Università degli Studi di Salerno

Dipartimento di Ingegneria dell'Informazione ed Elettrica e Matematica Applicata

Corso di Laurea Magistrale in Ingegneria Informatica

Data Science 30/01/2025
Canale A-H

Project Work

Quesito 1 – Regressione lineare

Gruppo n. **07 – AH**

Cognome e Nome	Matricola	e-mail
Apicella Antonio	0622702531	a.apicella97@studenti.unisa.it
Celano Benedetta Pia	0622702558	b.celano1@studenti.unisa.it
Cirillo Francesco Pio	0622702466	f.cirillo36@studenti.unisa.it
Fasolino Alessandra	0622702465	a.fasolino35@studenti.unisa.it

Sommario

1. Valutazioni preliminari sul modello

Al fine di osservare graficamente la relazione che sussiste tra tutti i regressori e il singolo predittore Y, si è deciso per prima cosa di produrre la scatterplot matrix (Figura 1). A causa dell'ingente numero di predittori (p=25), è subito emersa la necessità di procedere con uno studio analitico più approfondito.

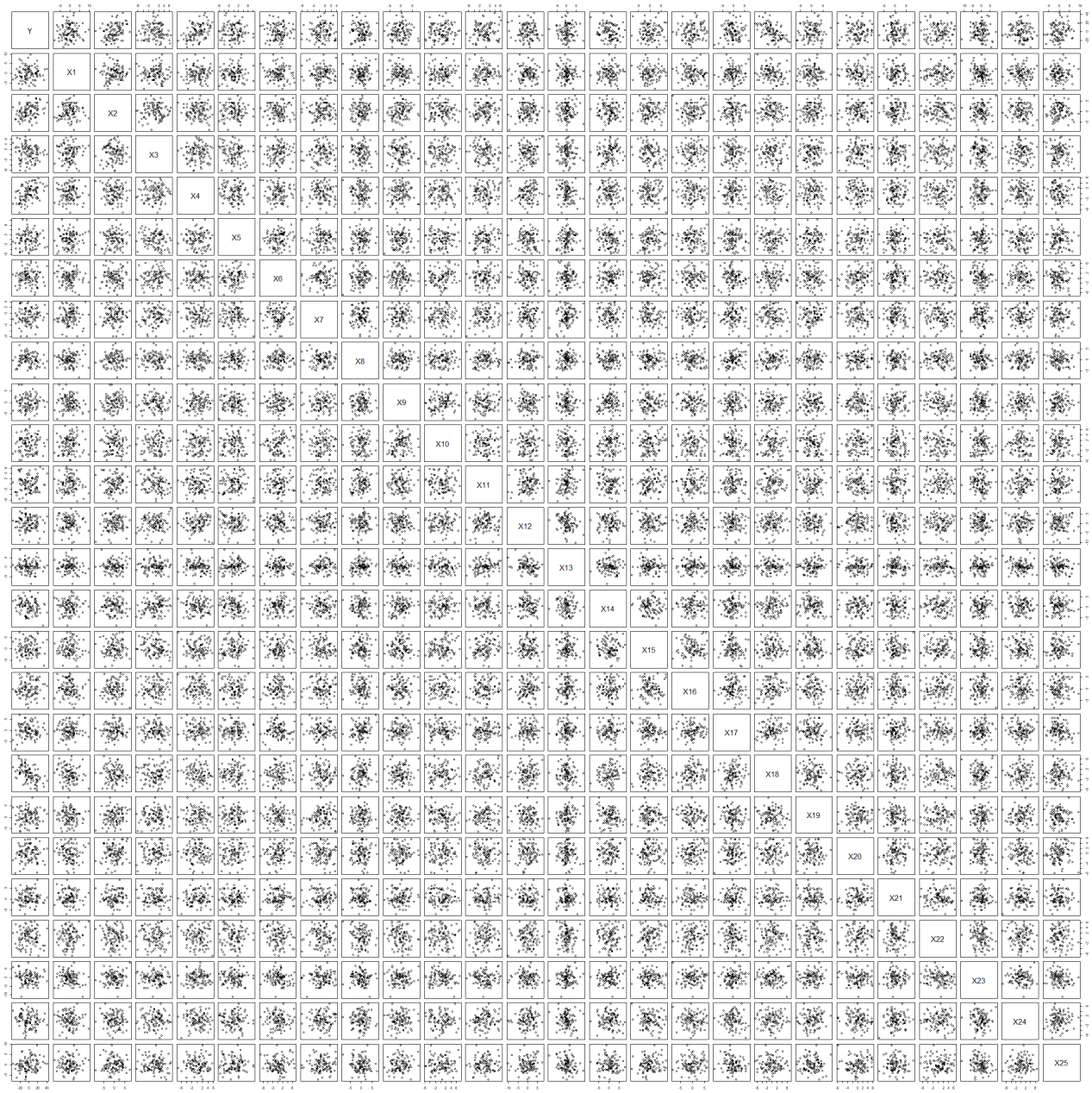

Figura 1: Scatterplot Matrix del modello

1.1. Studio della correlazione tra i regressori

Il grafico in Figura 3 mostra, mediante delle ellissi, la correlazione tra i vari regressori: le variabili più correlate presentano un colore più intenso; in particolare il blu è usato per la correlazione positiva, il rosso per quella