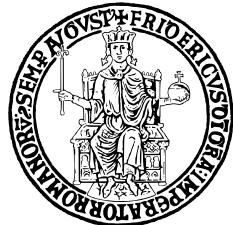


UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II



DIPARTIMENTO DI SCIENZE POLITICHE

CORSO DI LAUREA MAGISTRALE IN
SCIENZE STATISTICHE PER LE DECISIONI

TESI IN ANALISI MULTIVARIATA AVANZATA

ALGORITMI DI FEATURE SELECTION
PER MODELLI ORDINALI:
UN'ANALISI SULLA PERCEZIONE
DEGLI EVENTI SISMICI

Relatore

Ch.mo Prof. Alfonso Iodice d'Enza

Correlatore

Ch.ma Prof.ssa Maria Iannario

Candidato

Rosario Urso

M10/392

Anno Accademico 2022-2023

*Al mio migliore amico, Rudy,
alleggerivi il mondo.
Ovunque tu sia.*

Indice

Introduzione	2
1 Modelli ordinali	3
1.1 Modelli Lineari Generalizzati	3
1.2 Modelli di regressione logistica ordinale	8
1.2.1 Proportional Odds Model	9
1.2.2 Adjacent Category Model	13
1.2.3 Continuation Ratio Model	16
1.3 Verifica dei modelli	18
1.4 Interpretazione dei modelli	23
2 Metodi di Feature Selection	26
2.1 Subset Selection	27
2.1.1 Forward Selection	27
2.1.2 Backward Selection	28
2.2 Metodi di Shrinkage	29
2.2.1 Penalizzazione Elastic Net	30
2.2.1.1 Coordinate Descent Algorithm	32
2.2.2 Penalizzazione Ridge	37
2.2.3 Penalizzazione Lasso	39
2.3 Riduzione della dimensionalità	40
2.3.1 Analisi delle corrispondenze	40
2.3.1.1 Analisi delle corrispondenze su tabelle concatenate	42
3 Implementazione degli algoritmi	45
3.1 Raccolta dei dati	45

3.2	Analisi Esplorativa	47
3.3	Implementazione degli algoritmi	52
3.4	Risultati	60
4	Conclusioni	66
5	Appendice	68

Introduzione

In un contesto in cui la tecnologia ed il progresso avanzano e le informazioni a disposizione aumentano in misura esponenziale, risulta necessario adottare un criterio che ci consenta di determinare quali sono le informazioni utili e quali meno.

Ciò è ancora più accentuato se il tema di riferimento è particolarmente sensibile come quello legato alla percezione (paura, ansia, ecc.) verso l'attività sismica e a tutte le ripercussioni politiche, sociali ed economiche che essa comporta. Il seguente lavoro di tesi, quindi, è stato stilato allo scopo di presentare differenti approcci per la *model selection* applicati a modelli con risposta ordinale.

Nel primo capitolo, vengono illustrati i modelli utilizzati, che rientrano nella classe dei Modelli Lineari Generalizzati (*Generalized Linear Models, GLM*); in particolare sono stati stimati modelli quali *Proportional Odds Models*, *Adjacent Category Models* e *Continuation Ratio Models*. Vengono inoltre discussi anche varianti dei modelli sopracitati, in cui assunzione di proporzionalità non è rispettata, ovvero la cd. *non parallel form* e la *semi-parallel form*, considerando sia la parte inferenziale (stime di massima verosimiglianza) che quella puramente interpretativa.

Nel secondo capitolo vengono illustrati, da un punto di vista analitico, tre diversi approcci per la feature selection riferiti rispettivamente a metodi di *subset selection*, metodi di *shrinkage* e metodi di *dimensionality reduction*. In riferimento alla prima metodologia, sono stati approfonditi la *backward selection* e la *forward selection*, valutati in base alla metrica dell'*accuracy*.

Attraverso tale metodologia, se si considera la backward selection, sono stati

valutate tutte le combinazioni di covariate a partire dal modello completo e da cui viene eliminata la covariata il cui apporto, in termini di accuracy, risulta essere il più basso. Per la forward selection, si parte dal modello senza covariate e ad ogni passo si aggiunge il predittore il cui contributo (sempre basato su accuracy) risulta il più alto tra tutti gli altri.

Di seguito, per quanto concerne i metodi di shrinkage, sono stati approfonditi tre diverse penalizzazioni: la penalizzazione lasso, ridge ed elastic net. Quest'ultima rappresenta una combinazione delle prime due penalizzazioni sopracitate.

L'analisi è stata svolta nell'ambiente *R* utilizzando la libreria **OrdinalNet**.

Infine, è stata applicata un'analisi delle corrispondenze su tabelle concatenate allo scopo di ridurre la dimensionalità e ottenere in questo modo un numero limitato di predittori, con lo svantaggio però di perdere informazioni utili per quanto concerne l'interpretazione dei risultati.

Tutte queste tecniche sono state applicate ad un dataset raccolto in riferimento al fenomeno dell'attività sismica, a seguito delle recenti scosse di terremoto nell'area dei Campi Flegrei, con l'obiettivo di studiare quali variabili e in che misura tali impattano sull'aspetto emotivo (paura, ansia, preoccupazione, ecc) degli individui.

Capitolo 1

Modelli ordinali

1.1 Modelli Lineari Generalizzati

I Modelli Lineari Generalizzati (*Generalized Linear Models, GLM*) rappresentano un'estensione dei classici modelli di regressione lineare, modelli che vengono utilizzati per descrivere relazioni di tipo *lineare* tra la variabile di risposta ed un set di variabili esplicative. [24]

I Modelli Lineari Generalizzati riescono quindi a modellare la relazione tra la variabile dipendente e le variabili esplicative attraverso funzioni di legame non lineari e distribuzioni di errore non gaussiane. Essi presentano 3 componenti:

1. *Componente casuale*, che specifica la distribuzione di probabilità della variabile di risposta Y che può appartenere alla famiglia esponenziale;
2. *Componente lineare*, che identifica le variabili esplicative (x_1, x_2, \dots, x_p) e la loro combinazione lineare (espressa in forma matriciale), ovvero $\eta = \mathbf{X}\boldsymbol{\beta}$;
3. *Funzione legame*, che indica il legame tra le due componenti sopracitate e consente di modellare relazioni non lineari. Essa, specificata come funzione g , indica come il valore atteso della risposta sia legato alla combinazione lineare delle variabili esplicative, dove $g(\cdot)$ è una funzione monotona e differenziabile. [3]

Essa collega il valore medio $\mu_i = \mathbb{E}(y_i)$, $i = 1, 2, \dots, n$ ai predittori η_i :

$$\eta_i = g(\mu_i) \quad (1.1)$$

In generale, nella classe dei Modelli Lineari Generalizzati, è possibile trovare:

1. **modelli per dati binari**, che sono utilizzati nei casi in cui la variabile dipendente è dicotomica e che quindi assume due modalità: assumerà valore 1 se si verifica un successo, 0 altrimenti. La variabile dipendente alla quale ci riferiamo è una variabile casuale di Bernoulli o una variabile casuale Binomiale.

In questo caso, dato che la media di Y assume valori nell'intervallo $(0,1)$, la funzione legame $g : [0, 1] \rightarrow \mathbb{R}$. Dal momento in cui la funzione di ripartizione di una variabile casuale continua assume valori su supporto \mathbb{R} , viene utilizzata funzione di ripartizione inversa $F^{(-1)}$ che, come specificato, assume valori sull'intervallo $[0,1]$. Per cui:

$$g(\mu_i) = F^{-1}(\mu_i) = \mathbf{x}_i \beta \Leftrightarrow \mu_i = \pi_i = F(\mathbf{x}_i \beta) \quad (1.2)$$

dove $\pi_i = Pr(Y_i = 1)$.

Le principali *link function* sono:

- *Link logistico*, esplicitato come segue:

$$g(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) = \mathbf{x}_i \beta \quad (1.3)$$

applicando l'inversa:

$$\mu_i = \frac{e^{\mathbf{x}_i \beta}}{1 + e^{\mathbf{x}_i \beta}} \quad (1.4)$$

- *Link probit*:

$$g(\mu_i) = \Phi^{-1}(\mu_i) = \mathbf{x}_i \beta \quad (1.5)$$

applicando anche qui l'inversa:

$$\mu_i = \Phi(\mathbf{x}_i\beta) \quad (1.6)$$

dove Φ rappresenta la funzione di ripartizione della variabile casuale normale standardizzata. [26]

Altre funzioni legami sono il *complementary log-log*, *log-log* e *link cauchy*, che tuttavia non verranno illustrate in questo lavoro di tesi.

Il modello appena illustrato può essere esplicitato in termini di variabile latente $Y_i^* = \mathbf{x}_i\beta + \epsilon_i$, $i = 1, 2, \dots, n$, per cui diremo che:

$$Y_i = \begin{cases} 0 & \text{se } Y_i^* \leq 0 \\ 1 & \text{se } Y_i^* > 0 \end{cases} \quad (1.7)$$

In presenza della variabile latente, equivale a dire che:

$$\begin{aligned} \Pr(Y_i = 1) &= \Pr(Y_i^* > 0) = \Pr(\mathbf{x}_i\beta + \epsilon_i > 0) \\ &= \Pr(\epsilon_i > -\mathbf{x}_i\beta) = 1 - F(-\mathbf{x}_i\beta) = F(\mathbf{x}_i\beta) \end{aligned} \quad (1.8)$$

2. modelli per dati conteggio, utilizzati nel caso in cui è richiesto studiare fenomeni rappresentabili mediante variabili di conteggio.

Dato che la Y rappresenta un conteggio, la variabile casuale utilizzata è la variabile casuale di Poisson. Tale distribuzione, che presenta diverse generalizzazioni, costituisce un modello per descrivere il numero di occorrenze in un determinato lasso temporale, dove $\mu_i = var(y_i)$. [26]

Affinchè la variabile dipendente Y possa assumere solo valori non negativi rappresentando conteggi, la funzione che viene solitamente applicata la funzione logistica. [3]

Per cui, il modello log-lineare di Poisson (considerato un set di variabili esplicative) è così specificato:

$$\log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} \quad (1.9)$$

da cui, applicando l'esponenziale si ottiene:

$$\boldsymbol{\mu} = \exp(\mathbf{X}\boldsymbol{\beta}) \quad (1.10)$$

Oltre al modello di Poisson, esistono diverse generalizzazioni quali il Truncated Poisson, Negative Binomial, ecc. che non verranno discussi in questo progetto di tesi;

3. **modelli per dati nominali**, impiegati quando la variabile di risposta presenta un numero $m > 2$ di categorie non ordinate (con $m = 2$ categorie, si è in presenza di una variabile dipendente Y binaria). [3]

Il modello logistico multinomiale viene descritto andando a rapportare ogni categoria della Y con una categoria *baseline*, ad esempio la m -esima categoria:

$$\log \frac{\pi_{i1}}{\pi_{im}}, \log \frac{\pi_{ij}}{\pi_{im}}, \dots, \log \frac{\pi_{i,m-1}}{\pi_{im}} \quad (1.11)$$

Dato un set di variabili esplicative per l' i -esima osservazione, possiamo specificare il modello logistico multinomiale come:

$$\log \frac{\pi_{ij}}{\pi_{im}} = \mathbf{x}_i \boldsymbol{\beta}_j \quad j = 1, 2, \dots, m-1 \quad (1.12)$$

dove π_{ij} rappresenta la probabilità della risposta di essere nella categoria j -esima e può essere definito come:

$$\pi_{ij} = \frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_j)}{1 + \sum_{h=1}^{m-1} \exp(\mathbf{x}_i \boldsymbol{\beta}_h)} \quad (1.13)$$

4. **modelli per dati ordinali**, che trovano applicazione nei casi in cui la variabile dipendente Y presenta m categorie di risposta come il caso multinomiale, ma è necessario che essa segua un ordinamento naturale, crescente o decrescente che sia. A titolo d'esempio, una variabile ordinale potrebbe essere rappresentata dalla variabile $X = \text{Titolo di studio}$ disposta con 3 differenti modalità: *Licenza Elementare*, *Licenza Media* e *Diploma*.

Si vedranno nel dettaglio questa tipologia di modelli, per quanto concerne

la specificazione, l'interpretazione e la parte inferenziale dei modelli per dati ordinali.

Per concludere, nella classe dei GLM (più precisamente nei *Vector Generalized Linear Model*, **VGLM**) rientrano anche i cd. modelli **ELMO** (acronimo di **Elementwise Link Multinomial-Ordinal**), in cui sono inclusi i modelli di regressione logistica multinomiale e ordinale introdotti precedentemente, alle quali si farà riferimento nel capitolo 2 per gli algoritmi di penalizzazione.

5. I modelli ELMO sono composti essenzialmente da due funzioni:

- La prima funzione (**M0**) determina la famiglia del modello, ovvero il tipo di modello ordinale che si vuole utilizzare (considerando tra un *Proportional Odds Model*, un *Adjacent Category Model* oppure un *Continuation Ratio Model*), ritenendo valida o meno l'assunzione di proporzionalità che sarà illustrata di seguito.
- La seconda funzione (**EL**) determina la funzione legame, quale funzione di tipo *logit*, *probit* oppure *complementary log-log*.

Tale classi di modelli presentano la seguente forma:

$$g(p) = (g_{EL} \circ g_{MO})(p) \quad (1.14)$$

dove le classi di probabilità sono collegate ai predittori lineari attraverso $g(p_i)$, dove $g : S^K \rightarrow \mathbb{R}$ rappresenta una link function multivariata invertibile ed $S^K = p : p \in (0, 1)^K, \|p\| < 1$, indicando con le p_i probabilità alle K classi. [33] La funzione g è composta quindi da due funzioni $g_{EL} : (0, 1)^K \rightarrow \mathbb{R}^K$ e $g_{MO} : S^K \rightarrow (0, 1)^K$:

$$\begin{aligned} g_{MO}(p) &= \delta = (\delta_1, \dots, \delta_K)^\top \\ g_{EL}(\delta) &= (\tilde{g}_{EL}(\delta_1), \dots, \tilde{g}_{EL}(\delta_K))^\top \end{aligned} \quad (1.15)$$

in cui \tilde{g}_{EL} rappresenta la funzione g_{EL} applicata a un vettore di lunghezza K .

Condizione necessaria e sufficiente affinchè la funzione g_{MO} determini la famiglia dei modelli ordinali-multinomiali è che essa sia invertibile e che abbia un comportamento monotono. [33]

Medesima condizione deve essere rispettata per la funzione g_{EL} , che può essere qualsiasi link function, a patto che sia rispettata la proprietà della *simmetria*, ovvero che:

$$\tilde{g}_{EL}(\delta) = -\tilde{g}_{EL}(1 - \delta) \quad (1.16)$$

Considerando le link functions illustrate precedentemente, rispettano tale proprietà la funzione legame *logit* e *probit*, mentre non viene rispettata con il *c-log-log*.

Tale proprietà ci assicura l'equivalenza dei modelli se considerati con il segno opposto. [26]

1.2 Modelli di regressione logistica ordinale

Al giorno d'oggi in molteplici discipline, scientifiche e non, risulta essere sempre più comune analizzare da un punto di vista prettamente statistico e non solo, set di dati in cui le variabili osservate risultano essere di natura differente.

L'obiettivo principale del ricercatore è quello di comprendere e quantificare l'effetto che hanno le variabili esplicative \mathbf{X} sulla variabile dipendente Y . Particolare attenzione è stata posta sulle variabili ordinali, il quale, a differenza delle variabili nominali, dispongono di un preciso ordinamento e quindi possono essere disposti lungo una scala.

Nella classe dei Modelli Lineari Generalizzati, è possibile trovare i modelli di regressione logistica ordinali. [1]

A seconda dell' impostazione che viene applicata alla probabilità che la variabile di risposta Y assuma una determinata categoria m , si può individuare:

- Proportional Odds Model
- Adjacent Category Model

- Continuation Ration Model

Le differenze tra tali modelli riguardano principalmente l'approccio alla modellazione e l'interpretazione degli effetti delle variabili indipendenti, i vincoli imposti ai parametri e di conseguenza la scelta di un modello più parsimonioso di un altro.

A tali modelli, è applicata la trasformazione logistica alla probabilità cumulata e, come sottolineato precedentemente, si assume che ci sia una variabile latente soggiacente, che non è direttamente osservabile, che si tratti di un *Proportional Odds Model*, un *Adjacent Category Model* o di un *Continuation Ratio Model*.

Essa racchiude l'insieme delle caratteristiche dell'individuo sottoposto a questionario e, per tale motivo, varia da rispondente a rispondente ed è specificata come segue:

sia \mathbf{Y}^* la variabile latente e \mathbf{x} le relative variabili esplicative e supponiamo che Y^* vari attorno ad un parametro di posizione $\boldsymbol{\eta}$ (come la media) che dipende da \mathbf{x} attraverso la relazione $\boldsymbol{\eta}(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x}$. [1]

Da tale assunzione deriva che:

$$Y^* = \boldsymbol{\beta}'\mathbf{x} + \epsilon \quad (1.17)$$

dove ϵ è la componente stocastica di errore.

Si utilizza quindi un modello di regressione per tale variabile continua non osservata che si presume sia alla base della nostra risposta Y .

Nei prossimi paragrafi si osserverà nel dettaglio la formulazione e la relativa interpretazione di tali modelli.

1.2.1 Proportional Odds Model

Con i Proportional Odds Models, proposti da *McCullagh* nel 1980, si è interessati a calcolare le probabilità cumulate applicando la trasformazione logistica. [21]

In merito alla link function, risulta possibile applicare una funzione $G(\cdot)$ diversa dalla logistica ed applicare, ad esempio, $G(\cdot) = \Phi(\cdot)$ ovvero la funzione di ripartizione della variabile casuale normale standardizzata. [26]

Tuttavia, in questo lavoro, verrà utilizzata la funzione di ripartizione logistica $G(z) = e^z/(1 + e^z)$, $z \in \mathbb{R}$ per consentire una immediata ed opportuna interpretazione dei risultati.

Nel caso in cui la variabile dipendente assuma m modalità di risposta ordinate, si ottiene:

$$\begin{aligned} \text{logit}[P(Y \leq j)] &= \log \frac{P(Y \leq j|x)}{1 - P(Y \leq j|x)} \\ &= \log \frac{\pi_1(x) + \pi_2(x) + \cdots + \pi_j(x)}{\pi_{j+1}(x) + \pi_{j+2}(x) + \cdots + \pi_m(x)}, \quad j = 1, \dots, m-1. \end{aligned} \quad (1.18)$$

Considerando una variabile di risposta Y con $m = 5$ categorie di risposte, le probabilità cumulative del modello appena considerato sono riportate nel grafico seguente:

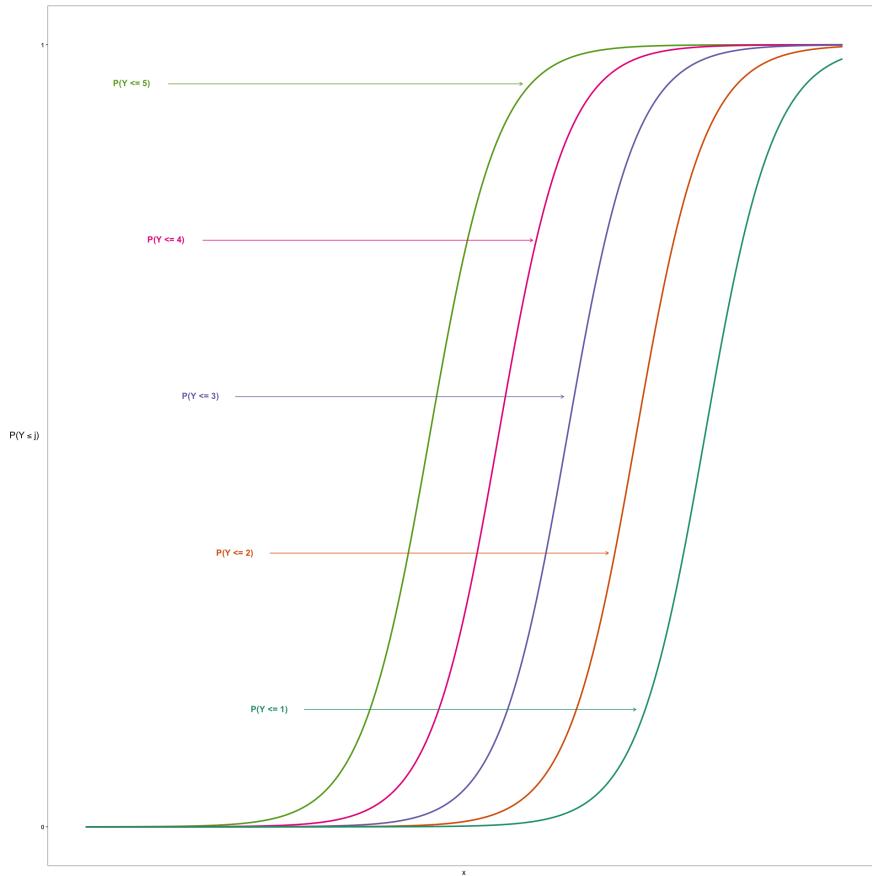


Figura 1.1: Probabilità cumulate nel Proportional Odds Model

dove ogni curva denota lo stesso effetto, ovvero il medesimo β .

Se si ipotizza che la funzione di ripartizione della variabile latente \mathbf{Y}^* illustrata in 1.17 sia

$$P(Y^* \leq y^* | x) = G(y^* - \eta) = G(y^* - \beta' x) \quad (1.19)$$

il valore medio ha una distribuzione che segue quella della funzione logistica con $\mathbb{E}(\epsilon) = 0$.

Essendo la variabile latente definita su un supporto continuo, esso viene diviso mediante dei *thresholds* o *cutpoints* in $m-1$ classi

$$-\infty = \tau_0 < \tau_1 < \dots < \tau_m = +\infty \quad (1.20)$$

Condizione essenziale è che i *thresholds* seguano un ordinamento crescente per poter poi procedere alla stima del modello. Con l'imposizione di tale vincolo è stata discretizzato il supporto della latente.

L'imposizione dei *thresholds* consente di calcolare la probabilità che la dipendente $Y = j$ quando $\tau_{j-1} < Y^* < \tau_j$, ovvero si osserva che la risposta Y cade nella j -esima categoria quando la latente è compresa tra il $j-1$ e il j -esimo threshold.

Sotto la struttura della variabile latente:

$$P(Y \leq j | x) = P(Y^* \leq \tau_j | x) = G(\tau_j - \boldsymbol{\beta}' \mathbf{x}) \quad (1.21)$$

dove G rappresenta la funzione di ripartizione della variabile casuale logistica

$$G(\epsilon) = \frac{e^\epsilon}{1 + e^\epsilon} \quad (1.22)$$

per cui, applicando l'inversa della funzione logistica G^{-1} a (1.21) si ricava:

$$\text{logit}[P(Y \leq j | x)] = \tau_j - \boldsymbol{\beta}' \mathbf{x} \quad (1.23)$$

In questo caso, dalla 1.23 si evince come i $\boldsymbol{\beta}$ siano gli stessi per tutte le probabilità cumulate e che quindi l'effetto sia il medesimo per ogni j .

Per determinare la probabilità che la Y assuma il j -esimo valore condizio-

natamente al valore dei regressori, si definisce:

$$\begin{aligned}
 \pi_j(x) &= P(Y = j) = P(\tau_j < \mathbf{x}\boldsymbol{\beta} + \epsilon \leq \tau_{j+1}) \\
 &= P(\tau_j - \mathbf{x}\boldsymbol{\beta} < \epsilon \leq \tau_{j+1} - \mathbf{x}\boldsymbol{\beta}) \\
 &= P(\epsilon \leq \tau_{j+1} - \mathbf{x}\boldsymbol{\beta}) - P(\epsilon \leq \tau_j - \mathbf{x}\boldsymbol{\beta})
 \end{aligned} \tag{1.24}$$

Tale probabilità equivale a calcolare la probabilità che $\mathbf{x}\boldsymbol{\beta} + \epsilon$ sia compreso tra i due *thresholds*, che si completa nell'espressione in cui, esplicitando la funzione rispetto ai termini di errore, si ottiene la differenza tra le due funzioni di ripartizione calcolate rispetto ai *cutpoints* di riferimento.

