Arrests 2010 NTA: Analisi

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).

Variabili considerate

Viene riportata la lista delle variabili considerate con una breve descrizione.

Le variabili considerate dal dataset arrests sono:

- KY_CD (fattore con 70 livelli): categoria granulare del crimine causa dell'arresto
- LAW_CAT_CD (fattore con 5 livelli): categoriea generale del crimine causa dell'arresto
- AGE_GROUP (fattore con 5 livelli): classe d'età dell'arrestato
- PERP_SEX (fattore con 2 livelli): sesso dell'arrestato
- PERP_RACE (fattore con 7 livelli): etnia dell'arrestato
- NTA2020 (fattore con 251 livelli): indicatore della specifica NTA del luogo dell'arresto
- MONTH (fattore con 12 livelli): indicatore del mese dell'arresto

Le variabili considerate dal censo sono: - Pop1 (numerica): popolazione per NTA - MaleP (numerica): percentuale di maschi per NTA - MdAge (numerica): età mediana per NTA - Hsp1P (numerica): percentuale di ispanici per NTA - WNHP (numerica): percentuale di bianchi non ispanici (NH) per NTA - BNHP (numerica): percentuale di neri non ispanici per NTA - ANHP (numerica): percentuale di asiatici non ispanici per NTA - OthNHP (numerica): percentuale di altre etnie non ispaniche per NTA - MIncome (numerica): reddito mediano per NTA

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 selezione delle variabili di stratificazione per zona da includere, benchè il censo fornisca molte variabili si è scelto di considerarne solo un esiguo sottoinsieme: ciò è dovuto a una mancanza di conoscenza di campo e al non poter dedicare eccessivo tempo alla selezione ed in particolare all'estrazione delle stesse dai varia database e fogli di calcolo.

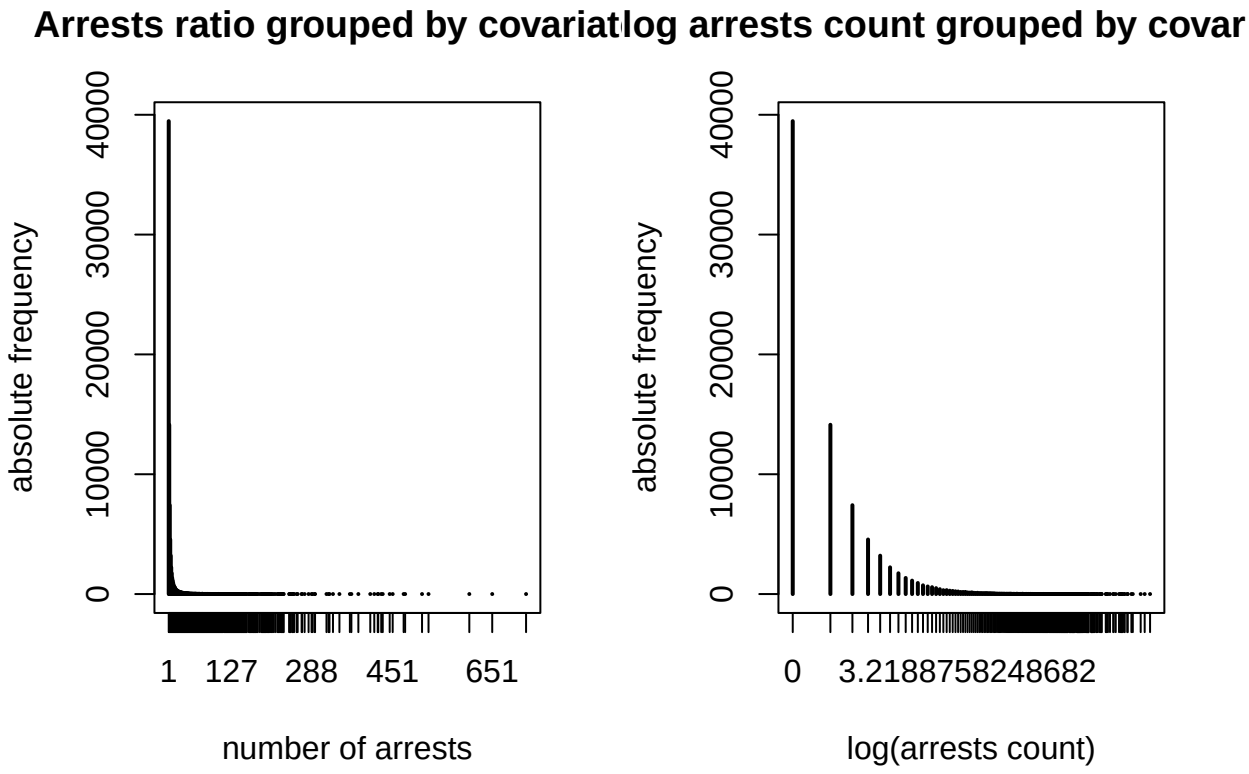
Per la definizione delle zone spaziali e degli intervalli temporali. Inizialmente si era pensato di eseguire le analisi sull'unità spaziale più piccola disponibile per cui sono presenti i dati del censo: i "Census Tracts" (CT) che per New York sono più di 2000 zone distinte (a differenza delle 250 degli NTA), si è però subito verificato che con i mezzi a disposizione lavorare su modelli con i CT era improponibile a causa delle eccessive

