfrac{1-R^2}{R^2}$$

$$= (0.3^2 + 0.5^2 + 0.6^2)\tfrac{1-0.8}{0.8}$$

$$= 0.175$$

Quindi ϵ_2 è generato da una N(0, 0.4183^2).

2.2 Confronto tra R^2_{PRESS} e R^2_{adj}

L'obiettivo di quest'analisi è confrontare le prestazioni di R^2_{PRESS} rispetto ad R^2_{adj} e di valutare come cambiano i rispettivi valori al variare della distribuzione dell'errore campionario. Sono state generate 10.000 simulazioni su 2 diversi modelli, entrambi di numerosità 80 (Appendice A):

$$\text{Y1} = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \epsilon_1$$

$$\text{Y2} = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \epsilon_2$$

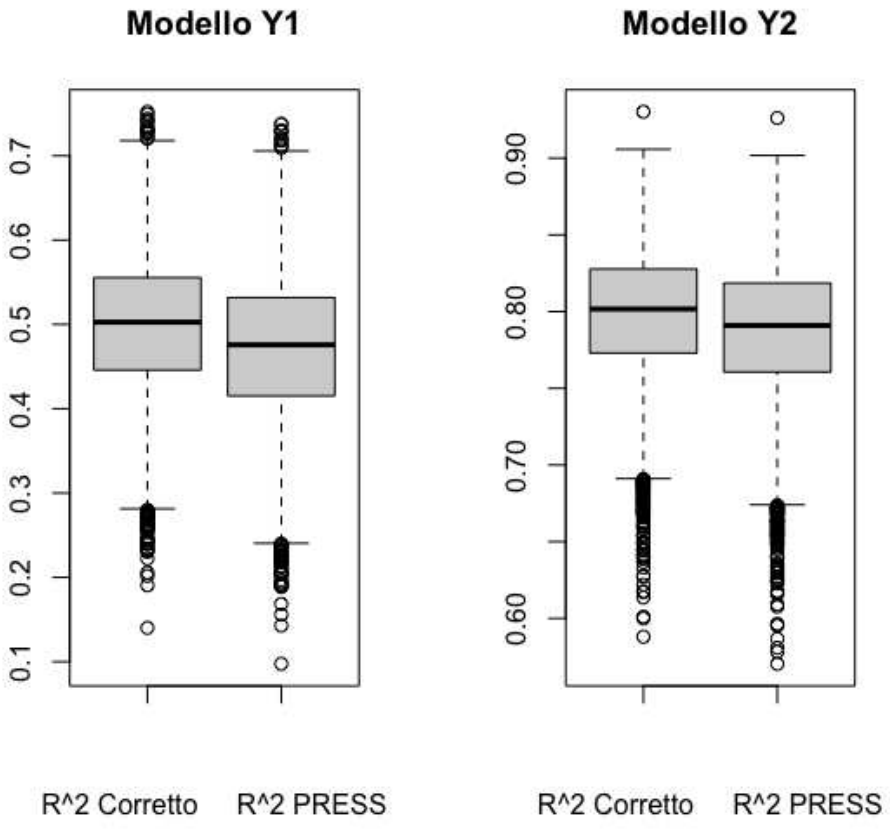

Figura 1: Boxplot di confronto: R^2_{adj} vs R^2_{PRESS}

R^2 target	R^2_{adj}	R^2_{PRESS}
$R^2 = 0.50$	0.4993	0.4722
$R^2 = 0.80$	0.7982	0.7873

Tabella 1: Valori medi di R^2_{adj} e di R^2_{PRESS}

Dai grafici riportati in Fig. 1 sono illustrati due boxplot che confrontano i 10.000 valori generati di R^2_{adj} ed R^2_{PRESS} rispettivamente per i modelli di regressione Y1 ed Y2 e nella tabella 1 sono riportati i valori medi di queste due misure.

Per il modello con R^2 imposto pari a 0.50, l'R^2_{adj} medio generato è di 0.4993, mentre l'R^2_{PRESS} medio generato è di 0.4722. Quindi mediamente c'è una differenza del 5.75%.

Per il modello con un obiettivo di R^2 pari a 0.80, l'R^2_{adj} medio generato è di 0.7982, mentre l'R^2_{PRESS} medio generato è di 0.7873. Quindi mediamente c'è una differenza dell'1.38%.

In generale, si osserva che la differenza tra i valori di R^2_{adj} e R^2_{PRESS} è più evidente nei dataset con un target di R^2 più basso.

I boxplot mostrano che i valori di R^2_{PRESS}, comparati con quelli dell' R^2_{adj} risultano mediamente minori.

2.3 Aggiunta di predittori non necessari al modello

Questa ulteriore analisi si concentra sull'effetto dell'aggiunta di predittori non necessari al modello e su come ciò influenzi le misure di R^2_{PRESS} e R^2_{adj}.
Inizialmente, viene considerato il modello vero, composto solo da tre predittori.
Successivamente, vengono aggiunti progressivamente fino a 70 predittori, al fine di studiare il comportamento di R^2_{PRESS} e R^2_{adj}.
Dato che i predittori aggiunti sono non necessari, l'R^2_{adj} dovrebbe diminuire indicando un peggiore adattamento del modello ai dati; in realtà, dopo le 10.000 simulazioni si nota come tale valore rimanga pressochè costante all'aumentare del numero dei predittori, indicando una limitazione di tale criterio (Ida Marie Alcantara, Joshua Naranjo, Yanda Lang, 2022). Al contrario, il valore di R^2_{PRESS} decresce in modo significativo, suggerendo un peggioramento della capacità del modello e indicando una maggiore adeguatezza del criterio R^2_{PRESS}.
In seguito viene fornita una panoramica dei risultati ottenuti dall'analisi di simulazione (Appendice A.1). Nella tabella 2 sono raffigurati i valori medi delle misure R^2_{PRESS} ed R^2_{adj} ricavati per alcuni valori dei t predittori aggiunti al modello.

	$R^2 = 0.50$		$R^2 = 0.80$	
t	R^2_{adj}	R^2_{PRESS}	R^2_{adj}	R^2_{PRESS}
0	0.4998	0.4724	0.7985	0.7876
1	0.4997	0.4656	0.7978	0.7856
2	0.5007	0.4626	0.7992	0.7838
3	0.5001	0.4504	0.7985	0.7785
4	0.5005	0.4440	0.7984	0.7766
5	0.5007	0.4350	0.7994	0.7730
10	0.5001	0.3907	0.7990	0.7541
20	0.5002	0.2795	0.7991	0.7094
30	0.5018	0.1142	0.7998	0.6426
40	0.5011	-0.1473	0.8007	0.5358
50	0.5001	-0.6219	0.8020	0.3465
60	0.5020	-1.7686	0.8026	-0.1123
70	0.5027	-8.7748	0.8030	-2.9257

Tabella 2: Valori medi di R^2_{PRESS} ed R^2_{adj} con l'aumentare di t

Si nota come l'R^2_{PRESS} può assumere valori negativi quando il numero di predittori non necessari diventa eccessivo. Ciò indica una sovra-parametrizzazione del modello, in cui la presenza di predittori non informativi o ridondanti causa una perdita di capacità predittiva.

La seguente figura illustra graficamente la separazione tra l'R^2_{PRESS} e l'R^2_{adj} al crescere dei predittori t non necessari aggiunti al modello.

In ogni grafico è rappresentata una diagonale che indica la linea in cui l'R^2_{adj} e l'R^2_{PRESS} hanno lo stesso valore. I punti in rosso rappresentano i valori di confronto tra l'R^2_{adj} e l'R^2_{PRESS} quando il target di R^2 è pari a 0.50, mentre i punti in blu rappresentano i valori corrispondenti quando il target di R^2 è pari a 0.80.

Questa visualizzazione permette di osservare la differenza tra le due misure al variare del numero di predittori non necessari e di apprezzare come tale differenza sia più pronunciata

per dataset con un obiettivo di R^2 più basso.

