Tesi di Laurea in Analisi Multivariata Avanzata

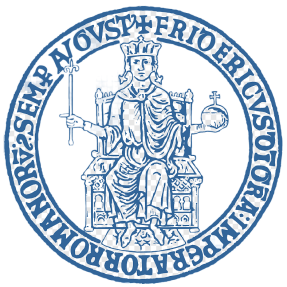
Algoritmi di Feature Selection per modelli ordinali: un'analisi sulla percezione degli eventi sismici

Università degli Studi di Napoli Federico II

Laurea Magistrale in Scienze Statistiche per le Decisioni

27 marzo 2024

Relatore:
Ch.mo Prof. Alfonso Iodice D'Enza
Correlatrice:
Ch.ma Prof.ssa Maria Iannario



Candidato:
Rosario Urso
M10000392

Indice

Introduzione

- In un contesto in cui le informazioni a disposizione aumentano in misura esponenziale, risulta necessario adottare un criterio che ci consenta di determinare le informazioni più utili.

- Il seguente lavoro di tesi è stato stilato allo scopo di presentare differenti approcci per la **feature selection** applicati a modelli appartenenti alla classe dei **GLM** (in particolare ai modelli *ELMO*), quali:
  - *Proportional Odds Model*
  - *Adjacent Category Model*
  - *Continuation Ratio Model*

- Gli algoritmi di *feature selection* (applicati in riferimento alla percezione degli eventi sismici) si riferiscono ad approcci differenti: *subset selection*, *metodi di shrinkage* e *dimensionality reduction*.

Proportional Odds Model

Alla base di tali modelli, si suppone ci sia una variabile latente $\mathbf{Y}^*$ non direttamente osservabile che, essendo definita un un supporto continuo, viene suddivisa attraverso dei *thresholds*:

$$-\infty = \tau_0 < \tau_1 < \ldots < \tau_m = +\infty$$

Dato un set di variabili esplicative, nel caso in cui la variabile dipendente assuma m modalità di risposta ordinate, si ottiene:

$$\begin{aligned}\text{logit}[P(Y \leq j)] &= \log \frac{P(Y \leq j|x)}{1 - P(Y \leq j|x)} \\ &= \log \frac{\pi_1(x) + \pi_1(x) + \cdots + \pi_j(x)}{\pi_{j+1}(x) + \pi_{j+2}(x) \cdots + \pi_m(x)}, \quad j = 1, \ldots, m-1.\end{aligned}$$

Sotto la struttura della variabile latente, applicando G^{-1}:

$$logit[P(Y \leq j|x)] = \tau_j - \boldsymbol{\beta}'\boldsymbol{x}$$

Adjacent Category Model/Continuation Ratio Model

Nell'**Adjacent Category Model**, si considera la probabilità che la Y sia esattamente uguale alla *j-esima* categoria rispetto alla probabilità che la Y sia uguale alla categoria immediatamente successiva:

$$logit[P(Y = j \mid Y = j+1, \boldsymbol{x})] = \log \frac{\pi_j(x)}{\pi_{j+1}(x)} = \tau_j - \boldsymbol{\beta'x}, \quad j = 1, \ldots, m-1.$$

Nel **Continuation Ratio Model**, invece, si considera la probabilità che la Y ricada nella *j-esima* categoria rispetto alla probabilità che la risposta ricada nelle categorie precedenti o in quelle successive.

$$logit[P(Y = j \mid Y \leq j, \boldsymbol{x})] = \log \frac{\pi_j}{\pi_j + \pi_{j+1} + \ldots \pi_{m-1}} = \tau_j - \boldsymbol{\beta'x}, \quad j = 1, \ldots, m-1$$

Proportional Assumption

I modelli specificati fino ad ora considerano il medesimo effetto delle covariate sulla Y. Per verificare sia rispettata l'**assunzione di proporzionalità** ci si avvale del **test di Brant**.

