Arrests 2010 NTA analysis

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 498168 26.7    1080107 57.7   686460 36.7
## Vcells 930808  7.2    8388608 64.0  1877429 14.4
```

Preprocessing

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

```
## 'data.frame':    419383 obs. of  15 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 ...
##  $ GeoID     : Factor w/ 251 levels "BK0101","BK0102",..: 91 28 126 6 68 231 46 140 155 217 ...
##  $ Pop1      : num  43885 76961 21140 24605 39214 ...
##  $ Male.P    : num  44.4 44.7 49.3 48.2 46.8 48.9 45.1 48.8 49.5 47.2 ...
##  $ MdAge     : num  46.1 33.2 39.1 34.8 29.3 35.3 34.1 34.1 34.7 37.8 ...
##  $ Hsp1P     : num  17 12 8.2 17.8 69.1 30.6 20.1 52.2 28.2 18.6 ...
##  $ WNHP      : num  68 9.2 65.1 47.8 1.5 25.3 8 10.9 47.2 6.4 ...
##  $ BNHP      : num  7.5 74.3 4.5 20.7 27.5 35.3 66.1 32.2 4.6 48.3 ...
##  $ ANHP      : num  5.3 1.9 19.8 10.1 0.8 5.2 3.2 2.2 16.4 17.1 ...
```

Check for NA

```
##      KY_CD LAW_CAT_CD  AGE_GROUP   PERP_SEX  PERP_RACE    NTA2020      MONTH
##          0          0          0          0          0          0          0
##      GeoID       Pop1     Male.P      MdAge      Hsp1P       WNHP       BNHP
##          0          0       2491       2489       2702       3111       3281
##       ANHP
##       4315
```

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] 0.9890935
```

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

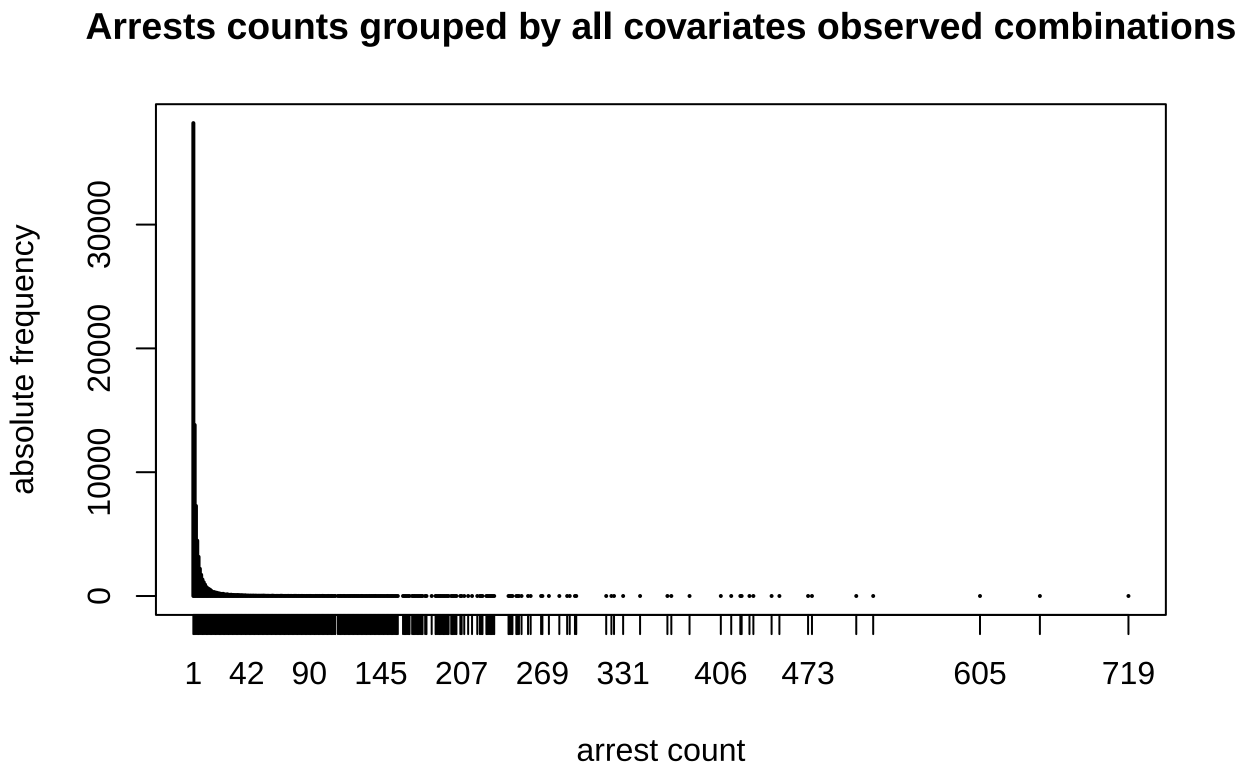
```
## ‘summarise()‘ has grouped output by ’KY_CD’, ’LAW_CAT_CD’, ’AGE_GROUP’,
## ’PERP_SEX’, ’PERP_RACE’, ’NTA2020’, ’GeoID’, ’Pop1’, ’Male.P’, ’MdAge’,
## ’Hsp1P’, ’WNHP’, ’BNHP’. You can override using the ‘.groups‘ argument.
```

```
## [1] 81946    16
```