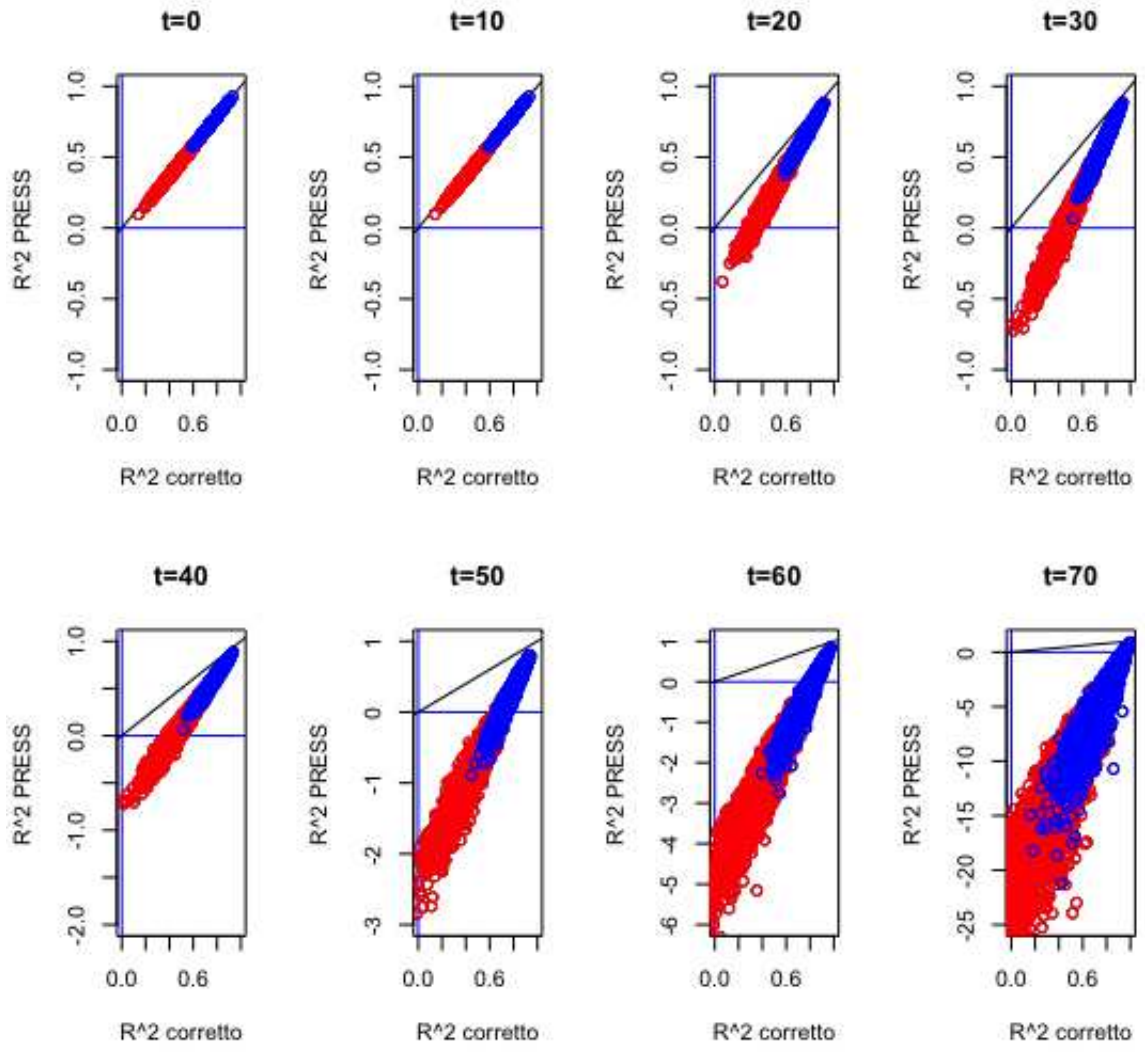

Figura 2: Scatterplot di R^2_{PRESS} e R^2_{adj} rispetto t

Dalla figura 2 si osserva quindi che all'aumentare di t che segue un range da 1 a 70, i punti dei grafici si allontanano progressivamente dalla linea diagonale, sia per il caso in cui il target di R^2 è pari a 0.50 che per il caso in cui è pari a 0.80.

Ciò esprime una grande differenza tra le due misure di adattamento.

Infatti, per un modello a cui non è aggiunto alcun predittrore i valori di R^2_{PRESS} ed R^2_{adj} sono molto simili e quindi i punti blu e rossi giacciono sulla diagonale, mentre per un modello in cui sono aggiunti fino a t = 70 predittori, i due valori risultano molto differenti e quindi i punti tendono a distaccarsi completamente dalla diagonale.

Da questo studio di simulazione si possono confermare e riassumere le diverse caratteristiche di R^2_{PRESS}:

- è sempre inferiore di R^2_{adj}, quindi tiene conto della complessità del modello e penalizza l'aggiunta di predittori non necessari;
- contrariamente a R^2_{adj}, R^2_{PRESS} decresce rapidamente all'aumentare dei predittori non necessari al modello. Ciò suggerisce che l'aggiunta di predittori non informativi o ridondanti riduce significativamente la capacità di generalizzazione del modello;
- R^2_{PRESS}, come l'R^2_{adj} può assumere valori negativi quando il numero di predittori non necessari diventa eccessivo, indicando un netto peggioramento delle prestazioni predittive del modello;
- R^2_{PRESS} decresce più velocemente quando predittori non necessari vengono aggiunti al modello per i dataset costruiti sulla base di un target di R^2 più basso (in questo caso, $R^2 = 0.50$). Questo indica che l'effetto negativo dei predittori non necessari è maggiormente evidente quando la relazione tra i predittori e la variabile di risposta è più debole.

3 Analisi comparativa tra R^2_{PRESS} e altre misure di adattamento

Per una migliore valutazione delle prestazioni di R^2_{PRESS}, si sono condotti ulteriori studi di simulazione che tendono ad esaminare la comparazione tra le sue performance oltre che con quelle dell'R^2_{adj} anche quelle con i seguenti criteri di selezione del modello :

- CP di Mallow
- AIC
- SBC

Lo scopo di questo studio è di calcolare le percentuali di identificazioni corrette, ovvero quante volte tra i vari modelli candidati viene selezionato il modello vero:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i$$

$$= 0.2 + 0.3X_1 + 0.5X_2 + 0.6X_3 + \epsilon_i$$

quindi quello contenente soli 3 predittori.
Lo studio di simulazione è svolto considerando numerosità pari a 15 e 80 per i seguenti scenari:

- modello con predittori ed errori distribuiti secondo una normale standard;
- modello con predittori distribuiti normalmente ed errori distribuiti come una T di Student con 4 gradi di libertà ;
- modello con predittori distribuiti normalmente ed errori distribuiti come una Normale asimmetrica con α=5.

Quindi, oltre alla distribuzione Normale standard, vengono scelte:

- una distribuzione che segue una Normale asimmetrica : distribuzione di probabilità continua utilizzata in statistica per modellare dati asimmetrici. È caratterizzata da una forma simile a una normale, ma con una coda allungata in una direzione, rendendo la distribuzione asimmetrica. In questo studio è definita da un parametro α pari a 5, significa che la distribuzione avrà una coda lunga nella direzione positiva;

- una distribuzione che segue una T di Student : distribuzione di probabilità continua utilizzata in statistica per modellare variabili casuali che seguono una distribuzione normale ma con maggiore incertezza o variabilità. In questo studio è definita da 4 gradi di libertà perciò la distribuzione T ha code più pesanti rispetto alla distribuzione normale.

Per ogni scenario, inoltre, vengono comparati due casi:

- predittori non correlati;
- predittori con correlazione pari a 0.25.

Vengono considerate, come nello studio precedente, 10.000 simulazioni ed ogni misura è implementata per valutare quale modello è il migliore.
Per R^2_{adj} ed R^2_{PRESS} viene selezionato come migliore, il modello con il valore più alto. Contrariamente, per AIC, SBC e Cp di Mallow viene selezionato come migliore il modello con il valore più basso (Appendice A1-A2-A3-A4-A5).
Tra i criteri di selezione utilizzati per il confronto, non fa parte il coefficiente di determinazione R^2 poichè il suo valore continuerebbe ad aumentare preferendo in ogni caso il modello con il numero di predittori più elevato, perciò la percentuale di identificazioni corrette risulterebbe 0.

3.1 Campione piccolo

L'attenzione della prima parte dello studio si concentra sulle simulazioni condotte su un campione di dimensione ridotta, n = 15. Sono state generate 10.000 simulazioni su modelli composti a partire da un solo predittore fino a un massimo di 8. Perciò sono aggiunti 5 predittori non necessari al modello.

modello	predittori nel modello
mod1	X_1
mod2	X_2
mod3	X_3
mod4	X_1, X_2
mod5	X_1, X_3
mod6	X_2, X_3
mod7	X_1, X_2, X_3
mod8	X_1, X_2, X_3, X_4
mod9	X_1, X_2, X_3, X_4, X_5
mod10	$X_1, X_2, X_3, X_4, X_5, X_6$
mod11	$X_1, X_2, X_3, X_4, X_5, X_6, X_7$
mod12	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$

Tabella 3: Lista di modelli candidati: campione piccolo

La tabella 3 rappresenta l'elenco dei modelli candidati esaminati in questo studio.
Questa lista include 12 modelli candidati, considerati durante la simulazione, comprendendo quelli composti da tutte le possibili combinazioni dei 3 predittori (X1, X2, X3) che fanno parte del modello vero.