Considerando un *Proportional Odds Model*, in caso di rifiuto dell'ipotesi H_0 di assunzione di proporzionalità, verrà considerata la sua forma non proporzionale o semi proporzionale:

- *Non Proportional Odds Model*:

$$logit[P(Y \leq j|x)] = \tau_j - \boldsymbol{\beta}_j' \boldsymbol{x} \quad j = 1, 2, \ldots, m-1$$

- *Partial Proportional Odds Model*:

$$logit[P(Y \leq j|x)] = \tau_j - \boldsymbol{\beta}' \boldsymbol{x} - \boldsymbol{\gamma}_j' \boldsymbol{u}, \quad j = 1, 2 \ldots, m-1$$

Subset Selection

Negli approcci di **Subset Selection**, partendo dai p predittori considerati inizialmente, viene scelta una combinazione di un numero $q < p$ di variabili esplicative. I metodi utilizzati sono:

- ➜ Backward Selection
- ➜ Forward Selection

- Con questo criterio vengono stimati $\mathbf{1 + \frac{p(p+1)}{2}}$ modelli e, partendo rispettivamento dal modello $\mathcal{M}_p$ ed $\mathcal{M}_0$, ad ogni passo, viene aggiunta/rimossa la variabile che apporta il maggior/minor contributo, in questo caso di **accuracy**, alla stima del modello.

- Al termine, saranno stati selezionati $\boldsymbol{p + 1}$ modelli con la relativa accuracy e tra questi verrà selezionato il modello migliore.

Metodi di Shrinkage

Un ulteriore approccio di feature selection è rappresentato dai **metodi di regolarizzazione**, che rappresentano una tecnica in grado di forzare e *regolarizzare* le stime dei coefficienti ad essere 0. I metodi utilizzati sono:

- *Penalizzazione Elastic Net*
- *Penalizzazione Ridge*
- *Penalizzazione Lasso*

Tali metodi sono applicabili ai modelli della classe *ELMO*, il quale sono composti da due funzioni:

- la prima funzione (`MO`) determina la famiglia del modello, ritenendo valida o meno l'assunzione di proporzionalità;
- la seconda funzione (`EL`) determina la funzione legame.

Tale classe di modelli presenta la seguente forma:

$$g(p) = (g_{EL} \circ g_{MO})(p)$$

Penalizzazione Elastic Net

Tale penalizzazione rappresenta una somma pesata tra la penalizzazione *ridge* e penalizzazione *lasso*, in cui vi sono parametri definiti $0 < \boldsymbol{\alpha} < 1$ e $\boldsymbol{\lambda} > 0$.
Le funzioni obiettivo sono le seguenti:

- *Forma parallela*:

$$\mathcal{M}(c, b; \alpha, \lambda) = -\frac{1}{N_+}\ell(c, b) + \lambda \sum_{j=1}^{p} \left(\alpha \left| b_j \right| + \frac{1}{2}(1-\alpha) b_j^2 \right)$$

- *Forma non parallela*:

$$\mathcal{M}(c, B; \alpha, \lambda) = -\frac{1}{N_+}\ell(c, B) + \lambda \sum_{j=1}^{p} \sum_{k=1}^{K} \left(\alpha \left| B_{jk} \right| + \frac{1}{2}(1-\alpha) B_{jk}^2 \right)$$

- *Forma semi parallela*:

$$\mathcal{M}(c, b, B; \alpha, \lambda, \rho) = -\frac{1}{N_+}\ell(c, b, B) +$$
$$+ \lambda \left(\rho \sum_{j=1}^{p} \left(\alpha \left| b_j \right| + \frac{1}{2}(1-\alpha) b_j^2 \right) + \sum_{j=1}^{p} \sum_{k=1}^{K} \left(\alpha \left| B_{jk} \right| + \frac{1}{2}(1-\alpha) B_{jk}^2 \right) \right)$$

Coordinate Descent Algorithm

L'algoritmo applicato per l'ottimizzazione della funzione obiettivo è il *Coordinate Descent Algorithm*, che prevede due cicli: uno **esterno** ed uno **interno**.

