Arrests 2010 NTA: Analisi

Contents

Analisi NTA 2010 - 2011 interazione tra variabili

Descrizione

L'obbiettivo di questa sezione di analisi è verificare se esiste un sottoinsieme di variabili esplicative particolarmente correlate con il numero di arresti, sia marginalmente che considerando l'interazione con ciascuna zona spaziale (NTA). Per vincoli computazionali si riduce l'insieme di stima al solo anno 2010: per quest'anno i dati del censo sono esatti e non si sono verificati eventi rari a differenza del 2020 (Covid); l'insieme di verifica scelto è l'anno 2011, in quanto è l'anno più vicino al 2010 (l'assunzione è che i due anni siano abbastanza simili per il fenomeno considerato).

Problematiche

Questi dati presentano diverse problematiche.

Le scelte fatte sono dovute a fattori computazionali, di tempo e al fatto che per permettere conteggi diversi da 1 è necessario considerare zone spaziali e intervalli temporali non eccessivamente ristretti.

Per quanto concerne gli intervalli temporali si è scelto di ignorare i possibili trend e considerare i singoli anni, per ciascun anno si sono utilizzati i mesi per la costruzione degli insiemi di convalida incrociata. Pur avendo a disposizione il giorno di ciascun arresto la selezione dei mesi è apparsa come un giusto compromesso per garantire che non tutti i conteggi fossero uguali a 1 che comunque è il conteggio minimo e più frequente molto superiore a tutti gli altri conteggi:

Questa "sovradispersione" di 1 è ragionevole data la costruzione dei conteggi per aggregazione di osservazioni con le stesse combinazioni di covariate; si dovrebbero inoltre aggiungere osservazioni con conteggi nulli per ogni combinazione di variabili per cui non si sono osservati arresti.

Mantenere tutti i conteggi unitari rende computazionalmente molto oneroso l'addattamenteo dei modelli e può creare problemi nella selezione degli stessi

La soluzione adottata è basata sul sottocampionamento: si stabilisce una soglia per i valori di conteggi oltre cui non sottocampionare, si conta la frequenza di conteggi osservati per tale soglia e si sottocampiona un sottoinsieme di grandezza uguale a quella frequenza da ciascun sottoinsieme di conteggi con valori inferiori alla soglia.

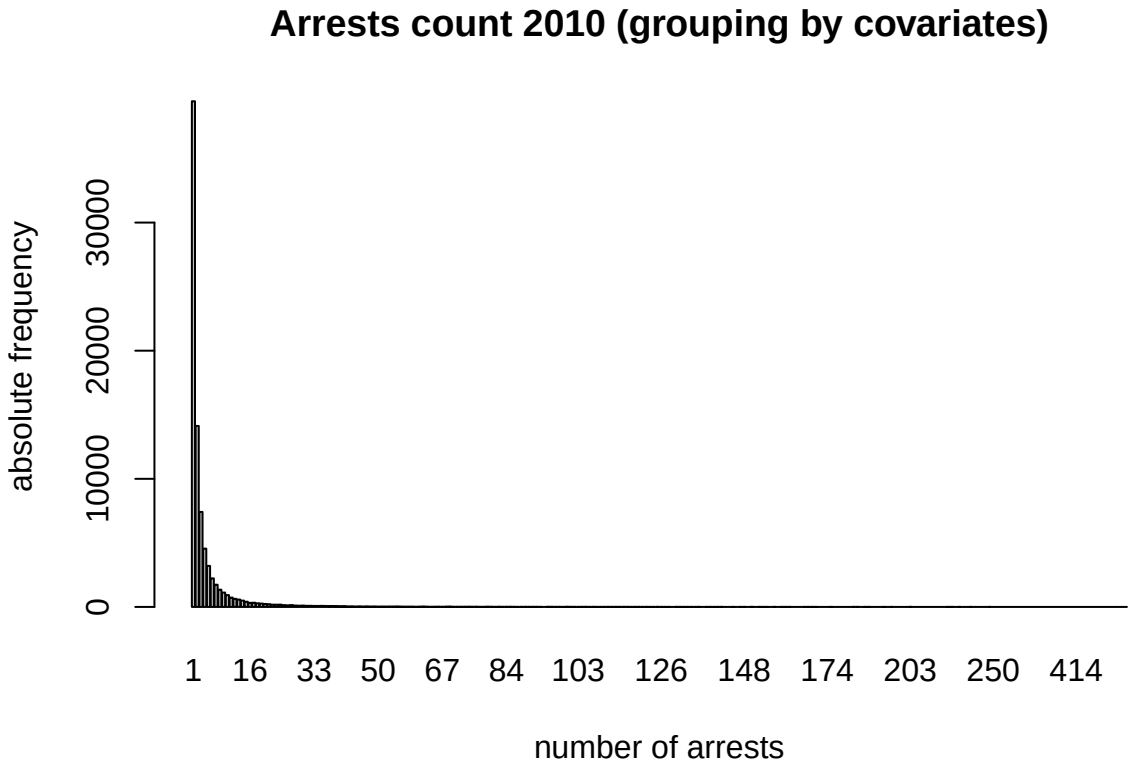


Figure 1: Tabella di frequenza assoluta per il numero di arresti raggruppando i dati del 2010 per tutte le variabili ad eccezione dei mesi

L'assunzione di fondo è che, almeno per i conteggi fino alla soglia considerata, la frequenza sia descrescente rispetto al valore degli stessi; un aspetto da sottolineare della metodologia proposta, in quanto compromesso, è che introduce distorsione nelle stime.

Si considerano due tipologie di dataset, entrambi impiegano il sottocampiomamento, ma in uno i conteggi nulli sono presenti e nell'altro sono assenti (i modelli per risposta continua sono adattati impiegando una trasformazione logaritmica dei dati senza conteggi nulli).

Per questo studio la soglia selezionata che è apparsa ragionevole in base alle considerazioni precedenti è di conteggi uguali a 10.

Il sottocampionamento è effettuato anche per i dataset completi relativi a 2010 e 2011 (escludendo i mesi come variabile di raggruppamento), ma non sono stati aggiunti gli zeri per permettere il confronto tra modelli per risposta continua.

Elevata dimensionalità

I dati presentano elevata dimensionalità considerando le interazioni tra la variabile spaziale (NTA) e le covariate (qualitative) di arrests. E' comunque interessante provare i metodi di selezione delle variabili anche sui dati senza interazioni.

Per avere delle misure quantitative si considera il dataset in cui si sono definiti i conteggi senza considerare i mesi: si riporta il rapporto tra il numero di osservazioni (righe) e il prodotto tra il numero di modalità di NTA e la somma delle modalità delle variabili qualitative di arrests.