3.1.1 Primo caso: predittori ed errori distribuiti secondo una Normale standard

Si analizzano i modelli composti da predittori (correlati e non correlati) ed errori entrambi distribuiti come una normale standard. Successivamente vengono calcolati i valori di tutti i criteri selezionati per essere confrontati.
Nella tabella 4 sono rappresentate le percentuali di identificazioni corrette per questo primo caso considerato.

criterio	X non correlate	X correlate
R^2_{PRESS}	16.98%	16.27%
SBC	12.10%	12.29%
CP di Mallow	11.56%	11.64%
AIC	10.90%	10.85%
R^2_{adj}	10.10%	10.70%

Tabella 4: campione piccolo : % di identificazioni corrette, X Normali e ϵ Normali

È evidente che, sia per predittori correlati sia per i non correlati, l'R^2_{PRESS} presenta la percentuale di identificazioni corrette più elevata rispetto agli altri criteri, pari al 16.98%. E' seguita dal criterio di informazione SBC, con il 12.10% e dalla statistica Cp di Mallow, con valore molto vicino, pari all'11.56%.
I criteri che risultano peggiori sono l'AIC con il 10.90% e l'R^2_{adj}, con il 10.10%.

3.1.2 Secondo caso: predittori distribuiti normalmente ed errori distribuiti secondo una T di Student

Nel presente caso, si esamina la stessa lista di modelli candidati del precedente, ma viene considerato uno scenario diverso in cui gli errori non seguono una distribuzione normale, bensì una distribuzione T di Student con 4 gradi di libertà.
Questa scelta di distribuzione degli errori consente di esplorare come la non normalità influisce sulle prestazioni dei modelli di regressione e sulla selezione del modello ottimale. La distribuzione T di Student, con un numero relativamente basso di gradi di libertà, è caratterizzata da code più pesanti rispetto alla distribuzione normale.

criterio	X non correlate	X correlate
R^2_{PRESS}	15.33%	14.22%
SBC	10.40%	10.91%
CP di Mallow	10.36%	10.71%
AIC	10.21%	10.10%
R^2_{adj}	9.11%	10.29%

Tabella 5: campione piccolo : % di identificazioni corrette, X Normali e ϵ distribuiti secondo una T di student con 4 gradi di libertà

Rispetto al caso precedente, si osserva che le percentuali di identificazioni corrette, sia per predittori correlati che non correlati, presentano mediamente un valore inferiore all'1.40%. Questo significa che i modelli candidati, nel contesto della distribuzione degli errori con distribuzione T di Student con 4 gradi di libertà, tendono a identificare correttamente i predittori con una frequenza più bassa rispetto al caso precedente.
Nonostante ciò, la classificazione per ordine decrescente delle misure di selezione del modello rimane la stessa del caso precedente.

3.1.3 Terzo caso: predittori distribuiti normalmente ed errori distribuiti secondo una Normale asimmetrica

Nel seguente caso, vengono considerati gli stessi modelli rappresentati nella Tabella 3, ma con una differenza: gli errori non seguono una distribuzione normale, ma una distribuzione Normale asimmetrica con un parametro α pari a 5.
La scelta di utilizzare una distribuzione Normale asimmetrica per gli errori consente di esaminare come la deviazione dalla normalità influisca sulle prestazioni dei modelli di regressione e sulla selezione del modello migliore.

criterio	X non correlate	X correlate
R^2_{PRESS}	21.70%	20.91%
SBC	16.32%	16.35%
CP di Mallow	14.40%	14.70%
AIC	13.36%	12.81%
R^2_{adj}	11.45%	12.23%

Tabella 6: campione piccolo : % di identificazioni corrette, X Normali e ϵ distribuiti secondo una Normale asimmetrica con $\alpha = 5$.

La Tabella 5 evidenzia che nel caso in cui gli errori seguano una distribuzione Normale asimmetrica, tutte le misure considerate presentano una percentuale di identificazioni corrette superiore rispetto ai due casi precedenti.
In particolare, l'R^2_{PRESS} mostra la percentuale di identificazioni corrette più elevata, pari al 21.70%. Questo valore è significativamente più alto rispetto ai casi precedenti, indicando che l'R^2_{PRESS} è particolarmente efficace nell'identificare correttamente i predittori significativi nel contesto della distribuzione Skew Normal degli errori.
Le misure SBC, CP di Mallow e AIC seguono in ordine decrescente, con percentuali di identificazioni corrette inferiori rispetto all'R^2_{PRESS} ma comunque superiori rispetto ai casi precedenti. L'R^2_{adj}, con una percentuale del 11.45%, si posiziona all'ultimo posto tra le misure considerate, indicando una minor capacità di identificazione dei predittori rispetto alle altre misure.

3.2 Campione grande

Dopo aver valutato le prestazioni delle misure di selezione del modello nel campione di piccole dimensioni, l'attenzione si sposta alla comparazione dei criteri in un campione di dimensione più elevata, composto da 80 osservazioni.
Il modello completo iniziale è composto da un totale di 20 predittori, dei quali 17 sono

predittori non necessari aggiunti al modello vero. Poiché il numero di possibili modelli è elevato, viene considerato solo un certo numero di modelli candidati per l'analisi.

Basandosi sulla regressione di tutti i sottoinsiemi effettuata nel campione di piccole dimensioni, si osserva che i modelli composti da meno predittori hanno maggiori probabilità di essere selezionati come migliori. Pertanto, vengono considerate anche in questo caso tutte le possibili combinazioni dei predittori del modello vero come modelli candidati. Successivamente, viene adattato un numero crescente di predittori non necessari aggiunti al modello vero.
I modelli candidati considerati nella sezione corrente sono riportati nella seguente tabella.

modello	predittori nel modello
mod1	X_1
mod2	X_2
mod3	X_3
mod4	X_1, X_2
mod5	X_1, X_3
mod6	X_2, X_3
mod7	X_1, X_2, X_3
mod8	X_1, X_2, X_3, X_4
mod9	X_1, X_2, X_3, X_4, X_5
mod10	$X_1, X_2, X_3, X_4, X_5 ... X_{10}$
mod11	$X_1, X_2, X_3, X_4, X_5 ... X_{20}$

Tabella 7: lista dei modelli candidati considerati per grandi campioni

3.2.1 Primo caso : predittori ed errori distribuiti secondo una Normale standard

Il caso analizzato in questa sezione riguarda il campione di dimensione grande in cui sia i predittori che gli errori dei modelli candidati seguono una distribuzione normale standard.

criterio	X non correlate	X correlate
R^2_{PRESS}	26.20%	25.04%
SBC	29.74%	29.37%
CP di Mallow	22.90%	23.35%
AIC	22.65%	22.64%
R^2_{adj}	17.15%	17.01%

Tabella 8: campione grande : % di identificazioni corrette, X Normali e ϵ Normali

I risultati ottenuti evidenziano che tutti i criteri di selezione del modello presentano, in media, una percentuale significativamente più elevata rispetto ai casi analizzati nel campione di dimensione inferiore.

In particolare, in questo caso l'R^2_{PRESS} mostra un incremento nella sua percentuale di identificazione corretta rispetto ai casi precedenti.

Tuttavia il criterio SBC supera l'R^2_{PRESS} con una percentuale più elevata, superiore del 3.17%. Seguono CP di Mallow e AIC, i quali presentano una percentuale simile tra loro. D'altra parte, R^2_{adj} mostra la percentuale più bassa di identificazione corretta del modello vero, raggiungendo il 17.15% dei casi.

3.2.2 Secondo caso : predittori distribuiti normalmente ed errori distribuiti secondo una T di Student

Il seguente caso riguarda gli stessi modelli rappresentati sulla lista della tabella 7 composti da predittori distribuiti secondo una normale standard ma considerando una diversa distribuzione degli errori.

In particolare, gli errori seguono una distribuzione T di Student con 4 gradi di libertà, anziché la distribuzione normale.

criterio	X non correlate	X correlate
R^2_{PRESS}	23.98%	22.24%
SBC	19.67%	19.31%
CP di Mallow	17.47%	17.30%
AIC	17.36%	17.85%
R^2_{adj}	16.17%	17.99%

Tabella 9: campione grande : % di identificazioni corrette, X Normali e ϵ distribuiti secondo una T di Student con 4 gradi di libertà

I risultati evidenziano che il criterio R^2_{PRESS} si conferma il migliore con la percentuale più elevata di identificazioni corrette, raggiungendo il 23.98%.
Seguono il criterio SBC, CP di Mallow, AIC ed R^2_{adj}, rispettivamente. Quindi, anche in questo contesto in cui gli errori seguono una distribuzione T di Student con 4 gradi di libertà, i criteri R^2_{PRESS} e SBC si confermano i migliori indicatori per la selezione del modello.

3.2.3 Terzo caso : predittori distribuiti normalmente ed errori distribuiti secondo una Normale asimmetrica

Come ultimo caso si analizzano i modelli rappresentati sulla lista della tabella 7, composti da predittori distribuiti secondo una Normale standard e con gli errori che seguono una distribuzione Normale asimmetrica con α pari a 5.

criterio	X non correlate	X correlate
R^2_{PRESS}	27.17%	27.26%
SBC	36.37%	36.96%
CP di Mallow	26.45%	25.50%
AIC	25.73%	26.03%
R^2_{adj}	14.19%	19.38%

Tabella 10: campione grande : % di identificazioni corrette, X Normali e ϵ distribuiti secondo una Normale asimmetrica con $\alpha = 5$.