negativa. Alternativamente è possibile vedere la direzione delle ellissi: se inclinate verso destra la correlazione è positiva, altrimenti negativa. Lo studio della correlazione è utile per valutare la relazione tra le variabili, a supporto di decisioni per un'eventuale model selection.

La **correlazione media calcolata** è pari a -0.004215385: trattandosi di un valore molto basso si può dedurre che le variabili sono indipendenti tra loro.
La **correlazione massima** presenta un coefficiente pari a 0.54 in valore assoluto per la coppia Y-X18. Dalla matrice di correlazione si evince che si tratta di una correlazione negativa.
Il fatto che la correlazione maggiore sia tra la risposta e il regressore X18 è sintomo di quanto questo predittore possa risultare rilevante nella stima di Y (tale ipotesi sarà in seguito valutata con il confronto tra le strategie).

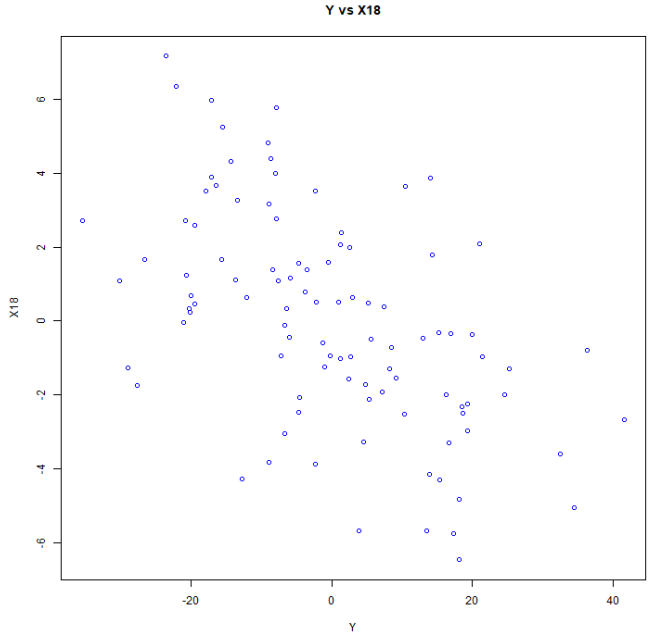

Figura 2: Scatterplot della coppia più correlata

Dallo scatter plot in Figura 2, che mostra le distribuzioni della coppia maggiormente correlata, si evince come i punti seguano un pattern ben definito.
La **correlazione con un coefficiente superiore in valore assoluto a 0.5** è soltanto quella sopra citata (Y-X18) a prova che i regressori risultano debolmente correlati tra loro.