Senza considerare interazioni tra KY_CD (esplicativa non spaziale di arrest con più modalità) con gli NTA il rapporto è (considerando i dati 2011):

```
##  NTA2020
## 3.816125
```

Considerando anche interazioni tra KY_CD e NTA rapporto è:

```
##   NTA2020
## 0.9178023
```

Modelli

Criterio di selezione dei parametri di regolazione

Come già accennato, i parametri di regolazione sono selezionati tramite convalida incrociata (CV) impiegando i mesi per la costruzione degli insiemi. La funzione di perdita scelta è il RMSE.

La procedura per la costruzione degli insiemi è la seguente:

- Selezione di k: il numero di insiemi di convalida (ad esempio k = 4)
- Ogni insieme di convalida è composto da osservazioni raggruppate di 12 / k (3) mesi e i mesi rimanenti (9) vengono utilizzati per adattare il modello.
- Per cercare di compensare e mediare le fluttuazioni stagionali, i mesi di validazione sono scelti il più distanziati possibile. Ad esempio, nel caso di k = 4, il primo insieme di validazione è (gennaio, maggio, settembre), il secondo set è (febbraio, giugno, ottobre), il terzo è (marzo, luglio, novembre) e il quarto è (aprile, agosto, dicembre).
- Per rendere ogni risposta comparabile avendo utilizzato un numero diverso di mesi, una nuova risposta è definita come il rapporto degli arresti diviso per il numero di mesi utilizzati nel raggruppamento (ovvero l'esponenziale dell 'offset nel modello di Poisson).

Matrice del modello

La matrice del modello considerata è quella con tutte le variabili e le interazioni tra tutte le variabili di arrests tranne KY_CD (per ragioni computazionali) e le zone spaziali degli NTA.

Poichè la matrice del modello dell'insieme di verifica e quella dell'insieme di stima non condividono tutte le colonne si considerare solo le colonne in comune alle due.

```
## [1] 17039  4094
```

```
## [1] 17164  4109
```

Esplicative quantitative

IL dataset presenta principalmente esplicative categoriali, benchè le esplicative quantitative permettano la specificazione di diverse forme funzionali qui, per ragioni computazionali ci si limita ad assumere una relazione monotona lineare con la risposta.

Modelli per risposta continua

Nell'impiego dei modelli con risposta continua (con errori gaussiani i.i.d) si è scelto di adattare il modello su una trasformazione logaritmica della risposta: `y = count / n_month_train` (per i dati senza introduzione di conteggi nulli) e calcolare l'errore di previsioni sulla trasformazione `count = exp(y) * n_month_test` rispetto al numero di conteggi osservati, in questo modo la previsione è sempre positiva.

Modelli e procedure considerati

I modelli considerati sono modelli normali con penalizzazioni LASSO, Elasticnet, SCAD ed MCP e modelli Poisson con penalizzazioni LASSO ed Elasticnet. Per per tutti i modelli si seleziona il parametro (eventualmente vettoriale) di regolazione che minimizza l'errore di convalida. Per i metodi per cui il parametro di regolarizzazione ha dimensione 2 si definisce una griglia di valori (di cui si riporta il grafico delle curve di livello dell'errore). Per i metodi SCAD e MCP, poichè "ncvreg" presenta dei problemi computazionali dovute alle dimensioni del dataset è impiegata la libreria "picasso" che però non fornisce indicazioni rispetto alle regioni non convesse.

Per il modello normale il λ minimo è molto vicino a zero (poichè la soluzione è sul bordo si dovrebbe provare a diminuire ulteriormente λ, ma già così i coefficienti sono quasi uguali alle stime non penalizzate). Per il modello Poisson il λ selezionato è maggiore. (Figura 2)

Per il modello continuo Elasticnet individua un α intermedio tra Ridge e LASSO, ma come sopra il λ è prossimo a zero (considerazioni uguali a sopra), nel modello Poisson invece è selezionata una Ridge con λ non prossimo a zero. (Figura 3)

Sia per SCAD che per MCP il λ selezionato è prossimo a zero, provando ad aumentare ulteriormente γ e diminuire λ la soluzione sostanzialmente non cambia. (Figura 4)

Per i modelli di Poisson adattati con conteggi nulli i λ ottimi tendono al metodo non penalizzato; nel caso di Elasticnet è selezionata una ridge (Figura 5).

Per confontare modelli per risposta continua e discreta nelle previsioni sui dati 2011 si approssimano le previsioni continue al primo intero

```
## NULL
```