Come nel primo caso, in cui sia predittori sia errori sono distribuiti secondo una normale standard, il criterio SBC risulta essere il migliore, con percentuale di identificazioni corrette pari al 36.37%, a discapito del criterio R^2_{PRESS} che risulta pari al 27.17%. Segue una classificazione simile a quella osservata nei casi precedenti, con le misure CP di Mallow, AIC e R^2_{adj} in ordine decrescente di performance.

3.3 Conclusione dell'analisi comparativa

Le simulazioni condotte per entrambe le dimensioni campionarie hanno evidenziato che l'R^2_{PRESS} ha identificato correttamente il vero modello in modo più frequente rispetto agli altri criteri considerati.

Per lo studio basato sul campione di dimensione ridotta, l'R^2_{PRESS} ha presentato una percentuale sempre superiore rispetto al criterio SBC e di conseguenza tutti gli altri indicando una maggiore capacità di selezionare il modello corretto.

Tuttavia per lo studio basato sul campione di dimensione più elevata, il criterio SBC ha mostrato una percentuale superiore dell' R^2_{PRESS} per due casi su tre.

Per quanto riguarda i criteri di informazione, SBC ha dimostrato prestazioni nettamente migliori rispetto all'AIC in tutti gli scenari considerati.

In conclusione, le misure di selezione R^2_{PRESS}, SBC e CP di Mallow si sono rivelati consistenti nell'identificare il modello vero come migliore. D'altra parte, i criteri R^2 ed AIC hanno quasi sempre identificato il modello completo come migliore risultando i metodi peggiori.

Le simulazioni evidenziano la competitività dell'R^2_{PRESS} in diverse condizioni e confermano la sua validità come criterio per la selezione del modello.

4 Analisi su dati reali

In questo capitolo viene illustrato come utilizzare la statistica R^2_{PRESS} nella selezione del modello confrontandola con altre misure di selezione utilizzando due dataset reali.

4.1 Dataset Air Quality

Il primo dataset è tratto dalla libreria "datasets" di Rstudio ed è denominato "Air quality". Questo dataset contiene informazioni riguardanti le misurazioni sulla qualità dell'aria di New York nel periodo da Maggio a Settembre del 1973. Il dataset ha numerosità n = 111 ed è composto da 6 variabili:

- ozone : concentrazione di ozono;
- solar r. : radiazione solare in langley (unità di energia solare);
- wind : velocità del vento in miglia all'ora;
- temp : temperatua in gradi Fahrenheit;
- month : mese dell'anno numerato da 5 (Maggio) a 9 (Settembre);
- day : giorno del mese, numerato da 1 a 31.

Si considera "ozone" come variabile risposta.
I valori per ciascun criterio di selezione del modello per ogni numero di predittori sono riassunti nella seguente tabella.

predittori	R^2	R^2_{adj}	R^2_{PRESS}	AIC	SBC	CP
temp	0.4883	0.4833	0.4795	1023.775	1031.904	36.348
wind, temp	0.5814	0.5736	0.5585	1003.416	1014.254	12.196
solar r., wind, temp	0.6059	0.5948	0.5783	998.717	1012.265	7.332
solar r., wind, temp, month	0.6199	0.6055	0.5873	996.712	1012.969	5.422
solar r., wind, temp, month, day	0.6249	0.6071	0.5836	997.219	1016.185	6.000

Tabella 11: Riepilogo dei valori dei criteri del modello ottenuti utilizzando i dati di Airquality.

Per trovare l'ordine con cui seguire la regressione lineare multipla, è stata calcolata la matrice di correlazione tra tutti i possibili predittori per osservare quali sono i più rilevanti per il modello (Appendice A6). In seguito è raffigurato il diagramma di dispersione a coppie.

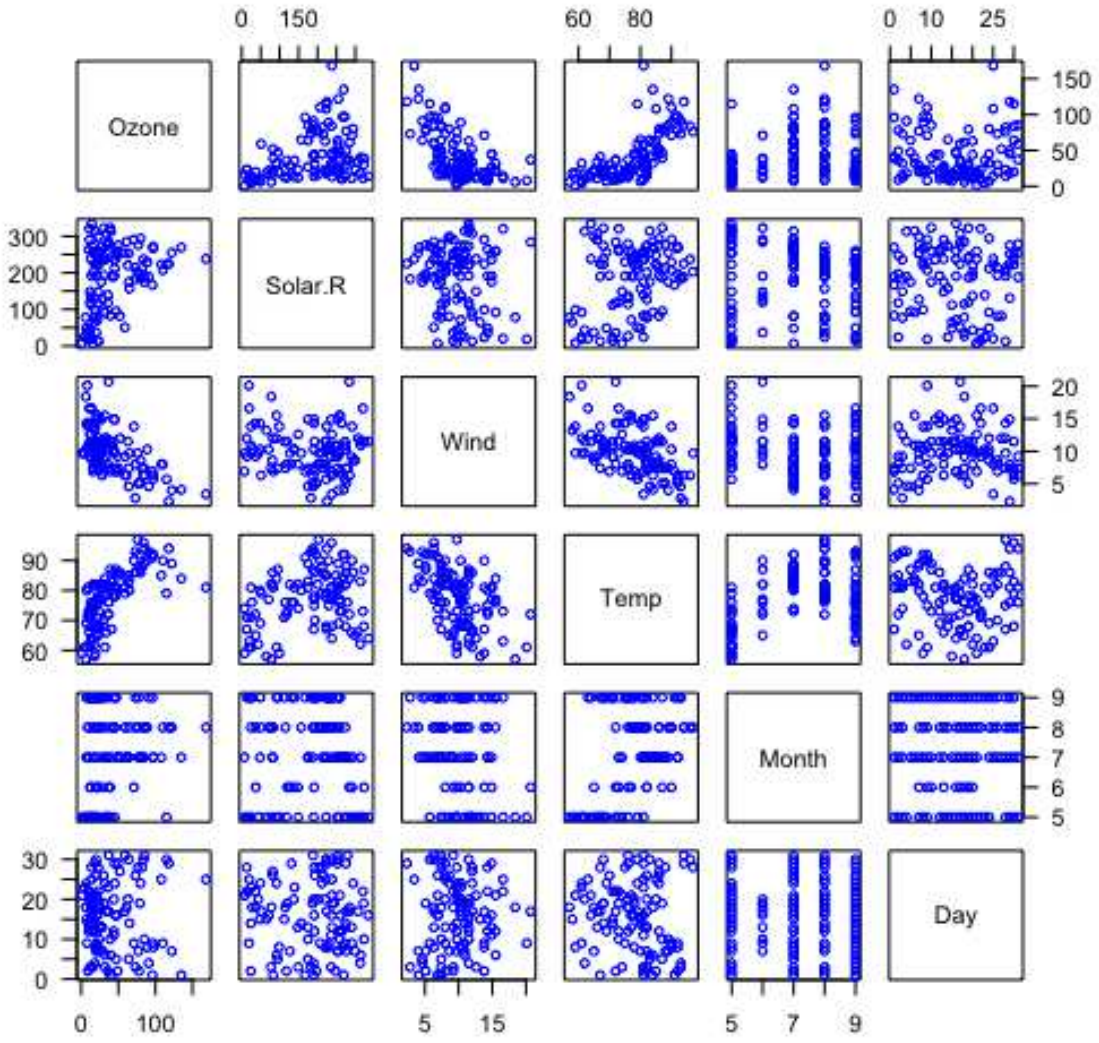

Figura 3: Diagramma di dispersione a coppie per il dataset "Airquality"

Si parte quindi dal modello contenente solamente dalla variabile "temp" che risulta la più correlata con la variabile risposta "ozone" e si arriva al modello completo composto da 5 predittori. Sulla base dei valori ottenuti, il modello con 4 predittori presenta il valore

più elevato di R^2_{PRESS}. Questa selezione è coerente anche con i criteri di selezione come AIC e CP di Mallow, che identificano lo stesso modello come il migliore.

Tuttavia, se si analizzano i valori di R^2_{PRESS} per il modello a 3 predittori e a 5 predittori, si può notare che le differenze tra di essi sono relativamente piccole (rispettivamente 0.0037 e 0.009). Entrambi i modelli sono quindi sufficientemente accurati per la predizione dei livelli di concentrazione di ozono nell'aria.

Analogamente, il metodo SBC seleziona come miglior modello quello composto da 3 predittori, ma il modello a 4 predittori risulta altrettanto valido, considerando la minima differenza tra i valori di R^2_{PRESS} (0.70).
Considerando la stima dei parametri e l'analisi diagnostica, si preferisce adottare il modello a 4 predittori in quanto presenta il valore più elevato di R^2_{PRESS}. Tuttavia, è importante considerare anche altri fattori come la complessità del modello, l'interpretabilità dei predittori e l'effetto dell'overfitting.

