Arrests 2010 NTA analysis

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 477488 25.6    1034058 55.3   660385 35.3
## Vcells 890729  6.8    8388608 64.0  1769844 13.6
```

Preprocessing

Remove date and NTA variables, convert to factor location, month and KY_CD.

```
## 'data.frame':    419383 obs. of  16 variables:
##  $ KY_CD     : Factor w/ 70 levels "101","102","103",..: 70 4 70 15 70 70 70 4 70 4 ...
##  $ LAW_CAT_CD: Factor w/ 5 levels "","F","I","M",..: 1 2 4 2 4 2 4 2 4 2 ...
##  $ AGE_GROUP : Factor w/ 5 levels "<18","18-24",..: 3 2 3 1 4 4 3 2 2 3 ...
##  $ PERP_SEX  : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ PERP_RACE : Factor w/ 7 levels "AMERICAN INDIAN/ALASKAN NATIVE",..: 6 3 3 3 3 7 6 7 7 3 ...
##  $ NTA2020   : Factor w/ 251 levels "BK0101","BK0102",..: 91 28 126 6 68 231 46 140 155 217 ...
##  $ MONTH     : Factor w/ 12 levels "1","2","3","4",..: 1 11 3 12 12 12 11 11 10 9 ...
##  $ Pop1      : int  43885 76961 21140 24605 39214 18774 67940 48520 39120 52734 ...
##  $ MaleP     : num  44.4 44.8 49.3 48.2 46.8 ...
##  $ MdAge     : num  46.1 33.2 39.1 34.8 29.3 35.3 34.1 34.1 34.7 37.8 ...
##  $ Hsp1P     : num  16.96 12.02 8.18 17.81 69.09 ...
##  $ WNHP      : num  68.01 9.17 65.12 47.84 1.46 ...
##  $ BNHP      : num  7.54 74.3 4.52 20.72 27.51 ...
##  $ ANHP      : num  5.31 1.86 19.82 10.15 0.8 ...
##  $ OthNHP    : num  2.19 2.65 2.35 3.48 1.14 3.66 2.54 2.46 3.57 9.53 ...
##  $ MIncome   : num  77817 34115 98132 85175 21477 ...
```

Check for NA

```
##      KY_CD LAW_CAT_CD  AGE_GROUP   PERP_SEX  PERP_RACE    NTA2020      MONTH
##          0          0          0          0          0          0          0
##       Pop1      MaleP      MdAge      Hsp1P       WNHP       BNHP       ANHP
##          0          0          0          0          0          0          0
##     OthNHP    MIncome
##          0          0
```

A possibility is to get rid of all NAs rows, the portion of deleted rows would be relatively small (of course we're introducing some bias here).

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

Other possibilities would be to impute values for numerical variables (using median, mean or more sofisticated methods). For simplicity we just delete missing values rows.

Description

Ideally 2010 data are our training set and 2011 data are the test set. The goal of the analysis is to identify is some covariates are correlated with the arrests rate: more specifically if the response is well explained by some non spatial covariates alone, some spatial alone or interaction between the two.

A reasonable response variable would be the count of arrests divided by the local (space zone) population, also grouping by any other covariates value.

To get an idea of the dataset used on which models are tested a