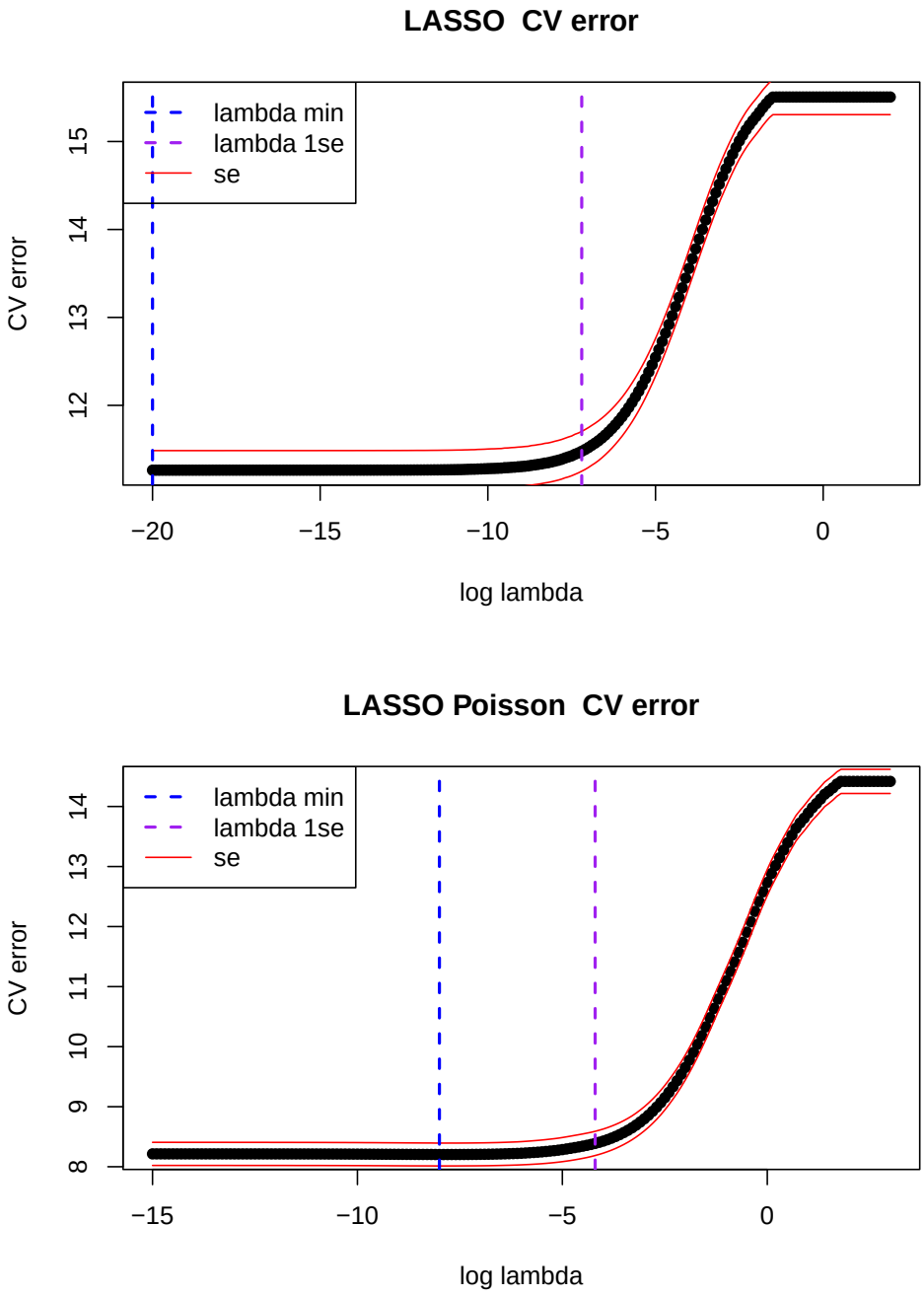


Figure 2: Grafici dell'errore di convalida incrociata in funzione del parametro di regolazione per LASSO per modelli lineare e Poisson

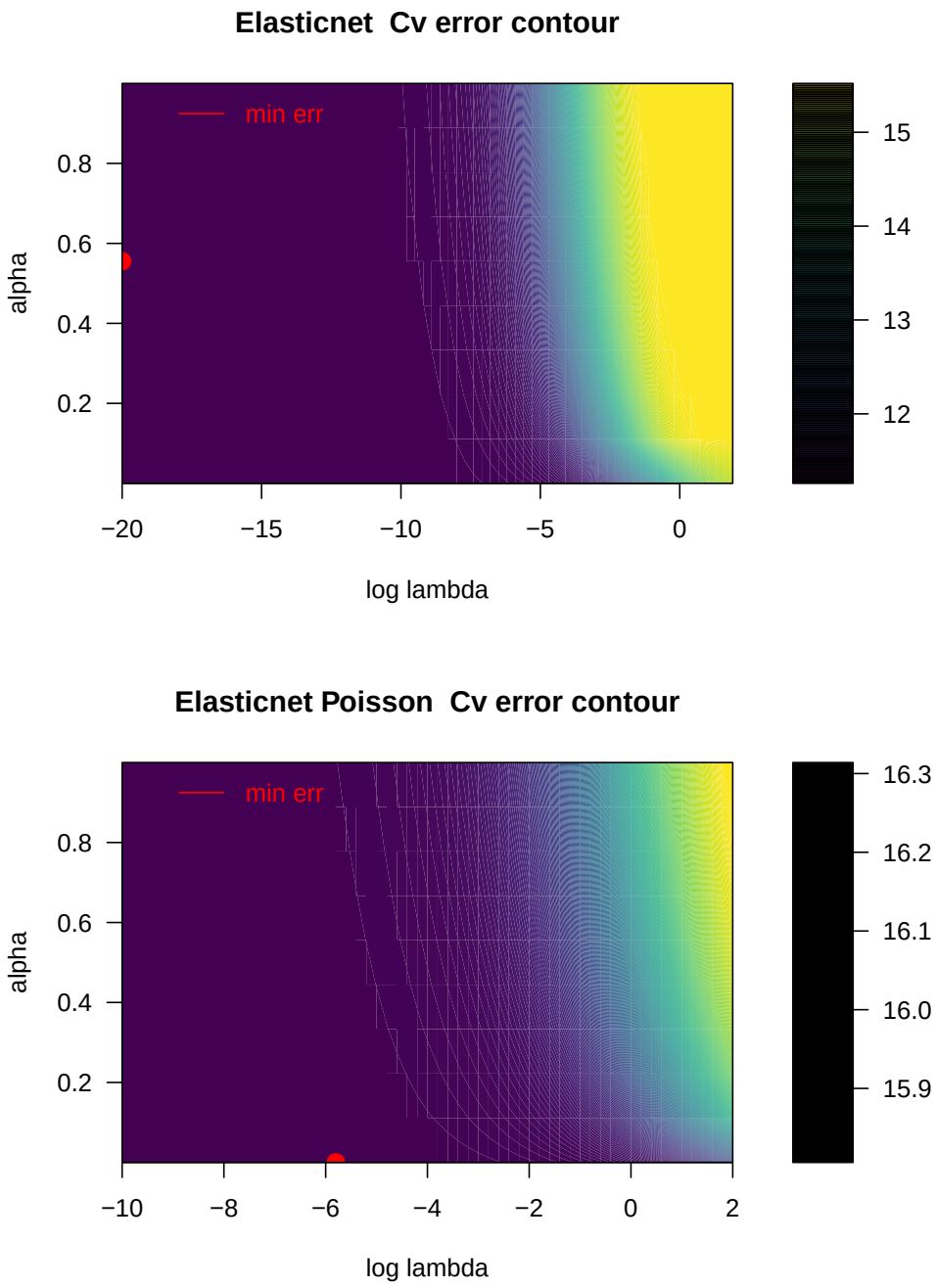


Figure 3: Curve di livello dell'errore di convalida incrociata in funzione dei parametri di regolazione per Elasticnet dei modelli lineare e Poisson