4.2 Dataset Gala

Il secondo dataset preso in considerazione è "Gala", proveniente dalla libreria "faraway" di R. Nel dataset sono presenti le 30 isole galapagos e 7 variabili che descrivono ciascuna isola:

- species : numero di specie di piante presenti sull'isola;
- endemics : numero di specie endemiche;
- area : superficie dell'isola espressa in chilometri quadrati;
- elevation : altitudine massima dell'isola espressa in metri;
- nearest: distanza dall'isola più vicina espressa in chilometri;
- scruz : distanza dall'isola di Santa Cruz espressa in chilometri;
- adjacent : area dell'isola adiacente espressa in chilometri quadrati.

In questo dataset l'obiettivo è determinare la relazione tra il numero di specie di piante trovate sull'isola ed alcune variabili geografiche. Quindi viene presa in considerazione "species" come variabile risposta e le altre come esplicative. Anche per questo dataset per trovare l'ordine con cui seguire la regressione lineare multipla, è calcolata la matrice di correlazione tra tutti i possibili predittori per osservare quali sono i più rilevanti per il modello (Appendice A7).

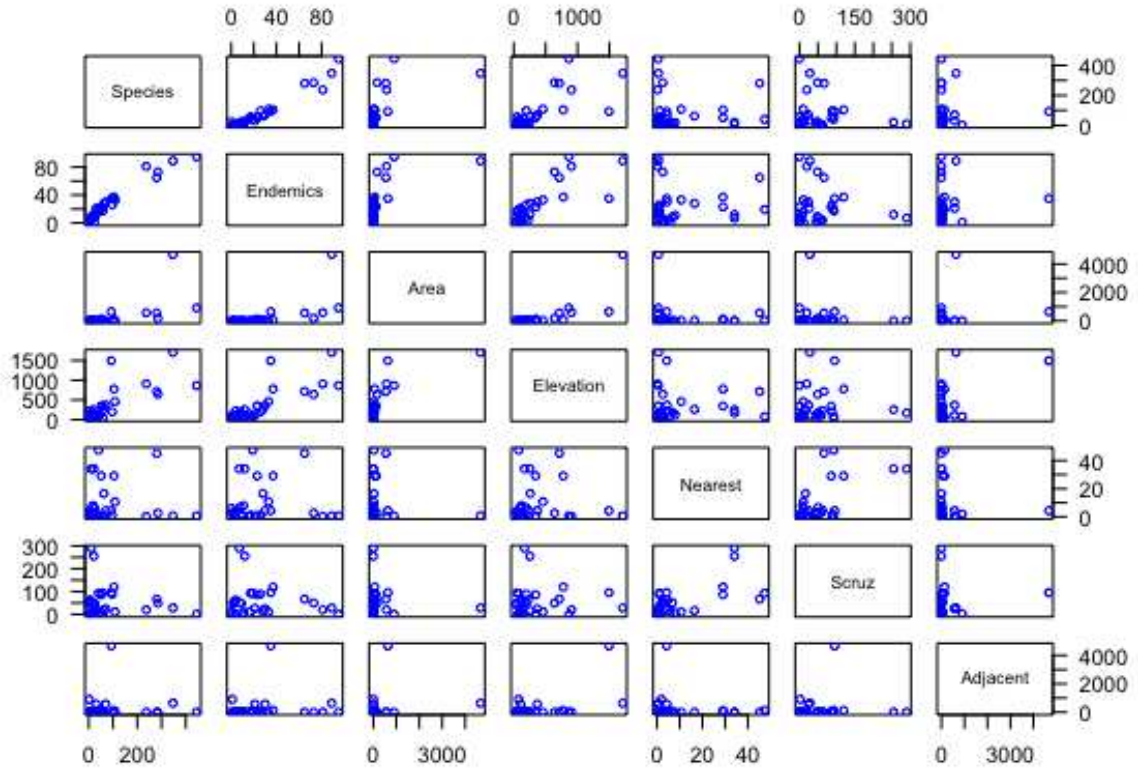

Figura 4: Diagramma di dispersione a coppie per il dataset "Gala"

Dal diagramma di dispersione a coppie si può notare quali variabili sono maggiormente correlate tra loro. In questo caso si può osservare una forte correlazione positiva tra la variabile risposta ed "endemics". Quindi si inizia a fare regressione in un modello composto dalla sola variabile "endemics" aggiungendo variabili (in ordine di correlazione) fino al modello completo, con i 5 predittori.

predittori	R^2	R^2_{PRESS}	R^2_{adj}	AIC	SBC	CP
endemics	0.9426	0.926	0.9406	288.9	293.1	0.074
endemics, elevation	0.9452	0.923	0.9412	289.5	295.1	0.874
endemics, elevation, area	0.9492	0.364	0.9433	289.2	296.2	1.088
endemics, elevation, area, adjacent	0.9493	0.136	0.9412	291.2	299.6	3.044
endemics, elevation, area, adjacent, nearest	0.9494	0.116	0.9388	293.1	302.9	5.012
endemics, elev, area, adjacent, nearest, scruz,	0.9494	0.106	0.9362	295.1	306.3	7.000

Tabella 12: Riepilogo dei valori dei criteri del modello ottenuti utilizzando i dati di Gala.

Dai valori dei criteri considerati nella tabella 12, il modello composto da 1 predittore risulta essere il modello migliore in base ai criteri R^2_{PRESS}, AIC, SBC e Cp di Mallow. R^2_{adj} ha selezionato il modello a 3 predittori come migliore in quanto presenta il valore più alto di questo criterio tra gli altri modelli. Dai valori di R^2_{PRESS} i modelli candidati sono quelli composti da uno o due fattori poiché i loro valori sono molto più alti rispetto agli altri modelli. I valori di R^2_{adj} per tutti i modelli sono vicini tra loro.
In questo caso, viene preferito il modello con un solo predittore, endemics che gioca quindi un ruolo cruciale nella spiegazione delle variazioni della variabile di interesse.
In statistica, la preferenza per un modello con un solo predittore può essere dovuta anche a ragioni di semplicità e interpretazione. In alcuni casi, un modello più semplice può essere preferito perché è più facile da interpretare e applicare.
Quindi, si può sottolineare che il modello composto da "endemics" come unico predittore offre una spiegazione accurata della variabile di risposta con una maggiore semplicità.

5 Conclusione

Lo studio condotto ha approfondito l'utilizzo della statistica PRESS e la sua misura di adattamento, R^2_{PRESS}, nella regressione lineare multipla. Lo studio di simulazione ha evidenziato una competizione tra la statistica R^2_{PRESS} e altri criteri di selezione del modello: R^2, R^2_{adj}, AIC, SBC e CP di Mallow.

In particolare, durante un confronto iniziale tra R^2_{PRESS} e R^2_{adj}, è stato osservato che R^2_{PRESS} offre una migliore misura della bontà di adattamento quando si lavora con un elevato numero di predittori che potrebbero non essere necessari per il modello.

Un ulteriore studio di simulazione ha dimostrato come l'R^2_{PRESS} abbia determinato il vero modello come migliore, ottenendo una percentuale di identificazioni corrette più elevata rispetto quella degli altri criteri.

Questo risultato si è ripetuto nel caso di un campione con numerosità pari a 15 per i seguenti casi:

- predittori (correlati e non correlati) ed errori distribuiti normalmente;
- predittori (correlati e non corrrelati) distribuiti normalmente ed errori distribuiti secondo una T di Student;
- predittori (correlati e non correlati) distribuiti normalmente ed errori distribuiti secondo una Normale asimmetrica.

Nel caso di un campione di numerosità più elevata invece le percentuali di identificazioni corrette del criterio SBC risultano simili e in due casi su 3 superiori rispetto quelle dell'R^2_{PRESS}. Perciò la misura SBC presenta performance chiaramente migliori rispetto gli altri criteri. Ponendo l'attenzione su dati reali si è confermato lo studio di simulazione condotto. Infatti la statistica R^2_{PRESS} , come atteso, stabilisce quali variabili del modello siano più rilevanti di altri per la regressione.

Sono in corso ulteriori ricerche per capire come la statistica PRESS possa essere estesa a modelli non lineari e come si confronti con altre metodologie di convalida al fine di determinare la piena portata dell'utilità di R^2_{PRESS} nella selezione dei modelli.

Appendice

In questa appendice vengono riportati dei pezzi delle righe di codice R utilizzate per ottenere i risultati riportati nel secondo, terzo e quarto capitolo.