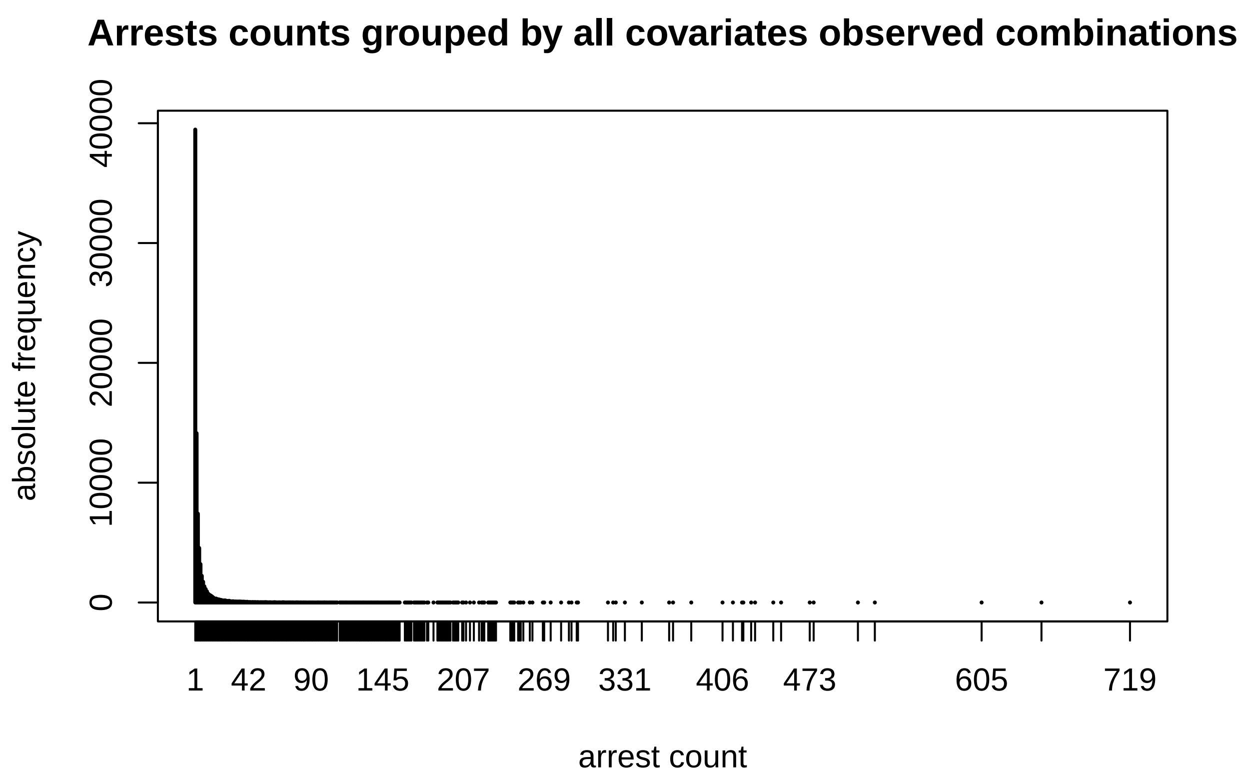
```
## ‘summarise()‘ has grouped output by ’KY_CD’, ’LAW_CAT_CD’, ’AGE_GROUP’,
## ’PERP_SEX’, ’PERP_RACE’, ’NTA2020’, ’Pop1’, ’MaleP’, ’MdAge’, ’Hsp1P’, ’WNHP’,
## ’BNHP’, ’ANHP’, ’OthNHP’. You can override using the ‘.groups‘ argument.
```

```
## [1] 83915    17
```

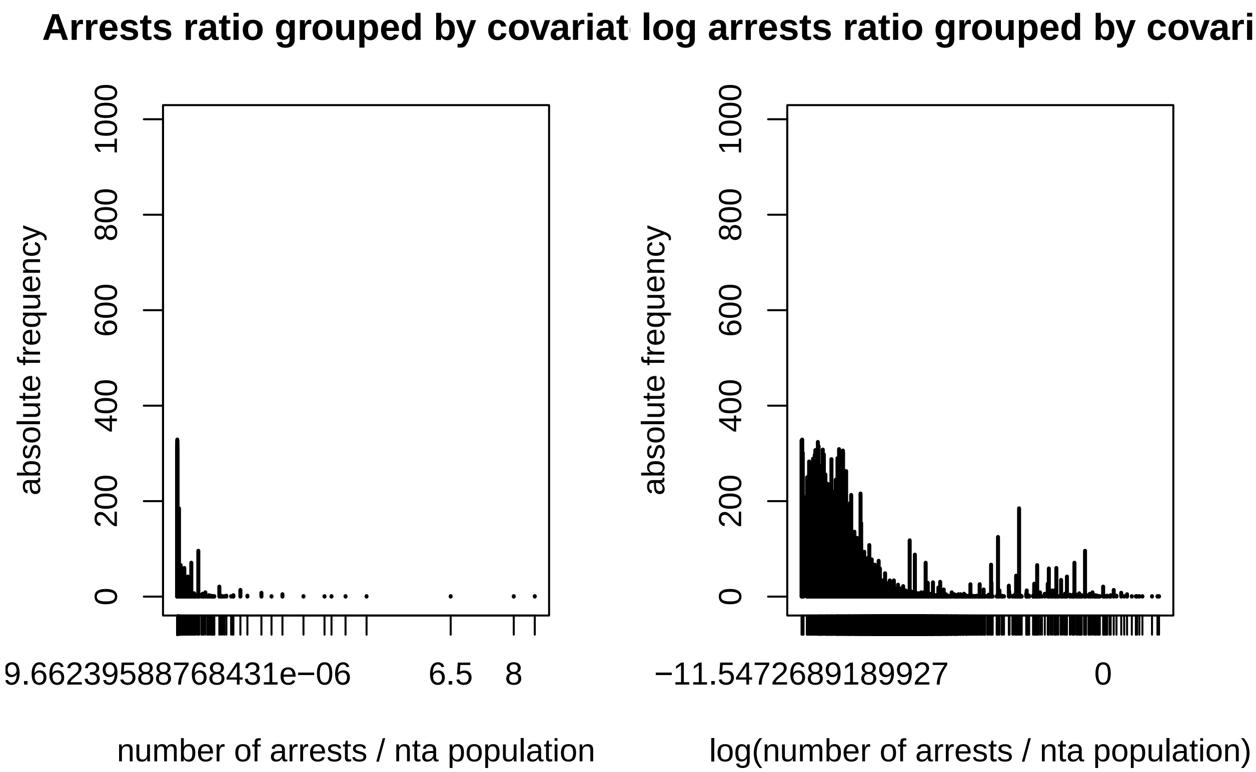
```
##  [1] "KY_CD"      "LAW_CAT_CD" "AGE_GROUP"  "PERP_SEX"   "PERP_RACE"
##  [6] "NTA2020"    "Pop1"       "MaleP"      "MdAge"      "Hsp1P"
## [11] "WNHP"       "BNHP"       "ANHP"       "OthNHP"     "MIncome"
## [16] "count"      "y"
```

Still a huge number of observations compared to the number of variables, but what if we add interactions?

Let’s look at the distribution of the counts.



We can see an inflation of ones. The ratios present a similar frequency table. Taking the logarithm of the ratio the distribution is still (a bit less) skewed.



Let's count the hypothetical number of interaction terms if ones considers only interactions between spatial zones and selected arrests covariates along with the obervations / number of parameter ratio (underestimate since there are other variables):

```
##      KY_CD LAW_CAT_CD  AGE_GROUP   PERP_SEX  PERP_RACE    NTA2020       Pop1
##         70          5          5          2          7        251        228
##      MaleP      MdAge      Hsp1P       WNHP       BNHP       ANHP     OthNHP
##        206        143        219        218        213        204        161
##    MIncome      count          y
##        214        248       5829
```

Not including KY_CD:

```
## NTA2020
##    4769
```

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

Including KY_CD

```
## NTA2020
##   22339

##  NTA2020
## 3.756435
```

We decide to not employ the MONTH time variable as a covariate but use it for a model selection method. This is because in order to get a count measure one aggregation criterion is needed, here aggregation by month is chosen.

