UNIVERSITÀ DEGLI STUDI DI NAPOLI
FEDERICO II

Dipartimento di Scienze Politiche
Corso di Laurea Magistrale in Scienze Statistiche per le Decisioni

Tesi di Laurea Magistrale in Modelli Statistici

Selezione e validazione di un modello statistico: controllo del rischio di modello con il delta BIC

Model selection and validation: controlling the model risk with the delta BIC

Relatore Prof. Domenico Piccolo
Correlatrice Dott.ssa Rosaria Simone

Candidato Paolo Francesco Griffo
Matricola M10239

Indice

Introduzione

Framework

Le tecniche disponibili per la selezione delle variabili di un modello statistico consentono la ricerca del modello migliore, per un insieme di dati, in modo automatico. I modelli stimati possono essere ordinati rispetto al criterio di informazione per la bontà di adattamento BIC e, successivamente, si possono misurare le distanze tra il modello migliore e i modelli vicini ad esso. Questo confronto avviene calcolando le differenze (ΔBIC_k) tra il BIC di ogni modello con il BIC del modello migliore. Una volta ottenute le differenze in termini di ΔBIC_k è possibile individuare i modelli sostanzialmente equivalenti al modello migliore creando una gerarchia tra i modelli. Questo approccio è utile quando vi sono più modelli equivalenti al modello migliore e non è scontato scartare tutti gli altri modelli a favore di uno solo. A partire dai criteri informativi per la bontà di adattamento si derivano misure aggiuntive circa le diverse performance dei modelli in discussione; una di queste è l'indice di stabilità che valuta la performance di un modello in termini di stabilità rispetto alle diverse tipologie di ricampionamento.

Il caso studio

La procedura è stata applicata al caso della regressione lineare multipla, analizzando il numero di accessi casuali al servizio di bike sharing nella citta di Washington D.C., US durante i weekend negli anni 2011 e 2012. La stabilità del modello è stata valutata in tre casi distinti; la prima volta si considera il campione di osservazioni dei weekend 2011-2012 e nella seconda e terza volta si considerano i sottocampioni separatamente. In ciascuna sottoprova si ricampiona rispettivamente il 90% e l'80% dei dati originali e dei dati senza valori anomali.

La *Best Subset selection*

Tra i principali metodi di selezione delle variabili la *Subset selection* merita particolare attenzione. Questa procedura considera l'insieme finito di modelli possibili ottenuti dalle combinazioni semplici dei p predittori disponibili procedendo alla ricerca esaustiva del miglior modello.

Table. Passi della procedura di selezione *Best subset.*

1. Sia M_0 il modello nullo senza predittori
2. Per $j = 1, 2, ..., p$ - Stimare tutti i possibili $\binom{p}{j}$ modelli che contengono esattamente j predittori - Scegliere il migliore tra i $\binom{p}{j}$ modelli indicandolo con M_j. La scelta del migliore è effettuata rispetto al minimo valore di RSS o il massimo R^2-*corretto* (nel caso di modelli lineari).
3. Selezionare il modello migliore tra $M_0, M_1, ..., M_p$ usando un criterio di riferimento: AIC, BIC, C_p, R^2-*corretto*

Procedura generale di derivazione dell'indice di stabilità

1.

Si individua tramite una tecnica di selezione delle variabili il miglior modello M* tra i k modelli possibili, M_1,..., M_k .

2.

Si definiscono i modelli equivalenti in termini di ΔBIC_k rispetto al modello migliore, individuando un insieme A.

3.

Si applica la procedura di bootstrap non-parametrico (totale o parziale) B volte e per ciascuna delle *i* repliche, con $i = 1,\ldots, B$, si ripetono le fasi 1 e 2. Si verifica se il modello migliore $M^*_i \in A$.

4.

La frequenza relativa dei successi: $I_s = \frac{\#M_i^* \in A}{B}$ sarà l'indice che misura la stabilità del modello scelto.