Per quanto concerne gli aspetti inferenziali, per le stime dei vettori dei parametri $\tau = (\tau_1, \tau_2, \dots, \tau_{m-1})$ e $\boldsymbol{\beta}$ si ricorre alla funzione di verosimiglianza. [26]

Per l'*i-esimo* soggetto, sia y_{ij} una variabile indicatrice che assume valore 1 se è osservata la *j-esima* categoria e 0 altrimenti, e siano x_i il set di variabili esplicative del soggetto *i*.

La funzione di verosimiglianza risulta essere:

$$\begin{aligned}
 &\prod_{i=1}^n \left[\prod_{j=1}^m \Pr(Y_i = j | \mathbf{X} = \mathbf{x}_i)^{y_{ij}} \right] \\
 &= \prod_{i=1}^n \left[\prod_{j=1}^m \pi_j(\mathbf{x}_i)^{y_{ij}} \right] \\
 &= \prod_{i=1}^n \left\{ \prod_{j=1}^m [\Pr(Y_i \leq j | \mathbf{x}_i) - \Pr(Y_i \leq j-1 | \mathbf{x}_i)]^{y_{ij}} \right\} \\
 &= \prod_{i=1}^n \left\{ \prod_{j=1}^m \left[\frac{\exp(\tau_j - \boldsymbol{\beta}' \mathbf{x}_i)}{1 + \exp(\tau_j - \boldsymbol{\beta}' \mathbf{x}_i)} - \frac{\exp(\tau_{j-1} - \boldsymbol{\beta}' \mathbf{x}_i)}{1 + \exp(\tau_{j-1} - \boldsymbol{\beta}' \mathbf{x}_i)} \right]^{y_{ij}} \right\}
 \end{aligned} \tag{1.25}$$

Di conseguenza, la funzione di log-verosimiglianza:

$$L(\boldsymbol{\tau}, \boldsymbol{\beta}) = \sum_{i=1}^N \sum_{j=1}^m y_{ij} \log [G(\tau_j - \boldsymbol{\beta}' \mathbf{x}_i) - G(\tau_{j-1} - \boldsymbol{\beta}' \mathbf{x}_i)] \tag{1.26}$$

dove, si ricorda, $G(\cdot)$ rappresenta la funzione di ripartizione logistica.

Dato che la funzione di log-verosimiglianza risulta non derivabile in forma chiusa

ovvero risulta non convessa, le relative stime di massima verosimiglianza dei parametri sono derivate mediante il metodo di **Newton-Raphson**. [1] [3]

Il metodo di Newton-Raphson è un metodo iterativo per determinare il punto di massimo di una funzione nei casi in cui le equazioni sono di tipo non lineare. Considerato tale modello (ed in generale la classe dei Modelli Lineari Generalizzati) e posto con $\boldsymbol{\theta}$ il vettore dei parametri del modello:

a) l'algoritmo ottiene una prima stima dei parametri considerando un valore iniziale dei parametri $\boldsymbol{\theta}$;

b) si costruisce la matrice Hessiana $H(\boldsymbol{\theta}_i^{(t)})$ ed il gradiente della funzione di log-verosimiglianza $G(\boldsymbol{\theta}_i^{(t)})$;¹

c) si calcola:

$$\boldsymbol{\theta}_i^{(t+1)} = \boldsymbol{\theta}_i^{(t)} - [H(\boldsymbol{\theta}_i^{(t)})]^{-1} G(\boldsymbol{\theta}_i^{(t)}) \quad (1.27)$$

d) se la stima ottenuta al passo $t + 1$ risulta molto vicino alla stima ottenuta al passo t (ovvero se la loro differenza è più piccola di una quantità infinitesimale definita con ϵ) l'algoritmo si ferma, altrimenti si torna al passo b) per l'iterazione $t + 1$.

$$\left\| \boldsymbol{\theta}_i^{(t+1)} - \boldsymbol{\theta}_i^{(t)} \right\| \leq \varepsilon \Rightarrow \begin{cases} \text{si ottiene } \hat{\boldsymbol{\theta}}_i \\ \text{altrimenti passo b) per l'iterazione t+1} \end{cases} \quad (1.28)$$

Con l'algoritmo di Newton-Raphson (in circa 4 o 5 iterazioni), nel caso di *Proportional Odds Model*, si raggiunge un'elevata accuratezza in tutti i parametri del modello. [21]

1.2.2 Adjacent Category Model

Nell'Adjacent Category Model, rispetto al Proportional Odds Model, si considera la probabilità che la Y sia esattamente uguale alla j -esima categoria rispetto alla

¹La matrice Hessiana rappresenta una matrice contenente le derivate seconde parziali della funzione di log-verosimiglianza, mentre il gradiente della funzione contiene tutte le derivate parziali della funzione di log-verosimiglianza rispetto ai parametri del modello.

probabilità che la Y sia uguale alla categoria immediatamente superiore. Tale modello viene utilizzato con lo scopo di operare confronti tra categorie di risposta per l'appunto *adiacenti*. Anche in questo caso, deve essere rispettata la 1.20.

Il logit tra le categorie adiacenti è dato da:

$$\text{logit}[P(Y = j \mid Y = j + 1, \mathbf{x})] = \log \frac{\pi_j}{\pi_{j+1}}, \quad j = 1, \dots, m - 1 \quad (1.29)$$

Così come nel Proportional Odds Model, considerato un set di predittori \mathbf{x} si ha:

$$\log \frac{\pi_j(x)}{\pi_{j+1}(x)} = \tau_j - \boldsymbol{\beta}' \mathbf{x}, \quad j = 1, \dots, m - 1. \quad (1.30)$$

dove, assumendo valida l'assunzione di proporzionalità, l'effetto di $\beta_1, \beta_2, \dots, \beta_k$ risulta costante al variare della variabile dipendente Y , mentre i relativi thresholds si modificano al variare di j .

Tale modello considera quindi un insieme di logit, al variare di j come si evince da 1.29, equivalenti ai modelli in cui la risposta è di tipo *nominale* (caso in cui l'ordine delle categorie della variabile dipendente Y è irrilevante) se si considera una categoria di base (cd. *baseline category*). L'assunzione di una categoria di base risulta rilevante ai fini inferenziali per poter procedere alla stima della massima verosimiglianza. [1]

Considerata la baseline category m , si ha:

$$\begin{aligned} \log \frac{\pi_j(\mathbf{x})}{\pi_m(\mathbf{x})} &= \sum_{k=j}^{m-1} \tau_k - (m-j)\boldsymbol{\beta}' \mathbf{x}, \\ &= \tau_j^* - \boldsymbol{\beta}' \mathbf{u}_j, \quad j = 1, \dots, m - 1 \end{aligned} \quad (1.31)$$

dove $\mathbf{u}_j = (m-j)\mathbf{x}$ rappresenta la differenza tra categoria considerata di base m e ciascuna categoria j .

Ad esempio, è possibile calcolare la stima di massima verosimiglianza di $\hat{\beta}_j$ considerando

$$\log \frac{\pi_j(x)}{\pi_{j+1}(x)} = \log \frac{\pi_j(x)}{\pi_m(x)} - \log \frac{\pi_{j+1}(x)}{\pi_m(x)}, \quad j = 1, \dots, m-1 \quad (1.32)$$

e calcolando le stime $\hat{\beta}_j^*$ e $\hat{\beta}_{j-1}^*$ otteniamo:

$$\hat{\beta}_j = \hat{\beta}_j^* - \hat{\beta}_{j+1}^* \quad (1.33)$$

imponendo, per l'assunzione di proporzionalità, che i β_j siano costanti.

Considerata la baseline category, è possibile procedere alla stima della funzione di log-verosimiglianza. Per ogni osservazione i , sia y_{ij} una variabile indicatrice che assume valore 1 se è osservata la j -esima categoria, 0 altrimenti, ed essendo x_{ik} il relativo predittore k per il soggetto i , la funzione di verosimiglianza risulta essere:

$$\begin{aligned} \log \left[\prod_{j=1}^m \pi_j(\mathbf{x}_i)^{y_{ij}} \right] &= \sum_{j=1}^{m-1} y_{ij} \log \pi_j(\mathbf{x}_i) + \left(1 - \sum_{j=1}^{m-1} y_{ij} \right) \log \pi_m(\mathbf{x}_i) \\ &= \sum_{j=1}^{m-1} y_{ij} \log \frac{\pi_j(\mathbf{x}_i)}{\pi_m(\mathbf{x}_i)} + \log \pi_m(\mathbf{x}_i) \end{aligned} \quad (1.34)$$

Determinati i parametri τ_j^* e β_j^* per la categoria considerata baseline, data la 1.31 si ha:

$$\begin{aligned} \log \prod_{i=1}^n \left[\prod_{j=1}^m \pi_j(\mathbf{x}_i)^{y_{ij}} \right] &= \sum_{i=1}^n \left\{ \sum_{j=1}^{m-1} y_{ij} (\tau_j^* - \boldsymbol{\beta}' \mathbf{u}_j) - \log \left[1 + \sum_{j=1}^{m-1} \exp(\tau_j^* - \boldsymbol{\beta}' \mathbf{u}_j) \right] \right\} \\ &= \sum_{j=1}^{m-1} \left[\tau_j^* \left(\sum_{i=1}^n y_{ij} \right) - \sum_k \beta_k^* \left(\sum_{i=1}^n u_{jk} y_{ij} \right) \right] \\ &\quad - \sum_{i=1}^n \log \left[1 + \sum_{j=1}^{m-1} \exp(\tau_j^* - \boldsymbol{\beta}' \mathbf{u}_k) \right] \end{aligned} \quad (1.35)$$

Sostituendo si ha:

$$L(\boldsymbol{\tau}, \boldsymbol{\beta}) = \sum_{j=1}^{m-1} \left[\sum_{k=j}^{m-1} \tau_k \left(\sum_{i=1}^n y_{ij} \right) - \sum_k (m-j) \beta_k \left(\sum_{i=1}^n x_{ik} y_{ij} \right) \right] \\ - \sum_{i=1}^n \log \left\{ 1 + \sum_{j=1}^{m-1} \exp \left[\sum_{k=j}^{m-1} \tau_k - (m-j) \left(\sum_k \beta_k x_{ik} \right) \right] \right\} \quad (1.36)$$

Data la 1.36, le statistiche sufficienti per i parametri τ_j e β_k risultano essere:

$$\tau_j = \sum_{i=1}^n \sum_{k=1}^j y_{ik} \quad (1.37)$$

$$\beta_k = \sum_{i=1}^n \sum_{k=1}^j x_{ik} (c - j) y_{ij} \quad (1.38)$$

La funzione di log-verosimiglianza stimata risulta essere concava così come nel Proportional Odds Model e le stime di massima verosimiglianza dei parametri sono derivate mediante lo stesso sviluppo descritto precedentemente.

1.2.3 Continuation Ratio Model

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

Questa tipologia di modello risulta particolarmente utilizzata nella ricerca medica ed anche nel campo dell'istruzione, considerando ad esempio la progressione scolastica degli studenti.

Il logit di ogni categoria rispetto alle categorie più basse è dato da:

$$\text{logit}[P(Y = j \mid Y \leq j, \mathbf{x})] = \log \frac{\pi_j}{\pi_j + \pi_{j+1} + \dots + \pi_{m-1}} \quad j = 1, \dots, m-1 \quad (1.39)$$

Considerato il set di predittori \mathbf{x} , tenendo valida l'assunzione di proporzionalità specificata in 1.20, si ha:

$$\log \frac{\pi_j(x)}{\pi_j(x) + \pi_{j+1}(x) + \dots + \pi_{m-1}(x)} = \tau_j - \boldsymbol{\beta}' \mathbf{x}, \quad j = 1, \dots, m-1 \quad (1.40)$$

Avendo sottolineato in 1.17 l'approccio variabile latente dei modelli di tipo cumulativo, si osserva che:

$$\begin{aligned} Y &= 1 \quad \text{se} \quad Y_1^* \leq \tau_1 \\ Y &= 2 \quad \text{se} \quad \tau_1 < Y_2^* \leq \tau_2 \\ &\vdots \\ Y &= j \quad \text{se} \quad \tau_{j-1} < Y_j^* \leq \tau_j \end{aligned} \quad (1.41)$$

Dalla specificazione del modello in 1.39, si evince che tale modello è stimato mediante regressioni logistiche binarie, in cui si confronta la j -esima categoria della variabile dipendente Y rispetto alle precedenti. [7]

Considerato ad esempio un numero di categorie m pari a 5, avremo il logit della probabilità che la Y sia uguale alla 5° categoria rispetto alla probabilità che sia minore, il logit della probabilità che la Y sia uguale alla 4° categoria rispetto alla probabilità che sia minore e via dicendo. Verranno, in questo caso, stimate $m-1$ regressioni logistiche binarie. Da ciò deriva l'analogia del Continuation Ratio Model con il modello multinomiale rispetto al quale, a fini inferenziali, la stima di massima verosimiglianza può essere ottenuta mediante una serie di regressioni logistiche binarie definite indipendenti. [9]

Pertanto, massimizzando questo set di funzioni di verosimiglianza separatamente, si massimizza anche l'intera verosimiglianza riferita al Continuation Ratio Model [1].

Si sfrutta quindi la fattorizzazione della funzione di verosimiglianza, mediante la quale si esplicita inizialmente la funzione di log-verosimiglianza per il modello logistico partendo dalla specificazione in 1.40:

$$L_j = \left(\sum_{\mathcal{S}_j} y_{ij} \right) \tau_j - \sum_k \left(\sum_{\mathcal{S}_j} y_{ij} x_{ik} \right) \beta_k - \sum_{\mathcal{S}_j} \log \left[1 + \exp \left(\tau_j + \sum_k \beta_k x_{ik} \right) \right] \quad (1.42)$$

dove \mathcal{S}_j rappresenta i soggetti la cui risposta è identificata nella j -esima categoria, e di conseguenza $\sum_{\mathcal{S}_j}$ rappresenta la somma degli individui i nel gruppo \mathcal{S}_j .

La log-verosimiglianza del modello completo (derivabile anche in questo caso attraverso l'algoritmo di *Newton-Raphson*) risulta essere la seguente:

$$L(\boldsymbol{\tau}, \boldsymbol{\beta}) = L_1 + L_2 + \dots + L_{m-1} \quad (1.43)$$

Le statistiche sufficienti per i parametri stimati sono:

$$\tau_j = \sum_{\mathcal{S}_j} y_{ij} = n_j \quad (1.44)$$

$$\beta_k = \sum_{\mathcal{S}_j} y_{ij} x_{ik} \quad (1.45)$$

dove n_j rappresenta il numero di osservazioni all'interno della j -esima categoria di risposta. La statistica sufficiente per β_k è ottenuta considerando la somma dei \mathcal{S}_{m-1} gruppi di rispondenti per le rispettive categorie sugli n soggetti.

1.3 Verifica dei modelli

Successivamente alla fase di specificazione e conseguente stima dei parametri risultanti del modello, è necessario valutare l'adeguatezza del modello attraverso indici di bontà di adattamento globale, test locali (per verificare sia rispettata l'assunzione di proporzionalità) e confronto tra diversi modelli.

In riferimento ai Proportional Odds Model, per quanto riguarda il test di adattamento globale (applicabili in presenza di tabelle di contingenza), si specifica l'ipotesi nulla H_0 di buon adattamento ai dati contro l'ipotesi alternativa H_1 che

denota un cattivo adattamento. [1]

La statistica Test per la bontà di adattamento globale è:

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (1.46)$$

dove questa Statistica test risulta essere una variabile Chi-quadro con gradi di libertà $(I - 1)(J - 1) - k$, dove I rappresenta il numero di covariate, J il numero di categorie della variabile di risposta Y e k il numero di covariate di natura categorica. [25]

Dato che è possibile eseguire il test di adattamento globale del modello solo in presenza di variabili di natura qualitativa, quando il modello contiene almeno una variabile esplicativa continua e quando i dati raggiungono una numerosità elevata, si utilizzano test per la bontà di adattamento alternativi come quello di *Lipsitz et. Al* del 1996. [19] [20]

Tale test, che non è sempre calcolabile per piccoli campioni, è basato sull'approssimazione di *Hosmer-Lemeshow* per dati binari ed è formulato definendo $g - 1$ variabili indicatrici binarie I_g che assumono i seguenti valori:

$$I_{ig} = \begin{cases} 1 & \text{se l'osservazione } i \text{ è nella cella } g \\ 0 & \text{altrimenti} \end{cases} \quad (1.47)$$

Definita tale variabile, viene stimato un nuovo modello L_{ij} che include la variabile indicatrice I_g , così specificato:

$$L_{ij} = \tau_j - \boldsymbol{\beta}' \mathbf{x}_i + \sum_{g=1}^{G-1} \boldsymbol{\gamma}' \mathbf{I}_{ig}, \quad g = 1, \dots, G-1 \quad i = 1, \dots, n. \quad (1.48)$$

dove $\gamma_1 = \gamma_2 = \dots = \gamma_{G-1} = 0$ se il modello risulta essere correttamente specificato. Tale tipologia di test però presenta diversi problemi: non tiene conto dell'overfitting, tende ad avere una bassa potenza e la scelta dei sottogruppi g risulta arbitraria. [10]

Pulkestenis e *Robinson* nel 2004 hanno ovviato al problema della presenza nel

modello di covariate continue andando a modificare la statistica χ^2 di Pearson. Il loro approccio parte raggruppando le osservazioni considerante le sole variabili categoriche presenti nel set di dati. Ciascun pattern creato viene reso binario mediante la mediana dei punteggi stimati. [25]

$$\chi^2 = \sum_{i=1}^I \sum_{h=1}^2 \sum_{j=1}^J \frac{(O_{ihj} - E_{ihj})^2}{E_{ihj}} \quad (1.49)$$

dove i indica il set di covariate, h la stratificazione ottenuta con i punteggi stimati della variabile dipendente e j rappresenta le categorie della variabile di risposta. La distribuzione di riferimento è ancora una volta una variabile casuale χ^2 con $(2I-1)(J-1) - k - 1$ gradi di libertà, seguendo le medesime specificazioni utilizzate per l'indice χ^2 illustrate precedentemente.

Analiticamente per poter verificare tale assunzione, si possono seguire approcci diversi: utilizzando il test di *Brant* (1990) o verificando l'ipotesi di proporzionalità mediante i residui.

Il test di Brant compara le stime $\hat{\beta}$ per ogni predittore utilizzando il Test di *Wald* [1].

Considerando un Proportional Odds Model, tale test verifica l'ipotesi di proporzionalità dell'intero modello e delle singole variabili esplicative, per cui si rifiuterà l'ipotesi nulla di assunzione di proporzionalità con una statistica test χ^2 con $(m-2)p$ gradi di libertà. Rifiuteremo H_0 (*assumption proportional odds* o anche indicata come *forma parallela*) se il *p-value* è < 0.05 , per cui si procederà poi a considerare una *partial proportional assumption* (forma semi-parallela) oppure una *non proportional assumption* (forma non parallela), con la specificazione dei modelli che verrà illustrata di seguito:

1. *Non Proportional Odds Model*, caratterizzato da un differente effetto delle variabili esplicative per ogni probabilità che la Y possa cadere nelle diverse categorie. [1]

$$\text{logit}[P(Y \leq j|x)] = \tau_j - \boldsymbol{\beta}'_j \mathbf{x} \quad j = 1, 2, \dots, m-1 \quad (1.50)$$

2. *Partial Proportional Odds Model*, quando l'assunzione di proporzionalità risulta essere parziale. [23]

Ciò significa che, considerata tale forma, il modello presenta un set di covariate \mathbf{x} che presenterà l'assunzione di proporzionalità ed un altro set indicato con \mathbf{u} in cui non sarà verificata l'assunzione di proporzionalità. Analiticamente:

$$\text{logit}[P(Y \leq y)] = \tau_j - \boldsymbol{\beta}'\mathbf{x} - \boldsymbol{\gamma}'_j\mathbf{u}, \quad j = 1, \dots, m-1 \quad (1.51)$$

Medesime considerazioni possono essere effettuate per l'Adjacent Category Model e per il Continuation Ratio Model, che presentano quindi entrambi la forma non parallela e semi-parallela così come il Proportional Odds Model appena illustrato.

Per quanto riguarda l'assunzione di probabilità di questi ultimi due modelli, essa può solo essere verificata globalmente (e non localmente come nel Proportional Odds Model), motivo per il quale si utilizzano verifiche alternative attraverso *Likelihood Ratio test* (che verrà esplicitato di seguito) oppure confronto tra indici quali AIC o BIC. [9]

Occorre sottolineare che, per quanto possa essere conveniente per motivi legati alla parsimonia e alla complessità del modello, sembra ragionevole utilizzare un modello sufficientemente flessibile tale da consentire una deviazione dall'ipotesi di parallelismo, in quanto essa non sempre risulta l'ipotesi giusta e ragionevole. [28]

Tornando al test di Brant, esso presenta due tipi di problemi: nel caso sia presente un elevato k o p , i gradi di libertà tenderanno ad aumentare notevolmente con conseguente diminuzione del livello di potenza del test; in secondo luogo, comprendere quali siano le cause della mancanza di adattamento non è agevole, in quanto, con la differenza delle stime di massima verosimiglianza $\tilde{\beta}_j - \tilde{\beta}_l$, non viene manifestata la causa del mancato adattamento. [5]

I motivi che portano al rifiuto del test sulla bontà di adattamento sono diversi. Si annovera: l'assenza di un importante covariata, di un effetto interazione o la

presenza di problemi legati alla dimensione campionaria. Risulta, quindi, sensato comparare due modelli che differiscono, ad esempio, per una o più covariate o per un effetto interazione attraverso il test sul rapporto della verosimiglianza (**Likelihood Ratio Test**), che è un test per modelli annidati. [1]

Le ipotesi alla base del test sono

$$\begin{aligned} H_0 : \beta_{k+1} &= \cdots = \beta_p = 0 \\ H_1 : \beta_{k+1} &= \cdots = \beta_p \neq 0 \end{aligned} \tag{1.52}$$

La statistica test sarà rappresentata da

$$D(\hat{\theta}) = 2 \left\{ \ell^S - \ell(\hat{\theta}) \right\} \tag{1.53}$$

dove ℓ^S rappresenta la *log-verosimiglianza* del modello saturo che presenta p parametri, mentre $\ell(\hat{\theta})$ rappresenta la *log-verosimiglianza* del modello annidato che presenta q parametri (con $p > q$).

La statistica test si confronta con il percentile di una variabile casuale χ^2 con $(q - p)$ gradi di libertà. Il test conduce al rifiuto dell'ipotesi nulla se il *p-value* < 0.05.

L'ipotesi H_0 impone che tutte le stime siano congiuntamente uguale a 0, e nel caso di rifiuto di tale ipotesi, verrà preferito il modello saturo: le stime $\beta_{k+1}, \beta_{k+2}, \dots, \beta_p$ erano statisticamente significative.

Nel caso in cui si volesse effettuare un confronto per valutare quale tra i modelli con *link* differenti sia il migliore, si calcolano indici quali AIC (*Akaike Information Criterion*) e BIC (*Bayesian Information Criterion*) per modelli non annidati. Il criterio di informazione di Akaike considera quanto i valori stimati dal modello siano vicini alle probabilità reali e viene quindi preferito il modello il cui AIC risulta più basso. [1] [24]

L'indice AIC è definito da:

$$\text{AIC} = -2l(\hat{\theta}) + 2(p + 1) \tag{1.54}$$

dove $l(\hat{\theta})$ è la funzione di log-verosimiglianza del modello, calcolata nel punto

di massimo stimato $\hat{\theta}$ e p sono il numero di parametri.

Tale indice tende a sovra-parametrizzare i modelli che seleziona, per cui per motivi legati alla parsimonia si tende a preferire il BIC. [24]

L'indice BIC di Schwarz è definito da:

$$\text{BIC} = -2l(\hat{\theta}) + (p + 1)\log(n) \quad (1.55)$$

Quest'ultimo risulta consistente ed asintoticamente non distorto rispetto all'indice AIC. Anche in questo caso, si sceglie il modello che restituisce valore di **BIC** più basso.