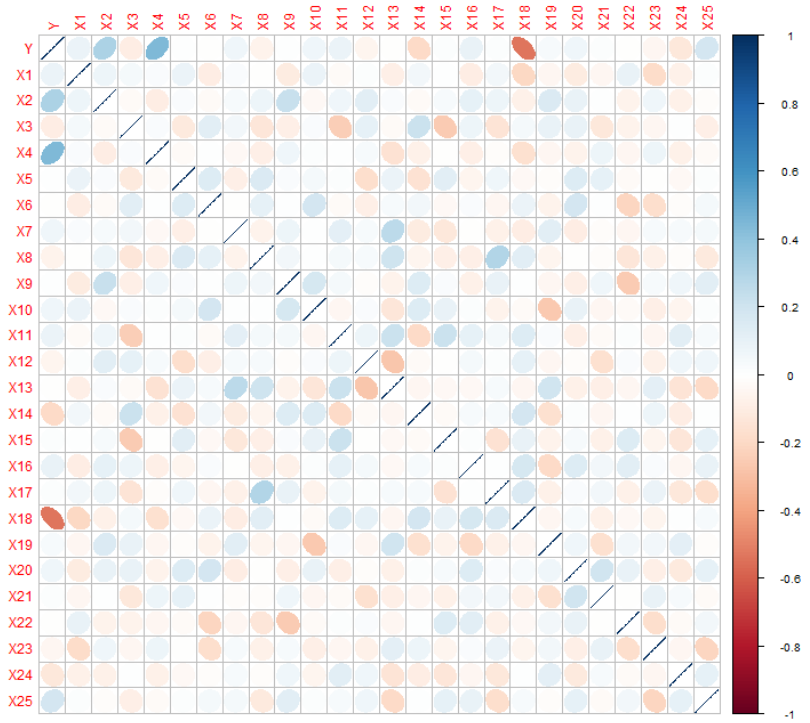

Figura 3: Matrice di correlazione

1.2. Studio della multicollinearità

Per valutare se sussistesse correlazione tra più regressori si è deciso di utilizzare il VIF. Il valore massimo ottenuto è 1.636007 per il regressore X13. Si tratta di un valore molto basso e ciò dimostra che la correlazione con gli altri regressori è estremamente ridotta e dunque non preoccupante. Ciò è confermato dalla VIF media che risulta: 1.303551.