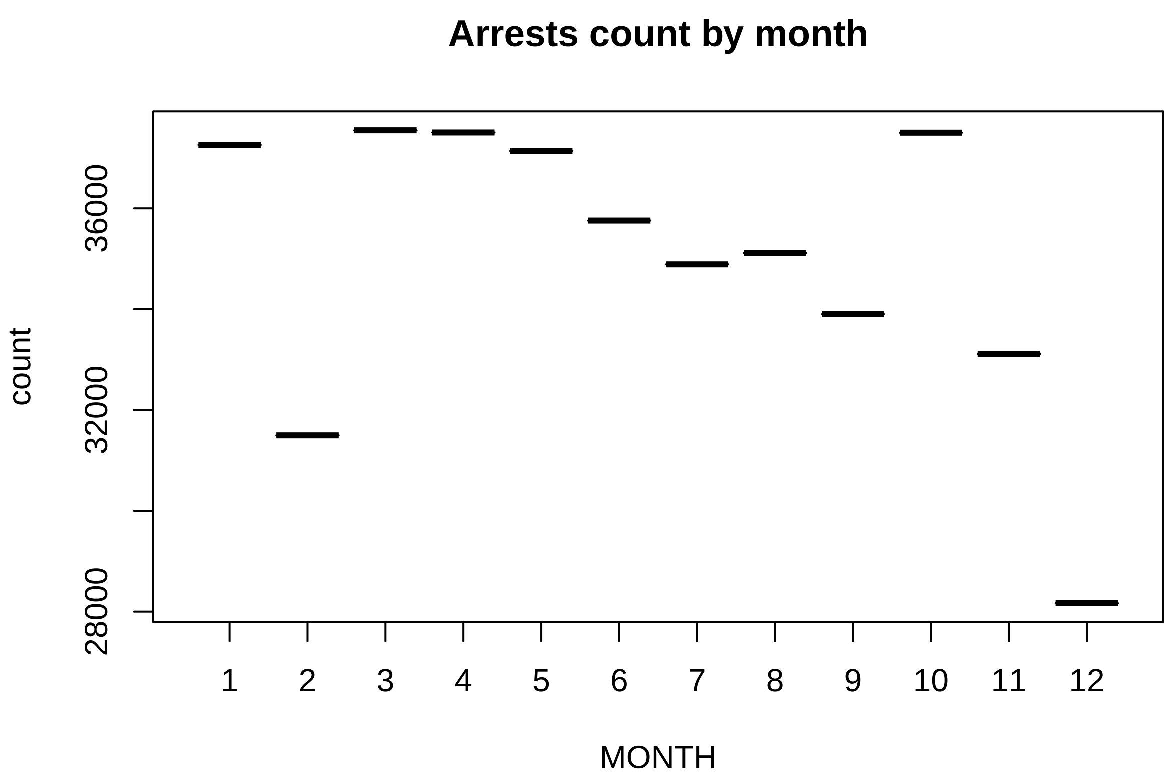
Variables description

Original dataset selected variables:

Census stratification variables:

Explorative analysis

Arrests count vs month



Arrests counts vs NTA

Arrest counts vs other covariates

Models

Model selection method

Given the previously described constraints, in order to be able to apply a cross validation (CV) selection method we choose to ignore the time (MONTH) factor using MONTH as index to create the CV folds as described below. Choose k: the number of validation sets (example k = 4) each validation set is made by grouped observations of 12 / k (3) months and the months left (9) are used to fit the model. To try to compensate and average for seasonal fluctuations the validation months are chosen as spaced as possible, for example, in the case k = 4 the first validation set is (january, may, september), the second set is (february, june, october), the third is (march, july, november) and the forth is (april, august, december); in order to make each response comparable having used a different number of months a new response is defined as the arrests ratio divided by the number of months used in the grouping. Note: this is also a way to reduce the computational burden compared to using many more months combinations.

Define Month indexes

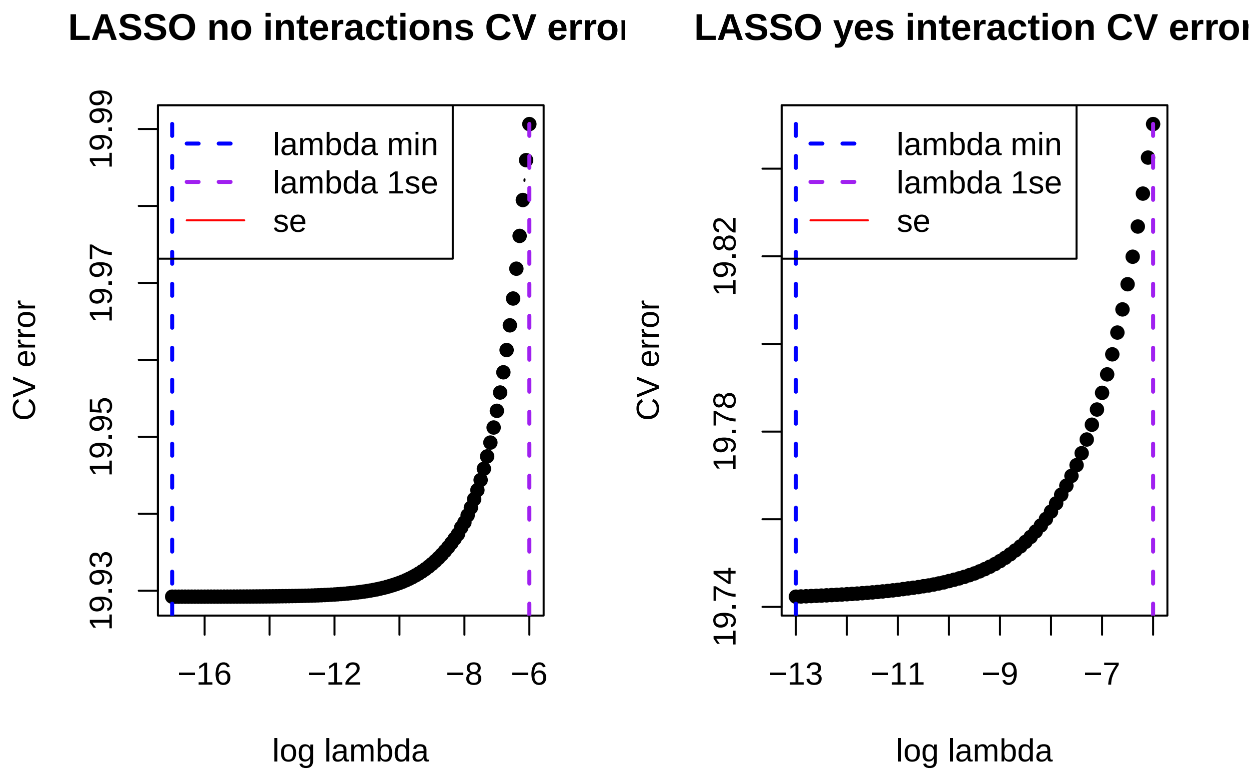
In order to simplify computations we remove the KY_CD variable (hoping LAW_CAT_CD will be sufficient to describe the crime type) when using a linear model (assuming gaussian errors) we consider the response as: $y = \log(\text{count}/\text{population})$ where each count is the events count obtained by grouping by all other covariates and each population is specific to each NTA.