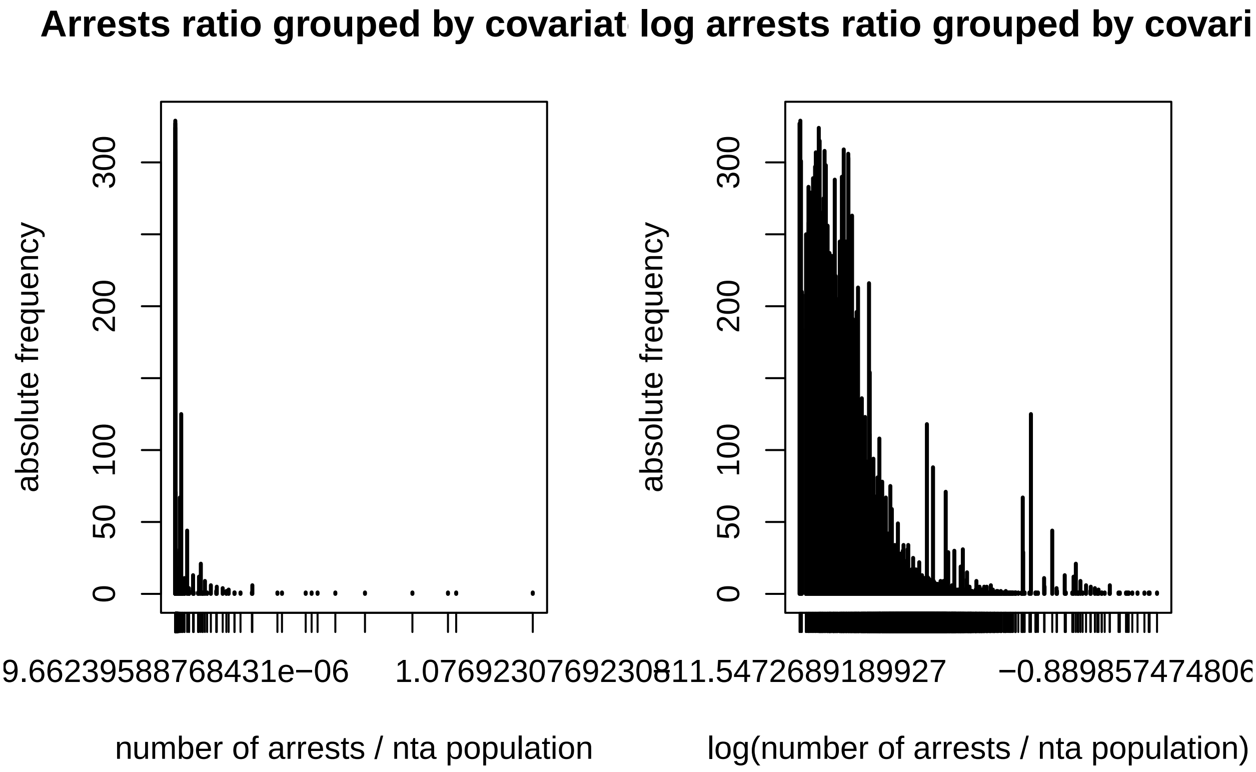
```
##  [1] "KY_CD"      "LAW_CAT_CD" "AGE_GROUP"  "PERP_SEX"   "PERP_RACE"
##  [6] "NTA2020"    "GeoID"      "Pop1"       "Male.P"     "MdAge"
## [11] "Hsp1P"      "WNHP"       "BNHP"       "ANHP"       "count"
## [16] "y"
```

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.