- Il ciclo esterno costruisce una approssimazione quadratica della funzione di log-verosimiglianza $\ell(\beta)$ come somma ponderata della funzione presentata di seguito, ottenuta grazie al polinomio di Taylor del secondo ordine:

$$\ell^{(r)}(\beta) = -\frac{1}{2}\left(z^{(r)} - X\beta\right)^{\top} W^{(r)}\left(z^{(r)} - X\beta\right)$$

- Il ciclo interno invece aggiorna le stime di coefficienti mediante la *funzione di verosimiglianza marginale* $\mathcal{M}^{(r)}$, aggiornando ognuno con il valore che ottimizza la funzione obiettivo *approssimata*:

$$\mathcal{M}_j^{(r,s)}(t) = \mathcal{M}^{(r)}\left(\hat{\beta}_1^{(r,s+1)}, \ldots, \hat{\beta}_{j-1}^{(r,s+1)}, t, \hat{\beta}_{j+1}^{(r,s)}, \ldots, \hat{\beta}_Q^{(r,s)}\right)$$

Dimensionality Reduction

L'**analisi delle corrispondenze** ha come scopo «*quello di individuare dimensioni soggiacenti alla struttura dei dati, dimensioni intese a riassumere l'intreccio di relazioni di interdipendenza tra le variabili originarie*».

In riferimento all'analisi delle corrispondenze, si definisce:

$$\mathbf{F} = \mathbf{Z}'_Y \mathbf{Z}_X$$

dove $\mathbf{Z}_Y$ e $\mathbf{Z}_X$ sono matrici di dimensioni $n \times k$ ed $n \times Q$.
Le coordinate degli **scores** identificati sono date da:

$$\mathbf{W} = \sqrt{\frac{n}{p}} \mathbf{M} \mathbf{Z}_X \mathbf{D}_X^{-\frac{1}{2}} \mathbf{B}^*$$

Determinati gli *scores*, si procederà (considerando un *Proportional Odds Model*) alla stima del modello:

$$logit[P(Y \leq j|x)] = \tau_j - \boldsymbol{\beta}' \boldsymbol{w}, \quad j = 1, \ldots, m-1$$

Struttura del dataset

Il dataset oggetto di analisi consta di **433** osservazioni ed è composto da **33** variabili. La variabile dipendente utilizzata è rappresentata da **paura**, così rappresentata:

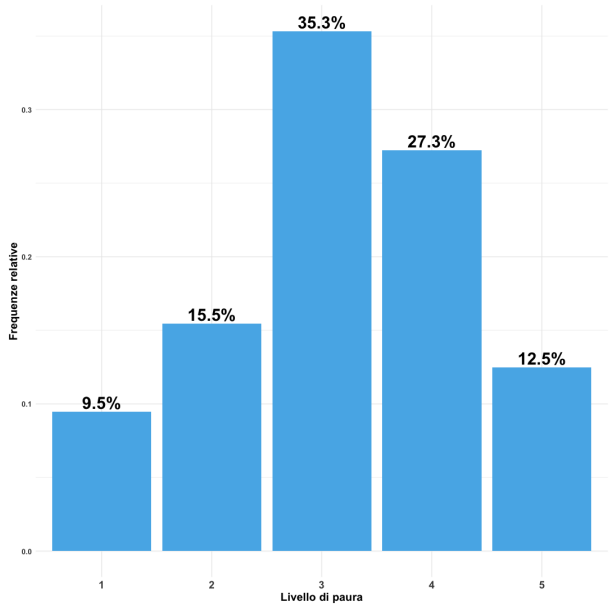


Figura 1: Distribuzione della variabile dipendente

Obiettivo dell'indagine è comprendere quali sono e in che misura impattano le variabili sulla *paura* in relazione agli eventi sismici della zona dei **Campi Flegrei**.

Subset Selection

Con la Subset Selection, sono stati stimati, considerato il numero di variabili, **529** modelli, dove per ogni modello con $1, 2, \ldots, p$ covariate viene considerato il set di variabili che restituisce l'accuracy più elevata.

Selezione	Modello	n	bic	accuracy
Backward Selection	*Adjacent Category Model*	**31**	698.2795	**0.7431**
	Continuation Ratio Model	**20**	700.3701	**0.7339**
	Proportional Odds Model	**31**	695.7759	**0.7339**
Forward Selection	*Adjacent Category Model*	**21**	699.0172	**0.7339**
	Continuation Ratio Model	**24**	667.1201	**0.7431**
	Proportional Odds Model	**22**	666.0391	**0.7339**

Tabella 1: Confronto tra modelli *acat*, *cratio* e *pom* in relazione alla subset selection

Il modello migliore, considerata l'accuracy, è il **Continuation Ratio Model** (*24*).