Un ulteriore criterio consiste nel confrontare i valori predetti dal modello con i valori osservati, andando quindi a considerare l'accuratezza delle performance del modello. Tale metrica viene calcolata nel seguente modo:

$$\text{Accuracy} = \frac{\text{Numero di predizioni corrette}}{\text{Numero totale di predizioni}} \quad (1.56)$$

L'*accuracy* è compresa tra un minimo 0 e un massimo 1. Essa sarà 0 quando il modello non avrà catturato alcuna informazione circa la reale distribuzione della variabile dipendente, quindi si è in presenza di un modello non correttamente specificato. Invece, sarà 1 quindi il modello riesce a predire la totalità delle osservazioni in modo corretto.

Spesso, in presenza di un accuracy molto vicino al massimo il modello potrebbe tendere all'*overfitting*, ovvero esso si adatta eccessivamente ai dati considerati senza però riuscire a generalizzare in presenza di nuove osservazioni.

1.4 Interpretazione dei modelli

Per quanto concerne l'interpretazione e per determinare l'impatto che hanno le covariate sulla variabile dipendente Y , quando il link è di tipo *logistico*, ci si avvale dell'ausilio degli odds.

Si definisce *odds* di un evento la probabilità che si verifichi l'evento successo diviso

la probabilità dell'evento insuccesso.

Considerando il Proportional Odds Models, l'odds è dato da

$$odds_j(\mathbf{x}) = \frac{P_r(Y_i \leq j | \mathbf{x})}{P_r(Y_i > j | \mathbf{x})} = \frac{P_r(Y_i \leq j | \mathbf{x})}{1 - P_r(Y_i \leq j | \mathbf{x})} \quad (1.57)$$

Ciò equivale a considerare la probabilità di avere una valutazione inferiore alla j -esima categoria rispetto ad averne una superiore.

In presenza del link logistico, l' $odds_j$ sarà pari a

$$odds_j(\mathbf{x}) = \exp(\tau_j + \boldsymbol{\beta}' \mathbf{x}), \quad j = 1, \dots, m-1 \quad (1.58)$$

Il legame logistico, inoltre, ci consente di calcolare i $\log(odds)$, che sono così definiti

$$\text{logit}[P(Y \leq j) | \mathbf{x}] = \log \frac{P(Y \leq j) | \mathbf{x}}{P(Y > j) | \mathbf{x}} = \tau_j + \boldsymbol{\beta}' \mathbf{x}, \quad j = 1, \dots, m-1 \quad (1.59)$$

Riguardo l'interpretazione:

- Se $\beta_k > 0$, è più probabile che Y cada delle categorie estreme della scala.
- se $\beta_k < 0$, si avrà una maggiore probabilità che la Y appartenga alle prime j categorie.

Rispetto agli *Adjacent Category Models* e dei *Continuation Ratio Models* così come nei modelli per dati multinomiali, i *Proportional Odds Models*, come già evidenziato, risultano più parsimoniosi: tale modello ci restituisce le stime con i valori relativi ad τ_j che rappresentano l'intercetta e i coefficienti di regressioni $\boldsymbol{\beta}_k$ associati alle relative variabili esplicative, che risultano costanti al variare di j , delle modalità di risposta.

Considerando gli *Adjacent Category Models*, il log-*odds* è rappresentato da:

$$\begin{aligned} \text{logit}[P(Y = j | Y = j+1, \mathbf{x})] &= \log \frac{P(Y = j | \mathbf{x})}{P(Y = j+1 | \mathbf{x})} \\ &= \tau_j - \boldsymbol{\beta}' \mathbf{x}, \quad j = 1, \dots, m-1. \end{aligned} \quad (1.60)$$

dove se $\beta_k > 0$ vi è una probabilità più alta che la variabile dipende possa cadere nella categoria j piuttosto nella $j - esima$ e se $\beta_k < 0$ vi è una maggiore probabilità che si trovi nella categoria inferiore invece che in quella superiore. Se il modello utilizzato fosse stato il *Continuation Ratio Model*, i parametri da stimare sarebbero stati $k \times m$, così come fatto presente nel caso di modelli multinomiali (a seconda se è verificata o meno l'assunzione di proporzionalità).

Per quanto riguarda gli *odds*, il ragionamento risulta analogo a quello presentato per i precedenti modelli, ricordando che gli odds rappresentano la probabilità di transizione da una categoria j all'altra definita $j + 1$ rispetto alla probabilità di rimanere nella $j - esima$ categoria.

Tuttavia, nella quasi totalità dei casi vengono trattati modelli con più di una variabile esplicativa, per cui risulta interessante calcolare il cd. *log(oddsratio)* (rapporto tra gli odds) per operare opportuni confronti.

Di seguito è indicato, per semplicità, il log-odds ratio e l'odds ratio relativo solo ai *Proportional Odds Models*, illustrato come segue:

$$\begin{aligned} & \text{logit}[P(Y \leq j | \mathbf{x}_1)] - \text{logit}[P(Y \leq j | \mathbf{x}_2)] \\ &= \log \frac{P(Y \leq j | \mathbf{x}_1) / P(Y > j | \mathbf{x}_1)}{P(Y \leq j | \mathbf{x}_2) / P(Y > j | \mathbf{x}_2)} = \boldsymbol{\beta}' (\mathbf{x}_1 - \mathbf{x}_2) \end{aligned} \quad (1.61)$$

da cui l'*odds ratio*

$$\frac{P(Y \leq j | \mathbf{x}_1) / P(Y > j | \mathbf{x}_1)}{P(Y \leq j | \mathbf{x}_2) / P(Y > j | \mathbf{x}_2)} = \exp(\boldsymbol{\beta}' (\mathbf{x}_1 - \mathbf{x}_2)) \quad (1.62)$$

Il log-odds ratio (e di conseguenza l'odds ratio) è proporzionale alla distanza tra x_1 e x_2 . Inoltre, nel caso usuale in cui si consideri di cambiare un solo parametro alla volta di una unità, e^β rappresenta il rapporto di probabilità, come al solito nella regressione logistica. La differenza, tuttavia, è che e^β ora rappresenta un rapporto di quote cumulative, denominato un odds ratio cumulativo.

Capitolo 2

Metodi di Feature Selection

In questo capitolo verrà trattato il tema della selezione delle covariate riferito ai modelli per variabile dipendente ordinale.

La selezione delle variabili rappresenta un tema importante nella modellistica, in quanto, in presenza di un numeroso set di variabili, risulta necessario e conveniente, da un punto di vista computazionale, selezionare un opportuno subset di predittori affinchè si possa ottenere un modello più parsimonioso ed efficiente rispetto al modello completo utilizzando tutte le covariate presenti nel set di dati che si utilizza.

Gli approcci utilizzati nel progetto di tesi sono diversi. Inizialmente, verranno analizzati due diversi approcci di *Subset Selection*, mediante il quale si individuerà un subset dei p predittori iniziali per spiegare la variabile di risposta Y . Il primo approccio è la *Forward Selection* ed il secondo è la *Backward Selection*, approccio attraverso la quale verranno selezionate tutte le possibili combinazioni delle covariate del set di dati utilizzato, rispettivamente in avanti e in indietro.

In secondo luogo, perseguiendo il medesimo obiettivo di riduzione dei predittori all'interno del modello, verranno illustrati i cd. *metodi di shrinkage* che rappresentano delle tecniche di regolarizzazione in grado di forzare le stime dei coefficienti stimati ad essere pari a 0. [18]

Tra tale metodologia rientra la regolarizzazione *Lasso*, *Ridge* ed *Elastic Net*. Infine, l'ultimo approccio si differenzia dai primi due sopracitati, in quanto non vengono considerati i p predittori originali X_1, X_2, \dots, X_p e riguarda la *dimensione*

sionality reduction, mediante la quale vengono definiti dei *fattori* che riassumono le relazioni tra le righe e le colonne del set di dati utilizzato. [13]

Tutti questi approcci saranno valutati secondo la metrica della *accuracy*.

2.1 Subset Selection

Il primo metodo di selezione dei predittori illustrato ed applicato è la *Subset Selection*, attraverso la quale, partendo dai p predittori considerati inizialmente, viene scelta la combinazione di un numero $q < p$ di variabili esplicative in base ad una determinata metrica.

Ciò allo scopo di semplificare il modello, ridurre il rischio di *overfitting* e di conseguenza migliorare le prestazioni generali del modello.

Tuttavia, a causa dell'elevata potenza computazionale richiesta quando i predittori tendono ad aumentare (i modelli da stimare sarebbero pari a 2^p), non è stata considerata la Best Subset Selection. [18]

Di conseguenza, i metodi definiti in questo lavoro di tesi sono *Forward Selection* e *Backward Selection*.

2.1.1 Forward Selection

La *Forward Selection* risulta essere computazionalmente più leggera rispetto alla Best Subset, soprattutto nelle situazioni in cui $p \gg n$. Infatti, rispetto alla Best Subset in cui vengono stimati 2^p modelli (ovvero viene stimato un modello per ogni p combinazione di variabili esplicative), con questo criterio vengono stimati $1 + \frac{p(p+1)}{2}$ modelli. Ad esempio, avendo a disposizione un numero di predittori p pari a 10, con la Best Subset sarebbero stati stimati 1024 modelli, invece con la Forward Selection solo 56. [18]

La Forward Selection considera il modello stimato senza nessun predittore ed, ad ogni passo, aggiunge la variabile che apporta il maggior contributo, considerata una determinata metrica, alla stima del modello.

La procedura è quindi la seguente:

1. Si parte con il modello nulla \mathcal{M}_0 ovvero senza predittori;

2. al primo passo vengono stimati p modelli (ovvero tanti modelli quanti sono le covariate) con 1 sola variabile esplicativa e, tra questi, viene selezionato il modello la cui accuracy risulta più alta;
3. al secondo passo, considerando la covariata scelta al primo passo, si stimano $p-1$ modelli con 2 covariate e si sceglie il modello con 2 covariate che riporta l'accuracy più alta;
4. l'algoritmo continua considerando i modelli a $3, 4, \dots, p$ covariate, terminando quindi con la stima del modello completo \mathcal{M}_p .

Al termine, saranno stati selezionati $p+1$ modelli con la relativa accuracy e tra questi verrà selezionato il modello migliore. [18]

Dal momento in cui la metrica utilizzata in questo lavoro di tesi è quella della *accuracy*, ogni $1 + \frac{p(p+1)}{2}$ modello viene stimato quindi due volte, essendo una volta allenato e poi testato su un set di dati differenti. Il difetto di questa procedura è che una variabile ritenuta importante in un certo momento possa poi risultare ai passi successivi inutile perché non più statisticamente significativa. [24]

Quindi, rispetto alla Best Subset, non vi è la garanzia di identificare il modello migliore.

2.1.2 Backward Selection

Approccio speculare a quello della *Forward Selection* è la *Backward Selection* e rappresenta anch'esso un'alternativa alla Best Subset Selection.

Tale criterio parte con un modello contentente p predittori e rimuove, ad ogni passo, il predittore la cui rimozione porta ad un miglioramento delle prestazioni secondo la metrica utilizzata.

I modelli stimati, analogamente al caso forward, sono $p+1$ e il processo di selezione segue le fasi illustrate di seguito:

1. Si parte con il modello completo \mathcal{M}_p che contiene p predittori;
2. al primo passo, vengono stimati $p-1$ modelli contenenti tutti i predittori escluso uno, considerando quindi $p-1$ predittori;

3. al secondo, avendo escluso il predittore risultato non significativo, si stimano $p - 2$ modelli con $p - 2$ predittori;
4. l'algoritmo continua considerando $p - 3, p - 4, \dots, 1$ modelli con rispettivamente $p - 3, p - 4, \dots, 1$ variabili esplicative.

Al termine viene scelto il miglior risultato tra i p modelli stimati attraverso la determinazione di un criterio di valutazione, in questo caso l'accuracy. Anche stavolta i modelli analizzati sono pari a $1 + \frac{p(p+1)}{2}$, tenendo presente la considerazione effettuata per la Forward Selection in riferimento alla metrica utilizzata. Il difetto di questa procedura è che una volta che la variabile è stata eliminata dal modello, essa non viene più reintrodotta, anche se poi possa risultare statisticamente significativa nel modello finale. [6]

Motivo per il quale una soluzione ottimale potrebbe riguardare un approccio definito ibrido attraverso la quale vengono combinati la selezione forward e quella backward. [18]

Tuttavia, in questo lavoro, ci limiteremo a trattare solo i metodi appena descritti.

2.2 Metodi di Shrinkage

Un ulteriore approccio di feature selection è rappresentato dai metodi di regolarizzazione, ovvero sono una tecnica utilizzata nell'ambito della selezione delle features da inserire nel modello, andando a forzare e *regolarizzare* le stime dei coefficienti ad essere 0. I vincoli imposti alle stime dei coefficienti (diversi a seconda del metodo che il ricercatore utilizza) hanno un impatto sul modello stimato, in quanto la varianza delle stime si riduce e di conseguenza migliora le performance del modello analizzato. [18]

Rispetto ai metodi di *Subset Selection* (dove invece è possibile trarre conclusioni mediante le stime dei coefficienti ed estrarre informazioni utili per il ricercatore), e dati i vincoli imposti, non è possibile ricavare un'accurata interpretazione dei risultati considerato che tali coefficienti stimati risultano penalizzati e di conseguenza valutabili solo attraverso il segno.

Tali metodi di shrinkage sono particolarmente utilizzati quando vi è la presenza di un notevole numero di predittori (talvolta di dimensione $p >> n$) e quindi si vuole utilizzare un set ridotto di variabili esplicative.

I metodi che verranno trattati in questo lavoro sono 3 e sono i seguenti:

1. *Penalizzazione Elastic Net*
2. *Penalizzazione Ridge*
3. *Penalizzazione Lasso*

Tali penalizzazioni sopracitate, inizialmente applicate a modelli di regressione lineare, attualmente sono ampiamente applicate anche alla classe dei Modelli Lineari Generalizzati, tra le quali alla famiglia dei modelli *ELMO*, che sarà approfondita nel prossimo paragrafo.

2.2.1 Penalizzazione Elastic Net

La penalizzazione Elastic Net rappresenta una tecnica di regolarizzazione capace di eseguire simultaneamente la selezione delle variabili del modello ed attuare uno shrinkage continuo, andando a selezionare gruppi di variabili correlate. [36]

Ciò è reso possibile poichè tale penalizzazione rappresenta una somma pesata tra la penalizzazione ridge e penalizzazione lasso, che saranno discusse nei prossimi paragrafi, e di conseguenza condivide con la penalizzazione Lasso la caratteristica di portare le stime dei coefficienti esattamente a 0, andando ad effettuare una selezione dei predittori.

Esso rappresenta una combinazione tra le due penalizzazioni sopracitate, dal momento in cui vi è la specificazione di un parametro definito $0 < \alpha < 1$ (*mixing parameter*), che rappresenta il peso della penalizzazione di tipo lasso rispetto alla penalizzazione di tipo ridge.

Applicabile ai modelli della classe ELMO (considerata la validità o meno dell'assunzione di proporzionalità), ogni modello presenterà le seguenti forme: forma parallela, non parallela e semi parallela, le cui funzioni obiettivo verranno illustrate in seguito.

Sia c il vettore contenente le intercette, sia b_j l'elemento j -esimo di b , dove b rappresenta il vettore dei coefficienti quando risulta verificata l'assunzione di proporzionalità e sia B_{jk} l'elemento corrispondente al coefficiente relativa alla j -esima riga e k -esima colonna della matrice \mathbf{B} , ovvero la matrice contentente $j \times k$ coefficienti quando non è verificata l'assunzione di proporzionalità. Sia $N_+ = \sum_{i=1}^N n_i$ ed $\ell(\cdot)$ la funzione di log-verosimiglianza di ogni modello.

La funzione obiettivo riferita alla penalizzazione elastic-net, per ogni forma di modello, è la seguente:

1. *Forma parallela:*

$$\mathcal{M}(c, b; \alpha, \lambda) = -\frac{1}{N_+} \ell(c, b) + \lambda \sum_{j=1}^p \left(\alpha |b_j| + \frac{1}{2}(1-\alpha)b_j^2 \right) \quad (2.1)$$

2. *Forma non parallela:*

$$\mathcal{M}(c, B; \alpha, \lambda) = -\frac{1}{N_+} \ell(c, B) + \lambda \sum_{j=1}^p \sum_{k=1}^K \left(\alpha |B_{jk}| + \frac{1}{2}(1-\alpha)B_{jk}^2 \right) \quad (2.2)$$

3. *Forma semi parallela:*

$$\begin{aligned} \mathcal{M}(c, b, B; \alpha, \lambda, \rho) = & -\frac{1}{N_+} \ell(c, b, B) + \\ & + \lambda \left(\rho \sum_{j=1}^p \left(\alpha |b_j| + \frac{1}{2}(1-\alpha)b_j^2 \right) + \sum_{j=1}^p \sum_{k=1}^K \left(\alpha |B_{jk}| + \frac{1}{2}(1-\alpha)B_{jk}^2 \right) \right) \end{aligned} \quad (2.3)$$

Il parametro di tuning, nei primi due casi, è rappresentato da λ , ovvero il parametro definito di regolarizzazione ed il cd. *mixing parameter* α .

Individuare il giusto valore di λ può essere computazionalmente costoso ed è per questo che si ricorre a criteri basati su cross-validation [34].

Invece, nella forma definita semi parallela, è necessario determinare anche il valore di ρ che rappresenta il livello di penalizzazione dei termini paralleli:

1. Un valore $\rho \rightarrow \infty$ in 2.3 porta i coefficienti paralleli pari a 0 e ciò equivale ad ottenere un modello non parallelo;

2. un valore $\rho = 0$ in 2.3 lascia i coefficienti paralleli non penalizzati e all'aumentare di λ si tenderà al modello parallelo.

In riferimento alle stime dei coefficienti, viene quindi minimizzata la cd. funzione obiettivo della *verosimiglianza penalizzata* definita come la somma di log-verosimiglianza (negativa) e dei termini di penalizzazione (proporzionale alla norma L_1 e alla norma quadrata L_2 , riferite a Lasso e Ridge) che è funzione dei vettori dei coefficienti. [33] L'algoritmo applicato è il *Coordinate Descent Algorithm*, specificato nel dettaglio nel prossimo paragrafo.

2.2.1.1 Coordinate Descent Algorithm

Esplicitata la funzione obiettivo per la classe dei modelli *ELMO* (proporzionali, non proporzionali e semi-proporzionali) visti nel paragrafo precedente, viene utilizzato il ***Coordinate Descent Algorithm***. [33]

Tale algoritmo è un metodo iterativo utilizzato per risolvere i problemi di ottimizzazione eseguendo una massimizzazione approssimata lungo le direzioni delle coordinate rispetto agli altri coefficienti. [31]

A differenza dell'algoritmo del *Gradient Descent*, dove i coefficienti del modello sono aggiornati insieme contemporaneamente, nell'algoritmo del *Coordinate Descent*, ad ogni iterazione, viene aggiornato un coefficiente, tenendo fissi gli altri. Tale approccio risulta maggiormente indicato per l'ottimizzazione della funzione obiettivo nei metodi di shrinkage perchè risulta semplice, stabile e veloce da un punto di vista computazionale. [32]

A conferma di ciò, è stato verificato che tale metodo rappresenta un valido approccio per risolvere problemi convessi con penalizzazioni L_1 , L_2 oppure penalizzazioni elastic net. [11]

Per quanto riguarda quest'ultima, *Wurl et Al.* hanno proposto il seguente algoritmo, che comprende due cicli: uno interno ed uno esterno.

Il ciclo esterno costruisce una approssimazione quadratica della funzione di log-verosimiglianza grazie all'espansione di Taylor del secondo ordine, mediante la quale viene sostituita la vera funzione di log-verosimiglianza presente nella funzione obiettivo. [33]

Il ciclo interno invece itera le stime di coefficienti, aggiornando ognuno con il valore che ottimizza la funzione obiettivo approssimata.

Affinchè si possa procedere alla costruzione dell'approssimazione quadratica della funzione di log-verosimiglianza, occorre riportare la *funzione score*, l'*Informazione di Fisher* e la *funzione di log-verosimiglianza* stessa che sono così definite:

$$L_i(\pi_i) = \sum_{j=1}^m y_{ij} \log(\pi_{ij}) + y_{i(m+1)} \log\left(1 - \sum_{j=1}^m \pi_{ij}\right). \quad (2.4)$$

dove $L_i(\pi_i)$ rappresenta la funzione di log-verosimiglianza dell'osservazione i -esima con probabilità π_i .

Assumendo $h = g^{-1}$ ovvero l'inversa della link function considerata, la funzione di log-verosimiglianza può essere scritta in funzione di β , che in questo caso rappresenta il vettore dei coefficienti stimati per qualsiasi modello considerato, tra parallelo, non parallelo e semi parallelo:

$$\ell_i(\beta) = L_i(h(X_i\beta)). \quad (2.5)$$

La funzione score, nonché la derivata della funzione di log-verosimiglianza rispetto ai parametri, può essere ottenuta mediante la *chain rule* ed è esplicitata come segue:

$$U_i(\beta) = \nabla \ell_i(\beta) = X_i^\top Dh(\eta_i)^\top \nabla L_i(p_i) = X_i^\top W_i(z_i - X_i\beta) \quad (2.6)$$

dove:

$$z_i = W_i^{-1} \left(Dh(\eta_i)^\top \nabla L_i(p_i) \right) + X_i\beta \quad (2.7)$$

$$\nabla L_i(p_i) = \left(\frac{y_{i1}}{p_{i1}}, \dots, \frac{y_{im}}{p_{im}} \right)^\top - \left(\frac{y_{i(m+1)}}{p_{i(m+1)}} \right) \cdot \mathbb{1} \quad (2.8)$$

dove $Dh(\cdot)$ rappresenta il *Jacobiano* dell'inversa della link function ed η rappresenta il vettore delle combinazioni tra X_i^T e β , ovvero $\eta_i = X_i^T \beta$

La matrice relativa all'informazione di Fisher risulta essere pari a:

$$\mathcal{I}_i(\beta) = E_\beta (U_i(\beta) U_i(\beta)^\top) = X_i^\top W_i X_i, \quad (2.9)$$

dove:

$$W_i = Dh(\eta_i)^\top \Sigma_i^{-1} Dh(\eta_i) \quad (2.10)$$

$$\Sigma_i^{-1} = n_i \left([\text{diag}(p_i)]^{-1} + \frac{1}{p_{i(m+1)}} \cdot \mathbb{1}^\top \right) \quad (2.11)$$

Essendo tutte le y_i indipendenti, si ha:

$$\ell(\beta) = \sum_{i=1}^N \ell_i(\beta) \quad (2.12)$$

$$U(\beta) = \sum_{i=1}^N U_i(\beta) = X^\top W(z - X\beta) \quad (2.13)$$

$$\mathcal{I}(\beta) = \sum_{i=1}^N \mathcal{I}_i(\beta) = X^\top W. \quad (2.14)$$

dove X risulta essere un vettore contenente tutte le variabili esplicative nelle varie forme già citate, z un vettore esplicitato in 2.7 e W una matrice diagonale che presenta N valori di W_i derivanti dal Jacobiano ed illustrati in 2.10. Si ottiene così l'approssimazione quadratica della funzione di verosimiglianza $\ell(\beta)$ come somma ponderata della funzione presentata di seguito:

$$\ell^{(r)}(\beta) = -\frac{1}{2} (z^{(r)} - X\beta)^\top W^{(r)} (z^{(r)} - X\beta) \quad (2.15)$$

dove r identifica l'iterazione r -esima del ciclo esterno in modo tale da ottenere le stime di $\hat{\beta}^{(r)}$

Nel ciclo interno lo scopo è massimizzare tale verosimiglianza che è definita penalizzata e dal momento in cui, così come già specificato per i modelli presentati nel capitolo precedente, non è possibile eseguire la massimizzazione attraverso i minimi quadrati ponderati, è necessario ricorrere ad altri approcci.