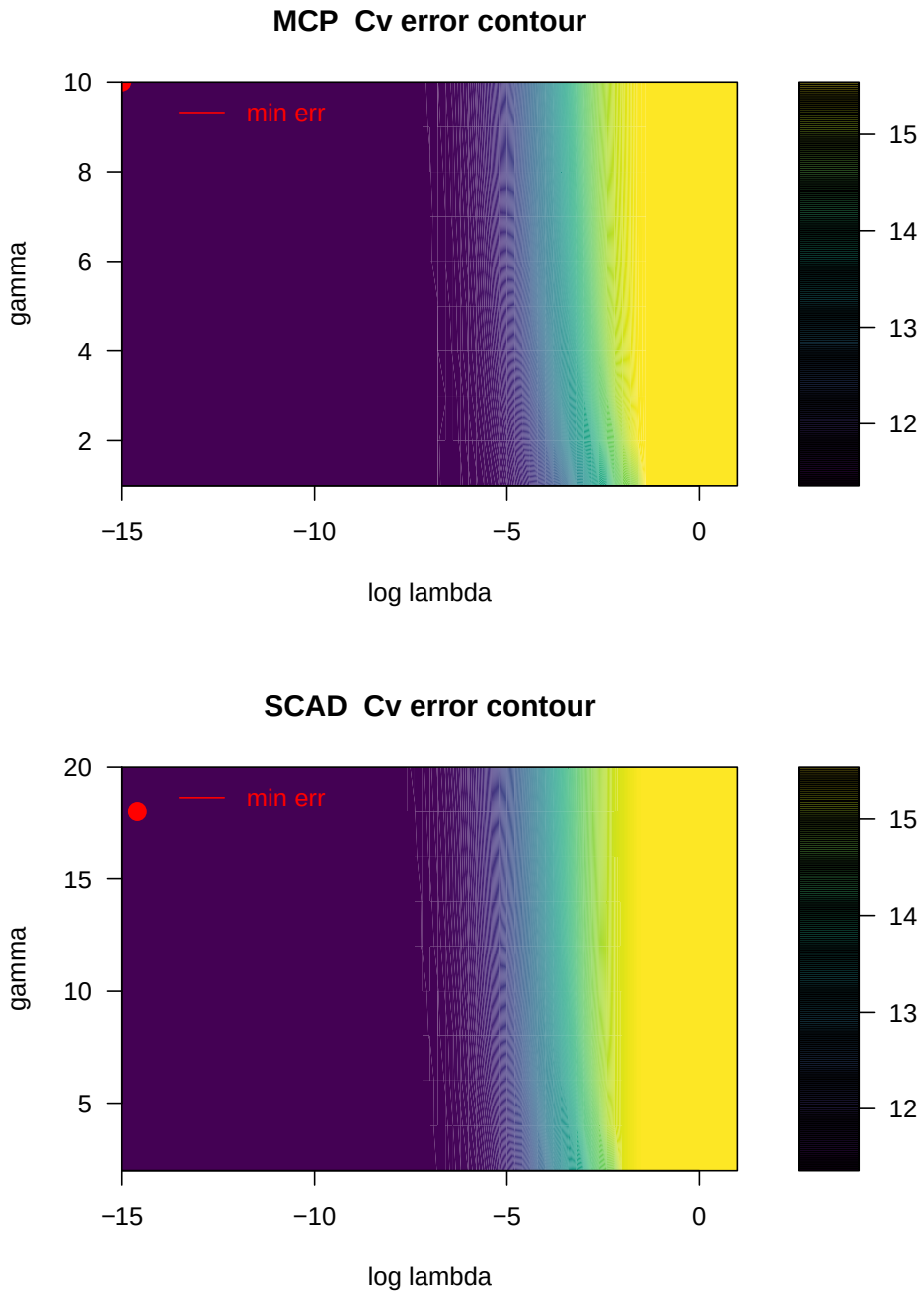


Figure 4: Curve di livello dell'errore di convalida incrociata in funzione dei parametri di regolazione per SCAD ed MCP del modello lineare

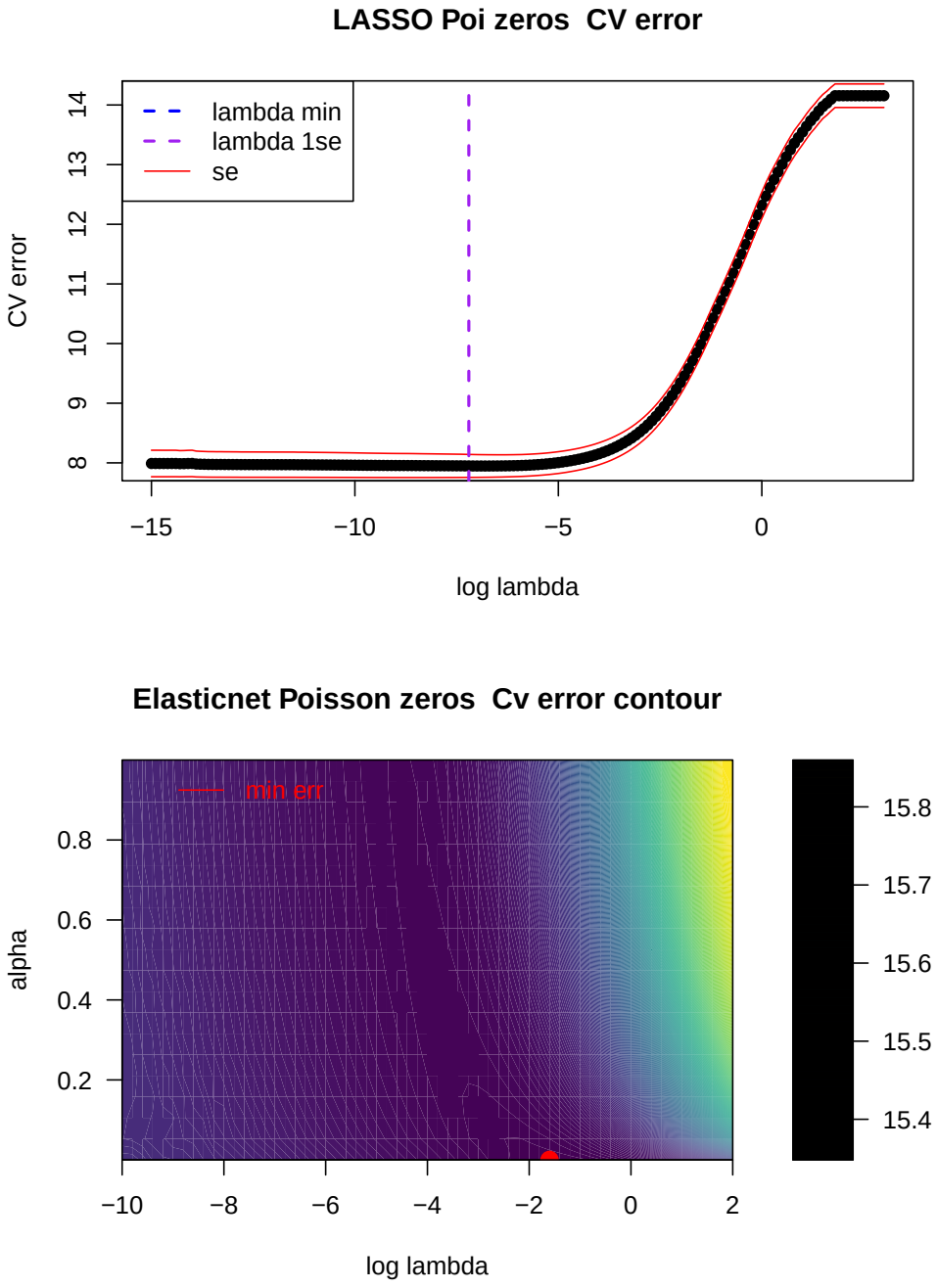


Figure 5: Grafici dell'errore di convalida incrociata in funzione dei parametri di regolazione per LASSO ed Elasticnet per modelli Poisson per dati con aggiunta di zeri

