Consegna Data Analysis

Antonio Sessa Angelo Molinario Pietro Martano
Massimiliano Ranauro

30 January 2025

Università degli Studi di Salerno

Name	**Surname**	**Serial**	**E-mail**
Antonio	Sessa	0622702305	a.sessa108@studenti.unisa.it
Angelo	Molinario	0622702311	a.molinario3@studenti.unisa.it
Massimiliano	Ranauro	0622702373	m.ranauro2@studenti.unisa.it
Pietro	Martano	0622702402	p.martano@studenti.unisa.it

Contents

1 Regressione

1.1 Divisione del Dataset

Le seguenti osservazioni sono effettuate sul dataset fornito *RegressionDSDA250130.csv*, il quale, come richiesto dalla traccia è stato diviso in un Train set contenente il 70% dei campioni e in un Test set contenente il restante 30%.

1.2 Osservazioni sul Dataset

Nella Figura 1 sono visualizzate le collinearità tra regressori, notiamo come non ci sia correlazioni evidenti tra i dati. Anche la verifica del VIF non ha fatto emergere potenziali problemi infatti tutti i valori sono inferiori a '2'
L'unica correlazione presente è tra la variabili dipendente *Y* e i regressori *X2*, *X4* e *X18*, che ci aspettiamo saranno significativi per la regressione lineare.
Le considerazioni seguenti sono fatte su un caso particolare di divisione del train set e del test set.

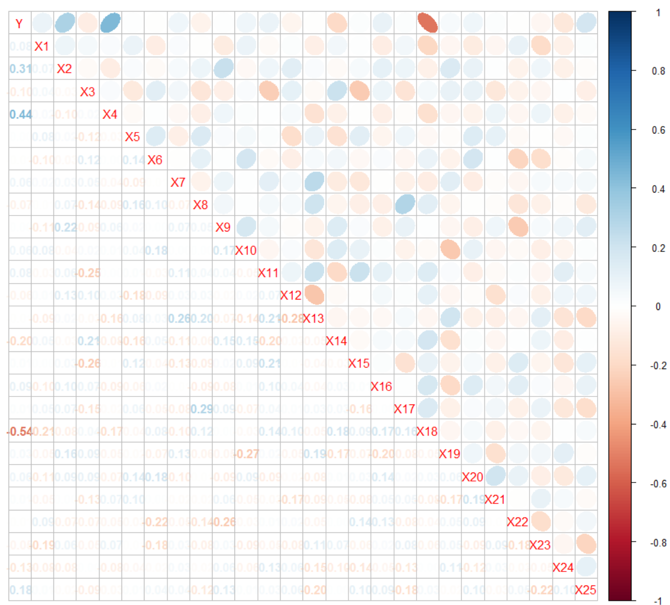

Figure 1: Collinearità

1.3 Best Subset Selection

Data l'analisi della collinearità ci aspettiamo che i regressori che sicuramente saranno presenti all'interno di subset selection siano quelli sopracitati. In verità dato che sono gli unici regressori che hanno una collinearità con Y apprezzabile rispetto agli altri che sicuramente rientraranno nel sub-set ottimo.
In figura 2 si può notare il grafico dell'errore relativo alla metrica BIC. Si può notare che l'errore minimo è quello avente 5 regressori; tuttavia, come ci aspettavamo, anche l'errore con solamente 3 regressori è comunque basso.
I valori dei coefficienti rispetto ai regressori sono:

```
(Intercept) -0.1547252
X2           1.5508302
X4           2.2984914
X16          0.9840383
X18         -2.5131398
X25          1.2350860
```