Ciò perchè, avendo stime dei coefficienti penalizzate, non è possibile ottenere le derivate parziale della funzione per tutti i coefficienti e quindi si ricorre al *Coordinate Descent Algorithm*, attraverso il quale vengono aggiornate le stime

dei coefficienti mediante la funzione obiettivo marginale. [33]

Considerata l'approssimazione quadratica del ciclo esterno e la verosimiglianza ottenuta in 2.15, viene definita una nuova funzione obiettivo $\mathcal{M}^{(r)}(\beta)$ ovvero la funzione obiettivo in 2.1, 2.2 e 2.3 a seconda della verifica dell'assunzione di proporzionalità, dove ad $\ell(\beta)$ viene sostituito $\ell^{(r)}(\beta)$, ricordando che β in questo caso rappresenta l'intero vettore di parametri da stimare.

Sia $\hat{\beta}_j(r, s)$ il valore di β_j stimato all' s -esima iterazione del ciclo interno riferito all' r -esima iterazione del ciclo esterno. Quindi:

$$\mathcal{M}_j^{(r,s)}(t) = \mathcal{M}^{(r)}\left(\hat{\beta}_1^{(r,s+1)}, \dots, \hat{\beta}_{j-1}^{(r,s+1)}, t, \hat{\beta}_{j+1}^{(r,s)}, \dots, \hat{\beta}_Q^{(r,s)}\right) \quad (2.16)$$

dove $\mathcal{M}^{(r)}$ rappresenta la funzione obiettivo marginale riferita al j -esimo coefficiente dato che l'iterazione avviene singolarmente per ogni coefficiente stimato posto gli altri coefficienti non iterati. [32]

Per $t \neq 0$:

$$\frac{d}{dt} \mathcal{M}_j^{(r,s)}(t) = -\frac{1}{N_+} X_{*j}^T W^{(r)} \left(z^{(r)} - X_{*-j} \hat{\beta}_{-j}^{(r,s)} - t X_{*j} \right) + \lambda(\alpha \cdot \text{sign}(t) + (1-\alpha) \cdot t) \quad (2.17)$$

dove X_{*j} indica la j -esima colonna della matrice delle covariate X ed X_{*-j} indica la stessa matrice alla quale viene rimossa la j -esima colonna e $\hat{\beta}_{-j}^{(r,s)} = \left(\hat{\beta}_1^{(r,s+1)}, \dots, \hat{\beta}_{j-1}^{(r,s+1)}, \hat{\beta}_{j+1}^{(r,s)}, \dots, \hat{\beta}_Q^{(r,s)} \right)$ dal momento in cui la j -esima colonna viene rimossa.

In questo caso, la funzione obiettivo risulta convessa ed è monotona crescente per tutti i valori di $t \neq 0$; nel punto $t = 0$ essa risulta essere $2\lambda\alpha\omega_j$, dove ω_j è un vettore che indica se e quali termini debbano essere penalizzati.

Può accadere quindi che se:

$$\left| \frac{1}{N_+} X_{*j}^T W^{(r)} (z^{(r)} - X_{*-j} \hat{\beta}_{-j}^{(r,s)} - t X_{*j}) \right| > \lambda\alpha\omega_j \quad \rightarrow \quad \begin{cases} \frac{d}{dt} \mathcal{M}_j^{(r,s)}(t) = 0 \\ \text{altrimenti cambia segno} \end{cases} \quad (2.18)$$

Quindi, il *Coordinate Descent Algorithm* aggiornato al passo $s + 1$ riferito al ciclo esterno può essere scritto nel modo seguente:

$$\hat{\beta}_j^{(r,s+1)} = \frac{S\left(\frac{1}{N_+} X_{*j}^\top W^{(r)} \left(z^{(r)} - X_{*-j} \hat{\beta}_{-j}^{(r,s)}\right), \lambda \alpha \omega_j\right)}{\frac{1}{N_+} X_{*j}^\top W^{(r)} X_{*j} + \lambda(1-\alpha)\omega_j} \quad (2.19)$$

dove $S(x, y) = sign(x)(|x| - y)_+$ è definito *operatore soft-thresholding*.

Tale operatore, usato in diverse tecniche di regolarizzazione (tra cui Lasso Regression), viene usato per ottimizzare la funzione obiettivo imponendo appunto un valore soglia (*threshold*), dove l'operatore restituisce valore 0 se il valore di x è inferiore al threshold, altrimenti il valore viene modificato aggiungendo o sottraendo il threshold stesso. [11] [12]

Da un punto di vista computazionale, risulta fondamentale scegliere dei validi criteri che riescano a portare a convergenza l'algoritmo. Questi criteri riguardano la scelta della sequenza dei termini di penalizzazione ed i valori di partenza per la procedura, così come la cd *stopping rule*.

Per la scelta della sequenza iniziale dei parametri, viene calcolata la verosimiglianza non penalizzata di un modello contenente solo i termini non penalizzati, si calcola l'approssimazione quadratica illustrata in 2.15 e si considera come valore di soglia più grande definito λ_{max} il più piccolo threshold dei coefficienti penalizzati grazie alla quale l'algoritmo continua ad iterare.

I valori iniziali sono scelti mediante la tecnica definita *warm start*. Con tale tecnica, in sostanza, viene usata una soluzione *ottimale* in un problema di ottimizzazione per trovare il valore iniziale per l'algoritmo che si sta utilizzando.

In riferimento ai metodi di shrinkage, data una sequenza di λ , ad ogni valore tale tecnica usa le stime dei $\hat{\beta}$ ottenute dal valore di λ fissato e le usa come valore iniziale per il prossimo valore di λ della sequenza. Ciò rende l'algoritmo più leggero computazionalmente perchè non risulta necessaria aggiornare tutte le stime dei coefficienti nel ciclo interno dell'algoritmo, in quanto essi sono già portati a 0. I valori ottenuti sono definiti in un sottoinsieme specificato come *set attivo*.

I coefficienti stimati, invece, che hanno valori diversi da 0 passano per il ciclo esterno:

1. se l'aggiornamento nell'algoritmo esterno è 0 per tutti i coefficienti il cui

valore era diverso da 0, l'algoritmo si arresta perchè giunto a soluzione ottimale

2. altrimenti, se il valore dei coefficienti dovesse cambiare e non essere più nullo, essi vengono aggiunti al set attivo e viene ripetuto il ciclo.

Per i criteri di stop, vengono definiti due diversi criteri usando la variazione relativa della funzione obiettivo in riferimento al ciclo esterno e a quello interno. In riferimento al ciclo esterno:

$$\left| \frac{\mathcal{M}(\hat{\beta}^{(r)}) - \mathcal{M}(\hat{\beta}^{(r-1)})}{\mathcal{M}(\hat{\beta}^{(r-1)})} \right| < \epsilon_{\text{out}} \quad (2.20)$$

si raggiungerà la convergenza se il risultato sarà più piccolo di un numero infinitesimale ϵ .

Invece, per il ciclo interno:

$$\left| \frac{\mathcal{M}^{(r)}(\hat{\beta}^{(r,s)}) - \mathcal{M}^{(r)}(\hat{\beta}^{(r,s-1)})}{\mathcal{M}^{(r)}(\hat{\beta}^{(r,s-1)})} \right| < \epsilon_{\text{in}} \quad (2.21)$$

l'algoritmo raggiungerà convergenza se la funzione obiettivo relativa al valore di β_j stimato all' s -esima iterazione del ciclo interno riferito all' r -esima iterazione del ciclo esterno rispetto alla funzione obiettivo relativa al valore di β_j stimato all'iterazione $s - 1$ sempre al passo r del ciclo esterno è minore di ϵ .

2.2.2 Penalizzazione Ridge

Un secondo approccio per la feature selection per i modelli ordinali risulta essere la penalizzazione di tipo ridge. Inizialmente applicata ai modelli di tipo lineare, nel 1984 e nel 1986 rispettivamente *Schaefer et Al.* e *Schaefer* hanno discusso la penalizzazione di tipo ridge a modelli di tipo logistico con risposte di tipo binario e nel 1991 *Nyquist* ha considerato tale approccio nell'ambito dei GLM. [34]

Tale penalizzazione rappresenta una tecnica di regolarizzazione mediante la quale viene considerato un termine di penalizzazazione proporzionale alla norma quadratica di L_2 del vettore dei coefficienti, riducendo così la loro varianza e limitando

l'impatto delle variabili più instabili.

Questo termine di penalizzazione ha come effetto principale quello di «*strizzare*» i coefficienti dei regressori, spingendoli verso lo zero senza mai annullarli completamente. [18]

Considerata la penalizzazione elastic net, quella ridge risulta applicabile, nel caso di modelli ELMO, andando ad impostare il *mixing parameter* della funzione obiettivo pari a 0 rispettivamente per la forma parallela in 2.1, per la forma non parallela in 2.2 e in quella semi-parallela in 2.3.

In conclusione, usando la stessa notazione espressa nel paragrafo precedente, la funzione obiettivo riferita alla penalizzazione Ridge, per ogni forma di modello, è la seguente:

1. Forma parallela:

$$\mathcal{M}(c, b; \lambda) = -\frac{1}{N_+} \ell(c, b) + \frac{1}{2} \lambda \sum_{j=1}^p b_j^2 \quad (2.22)$$

2. Forma non parallela:

$$\mathcal{M}(c, B; \lambda) = -\frac{1}{N_+} \ell(c, b) + \frac{1}{2} \lambda \sum_{j=1}^p \sum_{k=1}^K B_{jk}^2 \quad (2.23)$$

3. Forma semi parallela:

$$\mathcal{M}(c, b, B; \lambda, \rho) = -\frac{1}{N_+} \ell(c, b, B) + \lambda \left(\frac{\rho}{2} \sum_{j=1}^p b_j^2 \right) + \frac{1}{2} \sum_{j=1}^p \sum_{k=1}^K B_{jk}^2 \quad (2.24)$$

Il parametro di tuning risulta sempre λ , ovvero il parametro definito di regolarizzazione, mentre ρ nella forma definita semi parallela rappresenta il livello di penalizzazione dei termini paralleli.

L'algoritmo utilizzato per derivare la funzione obiettivo è il medesimo utilizzato per l'elastic net ovvero il *Coordinate Descent Algorithm*, prendendo atto di tutte le considerazioni effettuate nel paragrafo precedente. [11]

2.2.3 Penalizzazione Lasso

Uno dei difetti della penalizzazione ridge è che, non forzando le stime dei coefficienti esattamente a 0, il modello risultante che si ottiene non è un modello parsimonioso, includendo nel modello i p predittori considerati inizialmente. [27] La Regressione Lasso (acronimo di «*Least Absolute Shrinkage and Selection Operator*»), formulato nel 1996 da *Tibshirani* e poi adattato alla classe dei GLM dal 2007 da *Park et Al.*, consente di ovviare a tale problema imponendo un termine di penalizzazione differente, che è proporzionale alla norma L_1 delle vettore dei coefficienti, andando a «*strizzare*» alcune stime dei coefficienti e forzandone altre ad essere esattamente 0. [33] [34]

Partendo dalle funzioni obiettivo esplicitate per l'elastic net, si ottiene una penalizzazione di tipo LASSO impostando un valore di $\alpha = 1$, andando quindi ad annullare la parte relativa alla penalizzazione ridge.

A seconda della verifica della assunzione di proporzionalità, non assunzione ed assunzione parziale, le funzioni obiettivo saranno le seguenti:

1. Forma parallela:

$$\mathcal{M}(c, b; \lambda) = -\frac{1}{N_+} \ell(c, b) + \lambda \sum_{j=1}^p |b_j| \quad (2.25)$$

2. Forma non parallela:

$$\mathcal{M}(c, B; \lambda) = -\frac{1}{N_+} \ell(c, B) + \lambda \sum_{j=1}^p \sum_{k=1}^K |B_{jk}| \quad (2.26)$$

3. Forma semi parallela:

$$\mathcal{M}(c, b, B; \lambda, \rho) = -\frac{1}{N_+} \ell(c, b, B) + \lambda \left(\rho \sum_{j=1}^p |b_j| \right) + \sum_{j=1}^p \sum_{k=1}^K |B_{jk}| \quad (2.27)$$

Anche nella selezione dei predittori con metodo LASSO, al fine di ottimizzare la funzione obiettivo, per la stima dei parametri viene applicato il *Coordinate Descent Algorithm*. [11]

2.3 Riduzione della dimensionalità

Rispetto ai due metodi menzionati precedentemente, il seguente paragrafo tratta un approccio di riduzione della dimensionalità attraverso la quale, anzichè utilizzare nel modello i predittori originali X_1, X_2, \dots, X_P , viene definita una trasformazione di tali variabili esplicative mantenendo le informazioni più rilevanti. [18]

Ciò è particolarmente utile quando all'interno del set di dati sono presenti un numero elevato di variabili e che, di conseguenza, alcune di esse possano essere ridondanti, rumorose e poco informative. Anche per quanto concerne i tempi di calcolo, rispetto alla stima di un modello con i p predittori iniziali, tale processo risulta gestire più efficacemente la complessità dei dati e a migliorare l'efficienza computazionale.

2.3.1 Analisi delle corrispondenze

A partire dagli anni 60 con *J.P.Benzècri*, al fine di analizzare dati di natura categorica, viene formalizzato il nome *Analisi delle corrispondenze*. [13]

Essa viene talvolta definita un tipo di *analisi su componenti principali* (Principal Components Analysis, **PCA**) su dati categorici, dove viene considerata però la definizione geometrica di PCA piuttosto che quella prettamente statistica. [14] Entrambi i metodi cercano di identificare le dimensioni in grado di spiegare e catturare la massima percentuale di inerzia, ovvero la massima variabilità nella struttura dei dati.

Per quanto riguarda la notazione, considerando una cd. *tabella incrociata* come si vedrà più nel dettaglio nel prossimo paragrafo, indichiamo con I e J rispettivamente le righe e le colonne della tabella e con N il numero totale delle osservazioni (il generico elemento n_{ij} rappresenta il numero di osservazioni della categoria I rilevati nella variabile J). Si considera quindi una nuova matrice delle frequenze relative P , dove il generico elemento p_{ij} è ottenuto rapportando ciascuna cella n_{ij} considerata precedentemente rispetto al totale n . [14]

Sommando gli elementi ad ogni riga otterremo i marginali di riga $p_{i\cdot} = \frac{n_{i\cdot}}{n}$ (indi-

cati con r_i) e sommando per colonna si avranno i marginali di colonna $p_{\cdot j} = \frac{n_{\cdot j}}{n}$ (indicati con c_j); questi elementi prendono il nome di *masse*, utili per centrare e normalizzare la matrice P . Sotto ipotesi nulla di indipendenza tra i caratteri, si avrà che:

$$p_{ij} = r_i c_j \quad (2.28)$$

per cui, la centratura e la normalizzazione sarà data da:

$$\tilde{p}_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} \quad (2.29)$$

dove \tilde{p}_{ij} rappresentano i residui standardizzati.

In forma matriciale:

$$\tilde{\mathbf{P}} = \mathbf{D}_r^{-\frac{1}{2}} (\mathbf{P} - \mathbf{rc}') \mathbf{D}_c^{-\frac{1}{2}} \quad (2.30)$$

dove \mathbf{r} e \mathbf{c} rappresentano le masse e \mathbf{D}_r e \mathbf{D}_c le matrici diagonali delle masse r e c .

Usando la decomposizione in valori singolari (*Singular Value Decomposition, SVD*), si ottiene:

$$\tilde{\mathbf{P}} = \mathbf{U} \Sigma \mathbf{V}' \quad (2.31)$$

dove Σ è una matrice diagonale contenente, in ordine decrescente, i valori singolari sulla diagonale principale $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{\tilde{P}} > 0$, mentre \mathbf{U} e \mathbf{V} sono rispettivamente i vettori singolari di sinistra e di destra e sono ortonormali, ovvero $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$.

Considerata la SVD, si ottengono le coordinate per le righe e per le colonne:

$$\mathbf{A} = \mathbf{D}_r^{-\frac{1}{2}} \mathbf{U} \quad (2.32)$$

$$\mathbf{B} = \mathbf{D}_c^{-\frac{1}{2}} \mathbf{V} \quad (2.33)$$

Partendo dall'equazione in 2.29, l'inerzia totale è rappresentata dalla somma

degli elementi \tilde{p}_{ij} al quadrato, corrispondente a $\sum_i \sum_j \tilde{p}_{ij}^2 = \text{trace}(\tilde{P}\tilde{P}')$

2.3.1.1 Analisi delle corrispondenze su tabelle concatenate

L'analisi delle corrispondenze può essere derivata e presentata in diversi modi: il metodo utilizzato in questo lavoro di tesi si riferisce all'*analisi delle corrispondenze su tabelle concatenate* dove, all'interno di ogni cella della stessa, vengono registrate le osservazioni che hanno una determinata caratteristica. [30]

Si consideri una matrice \mathbf{Z}_X di dimensioni $n \times Q$, dove n rappresenta il numero di osservazioni e Q rappresenta la somma delle q categorie delle p variabili prese in considerazione, dove $Q = \sum_{j=1}^p q_j$, e sia \mathbf{Z}_Y una matrice di dimensioni $n \times k$ dove k sono il numero di categorie della variabile dipendente Y . [30]

Si costruisce così una tabella concatenata indicata con $\mathbf{F} = \mathbf{Z}'_Y \mathbf{Z}_X$ di dimensioni $k \times Q$. Rispetto alla notazione introdotta precedentemente, si indichi:

$$\mathbf{P} = \frac{1}{np} \mathbf{F} \quad (2.34)$$

e considerando $\mathbf{P} - \mathbf{r}\mathbf{c}'$ in 2.30, si ha:

$$\mathbf{P} - \mathbf{P}\mathbf{1}_Q\mathbf{1}'_K\mathbf{P} = \frac{1}{np} \left(\mathbf{F} - \frac{1}{np} \mathbf{F}\mathbf{1}_Q\mathbf{1}'_K\mathbf{F} \right) = \frac{1}{np} \left(\mathbf{Z}'_Y \mathbf{Z}_X - \frac{1}{n} \mathbf{Z}'_Y \mathbf{1}_n \mathbf{1}'_n \mathbf{Z}_X \right) = \frac{1}{np} \mathbf{Z}'_Y \mathbf{M} \mathbf{Z}_X, \quad (2.35)$$

dove $\mathbf{M} = \mathbf{I}_n - \frac{\mathbf{1}_n \mathbf{1}'_n}{n}$. [30]

Definita la matrice diagonale \mathbf{D}_X come $\mathbf{D}_X \mathbf{1}_Q = \mathbf{Z}'_X \mathbf{1}_n$, la funzione obiettivo dell'analisi delle corrispondenze è:

$$\max \phi_{ca} (\mathbf{Z}_Y, \mathbf{B}) = \frac{1}{np^2} \text{trace} \mathbf{B}' \mathbf{Z}'_X \mathbf{M} \mathbf{Z}_Y \mathbf{D}_Y^{-1} \mathbf{Z}'_Y \mathbf{M} \mathbf{Z}_X \mathbf{B} \quad (2.36)$$

soggetto al vincolo che:

$$\frac{1}{np} \mathbf{B}' \mathbf{D}_X \mathbf{B} = \mathbf{I}_k \quad (2.37)$$

Il parametro di tuning è individuato da k , ovvero il numero di dimensioni relativo all'analisi condotta. Il valore di k deve essere più piccolo del rango della matrice \mathbf{F} e deve essere minore o uguale rispetto al minimo di $(K - 1, Q - 1)$.

[30]

Considerato, in questo caso, che $K \ll n$, esso deve essere più piccolo del numero di categorie della variabile dipendente Y .

Definendo $\mathbf{B}^* = \frac{1}{\sqrt{np}} \mathbf{D}_X^{1/2} \mathbf{B}$, la funzione obiettivo in 2.36 è riscritta nel seguente modo:

$$\max \phi_{ca} (\mathbf{Z}_Y, \mathbf{B}^*) = \frac{1}{p} \text{trace } \mathbf{B}^{*'} \mathbf{D}_X^{-\frac{1}{2}} \mathbf{Z}'_X \mathbf{M} \mathbf{Z}_Y \mathbf{D}_Y^{-1} \mathbf{Z}'_Y \mathbf{M} \mathbf{Z}_X \mathbf{D}_X^{-\frac{1}{2}} \mathbf{B}^* \quad (2.38)$$

in questo caso, soggetto al vincolo che:

$$\mathbf{B}^{*'} \mathbf{B}^* = \mathbf{I}_Y \quad (2.39)$$

La soluzione di $\mathbf{B}^{*'} \mathbf{B}^*$ è ottenuta mediante la scomposizione degli autovalori (*eigenvalue decomposition*):

$$\frac{1}{p} \mathbf{D}_X^{-1/2} \mathbf{Z}'_X \mathbf{M} \mathbf{Z}_Y \mathbf{D}_Y^{-1} \mathbf{Z}'_Y \mathbf{M} \mathbf{Z}_X \mathbf{D}_X^{-1/2} = \mathbf{B}^* \boldsymbol{\Lambda}^2 \mathbf{B}^{*'} \quad (2.40)$$

da cui si ricava la soluzione per \mathbf{B} :

$$B = \sqrt{np} \mathbf{D}_X^{-1/2} \mathbf{B}^* \quad (2.41)$$

Le coordinate delle k dimensioni identificate (chiamati anche *scores*) sono dati da:

$$\mathbf{W} = \sqrt{\frac{n}{p}} \mathbf{M} \mathbf{Z}_X \mathbf{D}_X^{-\frac{1}{2}} \mathbf{B}^* \quad (2.42)$$

Determinati gli scores, si procederà alla stima dei seguenti modelli:

1. *Proportional Odds Model*

$$\text{logit}[P(Y \leq j|x)] = \tau_j - \boldsymbol{\beta}' \mathbf{w}, \quad j = 1, \dots, m-1 \quad (2.43)$$

2. *Adjacent Category Model*

$$\text{logit}[P(Y = j \mid Y = j + 1, \mathbf{x})] = \tau_j - \boldsymbol{\beta}' \mathbf{w}, \quad j = 1, \dots, m - 1 \quad (2.44)$$

3. *Continuation Ratio Model*

$$\text{logit}[P(Y = j \mid Y \geq j, \mathbf{x})] = \tau_j - \boldsymbol{\beta}' \mathbf{w}, \quad j = 1, \dots, m - 1 \quad (2.45)$$

dove w saranno le componenti derivate dall'analisi delle corrispondenze appena enunciata.

Capitolo 3

Implementazione degli algoritmi

3.1 Raccolta dei dati

L'indagine in questione è stata condotta da me, Rosario Urso, e dal mio collega Antonio Cola, presso il LASt¹ volto a comprendere la percezione degli individui (paura, ansia, preoccupazione, ecc) in relazione agli eventi sismici a seguito dell'intensa attività sismica dei **Campi Flegrei** riferita all'anno 2023.

La ricerca è stata condotta poco dopo l'inizio dell'attività sismica nell'area dei Campi Flegrei, somministrando l'indagine dal 31 ottobre al 16 novembre 2023 mediante tecnica di campionamento a valanga, campione selezionato all'interno del territorio italiano.

Le domande dell'indagine vertono su vari aspetti socio-demografici, oltre che a domande specifiche su percezione, mobilità, informazione, fiducia nelle istituzioni e aspetti economico-finanziari legati agli eventi sismici. Questo approccio ha permesso un'analisi complessa e dettagliata delle reazioni e degli atteggiamenti della popolazione in relazione agli eventi sismici.

I risultati analizzati riflettono una vasta gamma di percezioni e reazioni che variano significativamente in base a diversi fattori come l'età, il sesso e altri parametri

¹Laboratory for Statistical Data Analysis (LASt) - Università degli Studi di Napoli Federico II - Direttore scientifico: Maria Iannario - Domenico Vistocco

Attività principali: Consulenza individuale per la ricerca, formazione, strumenti di e-learning e di management system. Gestione degli aspetti legati alla ricerca e alla pubblicazione dei risultati delle ricerche condotte attraverso il Centro; attività di ricerca e consulenza (compresa la formazione) in materia di analisi e visualizzazione dei dati.

che verranno illustrati nel dettaglio nei prossimi paragrafi.