Il dataset bike sharing

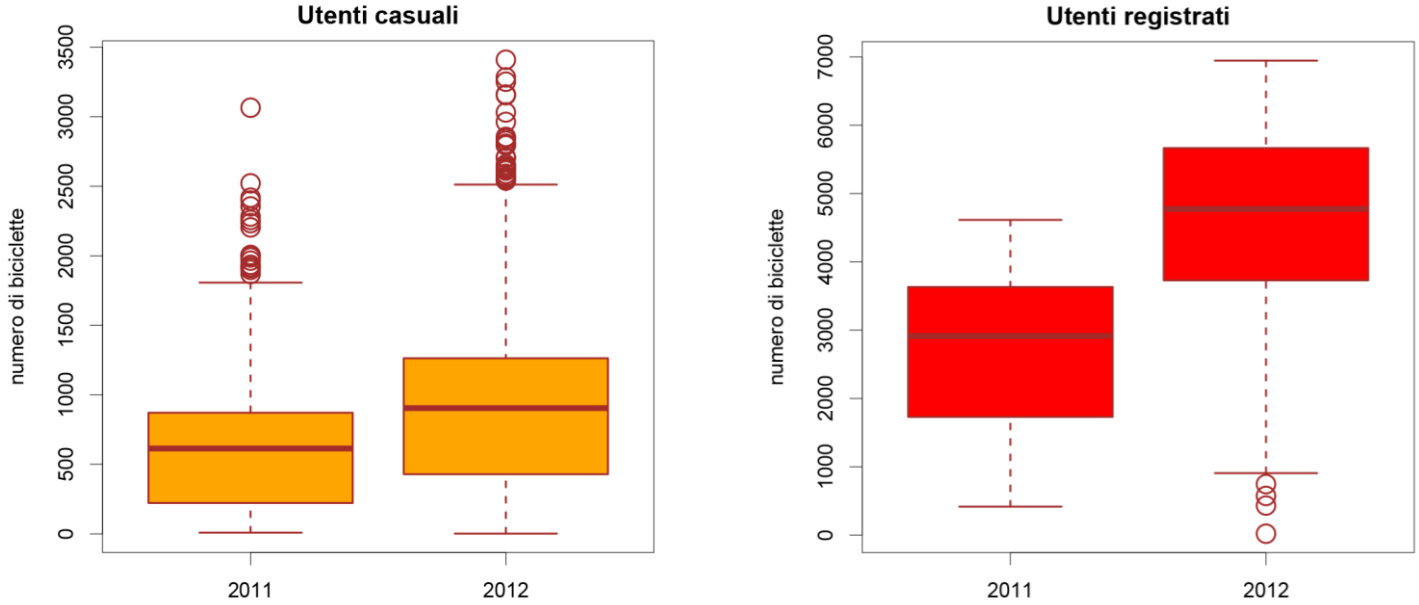
Il dataset contiene informazioni aggregate da varie fonti, a partire dagli accessi al servizio di bike sharing predisposto nella capitale degli Stati Uniti, Washington D.C., rispetto agli anni 2011 e 2012. Il servizio di bike sharing è legato alle condizioni meteorologiche e stagionali locali: giornate particolarmente calde e afose o al contrario fredde e/o piovose possono incentivare le persone a spostarsi con altri mezzi a discapito delle biciclette condivise. Tuttavia, nelle grandi città vi sono molte persone che prendono in prestito regolarmente le biciclette per andare al lavoro, affrontando spesso giornate piovose.

Table. Etichette e descrizioni delle variabili del dataset bike sharing con i livelli delle variabili qualitative e il range di variazione per quelle quantitative.

Etichetta	Descrizione	Livelli - Range
Weathersit	Situazione meteorologica	1: Chiaro, poche nubi, parzialmente nuvoloso 2: Nebbia + nuvoloso, nebbia + nuvole sparse, nebbia 3: Neve leggera, Pioggia leggera + temporale + nubi sparse 4: Pioggia forte + grandine + temporali + nebbia, neve + nebbia
atemp	Temperatura media giornaliera percepita in gradi Celsius standardizzata.	0.08 - 0.84
Hum	Umidità media giornaliera standardizzata.	0 - 0.97
Windspeed	Velocità del vento media giornaliera standardizzata.	0.02 - 0.51
Casual	Numero di utenti casuali giornalieri.	2 - 3410
Reg	Numero di utenti registrati superiore alla media	1: Presente, 0: Assente

Il servizio di bike sharing a Washington D.C., US

Figure. Accessi giornalieri al servizio di bike sharing nella capitale degli Stati Uniti da parte di utenti casuali e registrati al servizio nel 2011-2012.