risorse computazionali richieste. 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 dei conteggi superiori a 1.

Pur avendo notevolmente ridotto la potenziale complessità del problema il numero di osservazioni e il numero di modalità dei fattori alcune procedure impiegate possono comunque molto tempo (sulle piattaforme di calcolo impiegate) per essere eseguite: questi limiti rendono proibitive valutazioni dell'incertezza nelle stime, quali metodi bootstrap o bayesiani.

Poichè i conteggi sono costruiti raggruppando le osservazioni con le stesse combinazioni di modalità delle covariate, il conteggio minimo osservabile è uguale a 1. Ciò pone dei possibili problemi nella stima di modelli che comprendono lo zero nel supporto (ad esempio il modello di Poisson). Si è comunque provato ad adattare modelli (sia a risposta continua che discreta) su questa tipologia di dati, una delle possibili soluzioni al di fuori di impiegare modelli diversi è la metodologia "zero padding" in cui per ogni combinazione di variabili per cui non si verificano casi si aggiunge un'osservazione con conteggio nullo: la criticità qui è computazionale dovuta all'incremento notevole delle osservazioni.

A sinistra i conteggi di arresti raggruppati per valori di covariate e a destra il logaritmo della medesima quantità



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 KY_CD (esplicativa non spaziale di arrest con più modalità) il rapporto è:

```
##  NTA2020
## 17.59593
```

Considerando anche KY_CD il rapporto è:

```
##  NTA2020
## 3.756435
```

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 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 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 sceglie di ridurre le colonne di entrambe con l'intersezione delle colonne in comune.

```
## [1] 83915  4094
```

```
## [1] 83162  4109
```

Esplicative quantitative

Benchè le eslplicative quantitative permettano la specificazione di diverse forme funzionali qui, per ragioni computazionali ci si limita ad assumere (forzatamente) una relazione monotona lineare con la risposta.