2. Requisito 1: determinare i modelli che minimizzano l'MSE con diverse strategie

2.1. Best Subsets Selection basata su BIC

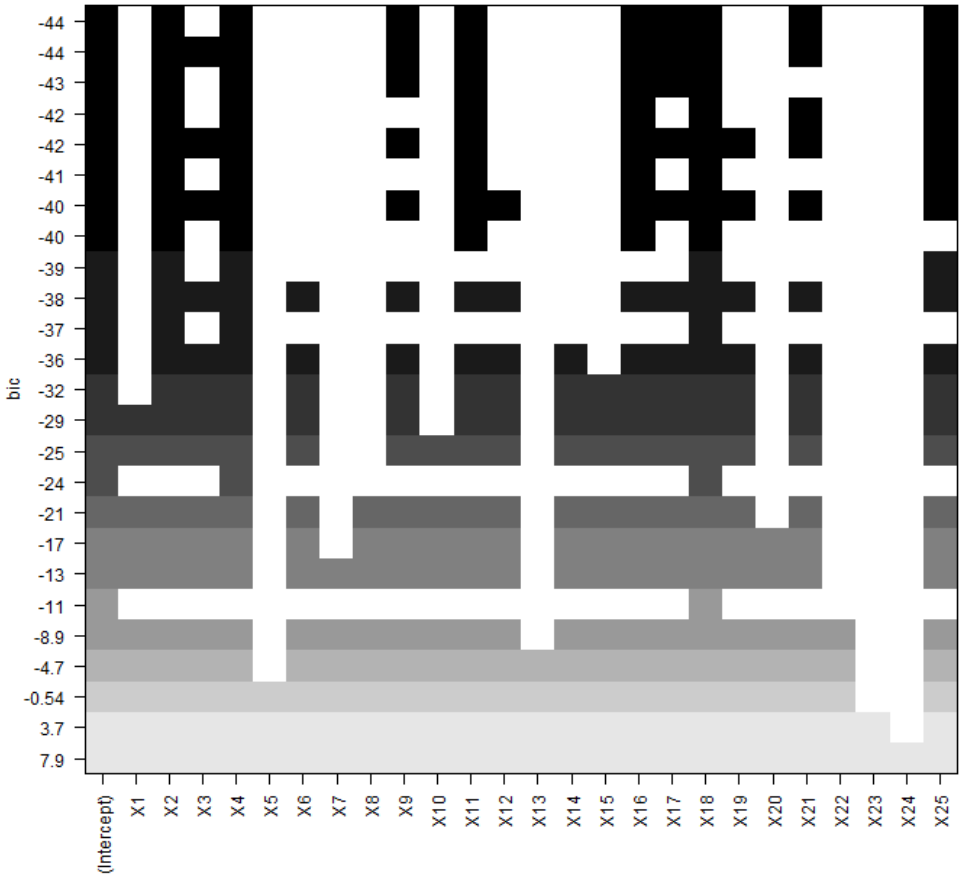

Figura 4: Plot coefficienti con BIC

Tramite il plot in Figura 4 è possibile osservare quali sono i regressori coinvolti nel modello a minimo BIC utilizzando Best Subset Selection. Il modello migliore è quello con 9 regressori ed include le seguenti variabili: X2, X4, X9, X11, X16, X17, X18, X21, X25. Nell'immagine ciò è mostrato nella parte alta del grafico, dove BIC assume valore minimo.
I coefficienti stimati ottenuti per ciascuno dei 9 regressori risultano:

Intercept	**X2**	**X4**	**X9**	**X11**
-0.6992200	1.9209904	2.3095899	-0.8919921	1.0367856
X16	**X17**	**X18**	**X21**	**X25**
1.0157244	0.8534967	-2.7230995	-0.9329403	1.1041542

2.2. Backward stepwise con 5-fold cross validation

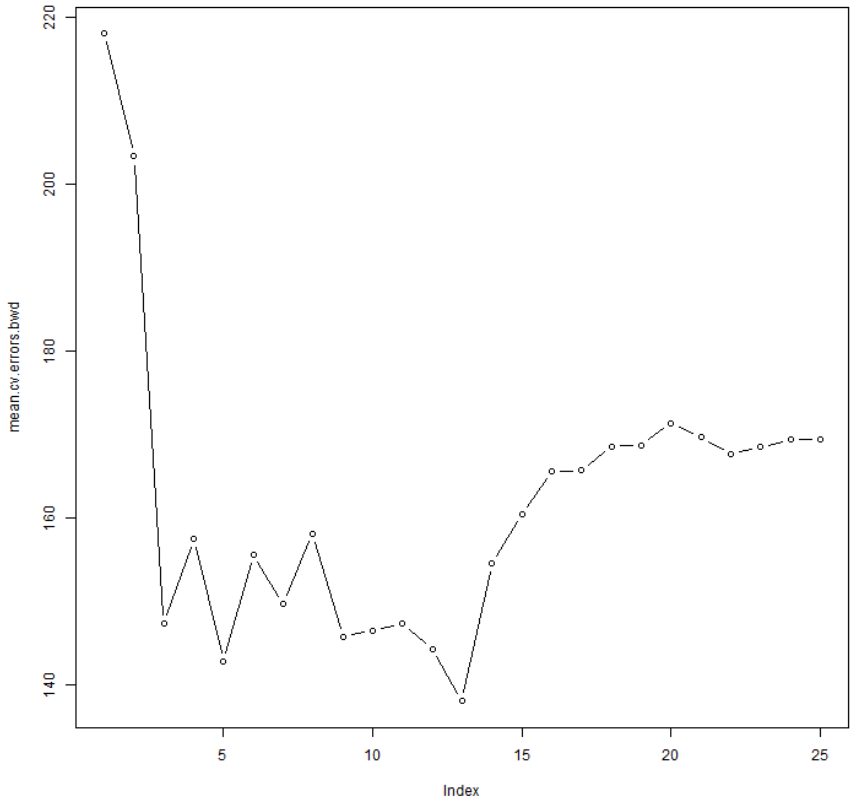

Figura 5: Confronto dimensione del modello ed MSE medio stimato con 5-fold cv

Il grafico in Figura 5 plotta l'estimated test error rate calcolato tramite 5-fold cross validation per ciascun numero di regressori. Dal grafico si evince che il modello migliore, e dunque con l'errore quadratico medio minimo, lo si ottiene con 13 regressori - il numero di variabili indipendenti coinvolte è visibile sull'asse delle ascisse.
Le stime dei coefficienti ottenuti sono presentate di seguito in formato tabellare:

Intercept	0.2065474	**X12**	-0.5579242
X2	2.0773848	**X16**	0.9990024
X3	-0.6224502	**X17**	0.8591479
X4	2.5109777	**X18**	-2.5099261
X6	0.4570145	**X19**	-0.6784999
X9	-1.0724819	**X21**	-1.1006516
X11	0.9601927	**X25**	0.9509704

2.3. Ridge Regression

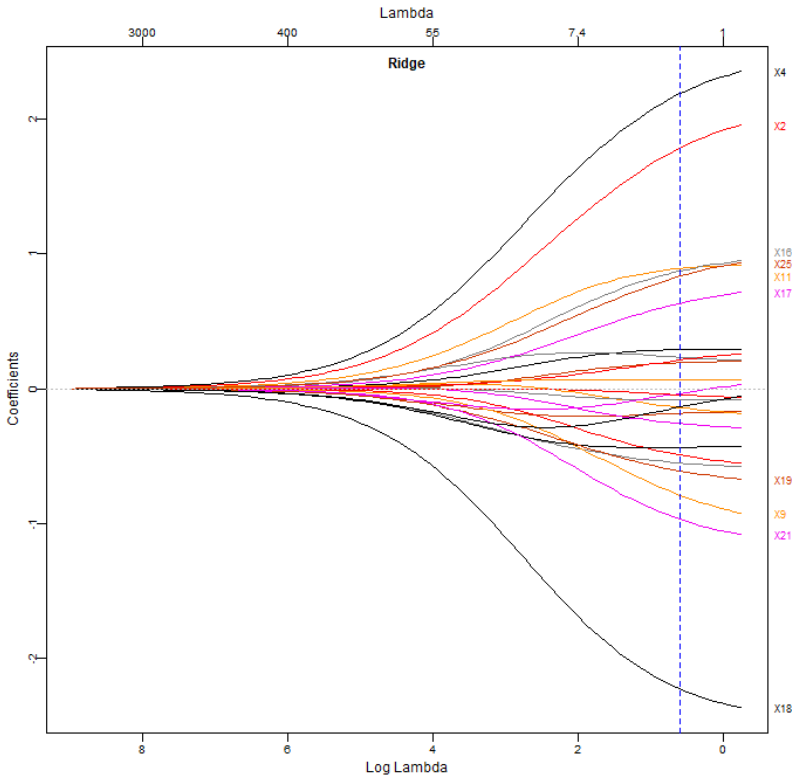

Figura 6: Variazione dei coefficienti in funzione di Lambda Ridge Regression

Nel grafico in Figura 6 ogni curva mostra l'andamento delle stime dei coefficienti per ciascuno dei 25 regressori, ottenute mediante la ridge regression. Si nota, come previsto teoricamente, che per valori di lambda più elevati - ed in misura più moderata del log del parametro di tuning, mostrato sull'asse orizzontale in basso - le stime dei coefficienti tendono a restringersi verso lo zero. Questo comportamento è coerente con quanto atteso, ridge regression non azzera mai completamente le stime dei coefficienti seppur riducendole significativamente.
Segue in formato tabellare l'elenco dei coefficienti stimati ottenuti per ciascuno dei regressori, scegliendo il valore di lambda che minimizza l'MSE di cross validation:

Intercept	0.09426679	**X13**	-0.03437904
X1	-0.13906773	**X14**	-0.43431712
X2	1.78666198	**X15**	-0.25892990
X3	-0.54993324	**X16**	0.87289619
X4	2.18713673	**X17**	0.63396137
X5	-0.08474565	**X18**	-2.22462763
X6	0.29253229	**X19**	-0.60986116
X7	-0.18260208	**X20**	0.23478775
X8	0.19361284	**X21**	-0.96757514
X9	-0.78797543	**X22**	-0.04415379
X10	0.21162096	**X23**	0.06818692
X11	0.89242508	**X24**	-0.13491788
X12	-0.49147085	**X25**	0.83522125