Eventi collegati alle anomalie nel fusso di noleggi nel 2011-2012

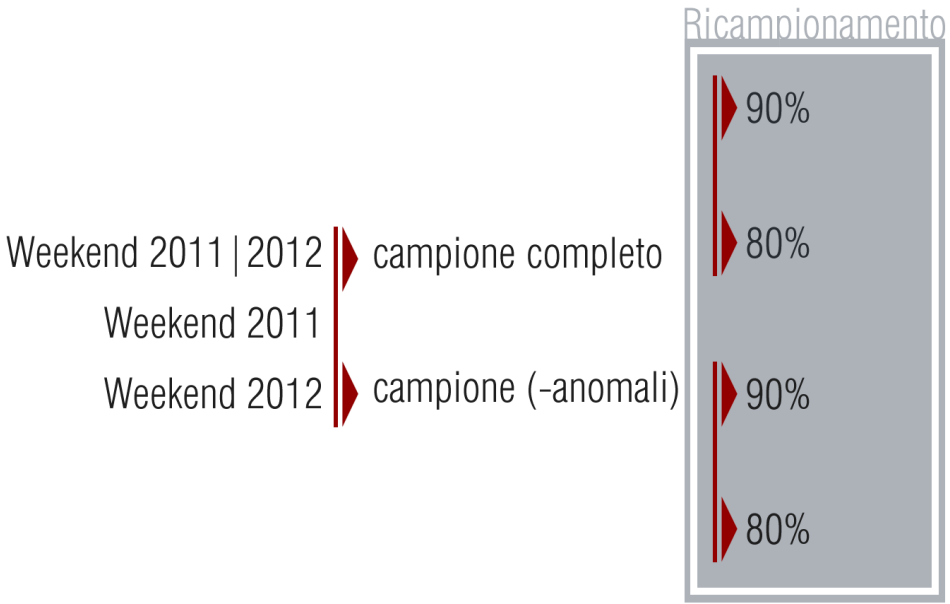
Gli eventi corrispondenti alle giornate con flussi di utenze anomale hanno caratteristiche e portata differente.

Table. Osservazioni anomale presenti nel dataset.

ID	Data	Evento
50	2011-02-19	Giornata più secca e ventosa nel 2011/2012
204	2011-07-23	Giornata di caldo record
239	2011-08-27	Allerta Uragano Irene
351	2011-12-17	Davis County windstorm
463	2012-04-07	D.C. United vs. Seattle Sounders FC
478	2012-04-22	D.C. United vs. Red Bull New York
548	2012-07-01	D.C. United vs. Montreal Impact
554	2012-07-07	Tempesta del 18 Luglio 2012
555	2012-07-08	Tempesta del 18 Luglio 2012

Struttura della sperimentazione

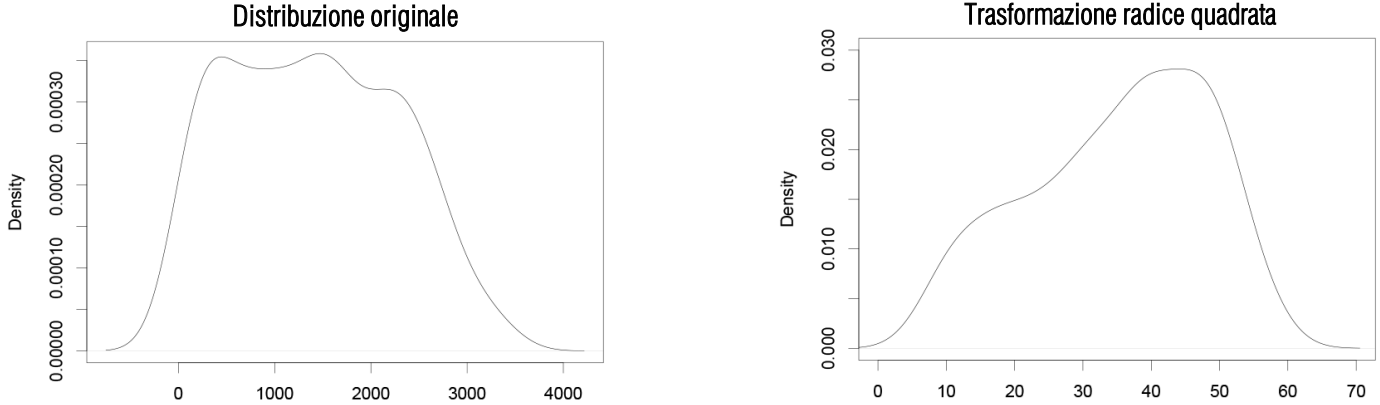
La derivazione dell'indice di stabilità per il modello migliore è stata eseguita sul campione completo e su diversi sottocampioni. La procedura è stata replicata rimuovendo i dati anomali individuati nel campione. In ciascuna replica è stato effettuato un ricampionamento casuale senza reimmissione.