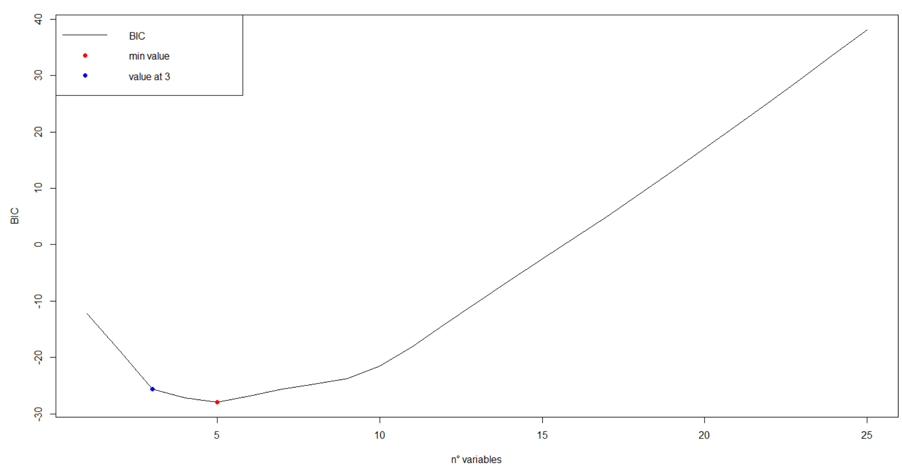

Figure 2: Best subset

In linea generale l'errore tende ad essere vicino o uguale al modello con tre regressori. A seconda della divisione del train set e del test set i risultati possono variare leggermente.

1.4 Stepwise Backward Selection

Utilizzando lo Stepwise Backward risulta che l'errore minimo è quello con 5 regressori. Analizzando i valori dei coefficienti possiamo notare che in questo caso particolare lo stepwise backward ha esattamente gli stessi valori di subset selection, infatti:

```
(Intercept) -0.1547252
X2           1.5508302
X4           2.2984914
X16          0.9840383
X18         -2.5131398
X25          1.2350860
```

Anche in questo caso l'errore con solamente tre regressori è comunque basso, come ci aspettavamo.

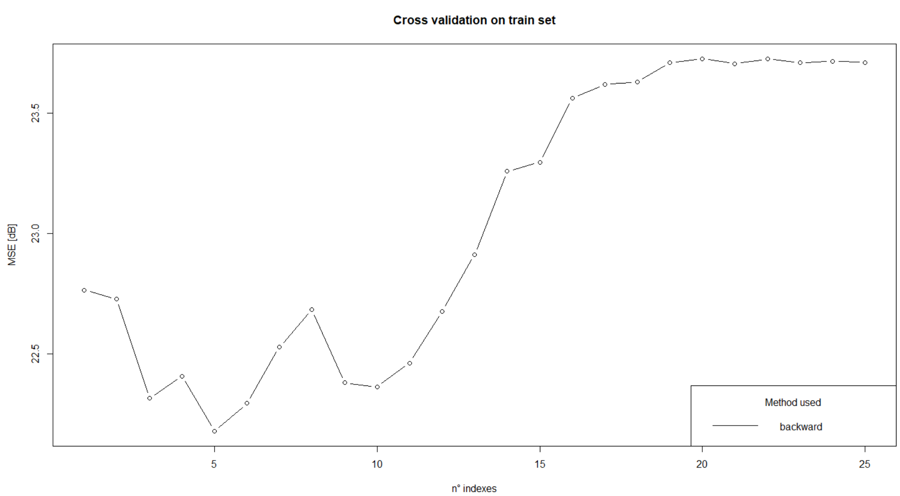

Figure 3: Stepwise backward

In linea generale l'errore tende ad essere vicino o uguale al modello con tre regressori. Come nel caso di subset selection, a seconda della divisione del train set e del test set i risultati possono variare leggermente. In questo caso particolare risulta che i due metodi restituiscano gli stessi risultati ma non è detto che ciò sia vero.

1.5 Ridge

Dall'analisi del metodo ridge di shrinkage possiamo notare come al variare del valore di lambda varia l'errore. Quello che scegliamo come errore è quello minimo, ovvero il primo valore che vediamo nella Figura 4.
In questo caso oltre ai valori dei coefficienti può essere utile anche specificare il valore di lambda utilizzato. In questo caso con λ pari a 9.754436 i coefficienti relativi ai regressori sono:

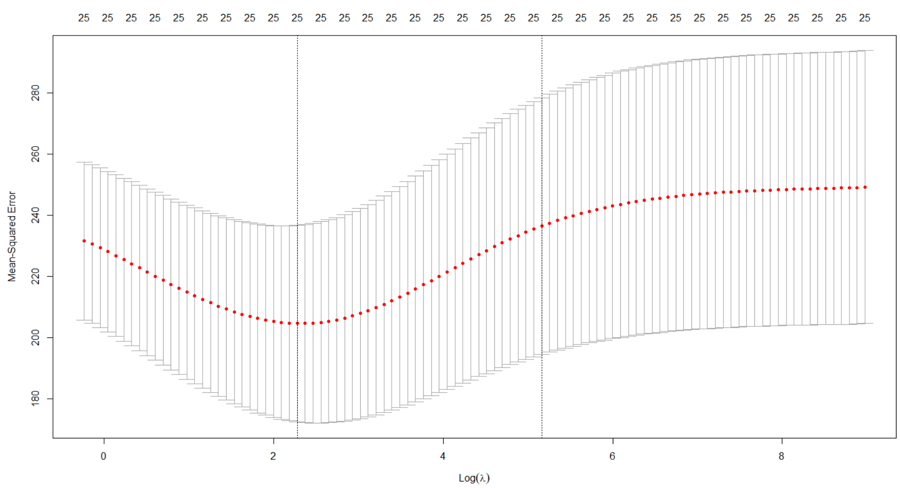

Figure 4: Ridge

(Intercept)	0.76864106	X18	-1.55410098
X1	-0.01855003	X19	0.01298946
X2	0.87230738	X20	0.11508614
X3	-0.22021159	X21	-0.07765226
X4	1.30771732	X22	-0.25023266
X5	-0.06347191	X23	-0.15492767
X6	0.03092281	X24	-0.25004111
X7	0.29665479	X25	0.59902942
X8	-0.21557158		
X9	-0.18071023		
X10	0.28223299		
X11	0.41395517		
X12	-0.08595923		
X13	0.05002504		
X14	-0.45551172		
X15	-0.22348573		
X16	0.47438346		
X17	0.21576803		

1.6 Lasso

In merito al metodo Lasso, in Figura 5 possiamo osservare i due riferimenti di lambda per lambda minimo e per lambda one-standard-error.
Possiamo notare, come possibile aspettarsi in seguito alle considerazioni precedenti, che il valore di lambda one-standard-error (linea tratteggiata destra) corrisponda all'utilizzo dei 3 coefficienti più significativi.
A seconda della divisione del train set e del test set i risultati possono variare leggermente. Quello che possiamo notare è che l'errore tende ad essere inferiore quando i regressori sono intorno al valore tre, anche se non sempre risulta tale.
Il valore di lambda scelto è quello minimo perché, come richiede la traccia, dobbiamo trovare i coefficienti che minimizzano l'errore MSE. Il valore di λ scelto in questo caso è 1.482588 con 7 regressori. I coefficienti associati sono:

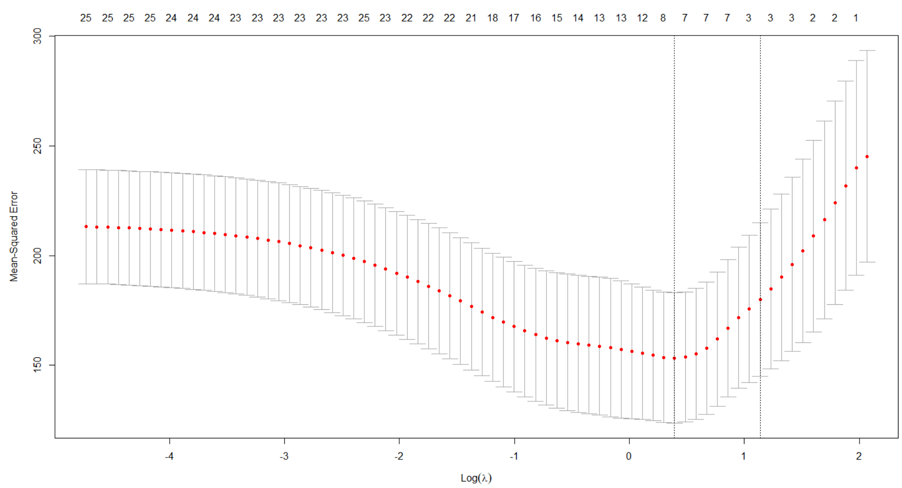

Figure 5: Lasso

```
(Intercept)  0.0009309972
X2           0.9931516964
X4           1.6117032132
X11          0.3287580447
X14         -0.1196036237
X16          0.3964145749
X18         -2.0010059551
X25          0.5498780684
```

1.7 Errori

L'MSE osservato sul Test set è il seguente:

```
Best subset con 5 regressori ha un errore di : 93.084882
Stepwise backward con 5 regressori ha un errore di : 93.084882
Ridge con lambda = 9.754436 ha un errore di : 129.368069
Lasso con lambda = 1.482588 ha un errore di : 109.409707
```

Come ci aspettavamo il miglior metodo è risultato essere Best Subset Selection, in questo caso anche Stepwise Backward ha ottenuto risultati ottimi. Ciò si verifica osservando i coefficienti dei due modelli mostrati negli output delle sezioni relative.