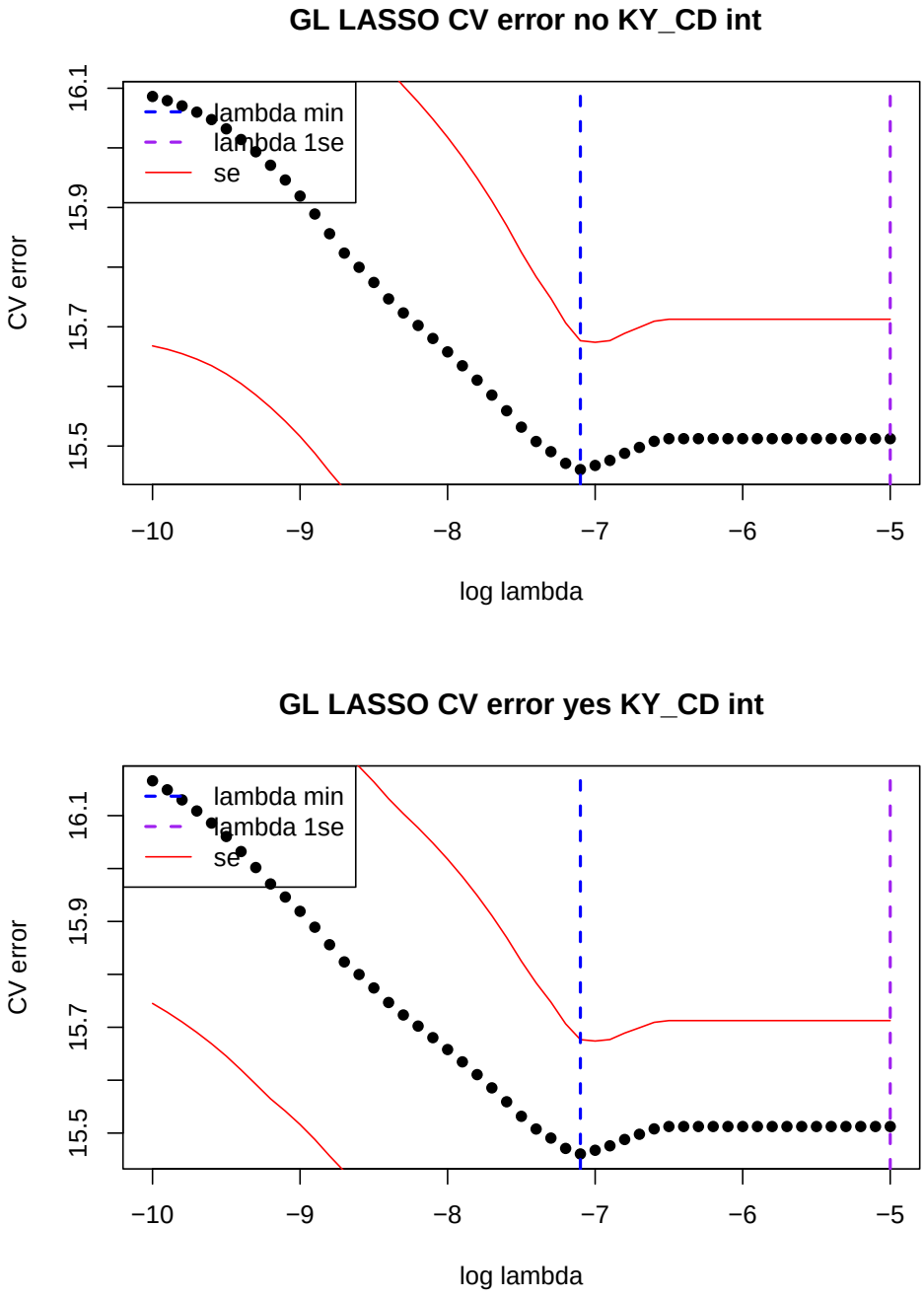


Figure 6: Grafici dell'errore di convalida incrociata in funzione del parametro di regolazione per Grouped LASSO per modelli lineari

Table 1: Errori di previsione sui dati 2011 dei migliori modelli selezionati tramite convalida incrociata

model	test_error
poisson_lasso.zeros.1se	16.76682
poisson_lasso.zeros	16.97896
poisson_lasso	17.07644
scad	20.24735
mcp	20.24735
lasso	21.27275
elasticnet	21.27275
lasso.1se	21.32034
poisson_elasticnet_zeros	30.24320
poisson_lasso1se	30.24875
poisson_elasticnet	30.25221

Modelli migliori

Si riportano (tabella (1)) gli errori di previsione sui dati del 2011 (senza zeri) dei vari modelli migliori stimati sui dati completi 2010 (senza zeri). La migliore previsione si ha per il modello Poisson con penalità LASSO (selezionato sui dati con gli zeri) per λ a errore a un errore standard, mentre il peggiore è sempre il modello di Poisson ma con penalità Elasticnet.

I grafici dei coefficienti stimati confermano, per LASSO ed Elasticnet non avviene selezione delle variabili. Le stime non sono sparse nemmeno con il criterio dell'errore a un errore standard. Anche per SCAD ed MCP non avviene selezione di variabili: le stime (non riportate) sono quasi uguali a quelle LASSO ed Elasticnet, non si reputa quindi necessario controllare l'eventuale regione non convessa delle stime (Figura 7).

I modelli Poisson adattati considerando gli zeri presentano una maggiore selezione di variabili rispetto ai precedenti (Figura 8).

La tabella (2) contiene per ciascun modello il rapporto tra il numero di elementi non nulli e il numero di elementi totali del vettore dei coefficienti: il modello Poisson con penalizzazione LASSO (criterio a un errore standard) e impiegando gli zeri è quello con il rapporto minore, questo modello è dunque l'oggetto delle successive analisi.