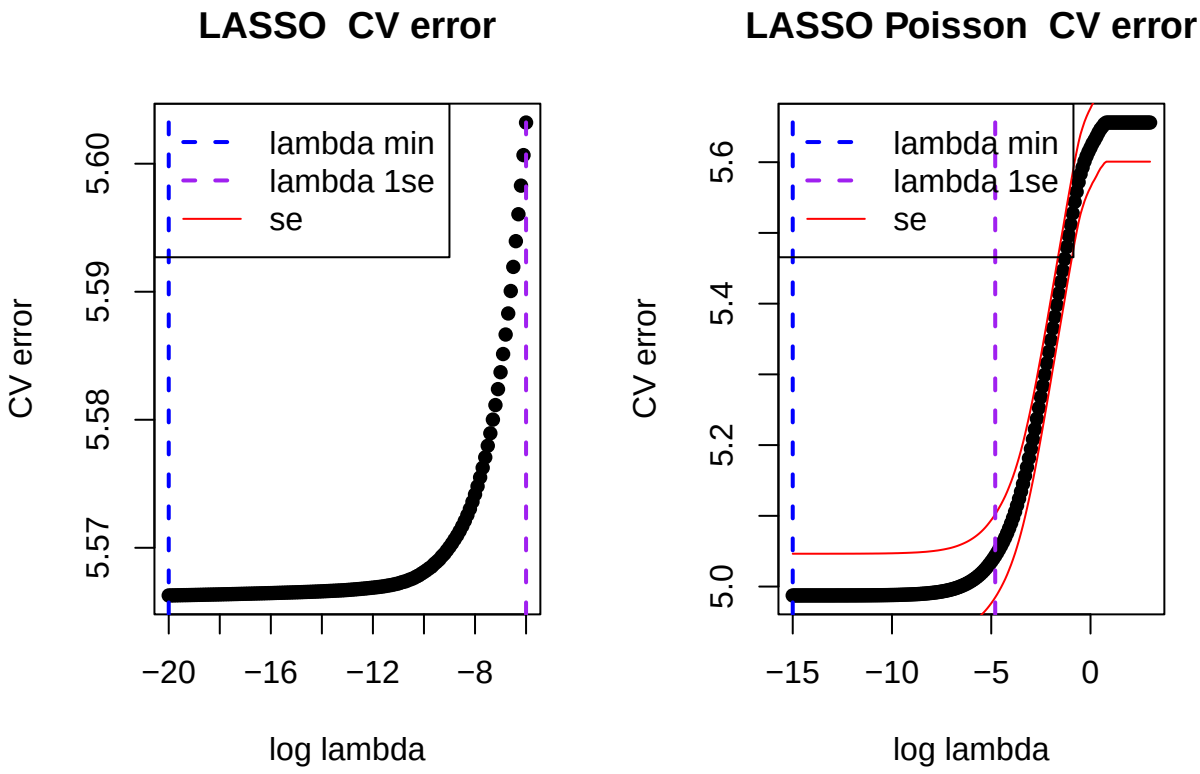
Modelli per risposta continua

Nell'impiego dei modelli con risposta continua (con errori gaussiani i.i.d) si è scelto di stimare il modello su una trasformazione logaritmica della risposta: "y = count / n_month_train" 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

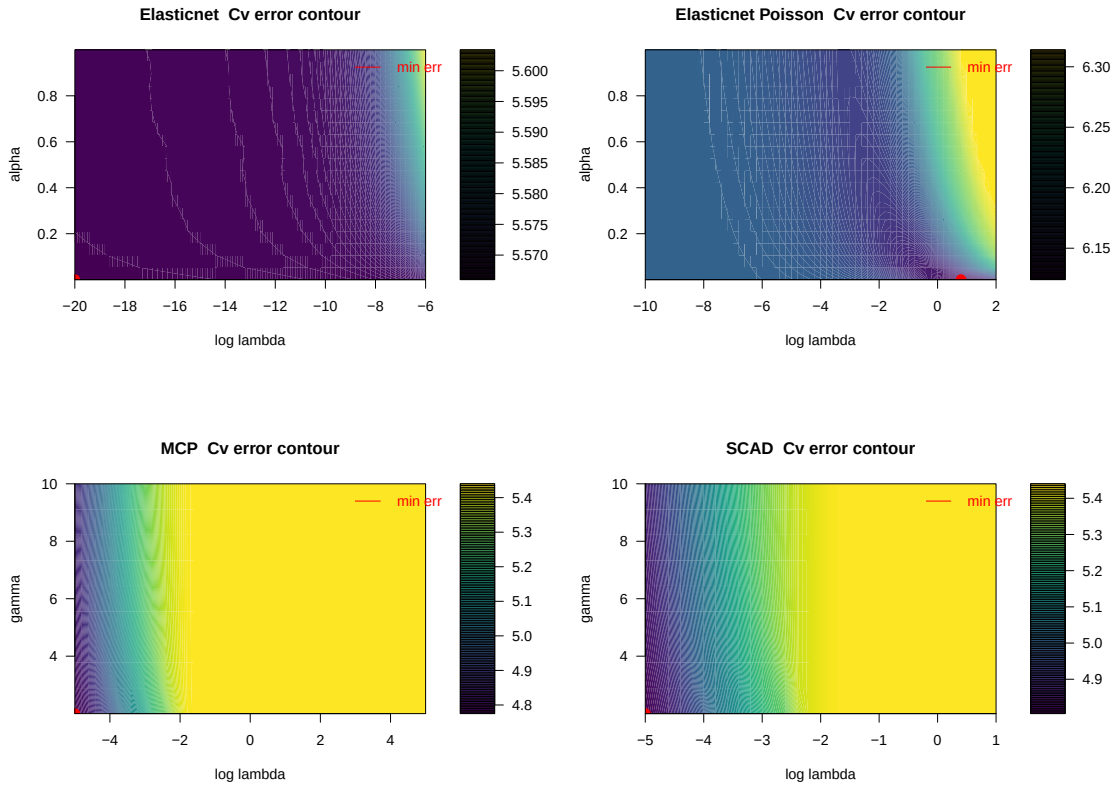
I modelli considerati sono modelli normali con penalizzazioni LASSO, Elasticnet, SCAD ed MCP e modelli Poisson con penalizzazioni LASSO ed Elasticnet. Per per tutti i modelli si seleziona il parametro (eventualmente vettoriale) di regolazione che minimizza l'errore di convalida. Per i metodi per cui 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.