La variabile risposta del modello: utenti casuali del servizio bike sharing

Il fenomeno indagato è il numero di attivazioni giornaliere da parte di utenti casuali.
La distribuzione delle attivazioni del servizio di bike sharing da parte di utenti casuali durante il week end è connotata da una forma asimmetrica positiva e tramite la trasformazione radice quadrata, scelta con il metodo Box-Cox, si è corretta rendendola prossima a quella della variabile casuale Normale standard.

Figure. Density plot della variabile *casual* e della trasformazione radice quadrata.



Il modello di regressione lineare migliore individuato

Il miglior modello per gli accessi giornalieri casuali nel weekend:

$$\sqrt{Y} = \beta_0 + \beta_1 atemp + \beta_2 hum + \beta_3 windspeed + \beta_4 reg$$

atemp (43,572)[1]: temperatura giornaliera media percepita standardizzata

hum (-13,618): umidita giornaliera media percepita standardizzata

windspeed (-17,379): velocita del vento media giornaliera standardizzata

reg (9,963): presenza di accessi di utenti casuali maggiori della media

[1] In parentesi le stime dei coefficienti delle variabili del modello

I modelli equivalenti a confronto

I modelli con ΔBIC$<$5 sono due e coincidono con i modelli ΔBIC$<$2. I modelli equivalenti hanno entrambi tre predittori, due dei quali in comune con il modello migliore.

	Dependent variable:		
		Utenti Casuali	
	Weekend 2011-2012	$\Delta BIC < 2$ (1)	$\Delta BIC < 2$ (2)
Atemp	43.572*** (3.493)	43.978*** (3.538)	41.124*** (3.477)
Hum	−13.618*** (3.469)	−11.500*** (3.418)	
Windspeed	−17.379** (6.703)		
Weathersit			3.730*** (1.058)
Reg	9.963*** (1.171)	10.587*** (1.161)	10.818*** (1.155)
Constant	21.787*** (3.098)	16.623*** (2.405)	8.156*** (1.606)
BIC	1445.8	1447.2	1446.2
Observations	210	210	210
Adjusted R^2	0.707	0.699	0.700
Note:			*p$<$0.1; **p$<$0.05; ***p$<$0.01

I modelli migliori a confronto

Risultati del modello migliore stimato per i dati 2011-2012 e i modelli migliori per gli anni 2011 e 2012.

	Dependent variable:		
		Utenti Casuali	
	Weekend 2011-2012	Weekend 2011	Weekend 2012
atemp	43.572*** (3.493)	38.494*** (5.165)	50.801*** (5.388)
hum	−13.618*** (3.469)		
windspeed	−17.379** (6.703)		
weathersit		3.743** (1.427)	
season			6.185*** (1.657)
reg	9.963*** (1.171)	7.721*** (1.795)	11.556*** (1.820)
Constant	21.787*** (3.098)	6.753*** (2.168)	5.080* (2.795)
BIC	1445.8	715.4	737.6
Observations	210	105	105
Adjusted R^2	0.707	0.681	0.696

Note: *p<0.1; **p<0.05; ***p<0.01

I modelli migliori a confronto

Risultati del modello migliore stimato senza dati anomali per i dati 2011-2012 e i modelli migliori per gli anni 2011 e 2012.

	Dependent variable:		
		Utenti Casuali	
	Weekend 2011-2012	Weekend 2011	Weekend 2012
atemp	50.957*** (3.312)	46.441*** (5.272)	57.413*** (5.221)
hum	−14.149*** (3.304)	−12.474*** (4.444)	
windspeed	−14.773** (6.437)	−20.204** (9.143)	
season			5.336*** (1.470)
reg	8.815*** (1.072)	5.818*** (1.830)	9.799*** (1.758)
Constant	19.325*** (2.896)	19.111*** (3.767)	3.789 (2.471)
BIC	1341.9	675	675.4
Observations	201	101	100
Adjusted R^2	0.760	0.731	0.767
Note:			*p<0.1; **p<0.05; ***p<0.01

Performance di stabilità per i modelli migliori

Table: Stabilità dei modelli con ΔBIC < 5.

		Indice di stabilità				
Modello	dati completi	90%	80%	- anomali	90%	80%
$M^*_{2011-12}$	(A, H, W, R)	0,972	0,928	(A, H, W, R)	0,992	0,980
M^*_{2011}	(A, WE, R)	0,900	0,900	(A, H, W, R)	0,992	0,908
M^*_{2012}	(A, S, R)	0,910	0,910	(A, S, R)	0,824	0,760

Table: Stabilità dei modelli con ΔBIC $= 0$.