A Codice R per generare i boxplot

```
rm(list=ls())
set.seed(123)
PRESS1 <- function(mod1) {
  i <- residuals(mod1)/(1 - lm.influence(mod1)$hat)
  return(sum(i^2))
}
PRESS2 <- function(mod2) {
  i <- residuals(mod2)/(1 - lm.influence(mod2)$hat)
  return(sum(i^2))
}
R2_PRESS1<-c()
R2_PRESS2<-c()
R2_ADJ1<-c()
R2_ADJ2<-c()

for (i in seq(1,10000)){
  n <- 80
  x1 <- rnorm(n)
  x2 <- rnorm(n)
  x3 <- rnorm(n)
  e_1 <- rnorm(n,0,0.8367)
  e_2 <- rnorm(n,0,0.4183)
  B0 <- 0.2
```

```
  B1 <- 0.3
  B2 <- 0.5
  B3 <- 0.6
  y1 <- B0+B1*x1+B2*x2+B3*x3+e_1
  y2 <- B0+B1*x1+B2*x2+B3*x3+e_2
  ybar1 <- mean(y1)
  ybar2 <- mean(y2)
  mod1=lm(y1 ~ x1+x2+x3)
  summary(mod1)
  mod2=lm(y2 ~ x1+x2+x3)
  summary(mod2)
  sse1 <- sum((y1-fitted(mod1))^2)
  sse2 <- sum((y2-fitted(mod2))^2)
  sst1 <- sum((y1-ybar1)^2)
  sst2 <- sum((y2-ybar2)^2)
  PRESS1(mod1)
  PRESS2(mod2)
  R2_PRESS1<- append(R2_PRESS1, 1-((PRESS1(mod1))/((n/(n-1))^2*sst1)))
  R2_PRESS2<- append(R2_PRESS2, 1-((PRESS2(mod2))/((n/(n-1))^2*sst2)))
  p <- 3
  R2_ADJ1 <- append(R2_ADJ1, 1- ((sse1/(n-p))/(sst1/(n-1))))
  R2_ADJ2 <- append(R2_ADJ2, 1- ((sse2/(n-p))/(sst2/(n-1))))
}

par(mfrow=c(1,2))
boxplot(R2_ADJ1,R2_PRESS1, main=c("Modello Y1"), xlab=c("R^2 corretto
R^2 PRESS"))
```

```
boxplot( R2_ADJ2,R2_PRESS2,main=c("Modello Y2"), xlab=c("R^2 corretto
R^2 PRESS"))
#R^2 target = 0.50
mean(R2_ADJ1)
mean(R2_PRESS1)
#R^2 target = 0.80
mean(R2_ADJ2)
mean(R2_PRESS2)
```

In seguito è riportata la stessa tabella presente nel capitolo 2 con l'aggiunta dei valori medi di R^2 ricavati dalle simulazioni per un ulteriore confronto con R^2_{PRESS}.

	$R^2 = 0.50$			$R^2 = 0.80$		
t	R^2_{adj}	R^2_{PRESS}	R^2	R^2_{adj}	R^2_{PRESS}	R^2
0	0.4998	0.4724	0.5119	0.7985	0.7876	0.8033
1	0.4997	0.4656	0.5187	0.7978	0.7856	0.8055
2	0.5007	0.4626	0.5266	0.7992	0.7838	0.8093
3	0.5001	0.4504	0.5327	0.7985	0.7785	0.8116
4	0.5005	0.4440	0.5387	0.7984	0.7766	0.8137
5	0.5007	0.4350	0.5451	0.7994	0.7730	0.8168
10	0.5001	0.3907	0.5766	0.7990	0.7541	0.8300
20	0.5022	0.2795	0.6413	0.7991	0.7094	0.8552
30	0.5048	0.1142	0.7058	0.7998	0.6426	0.8809
40	0.5071	-0.1473	0.7693	0.8007	0.5358	0.9067
50	0.5101	-0.6219	0.8326	0.8020	0.3465	0.9329
60	0.5220	-1.7686	0.8977	0.8076	-0.1123	0.9586
70	0.5420	-8.7748	0.9619	0.8240	-2.9257	0.9844

Tabella 13: Valori medi di R^2_{PRESS}, R^2_{adj} ed R^2 all'aumentare di t

A.1 Codice R aggiuntivo per generare lo scatterplot

```
par(mfrow=c(2,4))
a <- cbind(R2_ADJ1,R2_ADJ2)
b <- cbind(R2_PRESS1,R2_PRESS2)
plot(a, b, ylim=c(-1,1), xlim=c(0,1), col=1, main="t=0",
     xlab="R^2 corretto", ylab="R^2 PRESS")
abline(0,1)
abline(v=0,h=0, col="blue")
points(a[,1], b[,1], col="red")
points(a[,2], b[,2], col="blue")
```

Il codice procede lo allo stesso modo aggiungendo t = 2,3,4,5,10,20,30,40,50,60 e 70 predittori non necessari al modello vero.

Per analizzare il comportamento di R^2_{PRESS} tra i vari modelli viene utilizzato il seguente ciclo for:

```
count_RP <- 0
for (i in 1:length(R2_PRESS_3)) {
  if (R2_PRESS_3[i] > R2_PRESS_11[i] &
      R2_PRESS_3[i] > R2_PRESS_12[i] &
      R2_PRESS_3[i] > R2_PRESS_13[i] &
      R2_PRESS_3[i] > R2_PRESS_212[i] &
      R2_PRESS_3[i] > R2_PRESS_213[i] &
      R2_PRESS_3[i] > R2_PRESS_223[i] &
      R2_PRESS_3[i] > R2_PRESS_4[i] &
      R2_PRESS_3[i] > R2_PRESS_5[i] &
      R2_PRESS_3[i] > R2_PRESS_6[i] &
      R2_PRESS_3[i] > R2_PRESS_7[i] &
      R2_PRESS_3[i] > R2_PRESS_8[i]) {
    count_RP <-  count_RP + 1
```

```
  }
}
```

A.2 Codice per generare simulazioni del criterio SBC

```
SBC_11 <- c()
for (i in seq(1,10000)){
  n <- 15
  x1 <- rnorm(n)
  x2 <- rnorm(n)
  x3 <- rnorm(n)
  e_1 <- rnorm(n)
  B0 <- 0.2
  B1 <- 0.3
  B2 <- 0.5
  B3 <- 0.6
  y1 <- B0+B1*x1+B2*x2+B3*x3+e_1
  ybar1 <- mean(y1)
  mod1=lm(y1 ~ x1)
  summary(mod1)
  sse1 <- sum((y1-fitted(mod1))^2)
  p <- 1
  SBC_11 <- append(SBC_11, n*log(sse1)-n*log(n)+p*log(n))
}
```

A.3 Codice per generare simulazioni del criterio CP di Mallow

(Hebbali Aravind, 2017)

```
CP_1 <- c()
```

```
for (i in seq(1,10000)){
  n <- 15
  x1 <- rnorm(n)
  x2 <- rnorm(n)
  x3 <- rnorm(n)
  e_1 <- rnorm(n)
  B0 <- 0.2
  B1 <- 0.3
  B2 <- 0.5
  B3 <- 0.6
  y1 <- B0+B1*x1+B2*x2+B3*x3+e_1
  ybar1 <- mean(y1)
  mod0=lm(y1 ~ x1)
  summary(mod0)
  p <- 1
  sse0 <- sum((y1-fitted(mod0))^2)
  mse <- sse0/(n-p)
  CP_1 <- append(CP_1, 1/n * (sse0 + 2*p*mse))
}
```

A.4 Codice per generare simulazioni dell' R^2_{adj}

```
R2_ADJ_1 <- c()
for (i in seq(1,10000)){
  n <- 15
  x1 <- rnorm(n)
  x2 <- rnorm(n)
  x3 <- rnorm(n)
  e_1 <- rnorm(n)
```

```
  B0 <- 0.2
  B1 <- 0.3
  B2 <- 0.5
  B3 <- 0.6
  y1 <- B0+B1*x1+B2*x2+B3*x3+e_1
  ybar1 <- mean(y1)
  mod1=lm(y1 ~ x1)
  summary(mod1)
  sse1 <- sum((y1-fitted(mod1))^2)
  sst1 <- sum((y1-ybar1)^2)
  p <- 1
  R2_ADJ_1 <- append(R2_ADJ_1, 1-(sse1/(n-p)/(sst1/(n-1))))
}
```

A.5 Codice per generare simulazioni del criterio AIC

```
AIC_11 <- c()
AIC1 <- function(mod1){
  return (AIC(mod1))
}
for (i in seq(1,10000)){
  n <- 15
  x1 <- rnorm(n)
  x2 <- rnorm(n)
  x3 <- rnorm(n)
  e_1 <- rnorm(n)
  B0 <- 0.2
  B1 <- 0.3
  B2 <- 0.5
```

```
  B3 <- 0.6
  y1 <- B0+B1*x1+B2*x2+B3*x3+e_1
  ybar1 <- mean(y1)
  mod1=lm(y1 ~ x1)
  summary(mod1)
  AIC_11 <- append(AIC_11,AIC1(mod1))
}
```

Il codice precedente è replicato per generare simulazioni in cui i predittori sono correlati tra loro ed in cui il campione considerato è più grande.