Arrests counts grouped by all covariates observed combinations

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      GeoID 
##         70          5          5          2          7        209        209 
##       Pop1     Male.P      MdAge      Hsp1P       WNHP       BNHP       ANHP 
##        207         83        128        172        170        156        135 
##      count          y 
##        248       5754 
```

Not including KY_CD:

```
## NTA2020 
##    3971 
```

```
##  NTA2020 
## 20.63611 
```

Including KY_CD

```
## NTA2020
##   18601
```

```
##  NTA2020
## 4.405462
```

We decide to not employ the MONTH time variable as a covariate but use it for a model selection method.

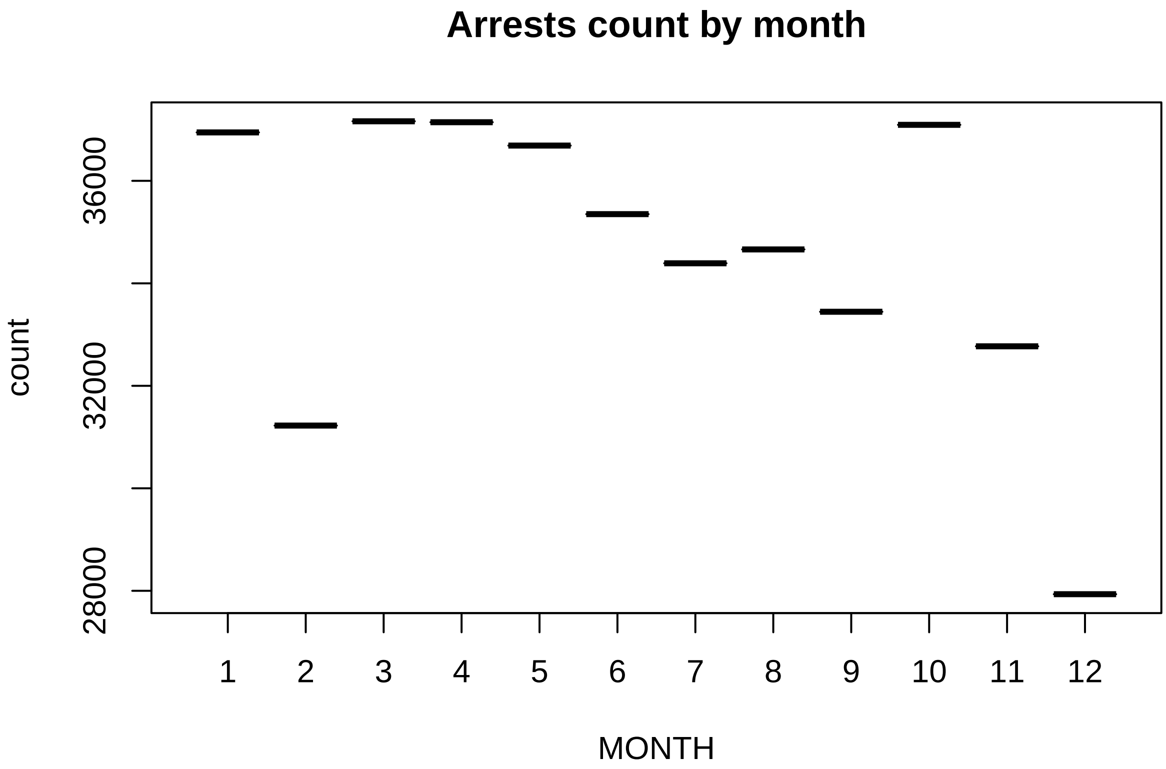
Variables description

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 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.

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: $y = \log(\text{count/population})$ 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’, ’GeoID’, ’Pop1’, ’Male.P’, ’MdAge’, ’Hsp1P’, ’WNHP’,
## ’BNHP’. You can override using the ‘.groups‘ argument.
```