		Indice di stabilità				
Modello	dati completi	90%	80%	- anomali	90%	80%
$M^*_{2011-12}$	(A, H, W, R)	0,438	0,352	(A, H, W, R)	0,330	0,244
M^*_{2011}	(A, WE, R)	0,304	0,304	(A, H, W, R)	0,392	0,270
M^*_{2012}	(A, S, R)	0,558	0,558	(A, S, R)	0,824	0,760

Conclusioni

Il modello di regressione migliore per il numero di noleggi di biciclette durante il weekend, presentato in questo caso studio, tiene conto di alcuni dei più significativi agenti atmosferici che condizionano gli utenti casuali del servizio. I risultati ottenuti alla fine della procedura di selezione e validazione del modello statistico fanno riflettere sia sull'influenza che hanno i fattori atmosferici sul flusso di noleggi di biciclette sia sull'impatto che hanno gli eventi anomali sul servizio in generale. In termini assoluti, la velocità del vento e la temperatura percepita registrata influenzano maggiormente il flusso di utenti casuali nel fine settimana.
Nella fase di model selection si può valutare la performance di un modello statistico in termini di bontà di adattamento tenendo conto della sua stabilità e di quella dei modelli sostanzialmente equivalenti ad esso, disponendo così di uno strumento determinante nella gestione del rischio di modello.

Nel caso del modello di regressione migliore $M^*_{2011-2012}$ l'analisi della stabilità circa il modello in questione e i modelli con $\Delta\text{BIC} < 5$ fornisce delle evidenze empiriche a favore del modello individuato. Quando si rimuovo i dati anomali presenti nel campione la stabilità del modello risulta migliorare, così come è positivo l'adattamento del modello ad uno specifico sottocampione di osservazioni.
In definitiva, le evidenze empiriche aggiuntive raccolte analizzando la stabilità di un modello statistico, rappresentano un criterio determinante per la decisione da assumere circa il modello migliore da adottare.

References

Agresti A. *Foundations of Linear and Generalized Linear Models*, 2015, Wiley, Harvard.

Bertaccini B. *Introduzione alla Statistica Computazionale con R*, 2018, Firenze University Press, Firenze.

Bishop, C.M. *Pattern Recognition and Machine Learning*, 2006, Springer, New York.

Buckland, S.T., Burnham K.P. and Augustin N.H. Model selection: an integral part of inference, Biometrics, Vol.53, No.2, pp. 603-618, 1997*, International Biometric Society*, New York.

Conner, J.S., Larimore, W., and Seborg, D.E. Analysis of the AIC Statistic for Optimal Detection of Small Changes in Dynamic Systems, Conference Paper, pp. 4408-4413, 2004, *American Control Conference*, Boston.

Efron, B.D. and Tibshirani, R.J., *An introduction to Generalized Linear Models, 2002,* Second Edition, Chapman & Hall, New York.

Efron, B.D. and Tibshirani, R.J., *An introduction to the Bootstrap*, 1993, Chapman & Hall, New York.

Fanaee-T, H. and Gama, J. *Event labeling; Event detection; Ensemble learning; Background knowledge*, 2013, Progress in Articial Intelligence, pp. 1-15, Springer Berlin Heidelberg, New York. 45

James, G. e Witten, D. and Hastie, T. and Tibshirani, R. *An introduction to Statistical Learning with application in R*, 2013, Springer, New York.

Kullback, S. *Information Theory and Statistics*, 1959, 2a edizione, John Wiley & Sons, New York.

Nelder, J.A. and Wedderburn, R.W.M. Generalized Linear Models, 1972, *Journal of the Royal Statistical Society*, A, 135, pp. 370-384.

Piccolo D. *Statistica*, 2010, (Terza ed.), Il Mulino, Bologna.

Sakamoto, Y. e Ishiguro, M. and Kitagawa, G. *Akaike Information Criterion Statistics*, 1986, KTK Scientic Publishers, Tokyo.

Sen, A. e Srivastava, M. Regression Analysis. *Theory, Methods, and Applications*. 1990, Springer-Verlag, New York.

Symonds, M.R.E. and Moussalli, A. *A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion.* 2010, Behavioral Ecology and Sociobiology, pp. 13-21, Springer Verlag, New York.

Wagenmakers, E.J. and Farrell, S. AIC model selection using Akaike weights, 2004, *Psychonomic Bulletin & Review*, pp. 192-196.

Zucchini, W. An Introduction to Model Selection, 2000, *Journal of Mathematical Psychology*, Volume 44, Issue 1, pp. 41-61, Academic Press.