L’analisi di questi dati offre una visione chiara ed approfondita delle preoccupazioni, delle paure e delle aspettative delle persone riguardo agli eventi sismici, evidenziando quanto sia indispensabile un intervento che abbracci tutti gli individui coinvolti sia per quanto riguarda la trasmissione di informazioni che per quanto concerne l’attuazione di provvedimenti atti a gestire l’emergenza.

Obiettivo principale di questa analisi è quindi quello di identificare e fornire efficaci ed efficienti indicazioni alle autorità e alle istituzioni utili per la gestione delle emergenze, in riferimento alla zona dei Campi Flegrei e a tutte quelle aree soggette a fenomeni sismici.

Il dataset² oggetto di analisi consta di **433** osservazioni ed è composto da **33** variabili, sia di tipo qualitativo che quantitative, discrete che continue. L’intero dataset è visionabile scansionando il *qr-code* presenti nell’appendice Sezione D. Allo scopo di verificare la validità ed il comportamento degli algoritmi di selezione (che rappresenta il fulcro centrale di questo progetto di tesi), sono state create due nuove variabili utili a sintetizzare la batteria relativa all’*informazione* e alla *fiducia*.

Le 2 nuove variabili sono state così riportate:

1. *batteria_fiducia*, che sintetizza la batteria relativa al livello di fiducia verso le istituzioni comunali, regionali, nazionali, INGV e livello di sicurezza nelle strutture pubbliche, così codificata:
 - *Alta_Info*, se, ad ogni rispondente, la somma dei punteggi è maggiore di 15;
 - *Bassa_Info*, altrimenti.
2. *batteria_info*, che sintetizza la batteria relativa al livello di utilità delle informazioni diffuse da radio, TV, social network, giornali e da app varie, dove si avrà:

²Dal dataset, che inizialmente conteneva 472 osservazioni e 47 variabili, data la presenza di diversi dati mancanti, sono state rimosse le variabili *orientamento_politico* e *ral*, ed infine sono stati rimossi 39 dati mancanti relativi alla variabile *disabili*.

- *Alta_Fiducia*, se, ad ogni rispondente, la somma dei punteggi è maggiore di 15;
- *Bassa_Fiducia*, altrimenti.

Tutte le variabili dell'intero set di dati sono elencate in Appendice Sezione A, dove è presente anche la relativa codifica utilizzata.

Inoltre, si specifica che, dal momento che l'analisi in questione è stata focalizzata tenendo in considerazione l'area dei Campi Flegrei, la variabile *residenza* è stata dicotomizzata in due categorie³:

- *Zona a Rischio*, in riferimento alle zone definite «gialle» e «rosse»;
- *Zona Bianca*, tutte le altre.

Ai fini dell'analisi, la variabile dipendente è rappresentata da *paura*, variabile qualitativa ordinale a 5 categorie, dove con 1 il rispondente indica un livello di paura molto basso in riferimento agli eventi sismici, mentre con 5 viene specificato un livello di paura molto alto.

Di seguito, verrà presentata l'analisi esplorativa effettuata sul set di dati a disposizione.

3.2 Analisi Esplorativa

Graficamente, la variabile di risposta *paura* che assume $m = 5$ categorie di risposta è così rappresentata:

³La categorizzazione è avvenuta considerando le mappe del piano nazionale di emergenza nell'area flegrea individuate nel Decreto del presidente del Consiglio dei ministri del 24 giugno 2016 presente sul sito del Dipartimento della Protezione Civile.

[https://mappe.protezionecivile.gov.it/it/mappe-e-dashboards-rischi/
piano-nazionale-campi-flegrei/](https://mappe.protezionecivile.gov.it/it/mappe-e-dashboards-rischi/piano-nazionale-campi-flegrei/)

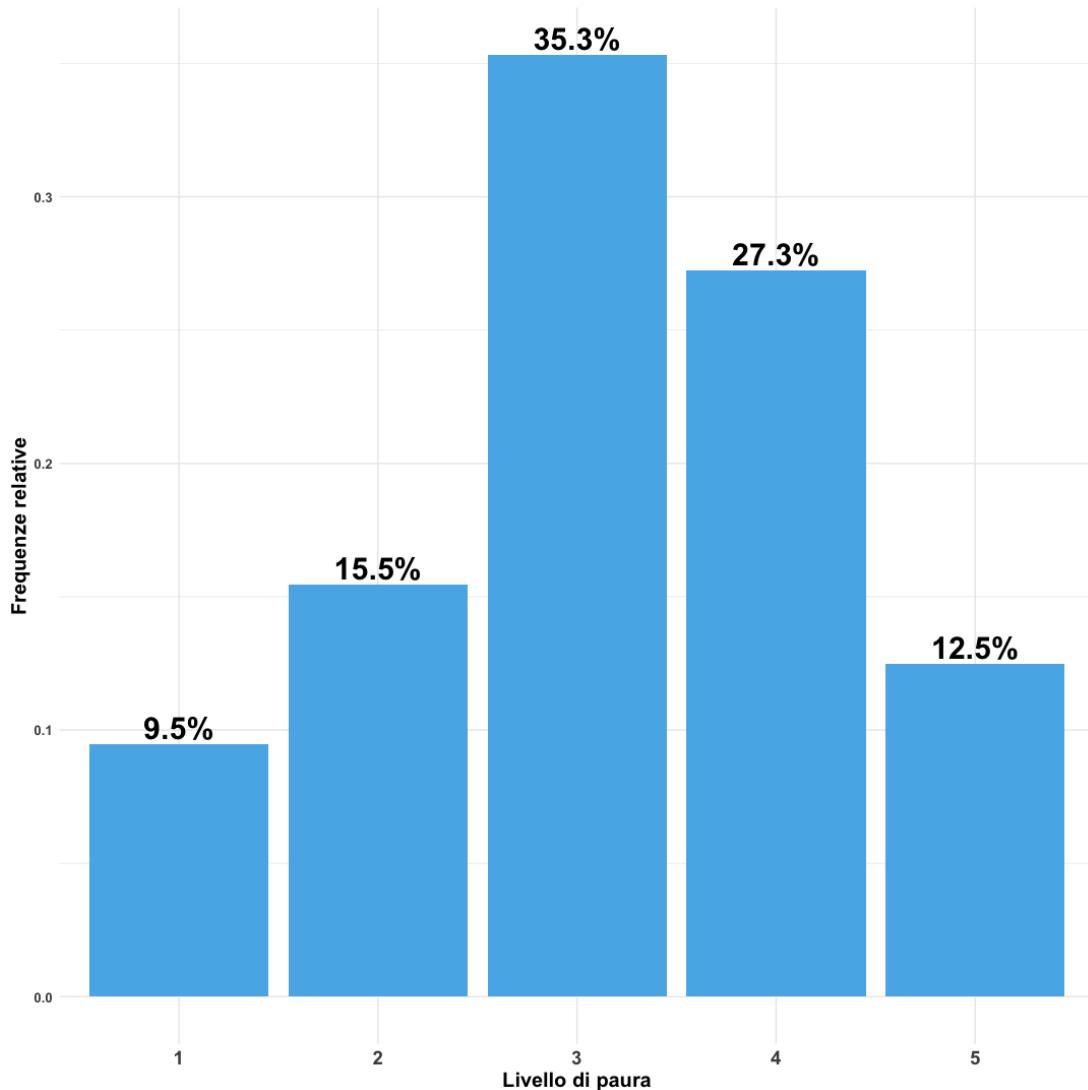


Figura 3.1: Distribuzione della variabile dipendente

Dal grafico in Figura 3.1 emerge che la distribuzione della variabile dipendente risulta essere leggermente asimmetrica negativa con picco sulla categoria 3, che rappresenta il livello medio di paura. Si può inoltre aggiungere che circa il 40% dei rispondenti ha un livello di paura piuttosto alto in relazione agli eventi sismici (si osserva altresì che più di 1 rispondente su 10 ha un livello di paura molto alto). Tuttavia, si è ritenuto opportuno osservare il comportamento della variabile dipendente rispetto al genere del rispondente ed emerge quanto segue:

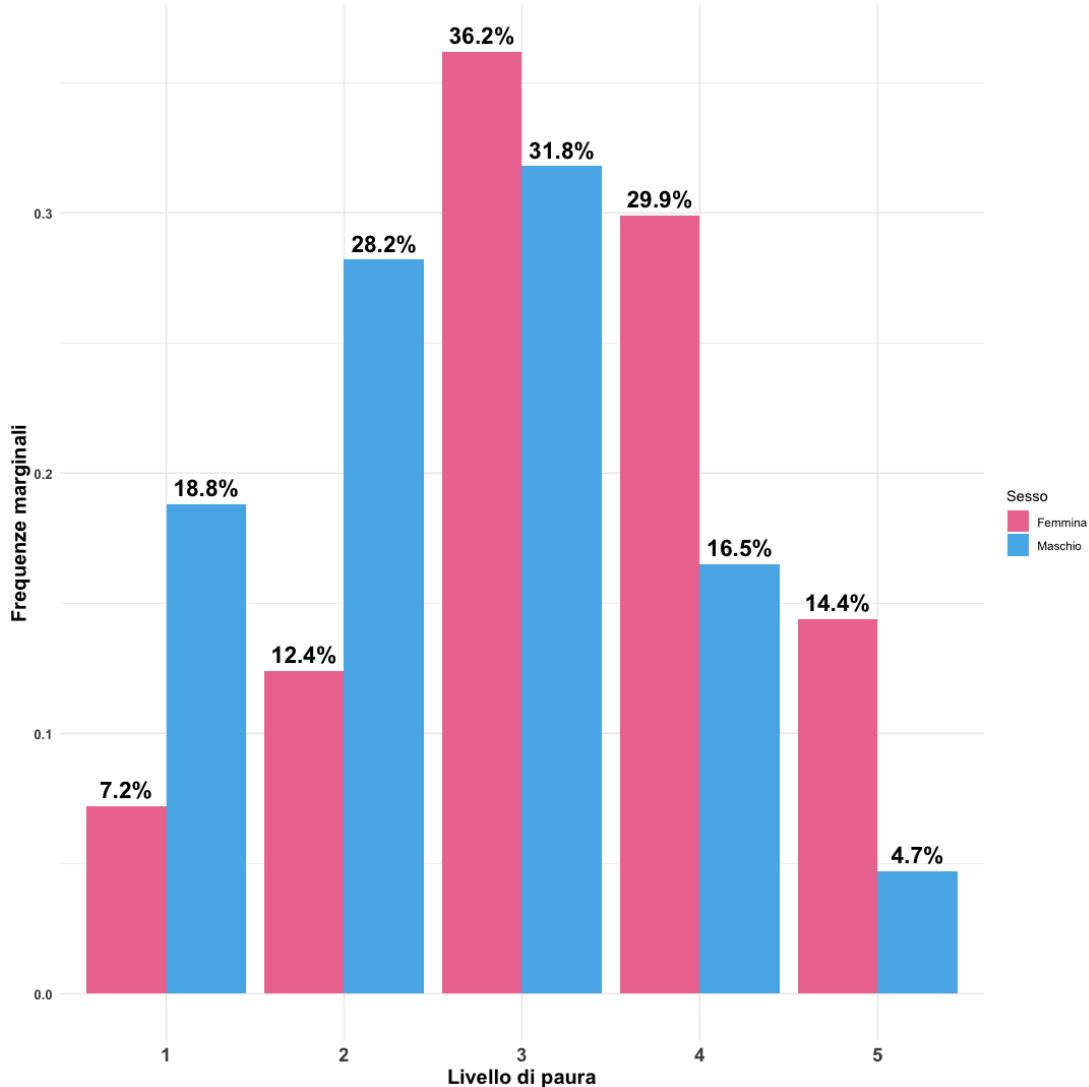


Figura 3.2: Distribuzione della variabile dipendente rispetto al genere

Considerato che la numerosità dei rispondenti di sesso femminile è maggiore della numerosità dei rispondenti di sesso maschile, è stato opportuno analizzare le *frequenze marginali* ottenute rapportando il numero di donne/uomini delle varie categorie della variabile di risposta rispetto al totale di ogni genere.

Dal grafico in questione risulta che i rispondenti di sesso femminile tendono ad avere un livello di paura molto più alto rispetto agli individui di sesso maschile e ciò lo si può evincere dal grafico in prossimità delle categorie 4 e 5, classi in cui è presente circa il 45% delle donne rispetto ad un circa 20% degli uomini, così come risulta dalla Figura 3.2. Inoltre, a conferma di ciò, si evidenza come, dalle prime categorie, emerga che gli uomini rilevino un livello più basso di paura rispetto

alle donne.

La medesima analisi è stata svolta in Figura 3.3 rispetto all'età dei singoli rispondenti, attraverso l'ausilio del *Raincloud Plot*, che rappresenta uno strumento di visualizzazione dei dati intuitivo e robusto.

Da tale rappresentazione grafica, si può desumere come l'età mediana aumenti all'aumentare del livello di paura in riferimento ai fenomeni sismici. Di fatti, anche rispetto alla media dei rispondenti (circa 39 anni), si osserva come la popolazione rappresentata dai giovani rilevi un livello di paura inferiore rispetto ai più anziani (così come evidenziabile dalla categoria 0 ed 1).

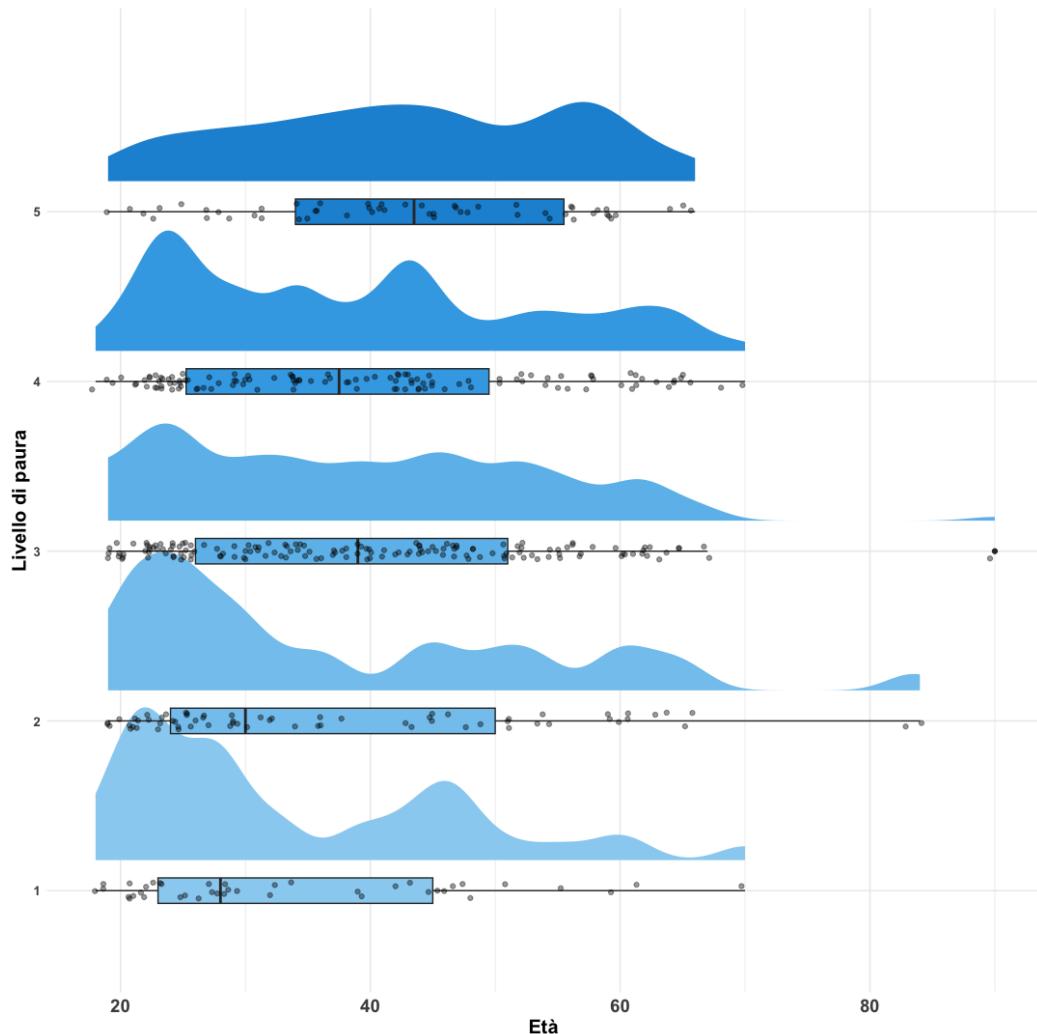


Figura 3.3: Raincloud plot della variabile dipendente rispetto all'età

Infine, in merito alla sezione relativa ai commenti presente all'interno del questionario, è stato effettuato un *Wordcloud* per comprendere le parole più frequenti in tale sezione.

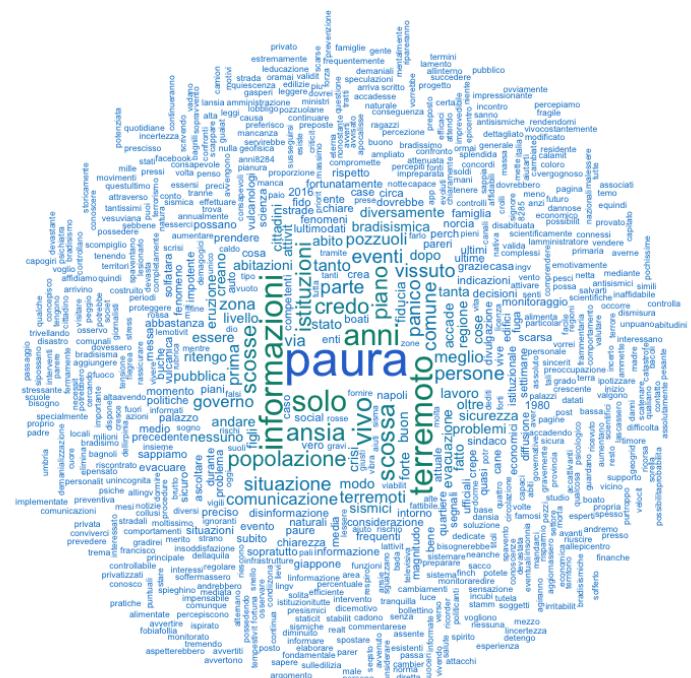


Figura 3.4: Wordcloud dei commenti

La parola che maggiormente ricorre è **paura**, seguita da *informazioni*, *terremoto* ed *ansia*; ciò dimostra che le emozioni come paura ed ansia, in relazione agli eventi sismici, si presentino e riaffiorino nella maggioranza degli individui sottoposti ad indagine.

Tutti gli altri grafici relativi alle distribuzione delle variabili sono allegati in Appendice Sezione A.

3.3 Implementazione degli algoritmi

Prima di procedere all’implementazione degli algoritmi di feature selection e successivamente alla stima dei modelli per risposta di tipo ordinale, è stato necessario verificare l’assunzione di proporzionalità mediante il test di Brant, considerando il modello con tutte le variabili esplicative, ovvero il modello completo. Il risultato del test è allegato in Appendice Sezione B per i Proportional Odds Models.

Dalla tabella 5.1 si può osservare come il test, che viene confrontato con una variabile χ^2 con gradi di libertà pari a $(k - 2)p$, conduca per tutte le covariate, compresa quella *Omnibus* che contempla l’assunzione di proporzionalità del modello complessivo, al non rifiuto dell’ipotesi nulla.

Si ricordi che nell’ipotesi nulla H_0 viene specificata la corretta assunzione di proporzionalità; per converso, l’ipotesi H_1 riflette la situazione in cui la proportional assumption non viene rispettata. Dal grafico si evince che il *p-value* risulta sensibilmente maggiore di 0.05 per tutte le variabili esplicative.

La verifica relativa all’assunzione di proporzionalità è stata effettuata anche per gli Adjacent Category Models e per i Continuation Ratio Models, andando a confrontare l’indice BIC per la forma parallela e per la forma non parallela. L’indice BIC, per entrambi i modelli, si presenta più basso nel caso di assunzione di proporzionalità nel modello rispetto all’ipotesi in cui essa non fosse rispettata. La medesima verifica sul modello completo è stata effettuata con il test di *Hosmer-Lemeshow*, *Pulkstenis-Robinson* ed infine con il test di *Lipsitz*.

L’ipotesi nulla H_0 per i test di bontà di adattamento identifica il buon adattamento del modello ai dati, mentre l’ipotesi alternativa H_1 evidenzia che ci sia qualche problema di adattamento non specificato nel modello.

Nonostante venga rifiutata l’ipotesi nulla di buon adattamento nel test di Hosmer-Lemeshow, il test di Lipstiz (che però denota difetti quali potenza piuttosto bassa e scelta arbitraria del numero dei gruppi g) ci suggerisce che i dati si adattano adeguatamente al modello stimato, come osservabile in Appendice Sezione B dalla figura 5.2. Il *p-value* associato al test risulta maggiore della soglia considerata del 5%, motivo per il quale si deduce il buon adattamento del modello ai dati

osservati.

Valutata l'assunzione di proporzionalità e la bontà d'adattamento tramite il test di Lipsitz, con l'obiettivo di valutare le performance del modello e di conseguenza ridurre al minimo il rischio di *overfitting*, è stato opportuno dividere l'intero dataset in training set (75%) e test set (25%) come segue:

Training set	Test set
324	109

Tabella 3.1: Divisione del dataset in training set e test set

Partendo dagli algoritmi *backward* e *forward* basati su accuracy, sono stati stimati per ogni tipologia di modello, considerato il numero di variabili, 529 modelli, dove per ogni modello con $1, 2, \dots, p$ covariate viene considerato il set di variabili che restituisce l'accuracy più elevata.

Considerando la totalità dei risultati, dalle stime dei modelli emerge come i due tipi di approcci abbiano fornito risultati pressocchè simili, come evidenziato dalla tabella 3.2:

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

Tabella 3.2: Confronto tra modelli *acat*, *cratio* e *pom* in relazione alla subset selection con accuracy più elevata.

Dalla tabella si può osservare inoltre come la selezione di tipo forward tenda ad essere più parsimoniosa rispetto alla backward, che tende a considerare modelli

con molteplici covariate. Ciò nonostante il modello migliore, considerata l'accuracy, è il *Continuation Ratio Model*, che tende a considerare 24 covariate. Le 24 variabili considerate da tale modello, in ordine di importanza secondo la selezione, sono *ansia*, *tipo_veicolo*, *lavoro*, *studio*, *occupazione_fuori_campagna*, *finemese*, *sesso*, *piano_abitazione*, *batteriainfo*, *disabili_famiglia*, *batteriafiducia*, *ascensore*, *n_veicoli*, *fuori_regione_sisma*, *preoccupazione_sismica*, *tempestivita_decisioni*, *insonnia*, *fuori_regione*, *cambio_residenza*, *tipo_abitazione* ed *estero*.

La variabile *ansia* appare essere la variabile esplicativa più importante considerando l'accuracy; ciò risulterebbe essere legato al fatto che tale variabile risulta essere la variabile maggiormente correlata alla variabile dipendente **paura** e quindi ritenuta statisticamente più importante rispetto alle altre.

L'interpretazione dei parametri legata alle variabili verrà poi discussa nel paragrafo relativo ai risultati.

Relativamente ai metodi di shrinkage, sono stati stimati **60903** modelli ed il tuning dei parametri è stato effettuato considerando, per ogni tipologia di modello ordinale, **20301** combinazioni di α e λ partendo dal valore minimo di *alpha* pari a 0 (penalizzazione ridge) fino al valore massimo 1 (penalizzazione LASSO) e dal valore minimo di *lambda* fissato a 0 (nessun effetto della penalizzazione) fino al valore massimo impostato pari a 2, aumentando di 0.01 ad ogni passo.