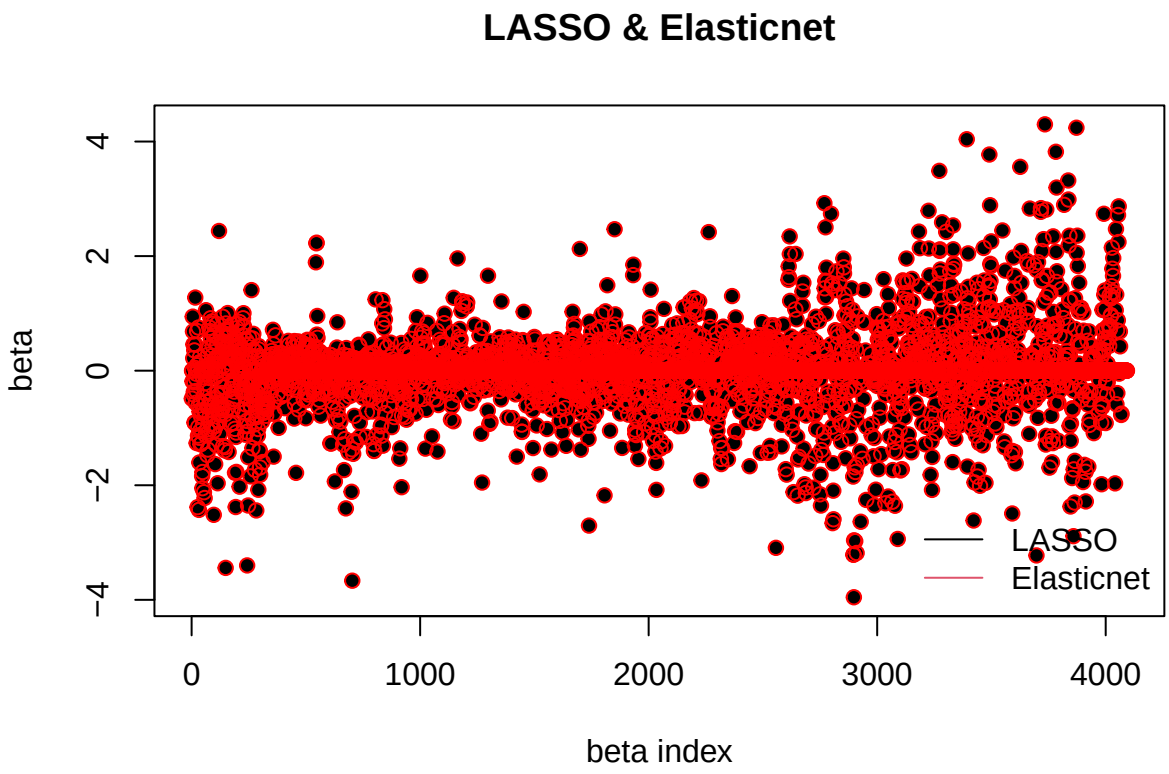
E' interessante notare come il modello con la migliore previsione sui dati 2011 sia quello con penalizzazione LASSO e non Elastic, come ci si sarebbe potuti invece aspettare data la natura di correlazione delle esplicative data dall'introduzione dei termini di interazione. Poichè le stime LASSO per esplicative correlate non sono stabili le conclusioni inferenziali ed intepretative devono essere prese con cautela.

Si ricorda che per GLM di Poisson con legame canonico, in assenza di interazioni, l'aumento unitario di una variabile fissate tutte le altre induce una modifica moltiplicativa nel parametro della media pari all'esponenziale del coefficiente associato alla variabile. Nel caso considerato il parametro è il numero di arresti medi mensili (poichè si è introdotto l'offset). Sono presenti delle ulteriori difficoltà interpretative dei coefficienti dovuti alla standardizzazione delle variabili e al sottocampionamento che di fatto sottostima il numero di conteggi nulli o bassi. In questa analisi si è inoltre più interessati alle variabili associate a un incremento degli arresti più che a una loro diminuzione.

E' riportata la mappa (Figura 9) degli NTA colorati in base al valore del corrispettivo coefficiente (marginale) (le zone grigie corrispondono a zone non presenti nei dati, e quindi coefficienti non stimati), per alcuni NTA sono presenti i nomi dei corrispettivi codici, questi sono in corrispondenza dei coefficienti più grandi relativi a termini di interazione in cui sono presenti tali NTA (vedasi sotto).

Si riportano anche le tabelle dei cofficienti per le variabili qualitative (marginali).

Tutti i gruppi di età presentao coefficienti positivi ad eccezione della fascia più anziana.



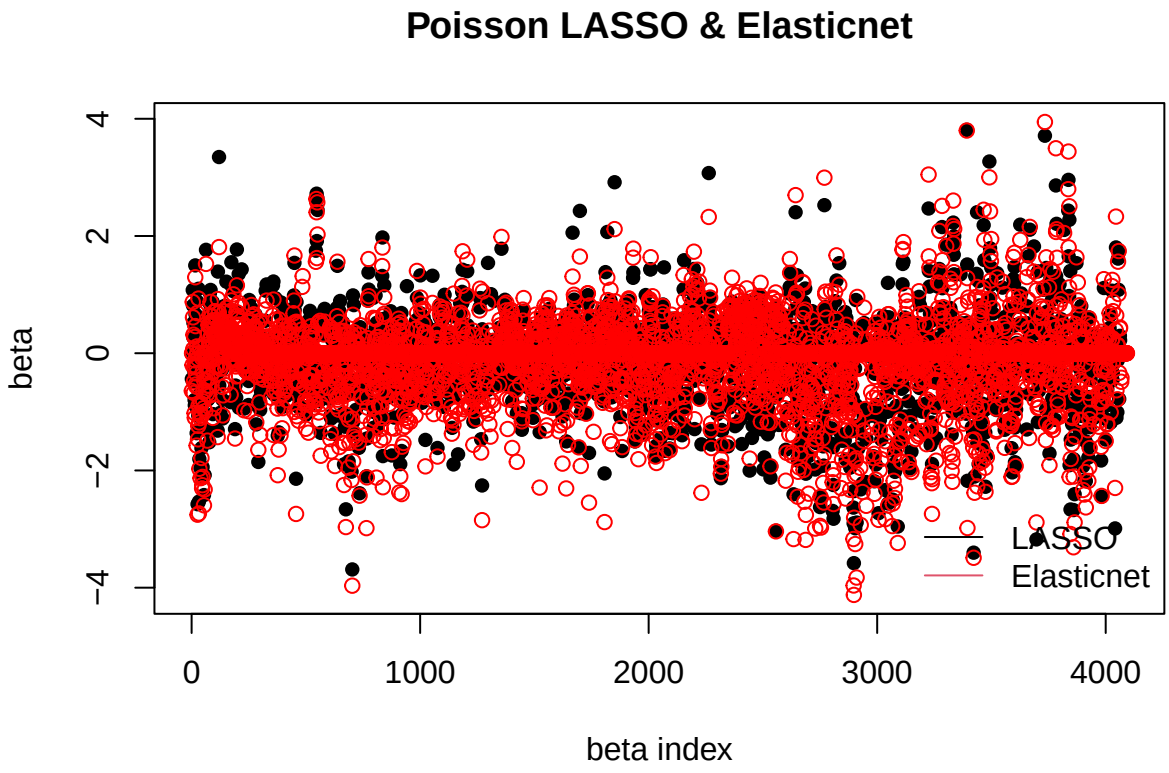


Figure 7: Grafici dei coefficienti stimati sui dati 2010 dei modelli selezionati precedentemente per LASSO ed Elasticnet per modelli lineare e Poisson