Error functions

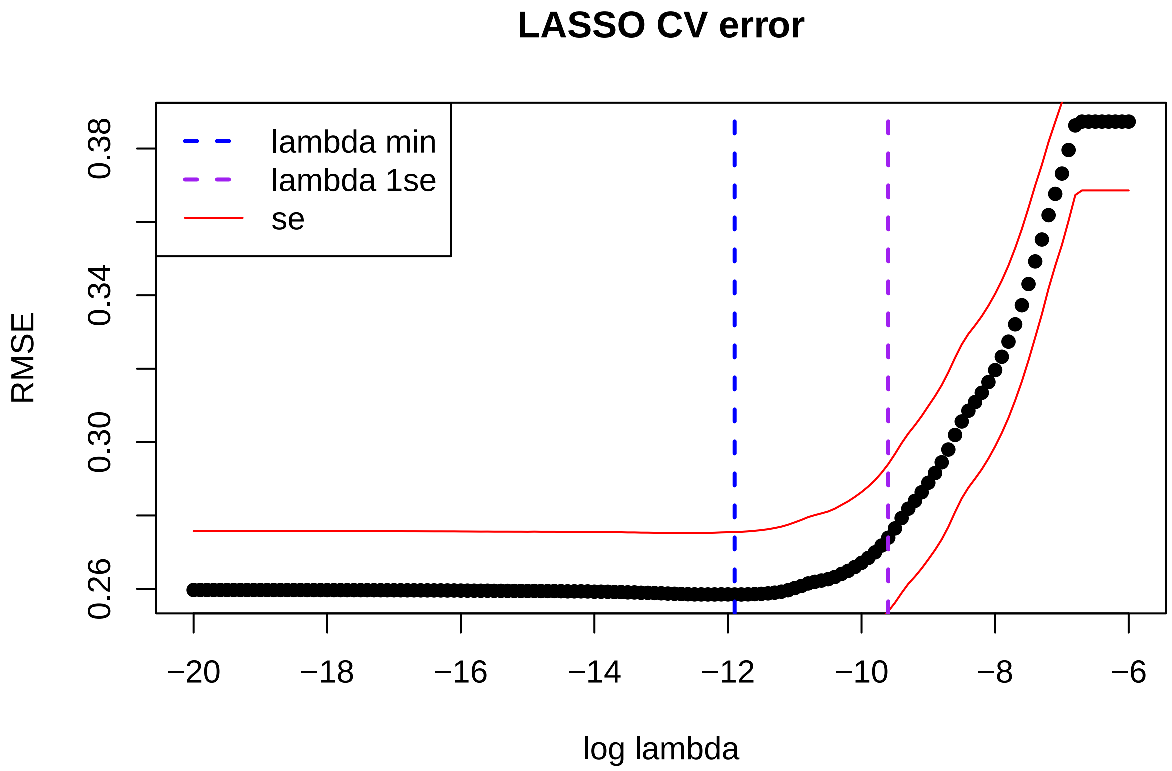
Model matrices: omit KY_CD variables due to computational issues. Make a list of fit and validation sets:

Note on quantitative covariates

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

LASSO

```
## Loaded glmnet 4.1-8
```



Best *lambda* is close to zero, so the best solution seems OLS classical solution.