La scelta di tali parametri è stata effettuata mediante la libreria **OrdinalNet** per i *Proportional Odds Models*, *Adjacent Category Models* e *Continuation Ratio Models* al fine di procedere al tuning di α e λ associati alla accuracy più elevata. Di seguito è riportata una tabella contenente le migliori 20 combinazioni di α e λ la cui accuracy risulta più alta.

alpha	lambda	family	n	aic	bic	loglik	accuracy
0.70	0.01	<i>acat</i>	27	679.9519	562.7489	-250.3744	0.7523
0.71	0.01	<i>acat</i>	27	679.9962	562.7932	-250.3966	0.7523
0.72	0.01	<i>acat</i>	27	680.0412	562.8382	-250.4191	0.7523
0.12	0.07	<i>cratio</i>	27	715.4220	598.2189	-268.1095	0.7523
0.13	0.07	<i>cratio</i>	27	715.8659	598.6629	-268.3314	0.7523
0.14	0.07	<i>cratio</i>	27	716.3282	599.1252	-268.5626	0.7523
0.12	0.08	<i>cratio</i>	27	716.1669	602.7446	-271.3723	0.7523
0.13	0.08	<i>cratio</i>	27	716.6789	603.2566	-271.6283	0.7523
0.10	0.08	<i>cratio</i>	27	720.9069	603.7039	-270.8519	0.7523
0.11	0.08	<i>cratio</i>	27	721.4172	604.2141	-271.1071	0.7523
0.12	0.09	<i>cratio</i>	27	722.3539	608.9316	-274.4658	0.7523
0.66	0.01	<i>acat</i>	28	685.4890	564.5052	-250.2526	0.7523
0.67	0.01	<i>acat</i>	28	685.5489	564.5651	-250.2825	0.7523
0.68	0.01	<i>acat</i>	28	685.6098	564.6260	-250.3130	0.7523
0.69	0.01	<i>acat</i>	28	685.6718	564.6880	-250.3440	0.7523
0.09	0.08	<i>cratio</i>	29	731.9323	607.1677	-270.5839	0.7523
0.03	0.09	<i>cumulative</i>	30	746.0739	617.5286	-274.7643	0.7523
0.04	0.09	<i>cumulative</i>	30	746.5792	618.0339	-275.0170	0.7523
0.00	0.15	<i>cratio</i>	34	782.9610	643.0735	-284.5367	0.7523
0.73	0.01	<i>acat</i>	27	680.0870	562.8840	-250.4420	0.7431

Tabella 3.3: Tuning degli iperparametri α e λ nell'elastic net per *acat*, *cratio* e *pom* considerata l'assunzione di proporzionalità.

Dalla tabella 3.3 (dove sono riportati solo i primi 20 modelli che riportano il valore di accuracy più alta) sono presenti tutti i modelli considerati e si può notare inoltre che i parametri di α e λ , che consentono di avere un accuracy più elevata, sono sostanzialmente diversi per ogni tipo di modello considerato.

Dal momento in cui i valori di α e λ sono stati determinati in maniera sequenziale aggiungendo 0.01 ad ogni passo, dalla tabella 3.3 si osserva come l'accuracy risulta la medesima in riferimento al range da 0.66 a 0.72 dei valori di α , considerando λ pari a 0.01 per il modello **acat**.⁴ Fissata l'accuracy più alta (0.7523,

⁴Ciò risulta essere anche una conseguenza della numerosità del test set la quale non appare significativamente elevata.

ovvero il modello ha predetto correttamente circa il 75% delle osservazioni), è stato selezionato il modello con il minor numero di variabili esplicative selezionate (**27**). Dal momento in cui più di un modello presentava 27 predittori, tra queste è stato scelto il modello con l'indice **BIC** più basso, ovvero il modello con α e λ rispettivamente impostati a 0.70 e 0.01, quindi una penalizzazione molto più vicina a quella lasso rispetto alla ridge. Visto che sono state considerate tutte le possibili combinazioni (penalizzazione \mathcal{L}_1 ed \mathcal{L}_2 compresa), dalla tabella si evince inoltre che le classiche penalizzazioni restituiscono risultati peggiori rispetto alle combinazioni riportate in tabella, a parte per una penalizzazione ridge su Continuation Ratio Model che presenta un valore di λ pari a 0.15.

A conclusione di quanto appena asserito, è stato realizzato il grafico presente di seguito allo scopo di analizzare, mediante un campione di 100 modelli estratti dai 60903 stimati, il comportamento degli stessi in relazione al numero di covariate nel modello e all'accuracy.

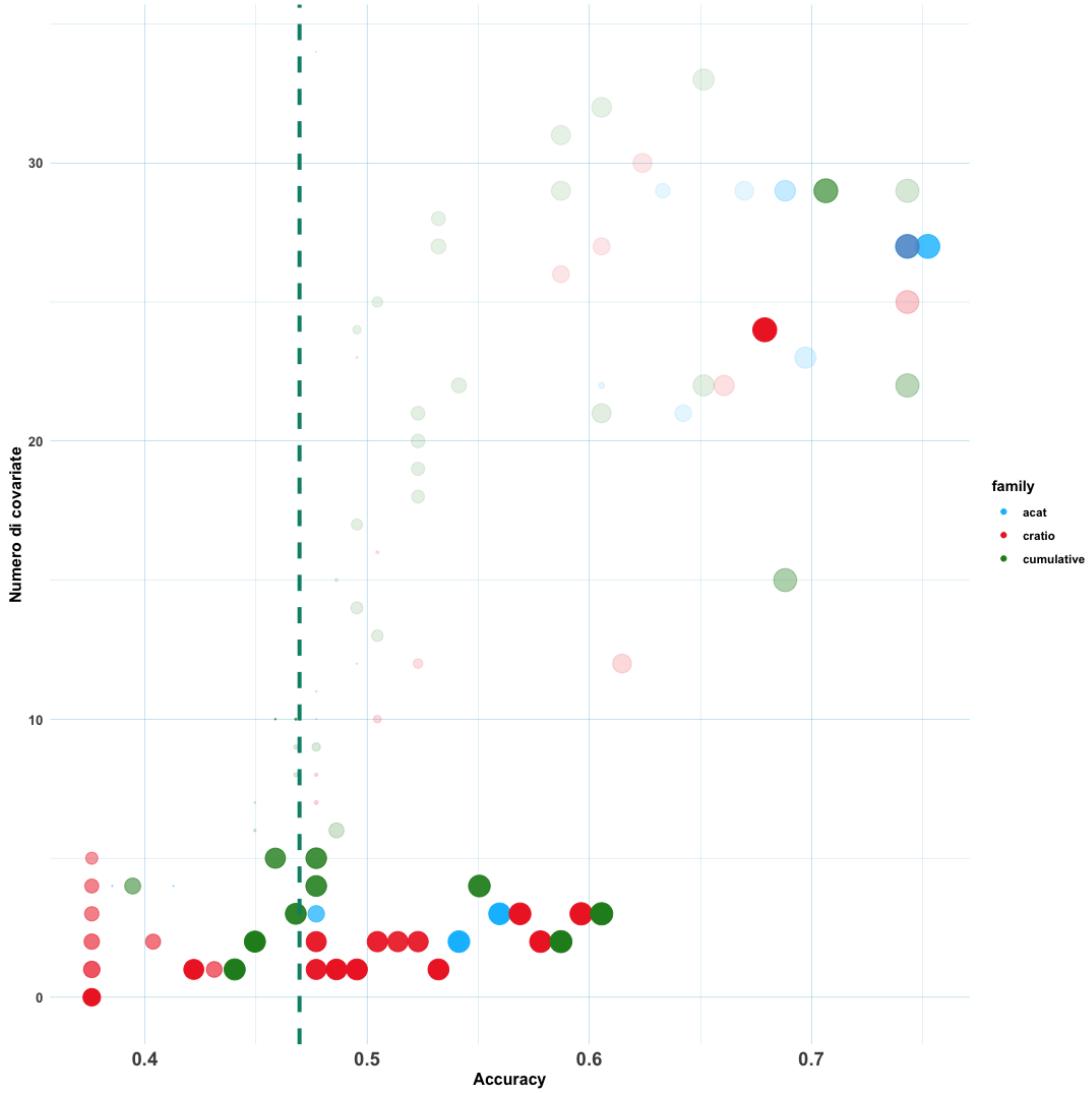


Figura 3.5: Scatter plot con numero di covariate ed accuracy in riferimento ad un campione di 100 modelli.

Relativamente alla figura 3.5, la linea verticale tratteggiata di colore verde riporta il livello medio di accuracy, per cui i modelli alla sinistra presentano un accuracy più bassa rispetto alla media mentre gli elementi a destra un accuracy più alta.

Inoltre, nello scatter plot in figura, è presente un ulteriore indicazione circa il valore di α e di λ , dove una dissolvenza più accentuata indica valore di α basso, mentre una dissolvenza meno marcata un valore di α alto. Quindi, tanto il più l'elemento è visibile tanto più il modello stimato è vicino ad una penalizzazione lasso piuttosto che ridge. Invece per quanto concerne λ , più basso è il termine

della penalizzazione più grande sarà la dimensione dell'elemento nel grafico ed, al contrario, più elevato sarà il valore di λ meno grande sarà la dimensione dell'elemento. Dal grafico, si evince come la maggioranza dei modelli (con accuracy sopra la media pari a 0.4696) è rappresentata da Proportional Odds Models, contraddisti da un valore piuttosto ridotto di α e di λ . I modelli peggiori risultano avere valori di α contenuti (quindi vicino ad una penalizzazione ridge) e di λ molto elevati.

In conclusione, dal grafico emerge, osservando il punto in alto a destra, il miglior modello indicato nella griglia riportata in 3.3, rappresentata dalla famiglia *acat*.

Infine, per quanto concerne l'algoritmo di *dimensionality reduction* utilizzato, data la quasi totalità di variabili di natura categorica, è stato opportuno considerare un'analisi delle corrispondenze su tabelle concatenate.

Le sole 2 variabili di natura quantitativa (età e *n_veicoli*) sottoposte a nuova codifica, osservata la loro distribuzione, sono state così rappresentate:

1. *età*, suddivisa in:

- *giovani*, per coloro con un'età compresa tra 18 e 30 anni;
- *adulti*, per coloro con un'età compresa tra 31 e 45 anni;
- *anziani*, per coloro con un'età superiore ai 46 anni.

2. *n_veicoli*, suddivisa nelle seguenti categorie:

- *Al massimo 1*, per coloro che sono in possesso di al massimo 1 veicolo;
- *Almeno 2*, per coloro che possiedono almeno 2 veicoli.

Codificate le variabili, prima di eseguire tale metodologia, è stato necessario convertire la variabile dipendente paura attraverso il metodo dell'*one-hot encoding* e poi concatenare la risultante matrice (trasposta) con la matrice contenente le variabili esplicative, anch'esse precedentemente sottoposte all'*one-hot encoding*. Data la divisione in training e test set illustrata in 3.1, la matrice risultante sarà:

$$F = \begin{matrix} Z'_Y \\ (5 \times 87) \end{matrix} \times \begin{matrix} Z_X \\ (5 \times 324) \end{matrix} \quad (3.1)$$

dove $k = 5$ saranno le categorie della variabile dipendente e $Q = 87$ l'insieme delle categorie dei predittori.

Applicata l'analisi delle corrispondenze sulla matrice F , sono stati calcolati gli *scores* mediante la formula in 2.42 ed utilizzati come predittori w per la stima dei modelli ordinali illustrati in questo lavoro di tesi.

Nonostante il numero massimo risulti essere pari a $\min(Q-1, k-1)$ (in questo caso studio quindi 4), il numero di componenti da utilizzare come variabili esplicative è stato opportunamente scelto mediante tuning, ovvero sono state ricavate le coordinate ed ogni modello è stato stimato con $0, 1, \dots, 4$ componenti allo scopo di ricercare il modello con accuracy più alta.

I risultati sono presentati di seguito:

family	componenti	aic	bic	loglik	accuracy
<i>acat</i>	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
<i>cratio</i>	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
<i>pom</i>	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 3.4: *Dimensionality reduction* applicata a modelli ordinali.

La tabella 3.4 (che riporta la percentuale d'inerzia spiegata utilizzando ciascun componente) evidenzia che tale approccio di riduzione della dimensionalità non fornisce risultati soddisfacenti se consideriamo i due metodi già applicati precedentemente.

Inoltre, si osserva che gli Adjacent Category Models risultano essere i modelli che

performano peggio in termini di accuracy. Invece, il modello migliore utilizzando le componenti dell'analisi delle corrispondenze come predittori è il Continuation Ratio con 3 componenti. Tuttavia, il modello a 4 componenti risulta essere ugualmente valido, ma per motivi legati alla parsimonia si preferisce il modello che utilizza un numero minore di variabili esplicative. Seguono poi i Proportional Odds Models con un accuracy superiore al 50% per i modelli con 3 e 4 componenti. Nel prossimo paragrafo sarà illustrata l'identificazione del miglior modello e l'interpretazione dei relativi risultati.

3.4 Risultati

Analizzati gli approcci di feature selection e identificati i relativi parametri di tuning, al fine di identificare il miglior algoritmo per i modelli considerati, è stato realizzato il seguente grafico per riassumere e ricapitolare quanto fin ad ora affermato.

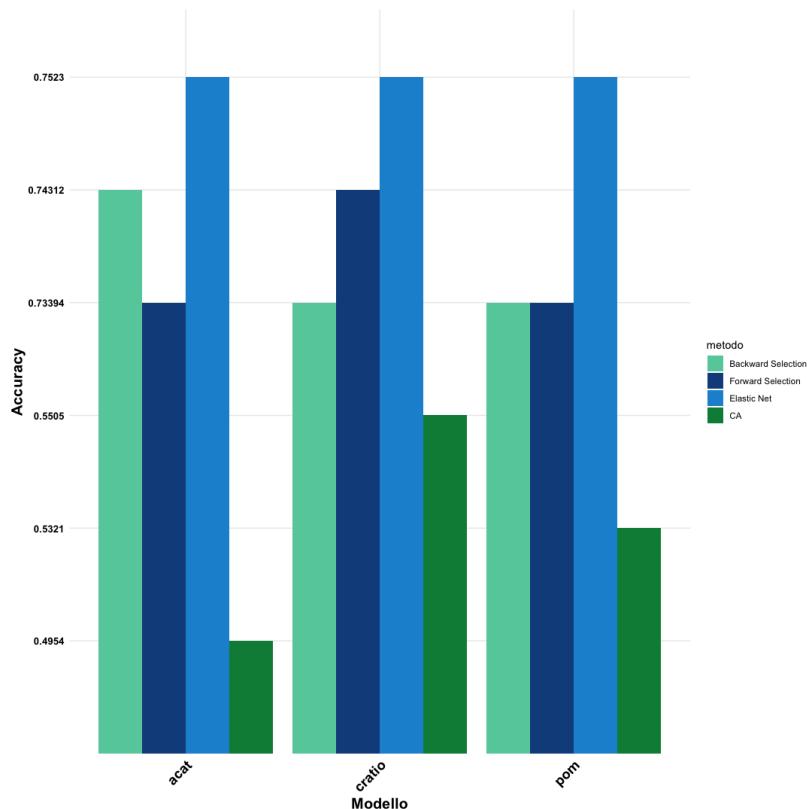


Figura 3.6: Accuracy su modelli *acat*, *cratio* e *pom* per metodi di shrinkage, subset selection e dimensionality reduction.

Dal grafico 3.6 si evince come i metodi di shrinkage (considerando anche le penalizzazioni classiche non presenti nel grafico) abbiano ottime performance in termini di accuracy. Medesima considerazione per gli approcci di subset selection, che riportano risultati piuttosto simili. Il metodo che restituisce prestazioni peggiori è quello legato alla riduzione della dimensionalità, che ha fornito risultati piuttosto bassi.

In conclusione quindi, il metodo migliore risulta essere l'elastic net con un accuracy pari a 0.7523 per tutti i modelli considerati. Dato che l'accuracy ed anche il numero di covariate considerato sono i medesimi, si tende a preferire il modello il cui BIC risulta più basso: il modello che detiene l'indice BIC più basso risulta essere appartenente alla famiglia *acat*.

Per avere una rappresentazione immediata ed intuitiva per valutare le performance del modello nel dettaglio, è stato opportuno costruire la matrice di confusione che riporta i valori osservati e i valori predetti dal modello (*acat* in questo caso).

		Valori osservati				
		1	2	3	4	5
Valori stimati	1	7	2	0	0	0
	2	1	10	3	1	0
	3	0	4	34	5	1
	4	0	0	4	22	4
	5	0	0	0	2	9

Figura 3.7: Matrice di confusione su modello *acat* con regolarizzazione elastic net.

Si può notare, dalla figura 3.7, che il modello ha predetto correttamente **85** osservazioni su un totale di 109. Si osserva come la classe 2 sia la classe con più osservazioni mal classificate, per l'esattezza il modello ha predetto correttamente 10 osservazioni rispetto ai 16 rispondenti che hanno dichiarato di avere un livello di paura *basso*.

Per quanto concerne le variabili selezionate dall'algoritmo, considerando i modelli stimati con regolarizzazione elastic net che presentano accuracy più alta, non è possibile dare una interpretazione strettamente numerica alle stime dei coefficienti, in quanto penalizzati. Ciò comporta, da un punto di vista interpretativo, non poter calcolare gli *odds ratio*, utili per poter ottenere una misura precisa circa il comportamento della variabile dipendente in relazione alle variabili esplicative.

x	acat
(Intercept):1	-5.8762
(Intercept):2	-7.9546
(Intercept):3	-11.5005
(Intercept):4	-14.0493
eta	0.0191
scosse	0.5792
ansia	1.822
sintomi_fisiologici	0.2515
tempestivita_decisioni	0
insonnia	0.0388
preoccupazione_sismica	0.2667
n_veicoli	0
finemese	0.0055
sesso_Maschio	0
stato_civile_Single	0.2268
residenza_Zona.Bianca	-0.0902
studio_Non.Laureati	-0.0803
lavoro_Non.Lavoratori	0.1739
occupazione_fuori.campania_Si	0.326
componenti_famiglia_Più.di.3	0.3888
disabili_famiglia_Si	0.3199
piano_abitazione_Oltre.il.1p	-0.0297
terremoto80_Si	0
estero_Si	-0.2461
fuori_regione_Si	0.3327
fuori_regione_sisma_Si	-0.2599
cambio_residenza_Si	0.0975
cambio_residenza_sisma_Si	-0.2776
frequenza_zone_rosse_Si	-0.0843
punti_accoglienza_Si	-0.0172
casa_proprieta_Si	-0.1095
tipo_abitazione_Appartamento	-0.2989
ascensore_Si	0.0752
tipo_veicolo_Auto	0.1433
tipo_veicolo_Nessuno	0
batteriainfo_BassaUtilitaInfo	0
batteriafiducia_BassaFiducia	0.1186

Tabella 3.5: Valori delle stime dei coefficienti del modello acat con penalizzazione elastic net

Dalla tabella 3.5 emerge che su 32 variabili considerate inizialmente nel modello (thresholds esclusi) solo le stime dei coefficienti di 5 variabili (*sesso*, *terremoto80*, *tempestivita_decisioni*, *n_veicoli* e *batteriainfo*) sono state azzerate, per cui l'algoritmo di selezione in questione non ha considerato tali variabili statisticamente significative per spiegare il fenomeno oggetto di studio.

Ciò risulta essere una conseguenza dei valori dei parametri di tuning: essendo il

parametro di α pari a 0.70, l'elastic net tende ad essere più parsimonioso rispetto alla penalizzazione ridge, ma in misura minore rispetto alla regolarizzazione lasso che tende a portare più coefficienti esattamente a 0.

Inoltre, λ è uguale a 0.01 (valore piuttosto basso), il che consente al modello di mantenere molte più variabili per la stima.

Avendo verificato globalmente l'assunzione di proporzionalità del modello completo attraverso l'indice BIC (*BIC parallel* 884.6752 < *BIC non parallel* 1346.38), relativamente all'interpretazione, un coefficiente positivo (verde in figura 3.5) indica che vi è un'associazione di tipo positivo tra la variabile esplicativa e la transizione della variabile dipendente di passare alla categoria successiva, ovvero un coefficiente positivo nel predittore è associato a un aumento del rapporto delle probabilità di passare dalla categoria k alla categoria $k+1$. Invece, un coefficiente negativo (rosso in figura) indica un'associazione negativa tra la x e il rapporto delle probabilità di transizione dalla k -esima categoria alla categoria $k+1$.

Considerando la tabella 3.5, i risultati maggiormente significativi sono i seguenti:

1. la probabilità di passare ad un livello superiore di paura aumenta, intuitivamente, all'aumentare del livello di intensità percepito delle scosse, all'aumentare del livello di ansia, sintomi fisiologici, insonnia e preoccupazione legata ai fenomeni sismici. Inoltre, essa aumenta all'aumentare dell'età, aumenta per coloro che non hanno un partner, per chi non lavora, per chi ha un'occupazione fuori dalla campania, per chi ha più di 3 componenti nel nucleo familiare, per chi ha disabili in famiglia, per chi vive in un'abitazione dotata di ascensore, per chi ha bassa fiducia nelle istituzioni ed infine aumenta all'aumentare della difficoltà di arrivare a fine mese;
2. la probabilità di passare ad un livello superiore di paura diminuisce, invece, per coloro che non sono in prossimità dei Campi Flegrei, che si trovano quindi in «zona bianca», per coloro che non sono in possesso di una laurea, per coloro che vivono in una abitazione situata al piano superiore al primo, per le persone che frequentano o lavorano abitualmente nelle «zone rosse»;

per coloro che conoscono quali sono i punti di prima accoglienza in caso di emergenza, per coloro che vivono in una casa di proprietà e per le persone che risiedono in un appartamento.

Capitolo 4

Conclusioni

A termine di questo lavoro di tesi, analizzati molteplici algoritmi di feature selection sui modelli ordinali, è stato possibile trarre conclusione sui fattori che incidono sulla paura in riferimento agli eventi sismici registrati nella zona dei Campi Flegrei.

La selezione delle variabili, soprattutto in contesti legati alla percezione del rischio, risulta essere quindi una fase cruciale nel processo di costruzione del modello statistico , soprattutto in uno scenario in cui vengono generati e trasformati (quotidianamente) flussi di dati caratterizzati da un volume enorme.

I metodi di shrinkage hanno senz'altro fornito un risultato soddisfacente in riferimento all'accuracy, andando sia a mantenere un indicazione circa la stima dei coefficienti, sia riuscendo ad identificare un subset ottimale di variabili utile per spiegare il fenomeno oggetto di studio.

In riferimento al miglior modello (*Adjacent Category Model con selezione elastic net*), esso ha rilevato **27** predittori come variabili idonee a spiegare la percezione degli individui legati al sisma, considerando la relativa forma parallela che, essendo verificata localmente, per quanto possa essere conveniente, non sempre risulta l'ipotesi giusta e ragionevole. [28]

In merito a ciò, occorrerebbe però sottolineare che, non solo l'elastic net, ma tutti i metodi applicati hanno suggerito modelli poco parsimoniosi e quindi modelli poco funzionali se fossimo stati in presenza di un dataset di dimensioni più elevate. Inoltre, in riferimento al caso studio, il numero di variabili, l'assunzione di pro-

porzionalità e il set di dati utilizzato hanno reso la selezione un processo che non ha fatto emergere una sostanziale differenza tra i metodi, se non per l'approccio basato sulla riduzione della dimensionalità che ha fornito risultati piuttosto bassi. Tuttavia, la selezione migliore ha fatto emergere spunti rilevanti da un punto di vista strettamente psicologico, quali il fatto che le persone con un livello di istruzione piuttosto basso e rispondenti che convivono con una persona diversamente abile nel nucleo familiare tendono ad avere una maggiore probabilità di riscontrare un livello di paura alto. Allora occorrerebbe fornire un piano di emergenza chiaro e alla portata di tutti, che consenta ai residenti in prossimità della zona rossa di usufruire di sostegni e aiuti nel caso in cui ce ne fosse bisogno e che aiuti coloro che non sono propriamente *nativi digitali* a filtrare le informazioni sicure ed attendibili emanate da fonti ufficiali da quelle *fake*.

Quindi, in relazione ai risultati e ai metodi applicati, come già ribadito, per un'analisi più dettagliata sul fenomeno del sisma e sulla percezione dei residenti in prossimità della zona rossa dei Campi Flegrei, occorrerebbe valutare modelli più complessi ed avanzati appartenenti alla classe degli *Item Response Theory* oppure utilizzare metodi di selezione *adattivi*, che consentirebbero di incrementare le performance sul dataset analizzato.