Metodi di Shrinkage

Relativamente ai metodi di shrinkage, sono stati stimati **60903** modelli ed il tuning dei parametri è stato effettuato con la tecnica della **Grid Search**.
I risultati sono presentati di seguito:

alpha	lambda	family	n	aic	bic	loglik	accuracy
0.70	***0.01***	*acat*	**27**	679.9519	562.7489	-250.3744	**0.7523**
0.71	***0.01***	*acat*	**27**	679.9962	562.7932	-250.3966	**0.7523**
0.72	***0.01***	*acat*	**27**	680.0412	562.8382	-250.4191	**0.7523**
0.12	***0.07***	*cratio*	**27**	715.4220	598.2189	-268.1095	**0.7523**
0.13	***0.07***	*cratio*	**27**	715.8659	598.6629	-268.3314	**0.7523**

Tabella 2: Tuning degli iperparametri α e λ nell'elastic net per *acat*, *cratio* e *pom*

dove emerge che il modello migliore è l'**Adjacent Category Model** con $\boldsymbol{\alpha} = 0.70$ e $\boldsymbol{\lambda} = 0.01$.

Dimensionality Reduction

Per quanto riguardo l'applicazione dell'AC, è stato opportuno procedere al tuning del numero di componenti k.

family	componenti	aic	bic	loglik	accuracy
acat	*2 (84.88%)*	673.7963	696.4808	-330.8982	0.4954
	1 (57.84%)	671.8789	690.7827	-330.9395	0.4862
	3 (94.73%)	671.1958	697.661	-328.5979	0.4862
	4 (100%)	673.1253	703.3712	-328.5626	0.4862
	-	982.0407	997.1637	-487.0203	0.3761
cratio	*3 (94.73%)*	664.8875	691.3527	-325.4437	0.5505
	4 (100%)	666.8826	697.1285	-325.4413	0.5505
	1 (57.84%)	664.215	683.1187	-327.1075	0.5229
	2 (84.88%)	665.5479	688.2323	-326.7739	0.5229
	-	982.0407	997.1637	-487.0203	0.3761
pom	*3 (94.73%)*	664.8429	691.3081	-325.4214	0.5321
	4 (100%)	666.8342	697.0802	-325.4171	0.5321
	1 (57.84%)	664.6927	683.5964	-327.3464	0.5138
	2 (84.88%)	666.1235	688.8079	-327.0617	0.5046
	-	982.0407	997.1637	-487.0203	0.3761

Tabella 3: *Dimensionality reduction* applicata a modelli ordinali.

Risultati

Allo scopo di valutare le **performance** dei diversi algoritmi di *feature selection*, è riportato il seguente grafico contenente le accuracy.