```
## `summarise()` has grouped output by 'LAW_CAT_CD', 'AGE_GROUP', 'PERP_SEX',
## 'PERP_RACE', 'NTA2020', 'Pop1', 'MaleP', 'MdAge', 'Hsp1P', 'WNHP', 'BNHP',
## 'ANHP', 'OthNHP'. You can override using the `.groups` argument.
```

Note on quantitative covariates

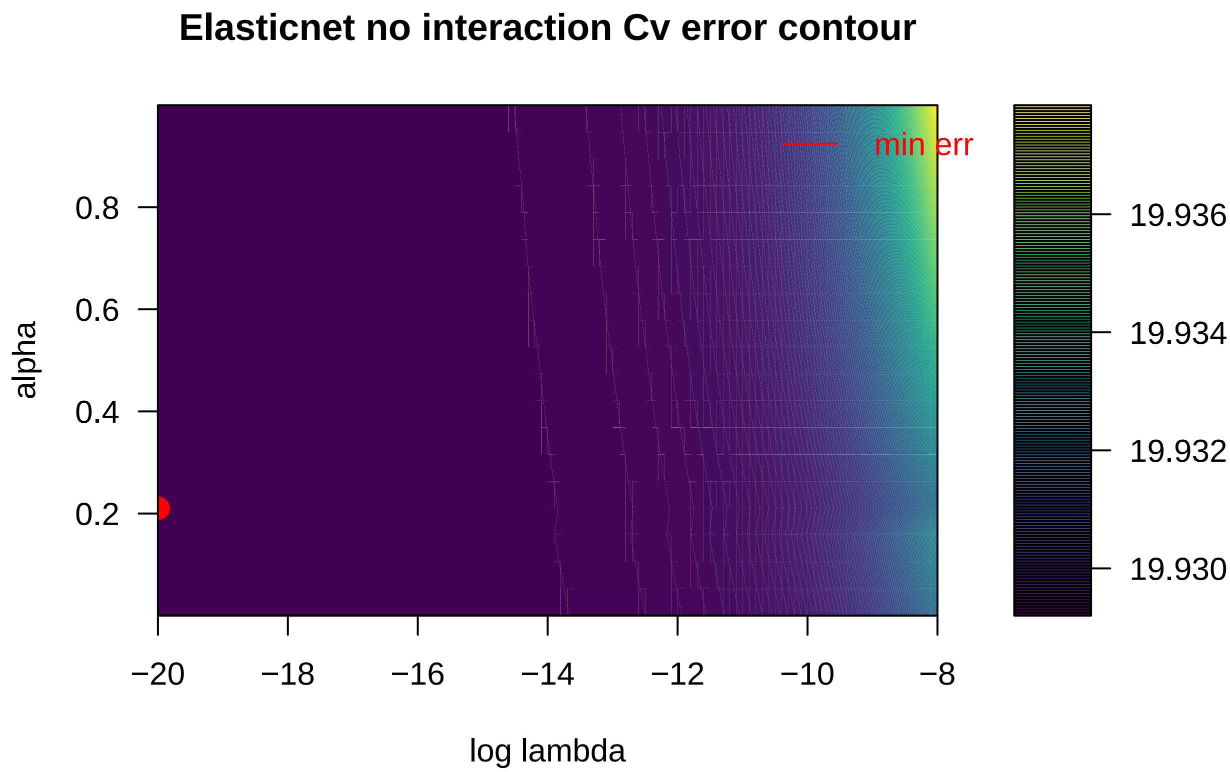
The simplest assumption is to assume a linear (monotone) trend of the response as a function of quantitative covariates.

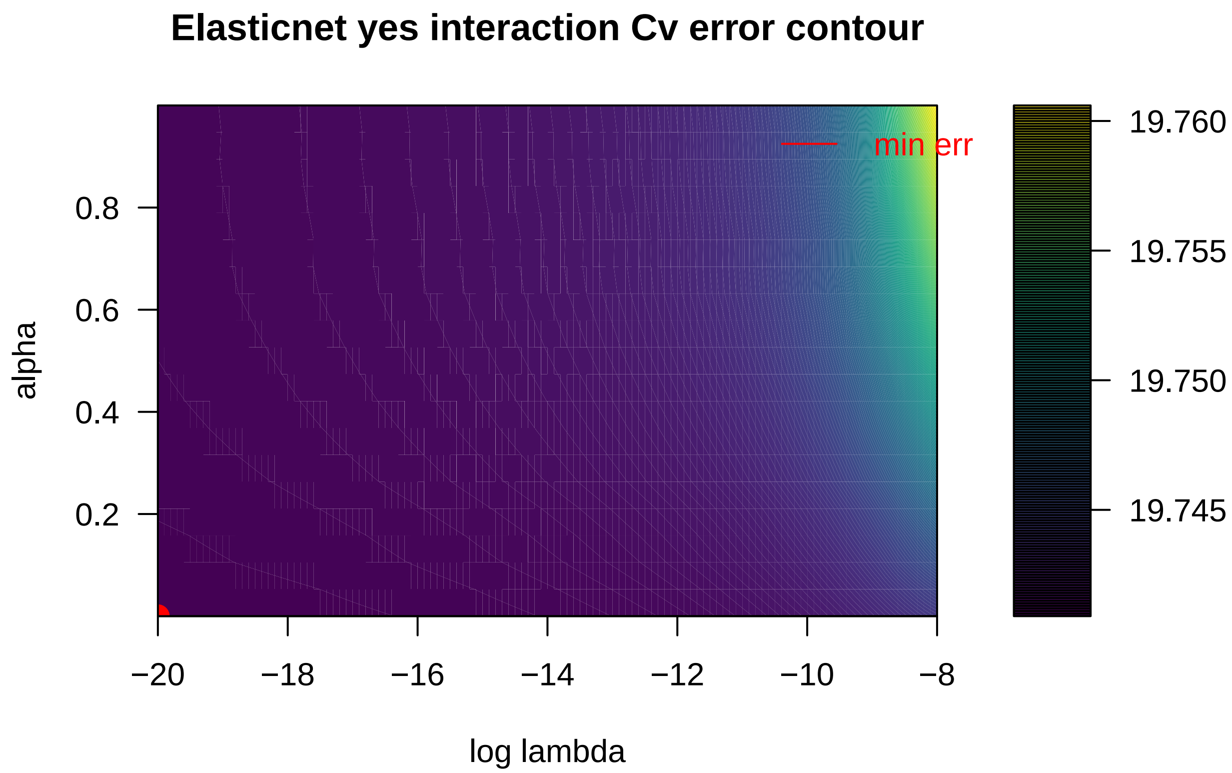
LASSO



LASSO no interactions CV error
CV error
19.93
19.95
19.97
19.99