Inoltre vengono considerati due casi : gli errori sono distribuiti secondo una T di Student con 4 gradi di libertà e quando sono distribuiti secondo una Skew Normal, in questo caso l'unica modifica al codice è la seguente:

```
#T di Student
  e_1 <- rt(n,4)
#Skew Normal
  e_1 <- rsn(n,alpha=5)
```

A.6 Codice per generare analisi per dati reali: dataset 1

```
rm(list=ls())
library(datasets)
dati <- datasets::airquality
dati <- na.omit(airquality)
plot(dati, cex = 0.75, col = "blue", las = 1)
cor(dati)
View(dati)
dim(dati)
attach(dati)
```

```
#MODELLO FULL
air.lm <- lm(Ozone ~ Solar.R + Wind + Temp + Month + Day, data=dati)
summary(air.lm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.11632   23.48249  -2.730  0.00742 **
Solar.R       0.05027    0.02342   2.147  0.03411 *
Wind         -3.31844    0.64451  -5.149 1.23e-06 ***
Temp          1.89579    0.27389   6.922 3.66e-10 ***
Month        -3.03996    1.51346  -2.009  0.04714 *
Day           0.27388    0.22967   1.192  0.23576

AIC(air.lm)
BIC(air.lm)
#R^2 PRESS
PRESS1 <- function(air.lm) {
  i <- residuals(air.lm)/(1 - lm.influence(air.lm)$hat)
  return(sum(i^2))
}
ozone_media <- mean(dati$Ozone)
differenze_quadrate <- (dati$Ozone - ozone_media)^2
SST <- sum(differenze_quadrate)
n <- 153
R2_PRESS <- 1-((PRESS1(air.lm))/((n/(n-1))^2*SST))
#CP DI MALLOW
library(olsrr)
```

```
full_model <- air.lm
ols_mallows_cp(air.lm, full_model)

#MODELLO CON UN PREDITTORE: TEMP
air.lm1 <- lm(Ozone ~  Temp, data=dati)
summary(air.lm1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -147.6461    18.7553  -7.872 2.76e-12 ***
Temp           2.4391     0.2393  10.192  < 2e-16 ***

AIC(air.lm1)
BIC(air.lm1)
#R^2 PRESS
PRESS1 <- function(air.lm1) {
  i <- residuals(air.lm1)/(1 - lm.influence(air.lm1)$hat)
  return(sum(i^2))
}
R2_PRESS <- 1-((PRESS1(air.lm1))/((n/(n-1))^2*SST))
#CP DI MALLOW
ols_mallows_cp(air.lm1, full_model)

#MODELLO CON DUE PREDITTORI: WIND E TEMP
air.lm2 <- lm(Ozone ~ Temp + Wind, data=dati)
summary(air.lm2)

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -67.3220     23.6210  -2.850  0.00524 **
Temp          1.8276      0.2506   7.294 5.29e-11 ***
Wind         -3.2948      0.6711  -4.909 3.26e-06 ***

AIC(air.lm2)
BIC(air.lm2)
#R^2 PRESS
PRESS1 <- function(air.lm2) {
  i <- residuals(air.lm2)/(1 - lm.influence(air.lm2)$hat)
  return(sum(i^2))
}
R2_PRESS <- 1-((PRESS1(air.lm2))/((n/(n-1))^2*SST))
#CP DI MALLOW
ols_mallows_cp(air.lm2, full_model)

#MODELLO CON 3 PREDITTORI: WIND, TEMP E SOLAR R.
air.lm3 <- lm(Ozone ~ Temp + Wind + Solar.R, data=dati)
summary(air.lm3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.34208   23.05472  -2.791  0.00623 **
Temp          1.65209    0.25353   6.516 2.42e-09 ***
Wind         -3.33359    0.65441  -5.094 1.52e-06 ***
Solar.R       0.05982    0.02319   2.580  0.01124 *

AIC(air.lm3)
```

```
BIC(air.lm3)
#R^2 PRESS
PRESS1 <- function(air.lm3) {
  i <- residuals(air.lm3)/(1 - lm.influence(air.lm3)$hat)
  return(sum(i^2))
}
R2_PRESS <- 1-((PRESS1(air.lm3))/((n/(n-1))^2*SST))
#CP DI MALLOW
ols_mallows_cp(air.lm3, full_model)

#MODELLO CON 4 PREDITTORI: WIND, TEMP, SOLAR R. E MONTH
air.lm4 <- lm(Ozone ~  Temp + Wind + Solar.R + Month, data=dati)
summary(air.lm4)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -58.05384   22.97114  -2.527   0.0130 *
Temp          1.87087    0.27363   6.837 5.34e-10 ***
Wind         -3.31651    0.64579  -5.136 1.29e-06 ***
Solar.R       0.04960    0.02346   2.114   0.0368 *
Month        -2.99163    1.51592  -1.973   0.0510 .

AIC(air.lm4)
BIC(air.lm4)
#R^2 PRESS
PRESS1 <- function(air.lm4) {
  i <- residuals(air.lm4)/(1 - lm.influence(air.lm4)$hat)
  return(sum(i^2))
```

```
}
R2_PRESS <- 1-((PRESS1(air.lm4))/((n/(n-1))^2*SST))
#CP DI MALLOW
ols_mallows_cp(air.lm4, full_model)
```

	ozone	solar r.	wind	temp	month	day
ozone	1.000000000	0.34834169	-0.61249658	0.6985414	0.142885168	-0.005189769
solar r	0.348341693	1.00000000	-0.12718345	0.2940876	-0.074066683	-0.057753801
wind	-0.612496576	-0.12718345	1.00000000	-0.4971897	-0.194495804	0.049871017
temp	0.698541410	0.29408764	-0.49718972	1.0000000	0.403971709	-0.096545800
month	0.142885168	-0.07406668	-0.19449580	0.4039717	1.000000000	-0.009001079
day	-0.005189769	-0.05775380	0.04987102	-0.0965458	-0.009001079	1.000000000

Tabella 14: matrice di correlazione tra le variabili presenti nel dataset Airquality