In conclusione, la comprensione della percezione degli eventi sismici nella zona dei Campi Flegrei non solo fornisce importanti spunti per migliorare la risposta ai terremoti, come fenomeno privo di un totale controllo, ma sottolinea anche l'urgente necessità di un dialogo continuo tra scienza, comunità locali e *decision makers* al fine di garantire la sicurezza e la resilienza di quest'area vulnerabile.

Capitolo 5

Appendice

Sezione A

Specificazione e Codifica Variabili

1. *data_ora*, ovvero in che giorno e a che ora il soggetto rispondente ha completato il questionario;
2. *accettazione*, variabile dicotomica che indica se il rispondente ha accettato o meno di proseguire alla compilazione del questionario;
3. *eta*, ovvero l'età del rispondente;
4. *sesso*, ovvero il sesso del rispondente;
5. *stato_civile*, ovvero se il soggetto è celibe/nubile, convivente, sposato/a, separato/a, divorziato/a oppure vedovo/a;
6. *residenza*, ovvero città in cui il soggetto risiede, se in una zona a rischio (in riferimento alle zone definite «gialle» e «rosse») o in zona «bianca».
7. *studio*, ovvero il titolo di studio del rispondente, dove le modalità sono licenza elementare, licenza media inferiore, licenza media superiore, laurea e post-laurea;
8. *lavoro*, ovvero lo stato occupazionale del rispondente tra studente, studente lavoratore, lavoratore autonomi, dipendente, pensionato o disoccupato;
9. *occupazione_fuori_campagna*, che indica se il soggetto lavora/studia fuori dalla regione Campania;
10. *componenti_famiglia*, ovvero il numero di componenti (rispondente compreso) presenti nel nucleo familiare;
11. *disabili_famiglia*, variabile binaria che indica la presenza/assenza di un soggetto diversamente abile nel nucleo familiare;
12. *piano_abitazione*, indicando a che piano si trova l'abitazione in cui si vive;
13. *terremoto80*, se il rispondente ha vissuto o meno il terremoto degli anni '80;
14. *orientamento_politico*, indicando l'orientamento politico del rispondente;
15. *scosse*, ovvero il livello di intensità delle scosse percepito durante gli eventi sismici;
16. *paura*, ovvero il livello di paura durante gli eventi sismici. Tale variabile rappresenta la variabile dipendente Y utilizzata per l'analisi;
17. *ansia*, ovvero il livello di ansia durante gli eventi sismici;

18. *sintomi_fisiologici*, ovvero il livello di sintomi fisiologici (vertigini, nausea, mal di testa, problemi gastrintestinali, ecc.) durante gli eventi sismici;
19. *tempestività_decisioni*, ovvero il livello di tempestività delle decisioni adottate durante gli eventi sismici;
20. *insonnia*, ovvero il livello di insonnia durante gli eventi sismici;
21. *preoccupazione_sismica*, ovvero il livello di preoccupazione durante gli eventi sismici;
22. *estero*, indicando se l'individuo è mai stato all'estero;
23. *fuori_regione*, se il rispondente ha mai considerato, per qualsiasi motivo, di lavorare/studiare fuori regione;
24. *fuori_regione_sisma*, se il rispondente ha mai considerato, per motivi legati ai fenomeni sismici, di lavorare/studiare fuori regione;
25. *cambio_residenza*, se il rispondente ha mai considerato, per qualsiasi motivo, di cambiare la sua residenza;
26. *cambio_residenza_sisma*, se il rispondente ha mai considerato, per motivi legati ai fenomeni sismici, di cambiare la sua residenza;
27. *frequenza_zone_rosse*, se il rispondente lavora o è solito trascorrere tempo in zone indicate come rosse;
28. *info_radio*, ovvero il livello di utilità delle informazioni diffuse dalla radio sulle attività sismiche;
29. *info_TV*, ovvero il livello di utilità delle informazioni diffuse dalla TV sulle attività sismiche;
30. *info_social*, ovvero il livello di utilità delle informazioni diffuse sui social (Whatsapp, Telegram, Facebook, Instagram, TikTok, Twitter) sulle attività sismiche;
31. *info_giornali*, ovvero il livello di utilità delle informazioni diffuse dai giornali sulle attività sismiche;
32. *info_app*, ovvero il livello di utilità delle informazioni diffuse da app sulle attività sismiche;
33. *fiducia_istituzioni_comunali*, ovvero il livello di fiducia nelle istituzioni comunali;
34. *fiducia_istituzioni_regionali*, ovvero il livello di fiducia nelle istituzioni regionali;
35. *fiducia_istituzioni_nazionali*, ovvero il livello di fiducia nelle istituzioni nazionali;
36. *fiducia_INGV*, ovvero il livello di fiducia nell'INGV¹;
37. *sicurezza*, ovvero il livello di sicurezza nelle strutture pubbliche (scuole, uffici, comune, ecc.);
38. *punti_accoglienza*, variabile dicotomica per indicare se l'individuo conosce o meno i punti di prima accoglienza nella sua area di residenza in caso di emergenza;
39. *casa_proprietà*, se l'individuo vive o meno in una casa di proprietà;
40. *tipo_abitazione*, ovvero la tipologia di abitazione nella quale il rispondente vive tra appartamento, villa/villetta ed altro;
41. *ascensore*, ovvero se l'abitazione nella quale il soggetto vive è dotata o meno di ascensore;
42. *n_veicoli*, ovvero il numero di veicoli (tra auto, moto, ecc.) che il rispondente possiede;
43. *tipo_veicolo*, ovvero la tipologia di veicolo posseduto tra auto, moto/scooter, ecc;
44. *finemese*, indicante la difficoltà con la quale il rispondente e la sua famiglia arrivano a fine mese da un punto di vista economico;
45. *ral*, ovvero il reddito annuo lordo del rispondente;
46. *commento*, sezione in cui i rispondenti hanno riportato un commento facoltativo sulle attività sismiche;

¹INGV, Istituto Nazionale di Geofisica e Vulcanologia.

Analisi Esplorativa

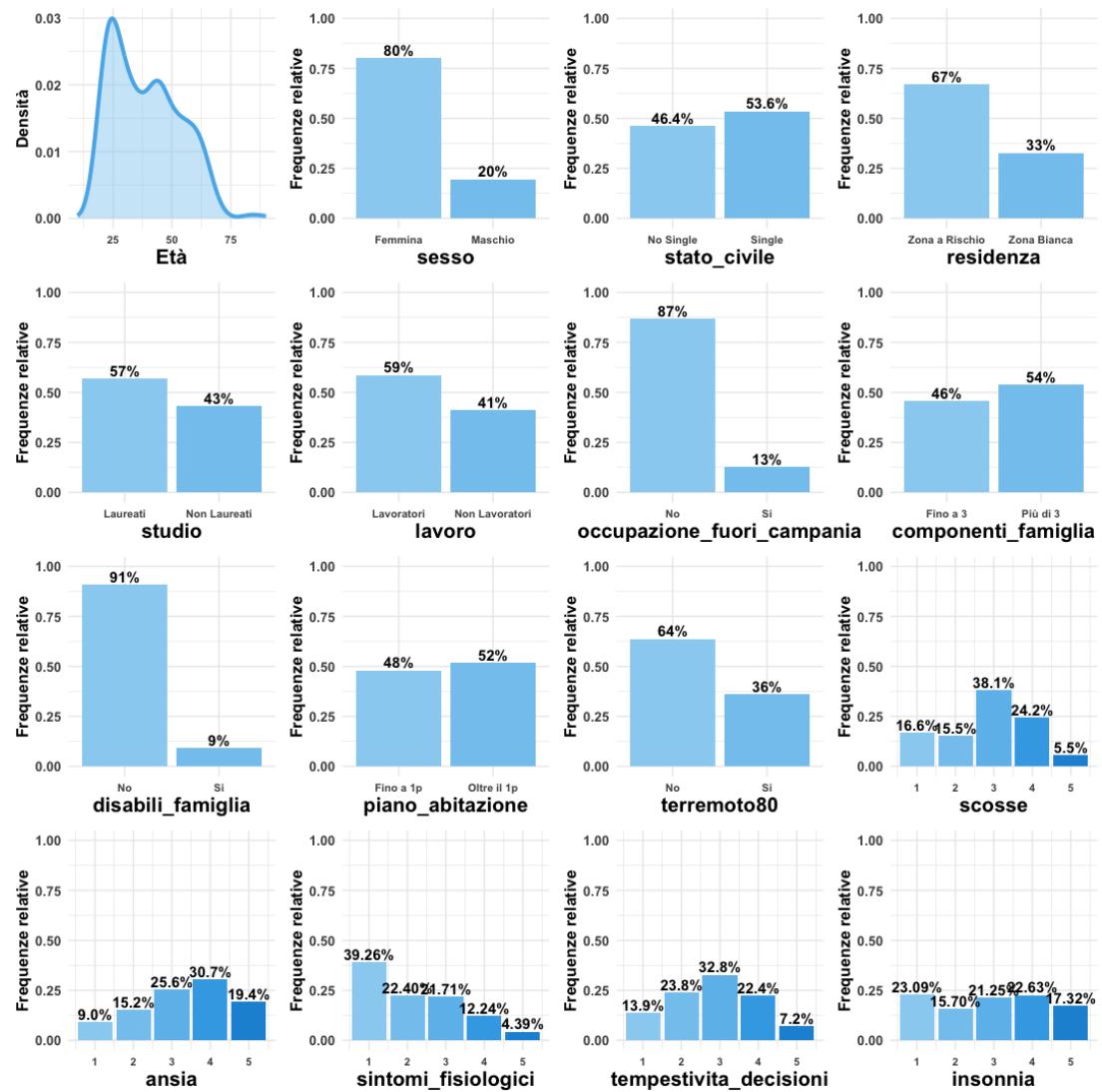


Figura 5.1: Distribuzioni variabili esplicative

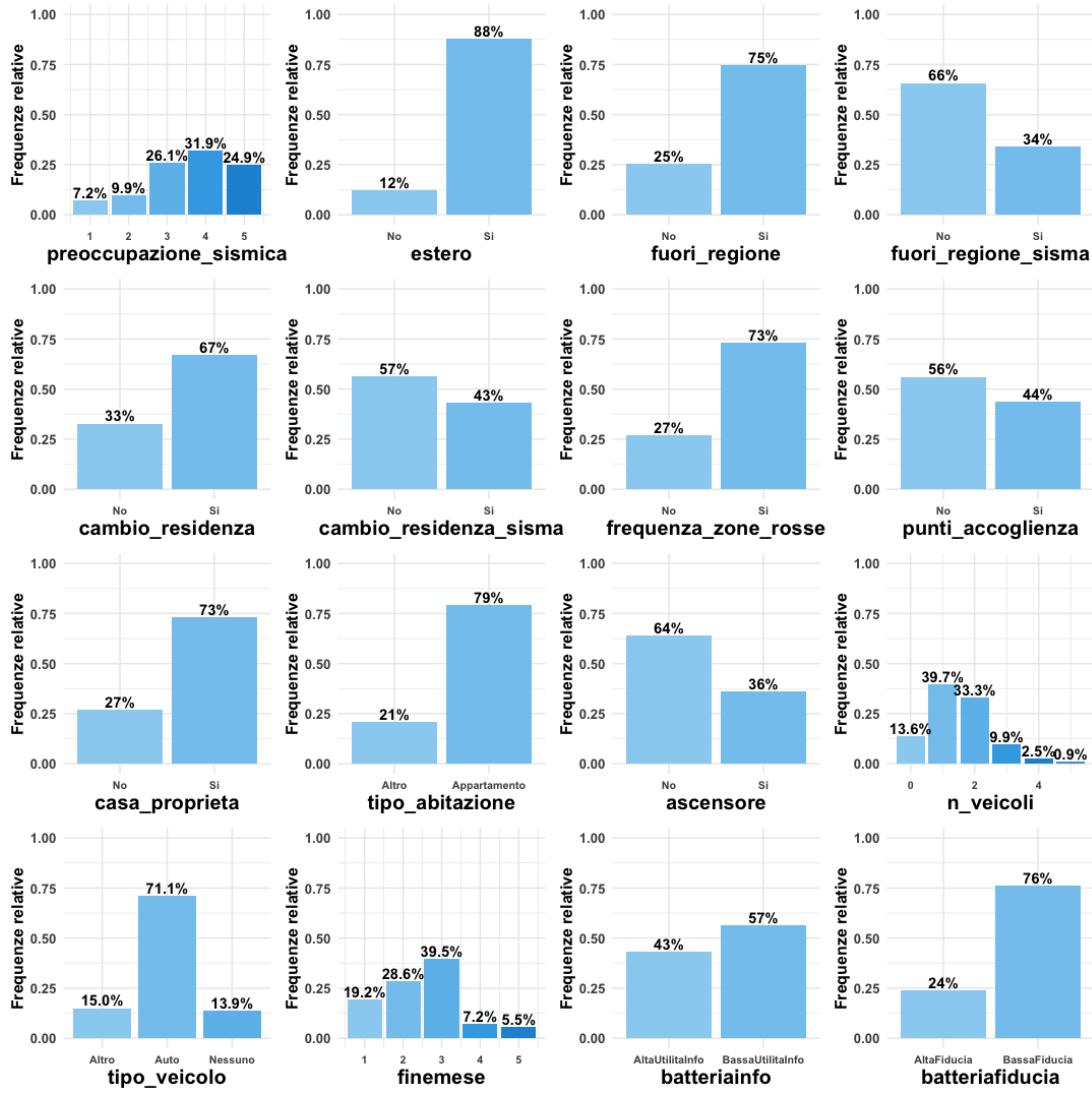


Figura 5.2: Distribuzioni variabili esplicative

Sezione B

Test di Brant

x	x2	df	probability
<i>Omnibus</i>	1.14677	99	1.00000
<i>eta</i>	0.10050	3	0.99178
<i>sessoMaschio</i>	0.40560	3	0.93908
<i>stato_civileSingle</i>	0.71307	3	0.87012
<i>residenzaZona_Bianca</i>	4.79990	3	0.18705
<i>studioNon_Laureati</i>	0.16102	3	0.98362
<i>lavoroNon_Lavoratori</i>	0.35649	3	0.94908
<i>occupazione_fuori_campagnaSi</i>	2.33069	3	0.50667
<i>componenti_famigliaPiù_di_3</i>	0.70379	3	0.87231
<i>disabili_famigliaSi</i>	0.10691	3	0.99100
<i>piano_abitazioneOltre_il_1p</i>	1.09569	3	0.77812
<i>terremoto80Si</i>	1.63165	3	0.65224
<i>scosse</i>	0.32993	3	0.95430
<i>ansia</i>	5.78956	3	0.12231
<i>sintomi_fisiologici</i>	2.66418	3	0.44635
<i>tempestività_decisioni</i>	1.43271	3	0.69789
<i>insonnia</i>	1.44018	3	0.69614
<i>preoccupazione_sismica</i>	1.02043	3	0.79631
<i>esteroSi</i>	0.56628	3	0.90411
<i>fuori_regioneSi</i>	0.86906	3	0.83289
<i>fuori_regione_sismaSi</i>	0.72042	3	0.86839
<i>cambio_residenzaSi</i>	0.98549	3	0.80476
<i>cambio_residenza_sismaSi</i>	0.04297	3	0.99766
<i>frequenza_zone_rosseSi</i>	1.78472	3	0.61827
<i>punti_accoglienzaSi</i>	0.25356	3	0.96851
<i>casa_proprietàSi</i>	1.15935	3	0.76277
<i>tipo_abitazioneAppartamento</i>	1.46260	3	0.69093
<i>ascensoreSi</i>	0.47644	3	0.92404
<i>n_veicoli</i>	1.04095	3	0.79135
<i>tipo_veicoloAuto</i>	2.25090	3	0.52199
<i>tipo_veicoloNessuno</i>	1.19634	3	0.75388
<i>finemese</i>	3.42869	3	0.33013
<i>batteriainfoBassaUtilitaInfo</i>	3.16646	3	0.36667
<i>batteriafiduciaBassaFiducia</i>	2.41540	3	0.49077

Tabella 5.1: Verifica dell'assunzione di proporzionalità tramite il test di Brant

Test di Lipsitz

Test	LR.statistic	df	p.value
<i>Lipsitz goodness of fit test for ordinal response models</i>	12.941	9	0.1653

Tabella 5.2: Test di Lipsitz sulla bontà di adattamento

Sezione C

Backward e Forward Selection

Numero covariante	x	aic	bic	loglik	accuracy
31	eta - sesso - stato_civile - residenza - studio - lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività决策 - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - finemese - batteriainfo - batteriafudicia	565.9535	698.2795	-247.9767	0.7431
24	componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività决策 - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	564.615	674.2566	-253.3075	0.7339
29	stato_civile - residenza - studio - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività决策 - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	571.3338	699.879	-251.6669	0.7339
30	sesso - stato_civile - residenza - studio - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività决策 - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	573.3206	705.6467	-251.6603	0.7339
20	scosse - ansia - sintomi_fisiologici - tempestività决策 - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	565.657	660.1756	-257.8285	0.7248
21	terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività决策 - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	563.0868	661.368	-255.5343	0.7248
22	piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività决策 - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	565.0677	667.1478	-255.5338	0.7248
25	occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività决策 - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	565.5265	678.9488	-252.7632	0.7248
27	studio - lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestivitàDecisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - finemese - batteriainfo - batteriafudicia	567.7996	688.7833	-251.8998	0.7248
28	residenza - studio - lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestivitàDecisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	571.1481	695.9126	-252.574	0.7248
26	lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestivitàDecisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	568.3868	686.0399	-253.4184	0.7156
32	eta - sesso - stato_civile - residenza - studio - lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestivitàDecisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - finemese - batteriainfo - batteriafudicia	568.5203	708.4078	-247.2602	0.7156
23	disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestivitàDecisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	564.9337	670.7945	-254.4669	0.7064
19	ansia - sintomi_fisiologici - tempestivitàDecisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	593.7776	684.5155	-272.8888	0.6972
18	sintomi_fisiologici - tempestivitàDecisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	753.5149	840.472	-353.7574	0.5688
17	tempestivitàDecisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	779.7994	862.9757	-367.8997	0.5413
16	insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	781.5138	860.9094	-369.7569	0.5046
15	preoccupazione_sismica - estero - fuori_regione - fuori_residenza - cambio_residenza - sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	820.4817	896.0966	-390.2409	0.4954
11	cambio_residenza - cambio_residenza_sisma - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	964.4342	1024.9261	-466.2171	0.4128
12	fuer_regione - fuer_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	963.3956	1027.6682	-464.6978	0.4128
13	fuer_regione - fuer_regione_sistema - cambio_residenza - cambio_residenza_sistema - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	964.4836	1032.5369	-464.2418	0.4037
4	n_veicoli - finemese - batteriainfo - batteriafudicia	983.1953	1013.4413	-483.5977	0.3945
5	ascensore - n_veicoli - finemese - batteriainfo - batteriafudicia	985.0992	1019.1259	-483.5496	0.3945
10	cambio_residenza_sistema - frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	963.5689	1020.28	-466.7844	0.3945
6	tipo_abitazione - ascensore - n_veicoli - finemese - batteriainfo - batteriafudicia	986.0407	1023.8482	-483.0204	0.3945
7	casa_proprieta - tipo_abitazione - ascensore - n_veicoli - finemese - batteriainfo - batteriafudicia	987.894	1029.4822	-482.947	0.3945
2	finemese - batteriainfo	980.262	1002.9465	-484.131	0.3853
3	tipo_veicolo - finemese - batteriainfo	983.2065	1013.4525	-483.6033	0.3853
1	batteriainfo	982.9492	1001.8529	-486.4746	0.3761
0	-	982.0407	997.1637	-487.0203	0.3761
8	casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafudicia	989.1452	1038.2948	-481.5726	0.3670
9	frequenza_zone_rosse - punti_acoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - finemese - batteriainfo - batteriafudicia	982.7741	1031.9238	-478.3871	0.3394

Figura 5.3: Backward Selection su Adjacent Category Model

Appendice

Numerico covariante	x	aic	bic	loglik	accuracy
20	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica - tipo_abitazione - casa_proprieta	605.8516	700.3702	-277.9258	0.7339
21	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica - tipo_abitazione - casa_proprieta	606.4146	704.7139	-277.2073	0.7339
22	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica - tipo_abitazione - casa_proprieta - tempesitvita_decisioni	608.4112	710.4913	-277.2056	0.7339
23	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica - tipo_abitazione - casa_proprieta - tempesitvita_decisioni_studio - punti_acoglienza	610.1762	716.037	-277.0881	0.7339
24	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica - tipo_abitazione - casa_proprieta - tempesitvita_decisioni_studio - punti_acoglienza - frequenza_zone_rosse	611.5307	721.1723	-276.7653	0.7248
25	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica - tipo_abitazione - casa_proprieta - tempesitvita_decisioni_studio - punti_acoglienza - frequenza_zone_rosse - batteriaffiducia	612.1613	725.5836	-276.0807	0.7248
26	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica - tipo_abitazione - casa_proprieta - tempesitvita_decisioni_studio - punti_acoglienza - frequenza_zone_rosse - batteriaffiducia - scosse	575.7837	692.9868	-256.8919	0.7156
19	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica - tipo_abitazione - casa_proprieta - tempesitvita_decisioni_studio - punti_acoglienza	604.6032	695.341	-278.3016	0.7156
31	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica - frequenza_zone_rosse - batteriaffiducia - scosse - residenza - estero - fuori_regione_sisma - sintomi_fisiologici - eta - terremotoB	567.3924	703.4992	-247.6962	0.7156
32	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica - frequenza_zone_rosse - batteriaffiducia - scosse - residenza - estero - fuori_regione_sisma - sintomi_fisiologici	568.5203	708.4078	-247.2602	0.7156
18	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica	605.8475	692.8046	-279.9238	0.7064
27	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica - tipo_abitazione - casa_proprieta - tempesitvita_decisioni_studio - punti_acoglienza - frequenza_zone_rosse - batteriaffiducia - scosse - residenza	577.1813	698.1651	-256.5907	0.7064
28	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica - tipo_abitazione - casa_proprieta - tempesitvita_decisioni_studio - punti_acoglienza - frequenza_zone_rosse - batteriaffiducia - scosse - residenza - estero	577.6206	702.3851	-255.8103	0.6972
29	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica - tipo_abitazione - casa_proprieta - tempesitvita_decisioni_studio - punti_acoglienza - frequenza_zone_rosse - batteriaffiducia - scosse - residenza - estero - fuori_regione_sisma - sintomi_fisiologici	578.1622	706.7075	-255.0811	0.6972
30	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione - preoccupazione_sismica - tipo_abitazione - casa_proprieta - tempesitvita_decisioni_studio - punti_acoglienza - frequenza_zone_rosse - batteriaffiducia - scosse - residenza - estero - fuori_regione_sisma - sintomi_fisiologici	575.7252	708.0512	-252.8626	0.6972
4	ansia - tipo_veicolo - finemese - lavoro	594.657	628.6837	-288.3285	0.6881
5	ansia - tipo_veicolo - finemese - lavoro - batteriainfo	596.657	634.4644	-288.3285	0.6881
7	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna	600.6346	646.0035	-288.3173	0.6881
9	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione	602.9703	655.9007	-287.4851	0.6881
10	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli	604.9653	661.6764	-287.4826	0.6881
14	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma	611.5858	683.42	-286.7929	0.6881
17	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore - insomnia - fuori_regione	607.0445	690.2209	-281.5223	0.6881
3	ansia - tipo_veicolo - finemese	592.6784	622.9244	-288.3392	0.6789
6	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso	598.6543	640.2425	-288.3272	0.6789
8	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia	601.0592	650.2089	-287.5296	0.6789
11	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza	606.8133	667.3052	-287.4067	0.6789
13	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile	609.6764	677.7298	-286.8382	0.6789
16	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma - ascensore	605.8645	685.2601	-281.9323	0.6789
15	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia - stato_civile - cambio_residenza_sisma	613.4159	689.0308	-286.708	0.6789
12	ansia - tipo_veicolo - finemese - lavoro - batteriainfo - sesso - occupazione_fuori_campagna - disabili_famiglia - piano_abitazione - n_veicoli - cambio_residenza - componenti_famiglia	607.9812	672.2538	-286.9906	0.6697
2	ansia - tipo_veicolo	591.284	617.7492	-288.642	0.6606
1	ansia	588.685	607.5887	-289.3425	0.6422
0	-	982.0407	997.1637	-487.0203	0.3761