log lambda
−16
−12
−8
−6
lambda min
lambda 1se
se
LASSO yes interaction CV error
CV error
19.74
19.78
19.82
log lambda
−13
−11
−9
−7
lambda min
lambda 1se
se

Elasticnet

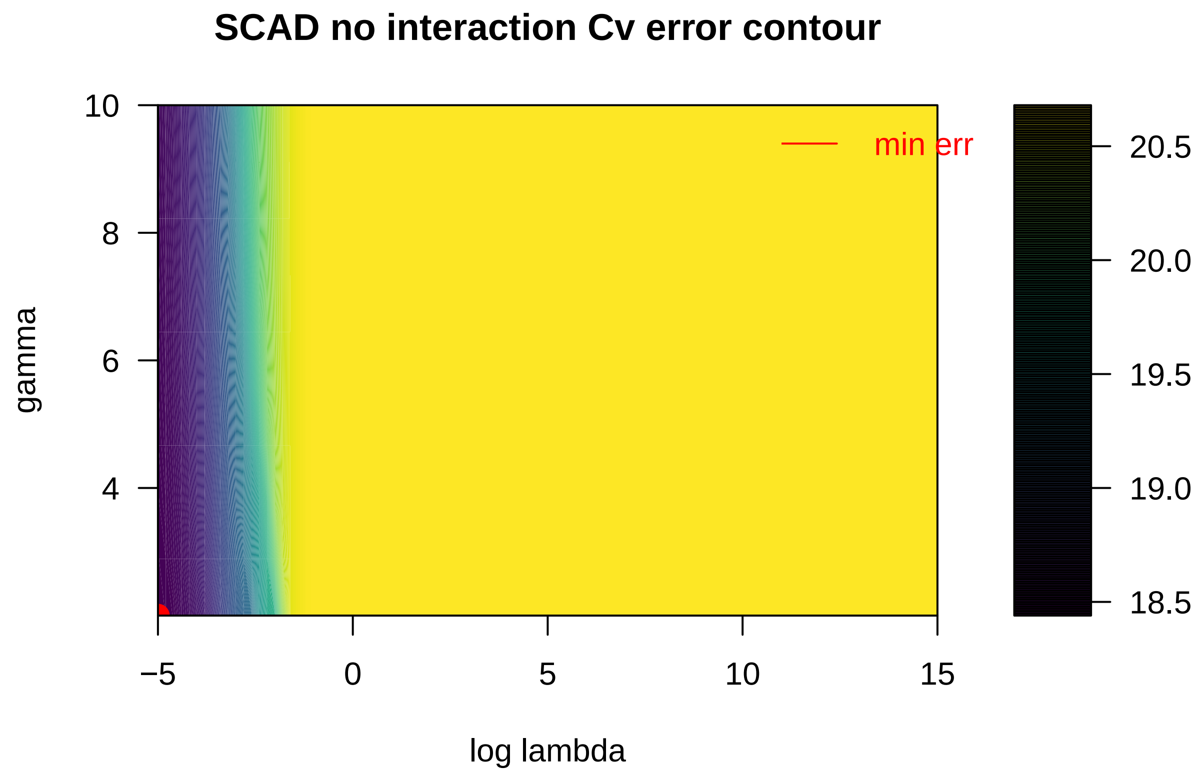

Elasticnet no interaction Cv error contour
min err
alpha
0.2
0.4
0.6
0.8
−20
−18
−16
−14
−12
−10
−8
log lambda
19.930
19.932
19.934
19.936



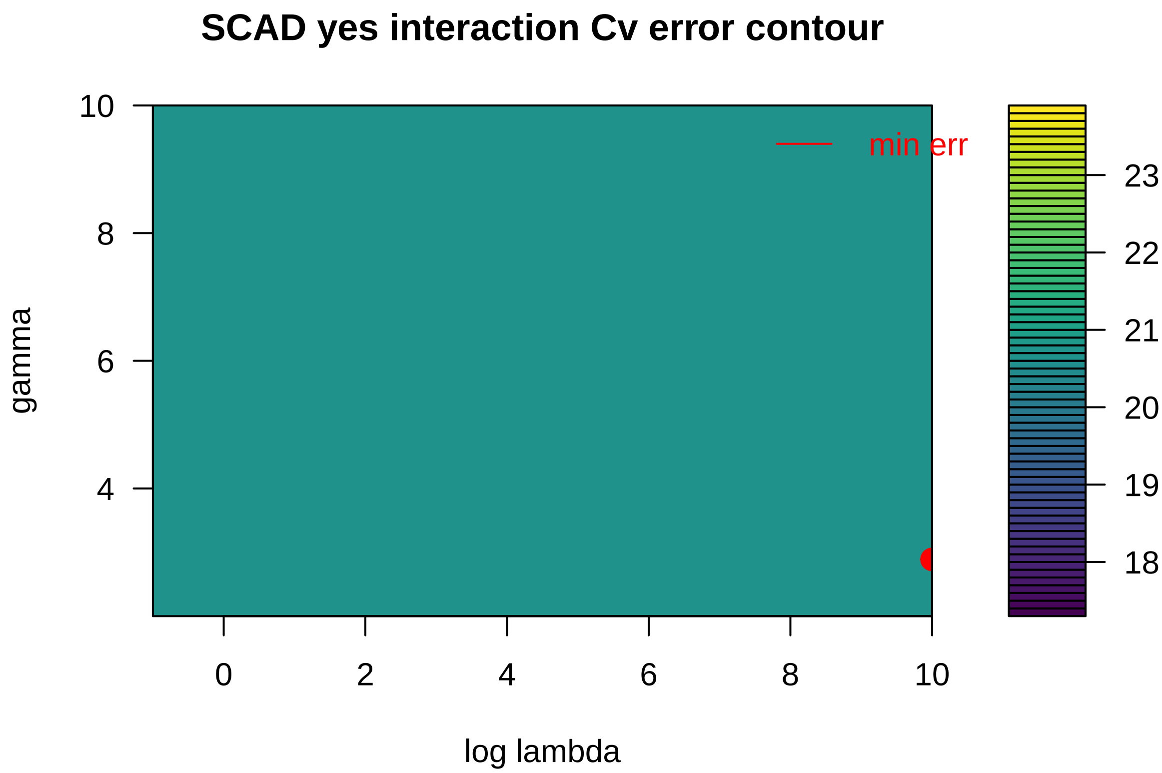
Grouped LASSO

Pur avendo provato a considerare solo il modello senza interazioni, a causa dell'eccessivo tempo computazionale richiesto non è qui riportato.