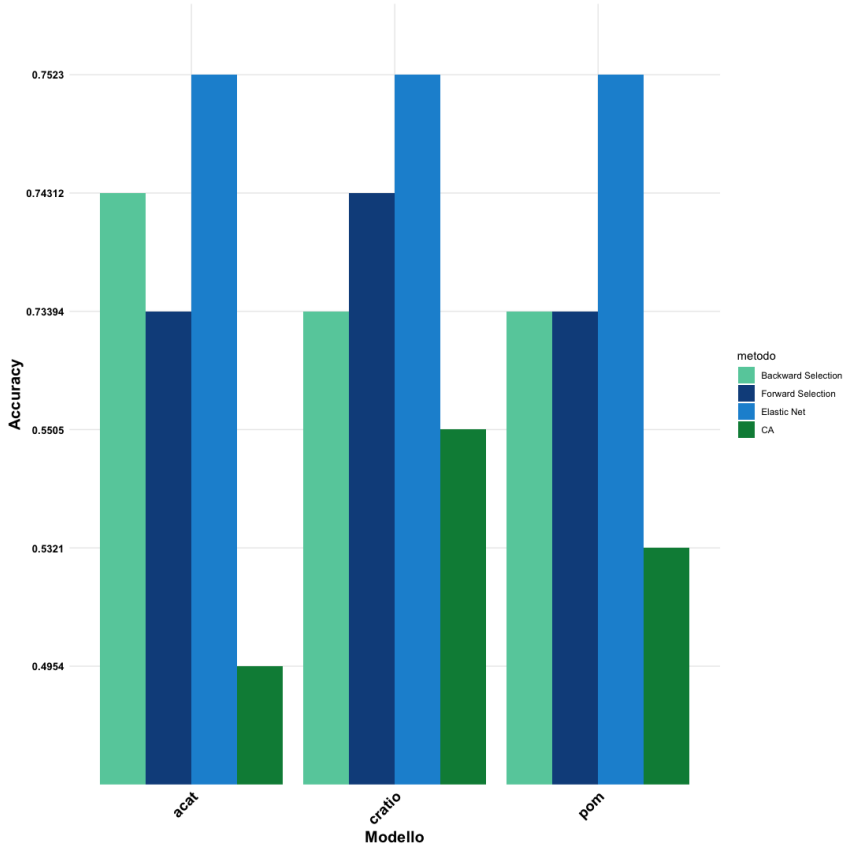


Figura 2: Accuracy su modelli *acat*, *cratio* e *pom* per metodi di shrinkage, subset selection e dimensionality reduction.

Conclusioni

In conclusione, il modello migliore risulta l'**Adjacent Category Model** con penalizzazione **elastic net** con *accuracy* pari a 0.7523.
Tuttavia, data la penalizzazione applicata, non è possibile dare una *interpretazione numerica* alle stime dei coefficienti.

Riportando solo alcune considerazioni:

- la probabilità di passare ad un livello superiore di paura aumenta all'aumentare dell'**età**, per i **single**, per chi non lavora e per chi ha difficoltà ad arrivare a fine mese;
- la probabilità di passare ad un livello superiore di paura diminuisce per **coloro che conoscono i punti di prima accoglienza**, per coloro che vivono in una casa di proprietà e per le persone che risiedono in un appartamento.

Bibliografia

[1] Alan Agresti. *Analysis of ordinal categorical data*. Vol. 656. John Wiley & Sons, 2010.

[2] Alan Agresti. *Categorical data analysis*. Vol. 792. John Wiley & Sons, 2012.

[3] Rollin Brant. ≪Assessing proportionality in the proportional odds model for ordinal logistic regression≫. In: *Biometrics* (1990), pp. 1171–1178.

[4] Jerome Friedman, Trevor Hastie e Rob Tibshirani. ≪Regularization paths for generalized linear models via coordinate descent≫. In: *Journal of statistical software* 33.1 (2010), p. 1.

[5] Michael Greenacre e Jorg Blasius. *Multiple correspondence analysis and related methods*. CRC press, 2006.

[6] Michael J Greenacre. *Biplots in practice*. Fundacion BBVA, 2010.

[7] Michael J Greenacre. ≪Theory and applications of correspondence analysis≫. In: *(No Title)* (1984).

[8] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[9] Stuart R Lipsitz, Garrett M Fitzmaurice e Geert Molenberghs. ≪Goodness-of-fit tests for ordinal response regression models≫. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 45.2 (1996), pp. 175–190.

[10] Peter McCullagh. ≪Regression models for ordinal data≫. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2 (1980), pp. 109–127.

[11] D. Piccolo. *Statistica*. Strumenti / il Mulino. Il Mulino, 2010.

[12] Robert Tibshirani. ≪Regression shrinkage and selection via the lasso≫. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

[13] Gerhard Tutz e Jan Gertheiss. ≪Regularized regression for categorical data≫. In: *Statistical Modelling* 16.3 (2016), pp. 161–200.

[14] Michel Van de Velden, A Iodice D'Enza e Francesco Palumbo. ≪Cluster correspondence analysis≫. In: *Psychometrika* 82 (2017), pp. 158–185.

[15] Tong Tong Wu e Kenneth Lange. ≪Coordinate descent algorithms for lasso penalized regression≫. In: (2008).

[16] Michael J Wurm, Paul J Rathouz e Bret M Hanlon. ≪Regularized ordinal regression and the ordinalNet R package≫. In: *Journal of Statistical Software* 99.6 (2021).

[17] Faisal M Zahid e Shahla Ramzan. ≪Ordinal ridge regression with categorical predictors≫. In: *Journal of Applied Statistics* 39.1 (2012), pp. 161–171.

[18] Hui Zou e Trevor Hastie. ≪Regularization and variable selection via the elastic net≫. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005), pp. 301–320.