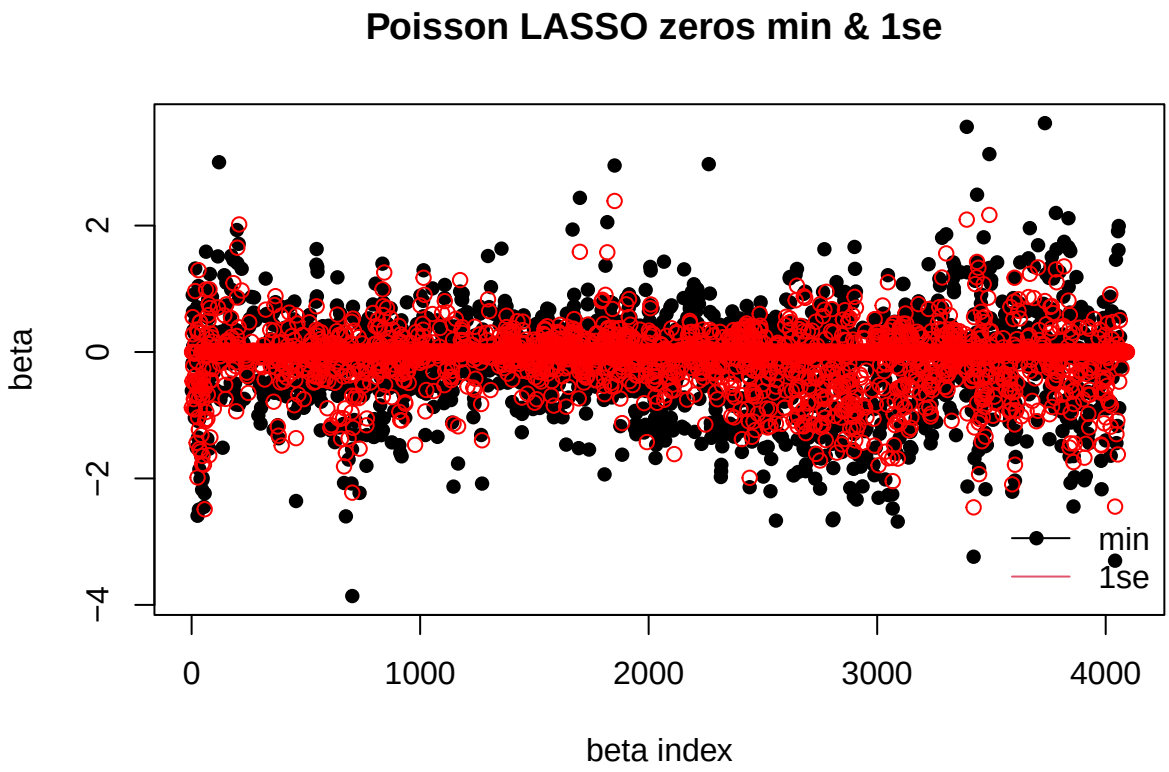


Figure 8: Grafici dei coefficienti stimati sui dati 2010 dei modelli Poisson (LASSO) selezionati precedentemente sui dati con aggiunta di zeri

Table 2: Rapporto del numero di coefficienti non nulli sul numero di coefficienti totali

model	not_null_ratio
poisson_lasso.zeros.1se	0.4273149
poisson_lasso1se	0.4348888
lasso.1se	0.5223552
poisson_lasso.zeros	0.5668214
poisson_lasso	0.6017591
lasso	0.7082824
elasticnet	0.7082824
scad	0.7082824
mcp	0.7082824
poisson_elasticnet	0.7082824
poisson_elasticnet_zeros	0.7082824

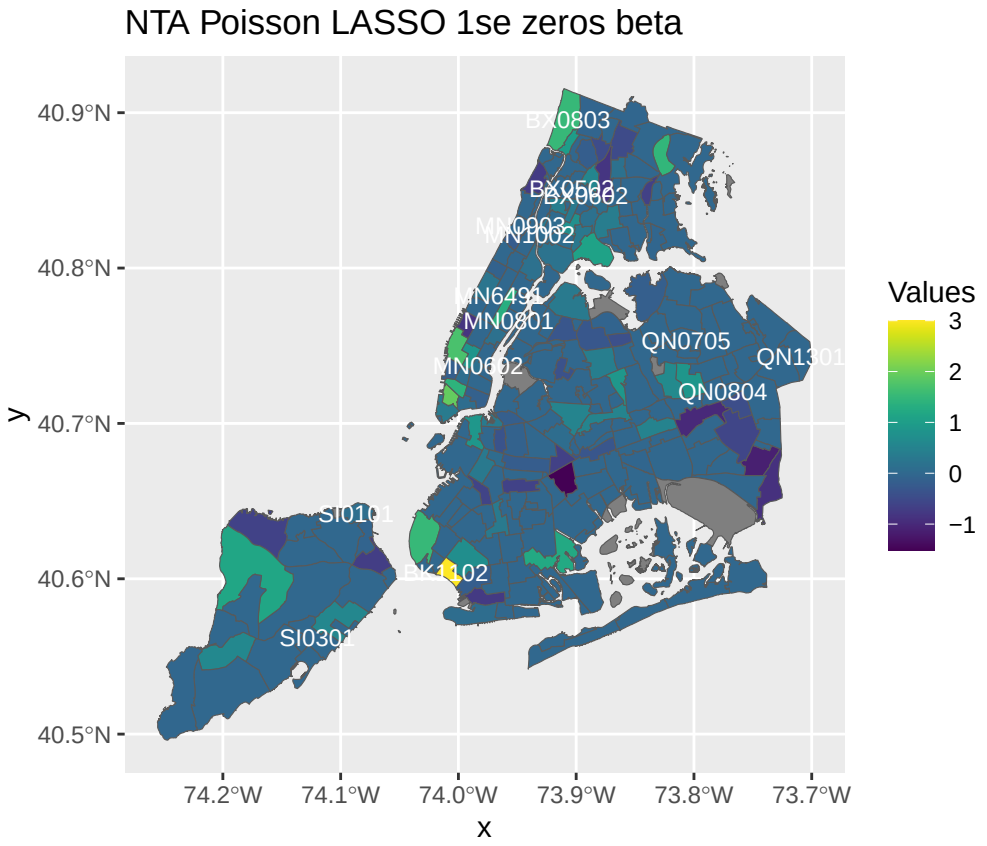


Figure 9: Grafici dei coefficienti stimati sui dati 2010 del modello Poisson (LASSO) selezionato con criterio a un errore standard sui dati con aggiunta di zeri relativi a ciascun NTA

AGE_GROUP	beta
18-24	0.5345937
25-44	0.7143626
45-64	0.1530545
65+	-1.5297436

La percentuale di maschi è associata ad un coefficiente positivo.

PERP_SEX	beta
M	0.992226

I coefficienti positivi sono relativi a etnia bianca, ispanica e nera.

PERP_RACE	beta
ASIAN / PACIFIC ISLANDER	0.0000000
BLACK	1.2340630
BLACK HISPANIC	0.0675111
UNKNOWN	-0.6569302
WHITE	0.8776179
WHITE HISPANIC	0.8468430

I coefficienti relativi alle variabili del censo sono relativamente piccoli rispetto a quelli per le altre variabili.

var	beta
Pop1	0.0000010
MdAge	0.0012548
MaleP	0.0000000
Hsp1P	0.0048905
BNHP	0.0000000
OthNHP	-0.0745797
WNHP	-0.0033235
ANHP	0.0106991
MIncome	0.0000064

Per le macro categorie di reato solo l'omicidio presenta una coefficiente positivo.