Figura 5.4: Forward Selection su Adjacent Category Model

Numerovariable	x	aic	bic	loglik	accuracy
24	componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insomnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	566.8436	676.4852	-254.4218	0.7523
25	occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insomnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	565.824	679.2463	-252.912	0.7523
27	studio - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insomnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	568.8769	689.8607	-252.4385	0.7523
28	residenza_studio - lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insomnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	571.3546	696.1191	-252.6773	0.7523
29	stato_civile_residenza_studio_lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insomnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	571.5325	700.0778	-251.7663	0.7523
26	occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insomnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	567.1362	684.3393	-252.5681	0.7431
31	eta_sesso - stato_civile_residenza_studio_lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insomnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	565.8767	701.9835	-246.9384	0.7431
30	eta_sesso - stato_civile_residenza_studio_lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insomnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	573.4563	705.7823	-251.7281	0.7431
32	eta_sesso - stato_civile_residenza_studio_lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insomnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	567.6245	707.512	-246.8123	0.7339
23	disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insomnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	564.5375	670.3984	-254.2688	0.7248
20	ansia - sintomi_fisiologici - tempestività_decisioni - insomnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	565.0256	659.5442	-257.5128	0.7156
21	terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insomnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	562.9131	661.2124	-255.4565	0.7156
22	piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insomnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	564.911	666.9911	-255.4555	0.7156
19	ansia - sintomi_fisiologici - tempestività_decisioni - insomnia - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	588.2533	678.9911	-270.1266	0.6972
18	sintomi_fisiologici - tempestività_decisioni - insomnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	758.7831	845.7402	-356.3915	0.5505
17	tempestività_decisioni - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	815.2944	898.4708	-385.6472	0.5413
15	preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	819.7908	895.4057	-389.8954	0.5229
16	preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	820.741	900.1367	-389.3705	0.5046
5	ascensore - n_veicoli - finemese - batteriainfo - batteriafiducia	983.2601	1017.2866	-482.6301	0.4037
11	cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	962.4665	1022.9584	-465.2333	0.4037
4	n_veicoli - finemese - batteriainfo - batteriafiducia	981.2898	1011.5357	-482.6449	0.3945
10	cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo	961.0923	1017.8034	-465.5461	0.3945
6	tipo_abitazione - ascensore - n_veicoli - finemese - batteriainfo - batteriafiducia	984.0702	1021.8776	-482.0351	0.3945
12	fuer_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	964.1399	1028.4126	-465.07	0.3945
2	finemese - batteriainfo	979.7025	1002.387	-483.8512	0.3853
14	estero - fuer_regione - fuer_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	956.4974	1028.3316	-459.2487	0.3853
13	fuer_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	962.3495	1030.4028	-463.1747	0.3853
1	batteriainfo	982.4445	1001.3482	-486.2222	0.3761
3	tipo_veicolo - finemese - batteriainfo - batteriafiducia	984.7378	1014.9837	-484.3689	0.3761
7	casa_proprieta - tipo_abitazione - ascensore - n_veicoli - finemese - batteriainfo - batteriafiducia	985.9853	1027.5735	-481.9926	0.3761
0	-	982.0407	997.1637	-487.0203	0.3761
8	punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - finemese - batteriainfo - batteriafiducia	984.6292	1029.9981	-480.3146	0.3578
9	frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finemese - batteriainfo - batteriafiducia	986.9106	1039.841	-479.4553	0.3578

Figura 5.5: Backward Selection su Continuation Ratio Model

Appendice

Numeri covariate	x	aic	bic	loglik	accuracy
27	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna - studio - frequenza_zone_rosse - batteriafiducia - sintomi_fisiologici - estero - stato_civile - preoccupazione_sismica - casa_proprieta - insomnia</i>	560.6538	677.8568	-249.3269	0.7431
28	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna - studio - frequenza_zone_rosse - batteriafiducia - sintomi_fisiologici - estero - stato_civile - preoccupazione_sismica - casa_proprieta - insomnia - batteriainfo - cambio_residenza_sisma</i>	562.6452	683.629	-249.3226	0.7431
31	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna - studio - frequenza_zone_rosse - batteriafiducia - sintomi_fisiologici - estero - stato_civile - preoccupazione_sismica - casa_proprieta - insomnia - batteriainfo - cambio_residenza_sisma - tipo_veicolo - lavoro</i>	565.8767	701.9835	-246.9384	0.7431
29	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna - studio - frequenza_zone_rosse - batteriafiducia - sintomi_fisiologici - estero - stato_civile - preoccupazione_sismica - casa_proprieta - insomnia - batteriainfo - cambio_residenza_sisma</i>	563.214	687.9785	-248.607	0.7339
32	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna - studio - frequenza_zone_rosse - batteriafiducia - sintomi_fisiologici - estero - stato_civile - preoccupazione_sismica - casa_proprieta - insomnia - batteriainfo - cambio_residenza_sisma - tipo_veicolo - finmeze</i>	567.6245	707.512	-246.8123	0.7339
18	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna</i>	563.6432	646.8196	-259.8216	0.7248
19	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna - studio</i>	565.6429	652.6	-259.8214	0.7248
23	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna - studio - frequenza_zone_rosse - batteriafiducia - sintomi_fisiologici - estero</i>	561.9439	664.024	-253.972	0.7248
26	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna - studio - frequenza_zone_rosse - batteriafiducia - sintomi_fisiologici - estero - stato_civile - preoccupazione_sismica - casa_proprieta</i>	559.0404	672.4627	-249.5202	0.7248
30	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna - studio - frequenza_zone_rosse - batteriafiducia - sintomi_fisiologici - estero - stato_civile - preoccupazione_sismica - casa_proprieta - insomnia - batteriainfo - cambio_residenza_sisma - tipo_veicolo</i>	565.2589	697.585	-247.6295	0.7248
17	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions</i>	562.2766	641.6722	-260.1383	0.7156
20	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna - studio - frequenza_zone_rosse</i>	566.7704	657.5082	-259.3852	0.7156
22	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna - studio - frequenza_zone_rosse - batteriafiducia - sintomi_fisiologici</i>	561.5404	659.8397	-254.7702	0.7156
21	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna - studio - frequenza_zone_rosse - batteriafiducia</i>	568.5213	663.0399	-259.2607	0.7156
24	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna - studio - frequenza_zone_rosse - batteriafiducia - sintomi_fisiologici - estero - stato_civile</i>	561.4561	667.3169	-252.7281	0.7156
25	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma - tempestività decisions - occupazione_fuori_campagna - studio - frequenza_zone_rosse - batteriafiducia - sintomi_fisiologici - estero - stato_civile - preoccupazione_sismica</i>	558.0791	667.7207	-250.0396	0.7156
13	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza</i>	558.7885	623.0612	-262.3943	0.7064
14	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore</i>	560.3663	628.4196	-262.1831	0.7064
15	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli</i>	562.3229	634.157	-262.1614	0.7064
16	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia - punti_accoglienza - ascensore - n_veicoli - fuori_regione_sisma</i>	561.3082	636.9231	-260.6541	0.7064
8	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione</i>	550.096	595.4649	-263.048	0.6972
9	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza</i>	552.0475	601.1972	-263.0238	0.6972
10	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza</i>	554.0472	606.9776	-263.0236	0.6972
11	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80</i>	555.9951	612.7063	-262.9976	0.6972
12	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione - piano_abitazione - residenza - cambio_residenza - terremoto80 - disabili_famiglia</i>	556.9106	617.4025	-262.4553	0.6972
7	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso - fuori_regione</i>	548.263	589.8512	-263.1315	0.6881
5	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia</i>	546.8071	580.8338	-264.4036	0.6789
6	<i>ansia - scosse - tipo_abitazione - eta - componenti_famiglia - sesso</i>	548.8044	586.6119	-264.4022	0.6789
3	<i>ansia - scosse - tipo_abitazione</i>	548.241	574.7062	-267.1205	0.6606
4	<i>ansia - scosse - tipo_abitazione - eta</i>	548.769	579.015	-266.3845	0.6606
2	<i>ansia - scosse</i>	548.748	571.4324	-268.374	0.6514
1	<i>ansia</i>	583.7096	602.6134	-286.8548	0.6422
0	-	982.0407	997.1637	-487.0203	0.3761

Figura 5.6: Forward Selection su Continuation Ratio Model

Numerovarivare	x	aic	bic	loglik	accuracy
24	componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	557.4786	667.1202	-249.7393	0.7431
31	eta - sesso - stato_civile - residenza - studio - lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	557.8683	693.975	-242.9341	0.7339
23	disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	556.2653	662.1261	-250.1326	0.7248
25	occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	556.9208	670.3431	-248.4604	0.7248
29	eta - sesso - stato_civile - residenza - studio - lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	562.3244	690.8697	-247.1622	0.7248
32	eta - sesso - stato_civile - residenza - studio - lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	558.9305	698.818	-242.4652	0.7248
26	occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	557.9418	675.1448	-247.9709	0.7156
27	studio - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	559.8343	680.8181	-247.9171	0.7156
28	residenza - studio - lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	561.3656	686.1301	-247.6828	0.7156
30	sesso - stato_civile - residenza - studio - lavoro - occupazione_fuori_campagna - componenti_famiglia - disabili_famiglia - piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	562.7044	695.0305	-246.3522	0.7156
21	terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	553.2877	651.587	-250.6438	0.7064
22	piano_abitazione - terremoto80 - scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	555.2869	657.367	-250.6435	0.7064
19	ansia - sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - finmese - batteriainfo - batteriafiducia	577.6073	664.5644	-265.8036	0.7064
20	scosse - ansia - sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	553.3803	647.8989	-251.6902	0.6972
18	sintomi_fisiologici - tempestività_decisioni - insonnia - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	745.5723	832.5294	-349.7862	0.5780
17	tempestività_decisioni - preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - finmese - batteriainfo - batteriafiducia	808.2455	891.4218	-382.1227	0.5413
15	preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	812.9563	888.5712	-386.4781	0.5229
16	preoccupazione_sismica - estero - fuori_regione - fuori_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	813.9475	893.3431	-385.9737	0.4954
12	fouir_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	965.321	1029.5936	-465.6605	0.4220
11	cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	963.1216	1023.6135	-465.5608	0.4128
14	estero - fouir_regione - fouir_regione_sistema - cambio_residenza - cambio_residenza_sistema - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	961.6142	1033.4483	-461.8071	0.4037
2	finmese - batteriainfo	980.8507	1003.5352	-484.4254	0.3945
4	n_veicoli - finmese - batteriainfo - batteriafiducia	982.5495	1012.7955	-483.2748	0.3945
5	ascensore - n_veicoli - finmese - batteriainfo - batteriafiducia	984.5238	1018.5505	-483.2619	0.3945
10	cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo	982.3773	1019.0884	-466.1886	0.3945
13	fouir_regione_sistema - cambio_residenza - cambio_residenza_sistema - frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	983.8095	1031.8629	-463.9047	0.3945
3	tipo_veicolo - finmese - batteriainfo	984.1457	1014.3917	-484.0729	0.3853
1	batteriainfo	982.729	1001.6327	-486.3645	0.3761
7	casa_proprieta - tipo_abitazione - ascensore - n_veicoli - finmese - batteriainfo - batteriafiducia	987.3895	1028.9777	-482.6947	0.3761
6	tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo - batteriafiducia	987.6479	1029.2361	-482.824	0.3761
0	-	982.0407	997.1637	-487.0203	0.3761
8	punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo	985.7501	1031.119	-480.8751	0.3578
9	frequenza_zone_rosse - punti_accoglienza - casa_proprieta - tipo_abitazione - ascensore - n_veicoli - tipo_veicolo - finmese - batteriainfo	986.5789	1039.5093	-479.2894	0.3486

Figura 5.7: Backward Selection su Proportional Odds Model

Appendice

Numerico covariante	x	aic	bic	loglik	accuracy
22	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni - fuori_regione - eta - occupazione_fuori_campagna - ascensore	563.9591	666.0391	-254.9795	0.7339
23	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni - fuori_regione - eta - occupazione_fuori_campagna - ascensore - estero	563.9701	669.831	-253.9851	0.7339
24	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni - fuori_regione - eta - occupazione_fuori_campagna - ascensore - estero - frequenza_zone_rosse	565.3138	674.9554	-253.6569	0.7339
25	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni - fuori_regione - eta - occupazione_fuori_campagna - ascensore - estero - frequenza_zone_rosse - lavoro	564.8075	678.2298	-252.4038	0.7339
26	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni - fuori_regione - eta - occupazione_fuori_campagna - ascensore - estero - frequenza_zone_rosse - lavoro - finemese	566.6483	683.8513	-252.3241	0.7339
20	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni - fuori_regione - eta	562.1988	656.7174	-256.0994	0.7248
21	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni - fuori_regione - eta - occupazione_fuori_campagna	562.3071	660.6064	-255.1535	0.7248
27	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni - fuori_regione - eta - occupazione_fuori_campagna - ascensore - estero - frequenza_zone_rosse - lavoro - finemese - terremoto80	568.1825	689.1663	-252.0913	0.7248
31	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni - fuori_regione - eta - occupazione_fuori_campagna - ascensore - estero - frequenza_zone_rosse - lavoro - finemese - terremoto80 - componenti_famiglia - fuori_regione_sisma - sintomi_fisiologici - casa_proprieta - cambio_residenza_sisma	558.9709	695.0777	-243.4855	0.7248
32	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni - fuori_regione - eta - occupazione_fuori_campagna - ascensore - estero - frequenza_zone_rosse - lavoro - finemese - terremoto80 - componenti_famiglia - fuori_regione_sisma - sintomi_fisiologici - casa_proprieta - cambio_residenza_sisma	558.9305	698.818	-242.4652	0.7248
11	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo	548.9974	609.4893	-258.4987	0.7156
12	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza	550.7832	615.0558	-258.3916	0.7156
13	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio	552.6191	620.6725	-258.3095	0.7156
14	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia	554.3545	626.1887	-258.1773	0.7156
15	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli	556.3433	631.9581	-258.1716	0.7156
16	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo	558.1755	637.5711	-258.0877	0.7156
17	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia	560.1298	643.3061	-258.0649	0.7156
19	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni - fuori_regione	562.9797	653.7175	-257.4898	0.7156
30	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni - fuori_regione - eta - occupazione_fuori_campagna - ascensore - estero - frequenza_zone_rosse - lavoro - finemese - terremoto80 - componenti_famiglia - fuori_regione_sisma - sintomi_fisiologici	558.1356	690.4617	-244.0678	0.7156
28	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni - fuori_regione - eta - occupazione_fuori_campagna - ascensore - estero - frequenza_zone_rosse - lavoro - finemese - terremoto80 - componenti_famiglia	565.7771	690.5417	-249.8886	0.7156
9	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza	544.084	593.2336	-259.042	0.7064
10	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia	545.3811	598.3115	-258.6906	0.7064
18	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni	561.52	648.4771	-257.76	0.7064
29	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza - cambio_residenza - disabili_famiglia - tipo_veicolo - punti_accoglienza - studio - insonnia - n_veicoli - batteriainfo - batteriafiducia - tempestività_decisioni - fuori_regione - eta - occupazione_fuori_campagna - ascensore - estero - frequenza_zone_rosse - lavoro - finemese - terremoto80 - componenti_famiglia	562.5799	691.1252	-247.29	0.7064
7	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione	540.3193	581.9074	-259.1596	0.6972
8	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile - piano_abitazione - residenza	542.1507	587.5196	-259.0754	0.6972
6	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso - stato_civile	538.8876	576.695	-259.4438	0.6881
5	ansia - scosse - preoccupazione_sismica - tipo_abitazione - sesso	537.4576	571.4843	-259.7288	0.6789
4	ansia - scosse - preoccupazione_sismica - tipo_abitazione	535.8773	566.1233	-259.9387	0.6697
3	ansia - scosse - preoccupazione_sismica	536.9968	563.462	-261.4984	0.6606
2	ansia - scosse	541.2054	563.8899	-264.6027	0.6514
1	ansia	580.4341	599.3378	-285.2171	0.6422
0	-	982.0407	997.1637	-487.0203	0.3761

Figura 5.8: Forward Selection su Proportional Odds Model

Qr-Code Questionario



Figura 5.9: Qr-code del questionario

Ringraziamenti

Al termine di questo incredibile viaggio sarebbero troppe le cose da dire e da raccontare.

Innanzitutto ringrazio il mio relatore il professore Iodice D'Enza e la mia corretrice la professoressa Iannario per tutti i consigli dispensati in questi mesi, per le indicazioni, per le dritte e soprattutto per la passione che hanno trasmesso nei loro fantastici corsi.

Ringrazio, in generale, tutto il corpo docenti per avermi coccolato dal primo all'ultimo giorno, perchè ho imparato che dietro ogni professore c'è una grande professionista, ma anche una persona. Vi ringrazio perchè non avrei potuto fare scelta migliore con questa magistrale.

Quante esperienze ed insegnamenti raccolti in questo percorso, quante opportunità, quante soddisfazioni e quante crisi prima di un esame. Quanti momenti condivisi insieme al mio gruppo, ai miei amici di percorso, le colonne portanti di questa incredibile e assurda avventura: Antonio e Stella. Amici miei, io da voi ho solo da imparare, e con questo non intendo di certo produttorie e cicli for. Io vi devo ringraziare per aver concesso ad una persona insicura ed impacciata nei confronti del mondo di aprirsi e combattere la sua sua guerra più grande: quella verso se stesso. Mi avete bacchettato quando c'era da bacchettare e sostenuto quando c'era da sostenere. Non avrei potuto incontrare persone migliori in questi 2 anni accademici, vi voglio bene.

Ringrazio Sabrina, per tutte le docce mancate e le ore spese a studiare inferenza; Martina, perchè in un modo o nell'altro c'eri quando avevo bisogno di te; Angela, perchè abbiamo capito che cane e gatto possono convivere nella stessa cuccia.

Ringrazio tutte le persone incontrate durante questo percorso: i ragazzi del corso,

i ragazzi dell'associazione, i ragazzi della triennale, tutti i ragazzi del tutorato e i ragazzi del BIP. Spero di non aver mancato nessuno.

Ringrazio mia mamma, che non ha mai distolto lo sguardo dal mio percorso accademico. Che io possa un giorno raggiungere la tua maturità e la tua forza nell'affrontare la vita.

Ringrazio la mia compagna, finalmente ce l'abbiamo fatta. Quanti «non ce la farò mai», «non sono all'altezza» e «vado al prossimo appello» ti sei dovuta sorbire. Grazie per essere stata sempre dalla mia parte.

Ringrazio il mio amico Salvo, un punto saldo da 11 anni a questa parte.

Ringrazio la mia famiglia, tutta. Spero che con questo traguardo possa io rendervi, anche se un minimo, orgogliosi di me.

Infine, ringrazio a chi questa tesi è dedicata: dal 2008, dal primo all'ultimo esame ci sei sempre stato ed una volta finito il tuo compito sei andato via senza far troppo rumore. Grazie bumbu mio, ti porto per sempre nel cuore, sulla pelle ed ora sulla tesi.

A chi è andato via, a chi è arrivato e a chi arriverà: a Tay, che è arrivato così dal nulla. Ti devo tanto, tu lo sai.

Infine voglio concludere con una frase di una canzone: *«È andata ma se ci ripensi: che razza di rischi ti sei preso? ed il fatto che non ti sei mai arreso è un miracolo e va difeso.»*

Grazie a tutti.

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] Alan Agresti. *Foundations of linear and generalized linear models*. John Wiley & Sons, 2015.
- [4] Alan Agresti e Claudia Tarantola. «Simple ways to interpret effects in modeling ordinal categorical data». In: *Statistica Neerlandica* 72.3 (2018), pp. 210–223.
- [5] Rollin Brant. «Assessing proportionality in the proportional odds model for ordinal logistic regression». In: *Biometrics* (1990), pp. 1171–1178.
- [6] Mohammad Ziaul Islam Chowdhury e Tanvir C Turin. «Variable selection strategies and its importance in clinical prediction modelling». In: *Family medicine and community health* 8.1 (2020).
- [7] Stephen R Cole e Cande V Ananth. «Regression models for unconstrained, partially or fully constrained continuation odds ratios». In: *International journal of epidemiology* 30.6 (2001), pp. 1379–1382.
- [8] Scott A Czepiel. «Maximum likelihood estimation of logistic regression models: theory and implementation». In: *Available at czep. net/stat/mlelr. pdf* 83 (2002).
- [9] NİMET Dolgun e Osman Saracbasi. «Assessing proportionality assumption in the adjacent category logistic regression model». In: *Statistics and its Interface* 7.2 (2014).

- [10] Morten W Fagerland e David W Hosmer. «How to test for goodness of fit in ordinal logistic regression models». In: *The Stata Journal* 17.3 (2017), pp. 668–686.
- [11] 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.
- [12] Jerome Friedman et al. «Pathwise coordinate optimization». In: (2007).
- [13] Marco Gherghi, Carlo Lauro et al. *Appunti di analisi dei dati multidimensionali*. RCE edizioni, 2008.
- [14] Michael Greenacre e Jorg Blasius. *Multiple correspondence analysis and related methods*. CRC press, 2006.
- [15] Michael J Greenacre. *Biplots in practice*. Fundacion BBVA, 2010.
- [16] Michael J Greenacre. «Theory and applications of correspondence analysis». In: *(No Title)* (1984).
- [17] Trevor Hastie, Robert Tibshirani e Ryan J Tibshirani. «Extended comparisons of best subset selection, forward stepwise selection, and the lasso». In: *arXiv preprint arXiv:1707.08692* (2017).
- [18] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.
- [19] Hyun Yung Lee. «Goodness-of-fit tests for a proportional odds model». In: 24.6 (2013), pp. 1465–1475.
- [20] 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.
- [21] Peter McCullagh. «Regression models for ordinal data». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2 (1980), pp. 109–127.

- [22] Stepan Nersisyan et al. «ExhauFS: exhaustive search-based feature selection for classification and survival regression». In: *PeerJ* 10 (2022), e13200.
- [23] Bercedis Peterson e Frank E Harrell Jr. «Partial proportional odds models for ordinal response variables». In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 39.2 (1990), pp. 205–217.
- [24] D. Piccolo. *Statistica. Strumenti* / il Mulino. Il Mulino, 2010.
- [25] Erik Pulkstenis e Timothy J Robinson. «Goodness-of-fit tests for ordinal response regression models». In: *Statistics in medicine* 23.6 (2004), pp. 999–1014.
- [26] Alessandra Salvan et al. *Modelli lineari generalizzati*. Springer, 2020.
- [27] 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.
- [28] Gerhard Tutz e Jan Gertheiss. «Regularized regression for categorical data». In: *Statistical Modelling* 16.3 (2016), pp. 161–200.
- [29] Gerhard Tutz e Torsten Scholz. «Ordinal regression modelling between proportional odds and non-proportional odds». In: (2003).
- [30] Michel Van de Velden, A Iodice D’Enza e Francesco Palumbo. «Cluster correspondence analysis». In: *Psychometrika* 82 (2017), pp. 158–185.
- [31] Stephen J Wright. «Coordinate descent algorithms». In: *Mathematical programming* 151.1 (2015), pp. 3–34.
- [32] Tong Tong Wu e Kenneth Lange. «Coordinate descent algorithms for lasso penalized regression». In: (2008).
- [33] 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).
- [34] Faisal M Zahid e Shahla Ramzan. «Ordinal ridge regression with categorical predictors». In: *Journal of Applied Statistics* 39.1 (2012), pp. 161–171.

- [35] Zhongheng Zhang. «Variable selection with stepwise and best subset approaches». In: *Annals of translational medicine* 4.7 (2016).
- [36] 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.