Check beta: proportion of non zero beta among all betas

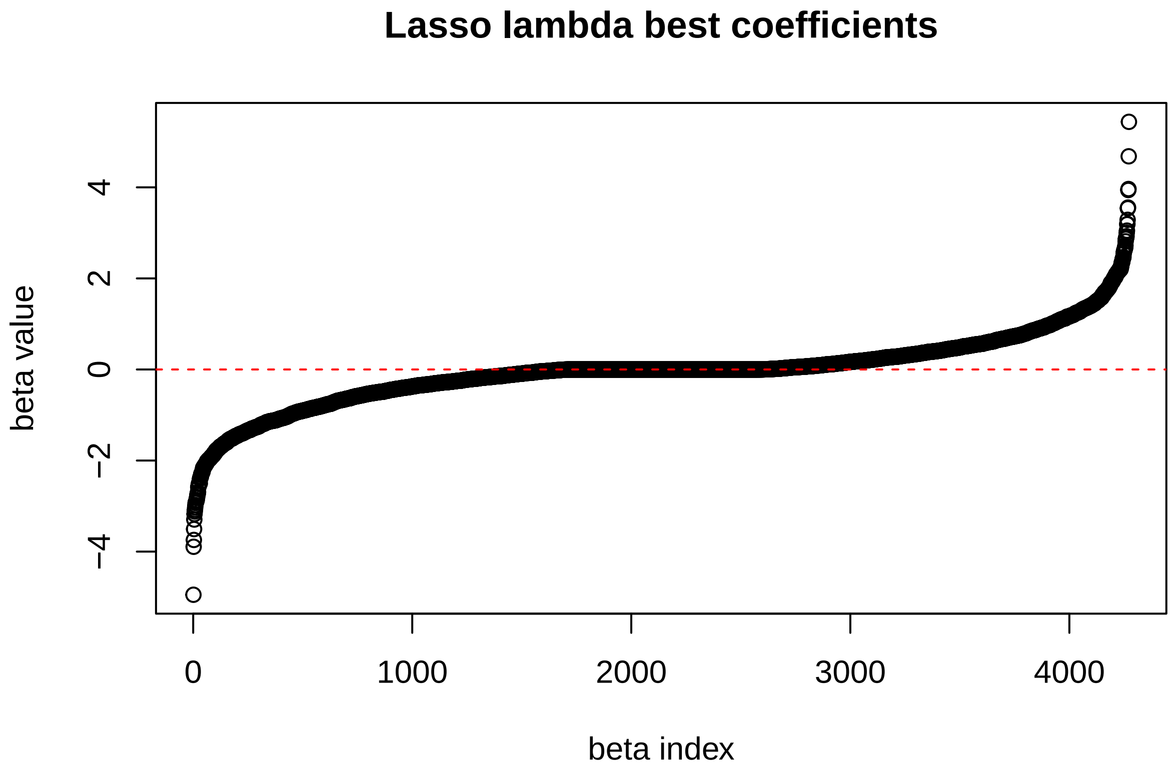
```
## [1] 0.8054775
```

So the solution is not sparse even thought many coefficients are close to zero.

While for the l1se

```
## [1] 0.7762172
```

Still not sparse.



Elasticnet

Selecting a grid of 20 $\alpha \in (0, 1)$.