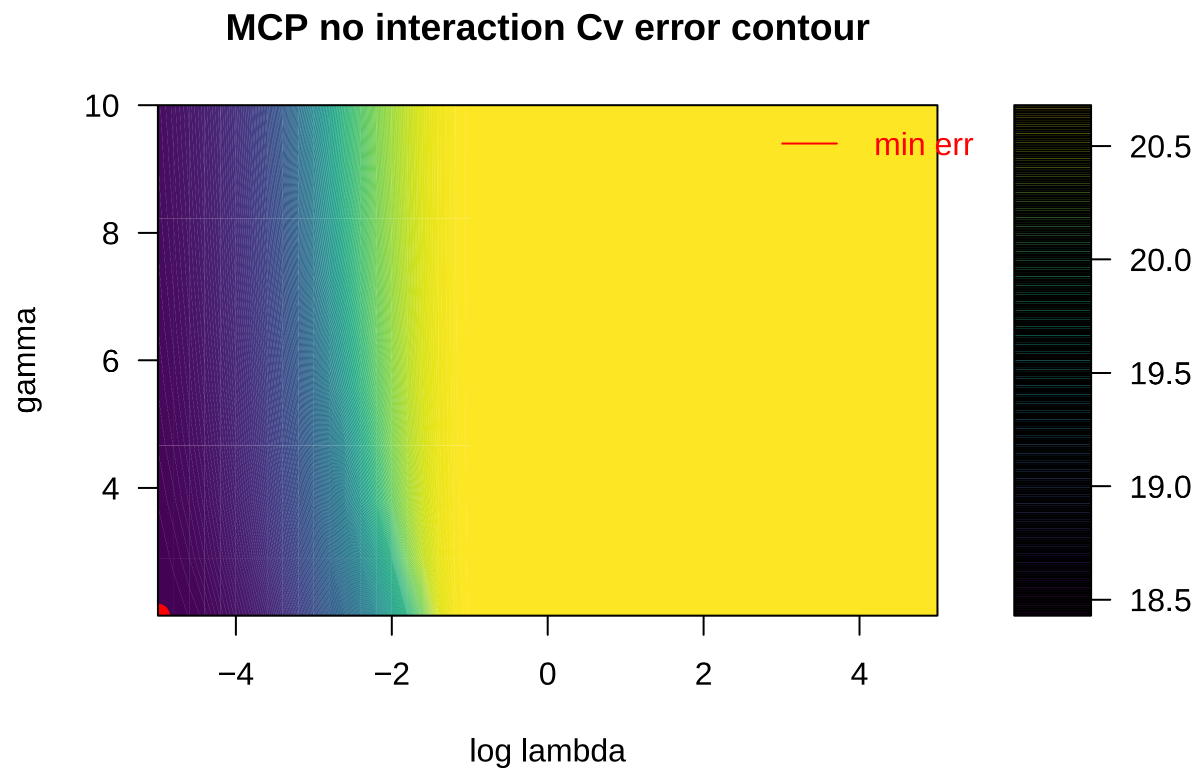
Scad



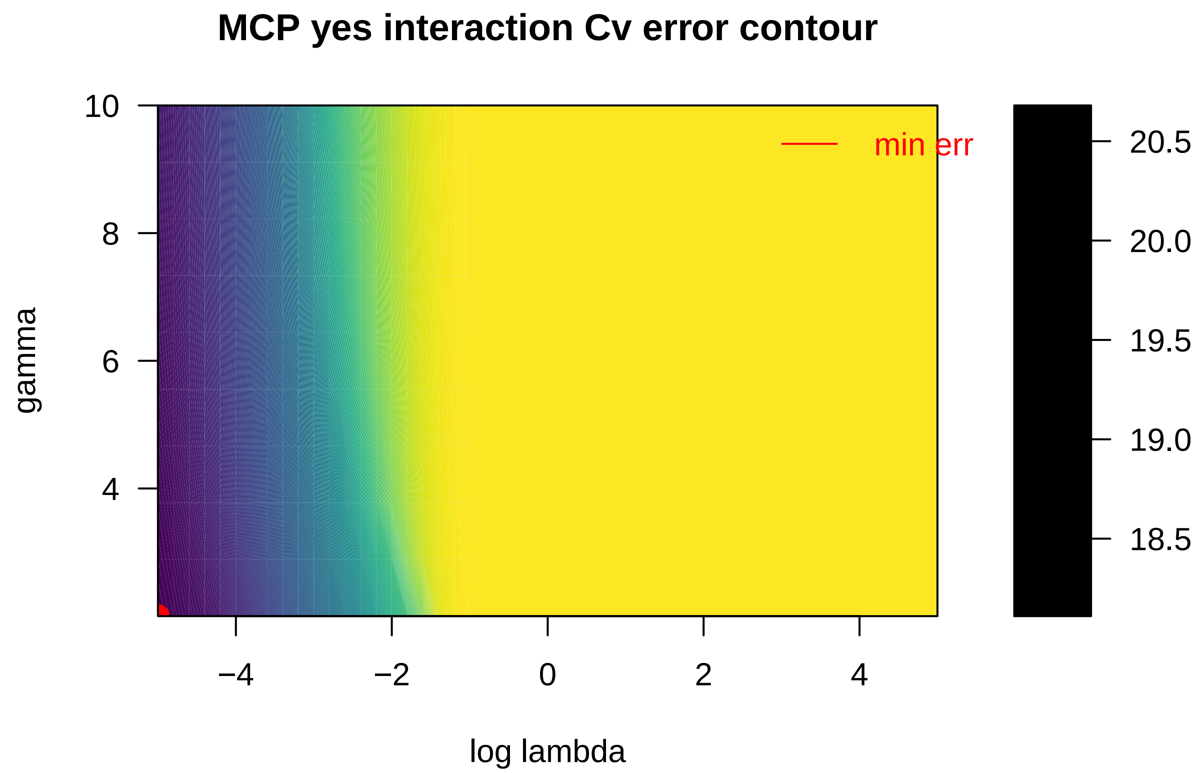
SCAD no interaction Cv error contour
10
8
6
4
gamma
−5
0
5
10
15
log lambda
min err
20.5
20.0
19.5
19.0
18.5


SCAD yes interaction Cv error contour
min err
gamma
log lambda
10
8
6
4
0
2
4
6
8
10
23
22
21
20
19
18

MCP



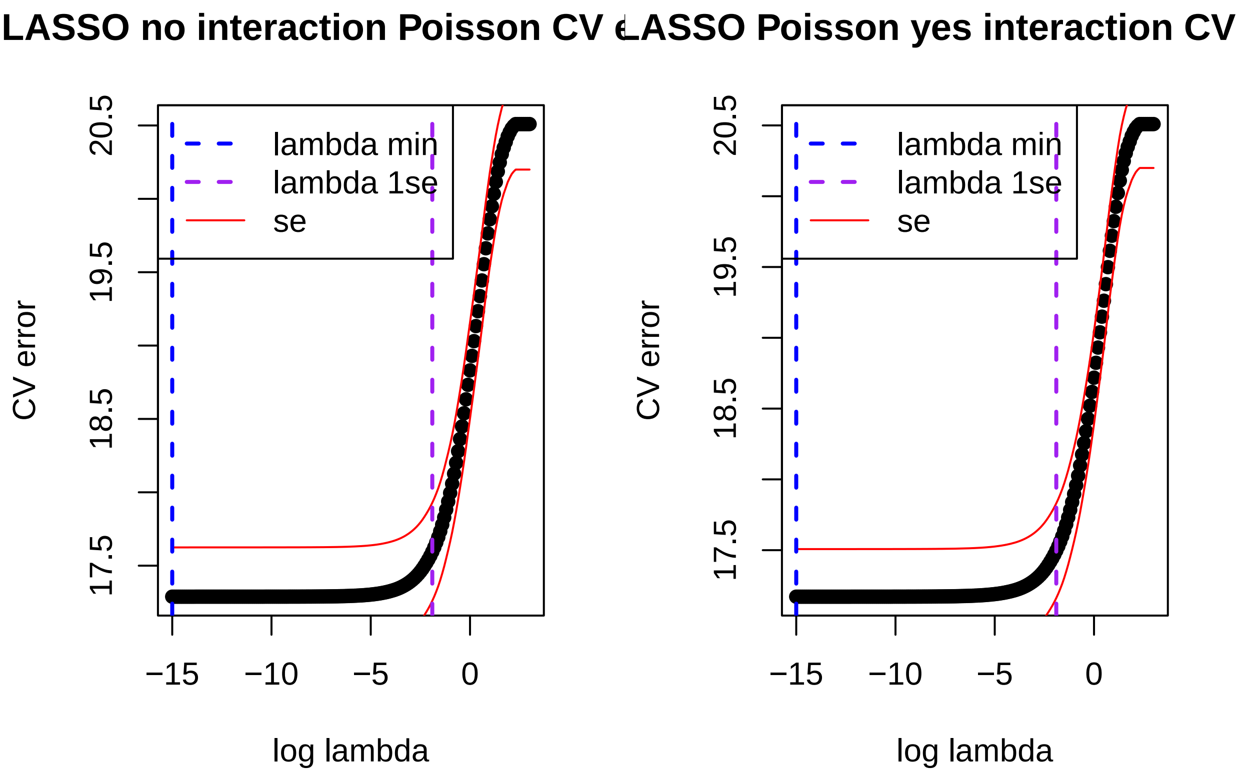
MCP no interaction Cv error contour
min err
gamma
10
8
6
4
−4
−2
0
2
4
log lambda
20.5
20.0
19.5
19.0
18.5

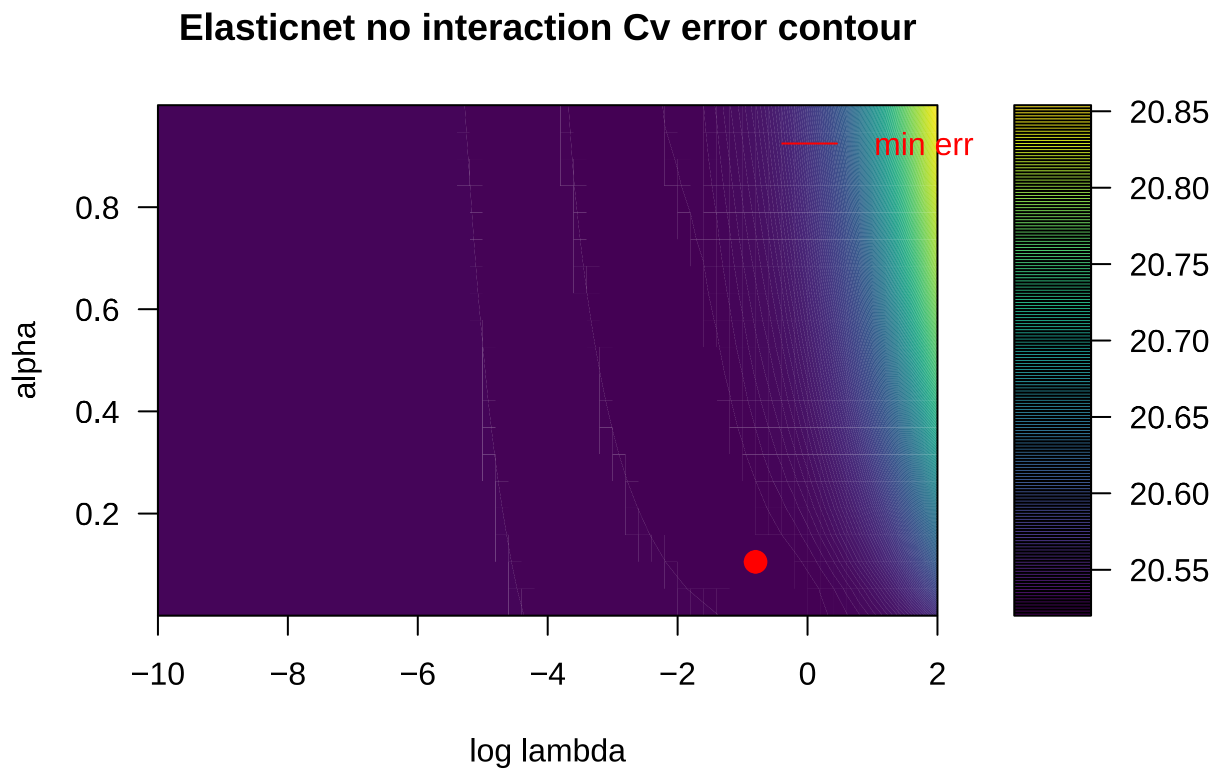