Sia per il modello normale che per quello Poisson 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).



Elasticnet individua in entrambi i casi un α prossimo a zero (ridge), nel modello normale presenta una soluzione molto vicina alle stime non penalizzate (λ è prossimo a zero), mentre nel modello Poisson il λ selezionato è

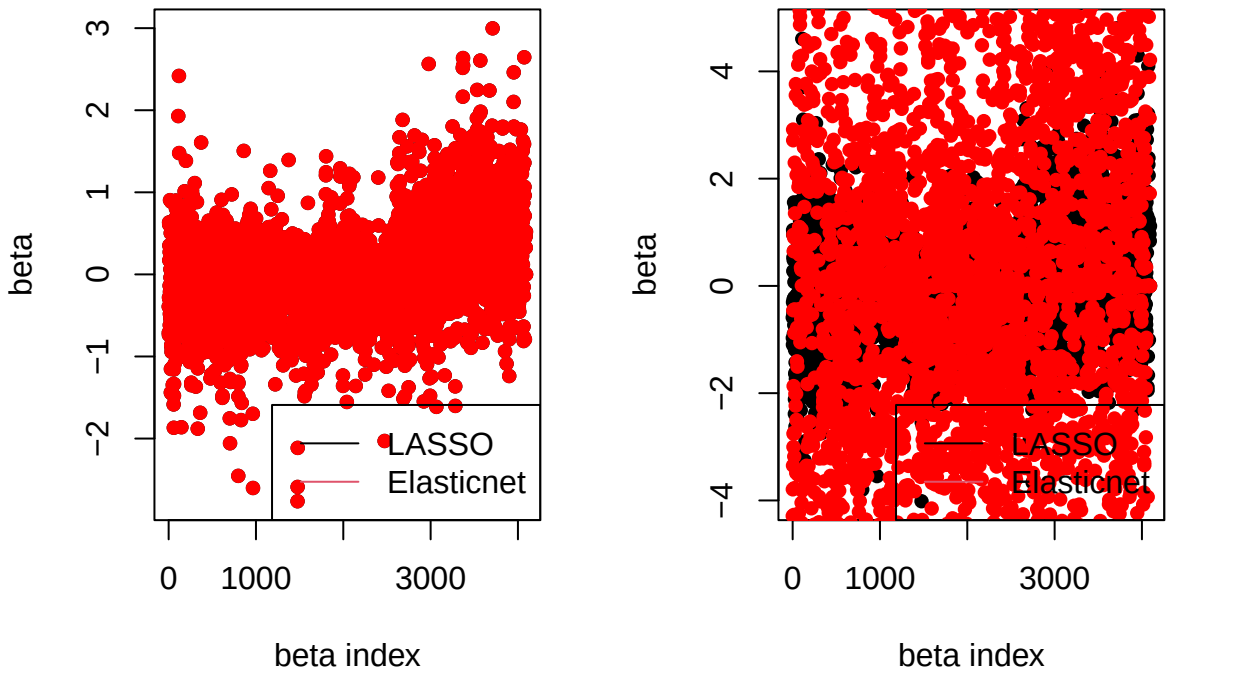
```
## [1] 2.225541
```

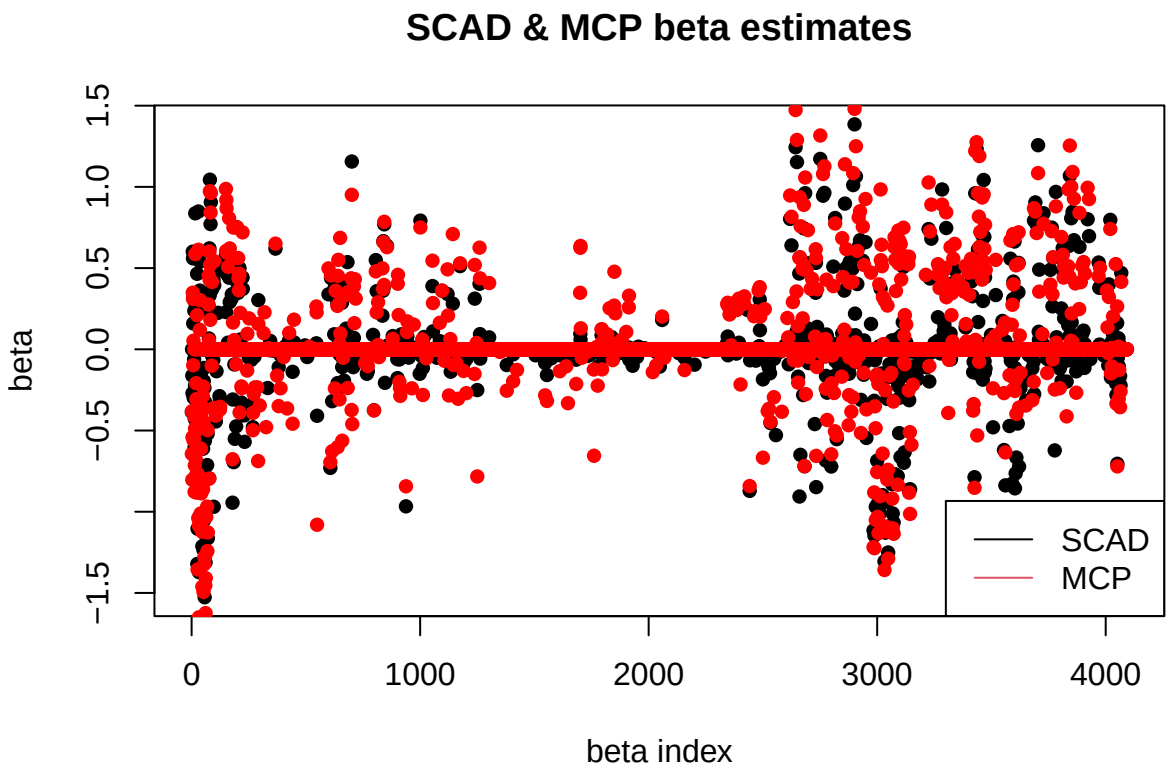


Modelli migliori

Si riportano gli errori di previsione sui dati del 2011 dei vari modelli migliori stimati sui dati completi 2010.

```
##                model test_error
## 1              lasso  12.390364
## 2         elasticnet  12.390363
## 3               scad  14.887738
## 4                mcp  14.947950
## 5      poisson_lasso   8.197242
## 6 poisson_elasticnet  12.174941
```


LASSO & Elasticnet beta estimat
isson LASSO & Elasticnet beta esti
beta
3
2
1
0
−1
−2
LASSO
Elasticnet
0
1000
3000
beta index
beta
4
2
0
−2
−4
LASSO
Elasticnet
0
1000
3000
beta index


SCAD & MCP beta estimates
beta
1.5
1.0
0.5
0.0
−0.5
−1.5
0
1000
2000
3000
4000
beta index
SCAD
MCP