LAW_CAT_CD	beta
F	-0.1766663
I	-0.7154991
M	0.8270686
V	-0.4761718

Si riporta il grafico dei coefficienti per le categorie più granulari di arresto (Figura 10).

Poichè si è più interessati ai termini correlati positivamente con il numero di arresti si riportano di seguito i 20 coefficienti più grandi del modello Poisson più sparso descritto sopra.

Il termine più grande e gli altri due che coinvolgono l'interazione tra NTA e la fascia d'età massima compensano il valore negativo del coefficiente marginale per quella fascia d'età per quelle specifiche zone; un ragionamento analogo vale per i termini di interazione che comprendono la modalità "etnia asiatica" (Figura 11)

Conclusioni

Modelli non impiegati

Non sono presenti SCAD ed MCP per verosimiglianza Poisson a causa degli eccessivi tempi computazionali richiesti.

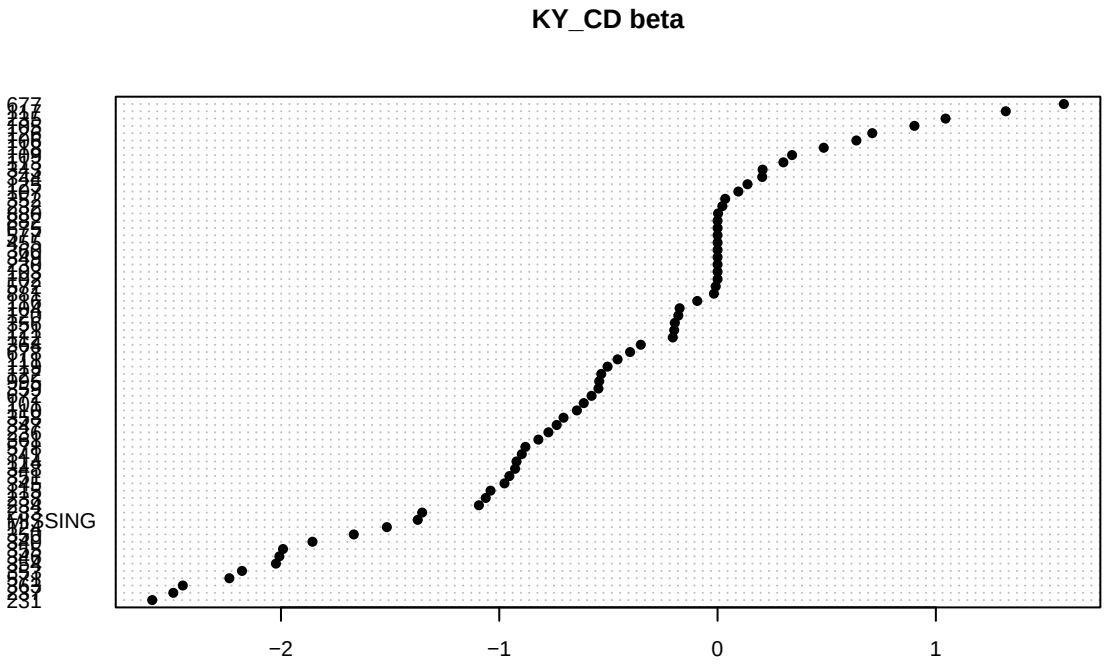


Figure 10: Grafici dei coefficienti stimati sui dati 2010 del modello Poisson (LASSO) selezionato con criterio a un errore standard sui dati con aggiunta di zeri relativi alle categorie granulare di arresto KYCD

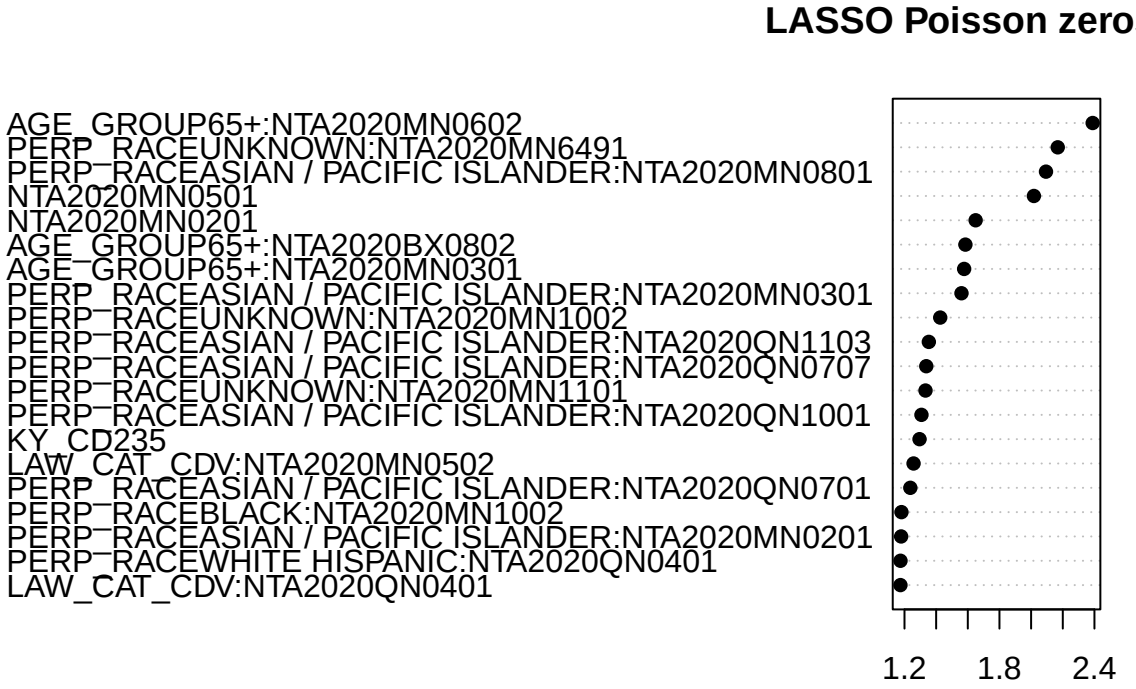


Figure 11: Grafici dei 20 più grandi coefficienti stimati sui dati 2010 del modello Poisson (LASSO) selezionato con criterio a un errore standard sui dati con aggiunta di zeri

Non si è considerato il Grouped LASSO per problemi computazionali.

Per i dati a disposizione un modello solitamente più appropriato di quello Poisson è la binomiale negativa in quanto permette di allentare le ipotesi sull'uguglianza tra media e varianza; provando metodi di stima penalizzati (per cui l'ulteriore parametro per la sovradispersione è un parametro di regolazione) si sono riscontrati vari problemi, ragion per cui tale modello non è presente.

Per i dati originali sarebbe stato opportuno o modello con inflazione sia di zeri che di uni, ma per adattarlo servirerebbero delle metodologie ad hoc dato l'eccessivo numero di righe e colonne.

Risultati