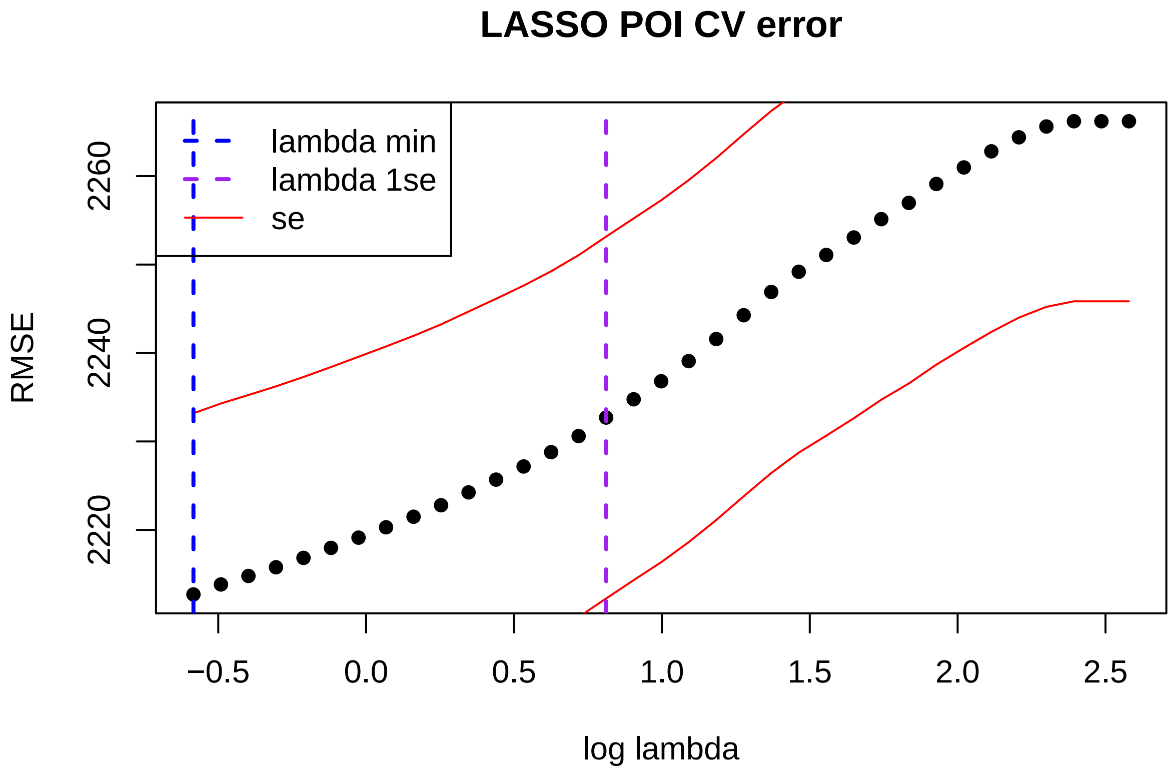
Grouped LASSO

Scad MCP

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 canoninical log link for a Poisson GLM)

Poisson LASSO



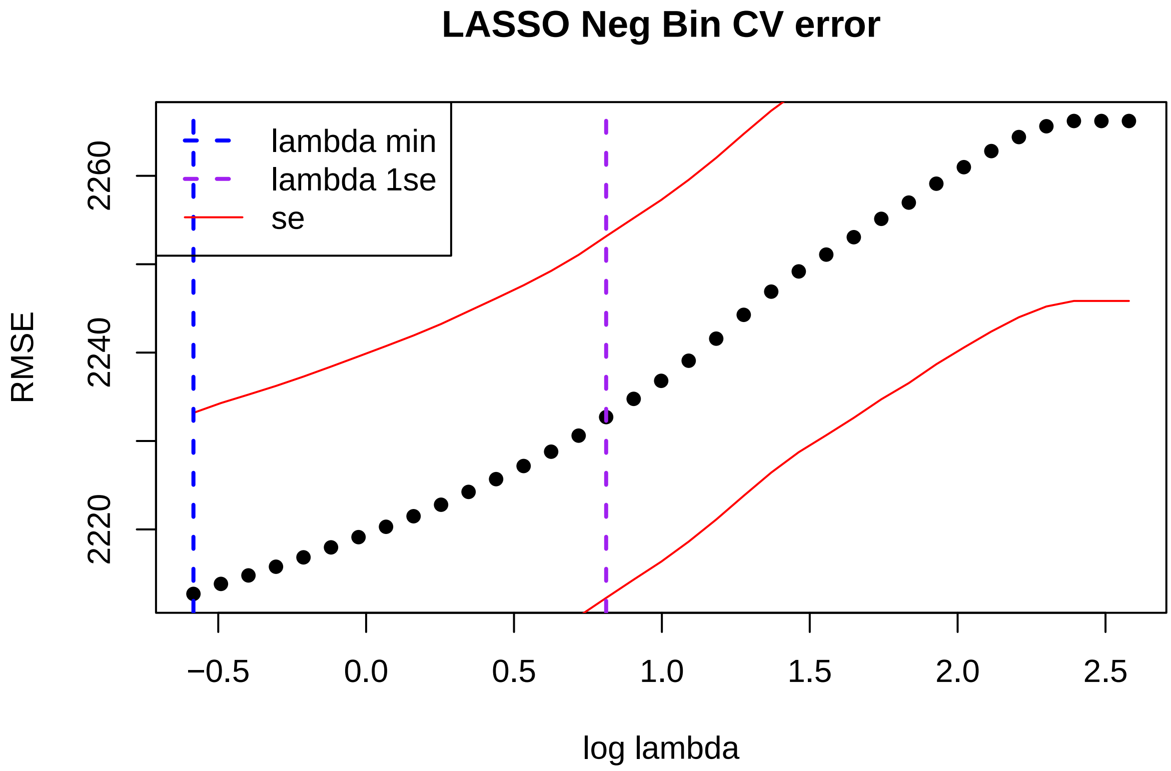
There's a problem here since for smaller λ values the algorithm does not converge.

Poisson Elasticnet

```
## [1] 12452
```

Neg bin Lasso

```
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
## Warning: glmnet.fit: algorithm did not converge
```


LASSO Neg Bin CV error
lambda min
lambda 1se
se
RMSE
log lambda