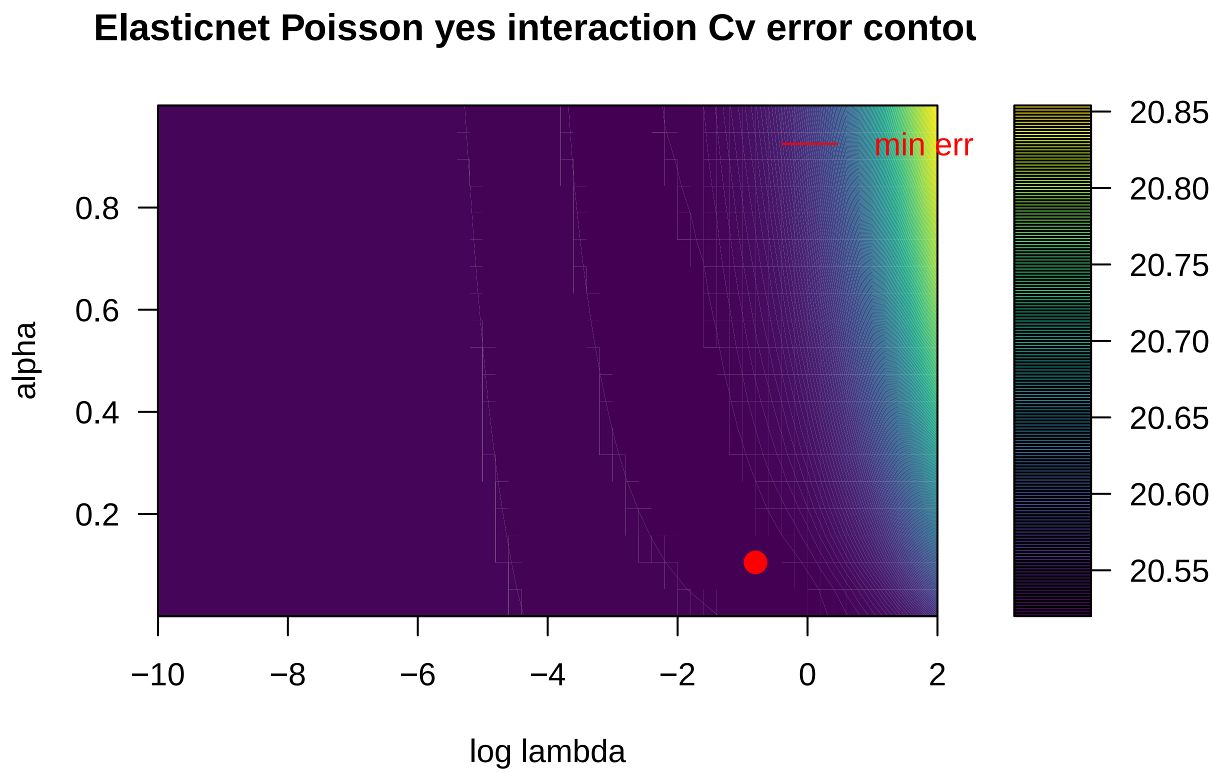
Discrete response models

In reality the counts are discrete so it seems reasonable to also try discrete response models such as Poisson, Negative Binomial and zero inflated Poisson. For all such cases, using the counts as response an offset has to be imposed: in analogy from what has been done assuming the continuous response the offset will be the product of the NTA Population by the number of months considered (in log scale using the canonincal log link for a Poisson GLM)

Poisson LASSO




Elasticnet no interaction Cv error contour
min err
0.8
0.6
0.4
0.2
alpha
−10
−8
−6
−4
−2
0
2
log lambda
20.85
20.80
20.75
20.70
20.65
20.60
20.55



Neg bin Lasso

Using a hierarchical specification we assume $Y_i \sim P(\mu_i \lambda_i)$ and $\lambda_i \sim Ga(\tau, \tau)$ so marginally the Y_i are negative binomials with variance $\mu_i(1 + \tau\mu_i)$. In the R parameterization adopted $\theta = 1/\tau$ which becomes another tuning parameter.

Modelli migliori