2.4. Lasso Regression

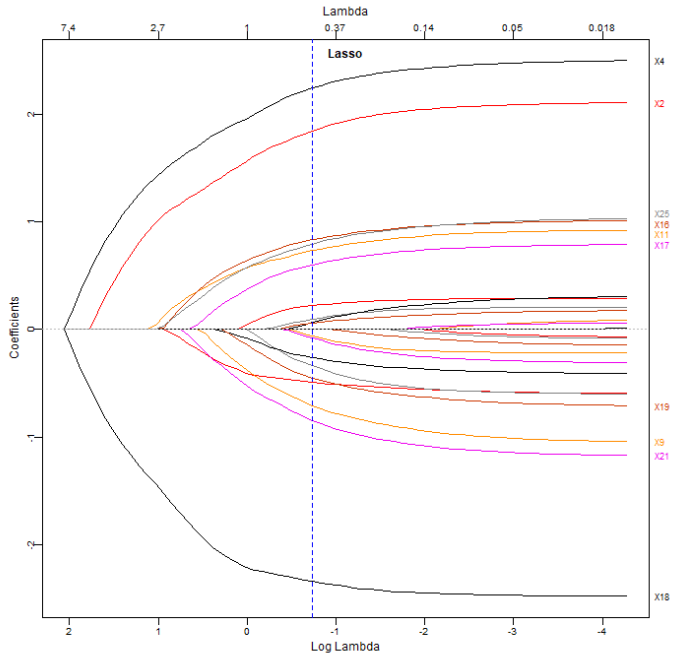

Figura 7: Variazione dei coefficienti in funzione di lambda Lasso Regression

Nel grafico in Figura 7 ogni curva mostra l'andamento delle stime dei coefficienti per ciascuno dei 25 regressori, ottenute mediante la regressione con Lasso. Si nota, come previsto teoricamente, che per valori di lambda più elevati - ed in misura più moderata del log del parametro di tuning, mostrato sull'asse orizzontale in basso - le stime dei coefficienti tendono a restringersi verso lo zero. Contrariamente alla norma euclidea, la l1 norm utilizzata da Lasso, per valori significativamente grandi di lambda, può comportare il setting di alcuni o di tutti i coefficienti direttamente a zero, effettuando di conseguenza variable selection.
Nel caso in esame, infatti, alcuni predittori sono stati scartati, comportando la restituzione da parte di Lasso di un modello sparso ossia con un sottinsieme delle variabili di partenza. I regressori scartati risultano: X5, X7, X13, X22, X23, X24.
Segue in formato tabellare l'elenco dei coefficienti stimati ottenuti per ciascuno dei regressori non nulli, scegliendo il valore di lambda che minimizza l'MSE di cross validation:

Intercept	-0.10069984	**X12**	-0.33951371
X1	-0.06196800	**X14**	-0.26137981
X2	1.84755630	**X15**	-0.08142343
X3	-0.49272583	**X16**	0.83243264
X4	2.24381639	**X17**	0.60162175
X6	0.21989513	**X18**	-2.34202280
X8	0.09442713	**X19**	-0.44984333
X9	-0.70709259	**X20**	0.05740034
X10	0.06381476	**X21**	-0.84669901
X11	0.73137052	**X25**	0.79468358

3. Requisito 2: valutare l'MSE di test e selezionare la strategia migliore

La strategia che minimizza l'MSE di test è Best Subsets Selection basata su BIC. Il risultato ottenuto era prevedibile in quanto BSS effettua la ricerca esaustiva: esso infatti valuta tutte le combinazioni dei predittori per ottenere i modelli (2^p) rispetto alla backward stepwise che valuta solo p(p+1)/2 modelli e dunque non assicura di trovare il modello migliore. Inoltre il fatto che la Ridge regression non performi in modo ottimale mostra come la risposta non sia influenzata da tutti i regressori, Ridge performa bene infatti in situazioni nelle quali quasi tutti i regressori sono utili alla previsione della response.