A.7 Codice per generare analisi per dati reali: dataset 2

```
rm(list=ls())
library(faraway)
dati <- faraway::gala
View(dati)
attach(dati)
plot(gala, cex = 0.75, col = "blue", las = 1)
cor(gala)

#MODELLO COMPLETO
gala.lm <- lm(Species ~ Endemics + Area + Elevation + Nearest + Scruz + Ad
summary(gala.lm)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -15.337942    9.423550  -1.628     0.117
Endemics      4.393654    0.481203   9.131 4.13e-09 ***
Area          0.013258    0.011403   1.163     0.257
Elevation    -0.047537    0.047596  -0.999     0.328
Nearest      -0.101460    0.500871  -0.203     0.841
Scruz         0.008256    0.105884   0.078     0.939
Adjacent      0.001811    0.011879   0.152     0.880
```

```
AIC(gala.lm)
BIC(gala.lm)
#R^2 PRESS
PRESS1 <- function(gala.lm) {
  i <- residuals(gala.lm)/(1 - lm.influence(gala.lm)$hat)
  return(sum(i^2))
}
specie_media <- mean(dati$Species)
differenze_quadrate <- (dati$Species - specie_media)^2
SST <- sum(differenze_quadrate)
n <- 30
R2_PRESS <- 1-((PRESS1(gala.lm))/((n/(n-1))^2*SST))
#CP DI MALLOW
library(olsrr)
full_model <- gala.lm
ols_mallows_cp(gala.lm, full_model)

#UN PREDITTORE : ENDEMICS
gala.lm1 <- lm(Species~Endemics, data=dati)
summary(gala.lm1)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.0480     7.1138  -2.959  0.00622 **
Endemics      4.0721     0.1899  21.443  < 2e-16 ***

AIC(gala.lm1)
BIC(gala.lm1)
#R^2 PRESS
PRESS1 <- function(gala.lm1) {
  i <- residuals(gala.lm1)/(1 - lm.influence(gala.lm1)$hat)
  return(sum(i^2))
}
specie_media <- mean(dati$Species)
differenze_quadrate <- (dati$Species - specie_media)^2
SST <- sum(differenze_quadrate)
n <- 30
R2_PRESS <- 1-((PRESS1(gala.lm1))/((n/(n-1))^2*SST))
#CP DI MALLOW
library(olsrr)
full_model <- gala.lm
ols_mallows_cp(gala.lm1, full_model)

#DUE PREDITTORI:ENDEMICS E ELEVATION
gala.lm2 <- lm(Species ~ Endemics + Elevation, data=dati)
summary(gala.lm2)

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.92862    7.14320  -2.790  0.00955 **
Endemics      4.35265    0.30997  14.042 6.29e-14 ***
Elevation    -0.02294    0.02009  -1.142  0.26366
```

```
AIC(gala.lm2)
BIC(gala.lm2)
#R^2 PRESS
PRESS1 <- function(gala.lm2) {
  i <- residuals(gala.lm2)/(1 - lm.influence(gala.lm2)$hat)
  return(sum(i^2))
}
specie_media <- mean(dati$Species)
differenze_quadrate <- (dati$Species - specie_media)^2
SST <- sum(differenze_quadrate)
n <- 30
R2_PRESS <- 1-((PRESS1(gala.lm2))/((n/(n-1))^2*SST))
#CP DI MALLOW
library(olsrr)
full_model <- gala.lm
ols_mallows_cp(gala.lm2, full_model)

#TRE PREDITTORI : ENDEMICS, ELEVATION E AREA
gala.lm3 <- lm(Species~ Endemics + Elevation + Area, data=dati)
summary(gala.lm3)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -15.891237    7.569210   -2.099    0.0456 *
Endemics      4.331791    0.304686   14.217 8.97e-14 ***
Elevation    -0.041439    0.023653   -1.752    0.0916 .
Area          0.012669    0.008936    1.418    0.1681
```

```
AIC(gala.lm3)
BIC(gala.lm3)
#R^2 PRESS
PRESS1 <- function(gala.lm3) {
  i <- residuals(gala.lm3)/(1 - lm.influence(gala.lm3)$hat)
  return(sum(i^2))
}
specie_media <- mean(dati$Species)
differenze_quadrate <- (dati$Species - specie_media)^2
SST <- sum(differenze_quadrate)
n <- 30
R2_PRESS <- 1-((PRESS1(gala.lm3))/((n/(n-1))^2*SST))
#CP DI MALLOW
library(olsrr)
full_model <- gala.lm
ols_mallows_cp(gala.lm3, full_model)
```

```
#QUATTRO PREDITTORI : ENDEMICS, ELEVATION, AREA E ADJACENT
gala.lm4 <- lm(Species~Endemics+Elevation+Area+Adjacent, data=dati)
summary(gala.lm4)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -15.683350    7.770240   -2.018    0.0544 .
Endemics      4.399807    0.439651   10.007 3.16e-10 ***
Elevation    -0.049342    0.043468   -1.135    0.2671
Area          0.013817    0.010511    1.315    0.2006
Adjacent      0.002394    0.010960    0.218    0.8288

AIC(gala.lm4)
BIC(gala.lm4)
#R^2 PRESS
PRESS1 <- function(gala.lm4) {
  i <- residuals(gala.lm4)/(1 - lm.influence(gala.lm4)$hat)
  return(sum(i^2))
}
specie_media <- mean(dati$Species)
differenze_quadrate <- (dati$Species - specie_media)^2
SST <- sum(differenze_quadrate)
n <- 30
R2_PRESS <- 1-((PRESS1(gala.lm4))/((n/(n-1))^2*SST))
#CP DI MALLOW
library(olsrr)
full_model <- gala.lm
ols_mallows_cp(gala.lm4, full_model)

#CINQUE PREDITTORI : ENDEMICS, ELEVATION, AREA, ADJACENT E NEAREST
gala.lm5 <- lm(Species ~ Endemics + Elevation + Area + Adjacent + Nearest,
summary(gala.lm5)

Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.056413   8.522230  -1.767    0.090 .
Endemics      4.383999   0.455268   9.629 1.02e-09 ***
Elevation    -0.046921   0.045953  -1.021    0.317
Area          0.013207   0.011146   1.185    0.248
Adjacent      0.001763   0.011615   0.152    0.881
Nearest      -0.077612   0.388341  -0.200    0.843
```

```
AIC(gala.lm5)
BIC(gala.lm5)
#R^2 PRESS
PRESS1 <- function(gala.lm5) {
  i <- residuals(gala.lm5)/(1 - lm.influence(gala.lm5)$hat)
  return(sum(i^2))
}
specie_media <- mean(dati$Species)
differenze_quadrate <- (dati$Species - specie_media)^2
SST <- sum(differenze_quadrate)
n <- 30
R2_PRESS <- 1-((PRESS1(gala.lm5))/((n/(n-1))^2*SST))
#CP DI MALLOW
library(olsrr)
full_model <- gala.lm
ols_mallows_cp(gala.lm5, full_model)
```

	species	endemics	area	elevation	nearest	scruz	adjacent
species	1.0000	0.9708	0.6178	0.7384	-0.0140	-0.1711	0.0261
endemics	0.9708	1.0000	0.6169	0.7929	0.0059	-0.1542	0.0826
area	0.6178	0.6169	1.0000	0.7537	-0.1111	-0.1007	0.1800
elevation	0.7384	0.7929	0.7537	1.0000	-0.0110	-0.0154	0.5364
nearest	-0.0140	0.0059	-0.1111	-0.0110	1.0000	0.6154	-0.1162
scruz	-0.1711	-0.1542	-0.1007	-0.0154	0.6154	1.0000	0.0516
adjacent	0.0261	0.0826	0.1800	0.5364	-0.1162	0.0516	1.0000

Tabella 15: matrice di correlazione tra le variabili presenti nel dataset Gala

Ringraziamenti

Vorrei dedicare questo spazio finale della mia tesi ai ringraziamenti verso tutti coloro che hanno contribuito con il loro supporto alla realizzazione della stessa.

In primo luogo desidero ringraziare la professoressa Luisa Bisaglia per la disponibilità dimostrata durante tutte le analisi svolte e la stesura del lavoro.

Dedico questa tesi alle persone più importanti della mia vita, coloro che con sacrificio mi hanno permesso di arrivare alla fine di questo percorso, coloro che fin dall'inizio hanno sempre creduto in me supportandomi in ogni mia scelta, coloro ai quali devo tutte le mie più grandi vittorie, ai miei genitori.

Ai miei nonni per avermi sempre trasmesso il valore e l'importanza della famiglia.

A mio fratello Edoardo, il regalo più grande che abbia mai ricevuto dai miei genitori.

Ad Andrea che ha da sempre creduto in me. A te dico grazie per avermi insegnato l'arte di amare come solo un vero uomo è in grado di fare.

Alla mia coinquilina e ai coinquilini "acquisiti" di Camplus, i migliori con cui poter vivere questi 3 bellissimi anni. La vostra allegria ha reso tutto più bello.

Ai ragazzi del mio corso di laurea, eccellenze naturali. Sono grata e fortunata di avervi incontrato, è in parte grazie a voi che sono arrivata fin qui.

A Padova, al suo caos del mercoledì e alla sua tranquillità nei giorni di festa, la città più bella e viva in cui potessi capitare.

A me stessa, per aver dimosrato una tenacia che non credevo di avere e per non aver mai rinunciato a nulla.

Sara De Tommaso

Bibliografia

Alcantara I. M., Naranjo Joshua, and Lang Yanda. Model selection using press statistic. *Computational Statistics (original paper)*, 38:285–289, 2022.

Allen D. M. Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3):469–475, 1971.

Brewer, M. J., Butler A., Cooksley, and S. L. Performance of aic and bic in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, 7(6):679–692, 2016.

Burnham Kenneth P. and Anderson David R. Understanding aic and bic in model selection. *Colorado Cooperative Fish and Wildlife Research Unit*, pages 261–304, 2004.

Burnham K. P. and Andersib D. R. A pratical information-theoretic approach. *Model selection and multimodel inference*, 2, 2002.

Chen S., Hong X., Harris C. J., and Sharkey P. M. Sparse modeling using orthogonal forward regression with press statistic and regularization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 34(2):898–911, 2004.

Columbu Silvia. Regressione multipla: selezione dei modelli. *modelli statistici*, pages 1–12, 2020.

Friedman J., Hastie T., and Tibshirani R. The elements of statistical learning. *New Yotk: Springer series in statistics*, 1(10), 2001.

Gilmour and S. G. The interpretation of mallow's cp-statistic. *Journal of the Royal Statistical Society: Series D*, 45(1):49–56, 1996.

Hebbali Aravind. olsrr : Tools for teaching and learning ols regression. *R package version*, 3, 2017.

Information criteria and press. *regression methods*, 10:285–289, 2022.

Kumar Ajitesh. R-squared adjusted r-squared: Differences. *Data Analytics*, 2015. URL `https://vitalflux.com/r-squared-adjusted-r-squared-differences-examples/`.

Kurter M. H., Nachtsheim C. J., Neter J., and Li W. Applied linear statistical models. *McGraw-Hill Irwin*, 5, 2005.

Landram F. G., A. Abdullat, and Shah V. The coefficient of prediction for model specification. *South-west Economic Rev*, 32:149–156, 2008.

McQuerrie A. D. and Tsai C. L. Regression and time series model in small sample. *World Scientific, Singapore*, 76:297–307, 1998.

Mediavilla F., Landram F., and Shah V. A comparison of the coefficient of predictive power, the coefficient of determination and aic for linear regression. *Journal of Applied Business and Economics*, 8(4):44, 2008.

Murtaugh Paul. Methods of variable selection in regression modeling. *Communications in Statistics - Simulation and Computation*, 27(3):711–734, 1998.

Searle S. R. and Gruber M. H. Linear models. *Wiley Series in Probability and Statistics*, 2016.

Shalabh. Regression analysis: Model adequacy checking. 4. URL `https://home.iitk.ac.in/shalab/regression/Chapter4-Regression-ModelAdequacyChecking.pdf`.

Tamhane A. and Dunlop D. Statistics and data analysis: from elementary to intermediate. *Prentice Hall, New Jersey*, 2000.

Walter A. Shewhart and Samuel S. Wilks. Applied linear regression-fourth edition. *Sanford Weisberg*, 10:235–245, 2014.