La matrice dei 4 plot in Figura 8 mostra l'andamento di RSS, CP(AIC), BIC ed adjusted R^2 rispetto al numero di regressori con la tecnica BSS. In ciascun grafico, il punto evidenziato in rosso mostra il numero di regressori nel modello che garantisce - procedendo dall'alto verso il basso e da sinistra a destra - il minimo valore di RSS, CP (equivalente ad AIC poiché gli errori sono gaussiane e di conseguenza ordinary least squares coincide con maximum likelihood) e BIC ed il massimo valore di adjusted R^2. Diverse metriche avrebbero condotto alla scelta di modelli diversi, ad esempio Cp avrebbe previsto la scelta di un modello a 10 variabili, tuttavia è stato richiesto di utilizzare BIC che, come è possibile riscontrare dai grafici, ne sceglie 9. In ultima analisi si evidenzia il fatto che RSS sceglie il modello con tutti i regressori, è infatti noto che RSS non è una metrica adeguata al confronto tra modelli con dimensioni diverse infatti sceglie sempre il modello con la dimensione massima disponibile.

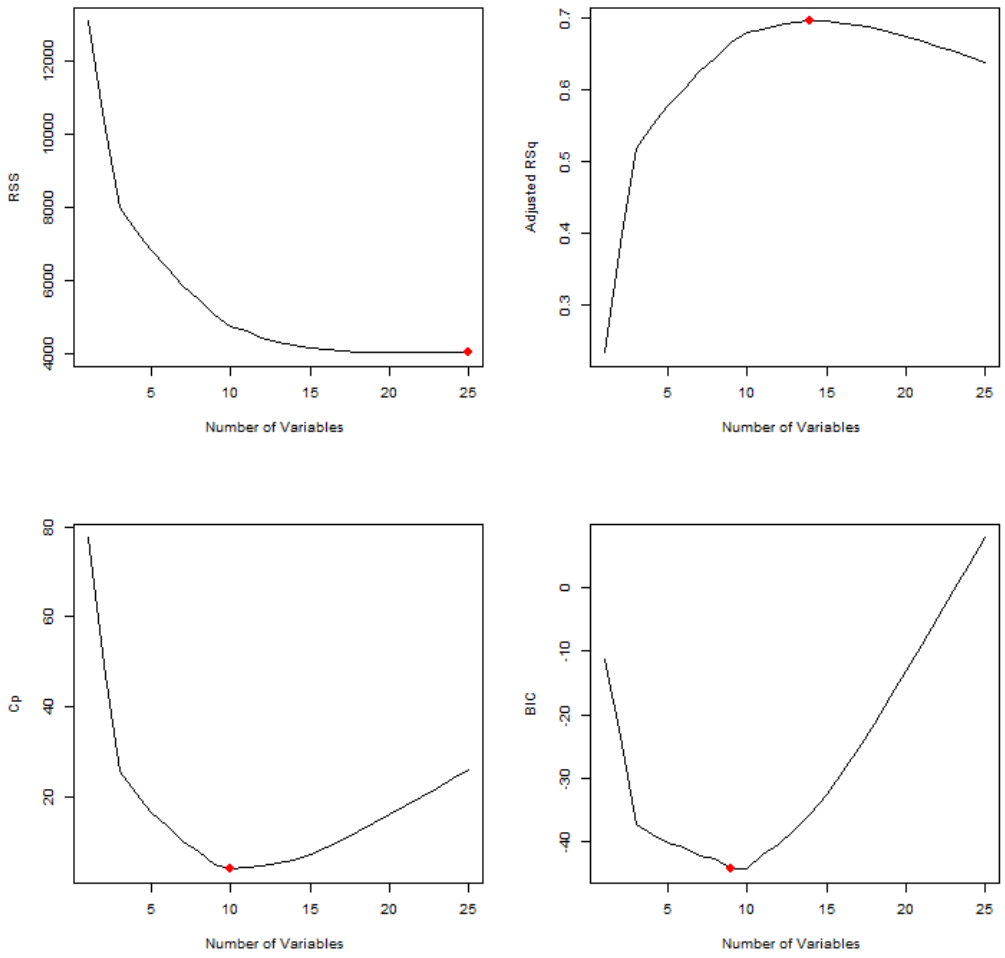

Figura 8: Andamento di RSS, CP, BIC ed AdjR2 per BSS

Indice delle figure