# ST404 Assignment 2 Modeling Part

Group K

## Data cleaning

Before creating and iterating our final models, we must take the preliminary step of applying the results of the EDA for the USACrime data.

Firstly, we omit the observations that contain the 52 and 21 missing values of medIncome and pctEmploy respectively, as these have been determined to be MCAR (Missing Completely at Random). Futhermore, there are an additional 92 entries (coming from ownHouseMed, rentMed, ownHouseQrange, and rentQrange) which appear to be entered in place of a missing value.

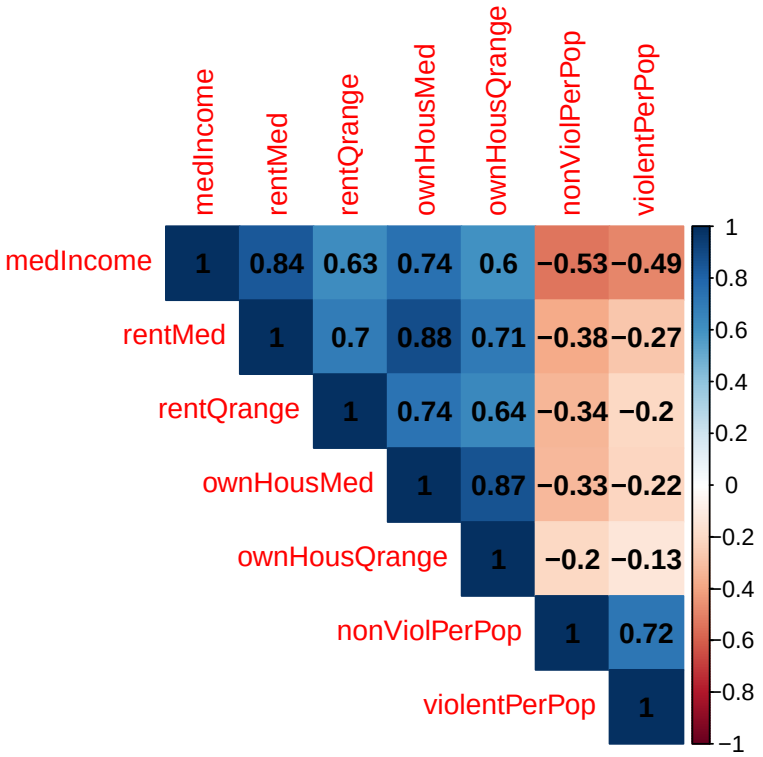
Additionally, the variable State had some empty levels and those are dropped at the beginning to clean up the variable. The variable region has a level "Pacific" which only contains three observations are all from Alaska. To fix this, the "Pacific" level was merged with "West" as this made the most sense geographically. In the end, region has four levels: Midwest, Northeast, West and South.

Finally, from the EDA, it is recommended that a log2 transformation is applied to both of the outcome variables violentPerPop and nonViolentPerPop, as they are positively skewed. This will improve the robustness of the model. To keep our model, these will be the only transformations we use for now.

## Variable choice

The first variables we will leave out of the model are the State and region variables. We suggest dropping the State variable, as it has a very large number of levels and may be difficult to interpret. There are also several states with a small number of observations, so the estimates may not be very reliable.

We can see that all of the variables medIncome, ownHouseMed, ownHousQrange, rentMed and rentQrange, which are related to measuring income/wealth, exhibit strong positive linear correlation with each other (correlation coefficients in excess of 0.6). Therefore, only one income related variable will be needed in the model. The relationship between income related variables is stronger than any individual income related variable to the outcomes. This can be seen in the following correlation matrix:




Therefore, to avoid problems related to multicollinearity, we should only include one of the 5 income related variables in the model. Since medIncome has the strongest correlation with both outcome variables, we would suggest this would be the variable to include.

Indeed, by looking at multicollinearity before and after we make these changes to the variable selection by calculating the VIF (Variance Inflation Factor), no variable has a VIF>5 in the latter case.

*VIF with all variables included*

```
##           pctUrban          medIncome            pctWdiv          pctLowEdu
##           1.595624          16.604910           7.212361           8.213528
##       pctNotHSgrad         pctCollGrad        pctUnemploy          pctEmploy
##          11.929074           4.588640           4.396150           5.087223
##        pctKids2Par pctKidsBornNevrMarr       pctHousOccup   pctHousOwnerOccup
##          10.028235           7.109926           1.656554           4.378956
##   pctVacantBoarded        pctVacant6up         ownHousMed       ownHousQrange
##           1.986586           1.889462          12.641230           6.388387
##           rentMed          rentQrange         popDensity      pctForeignBorn
##           9.727877           2.870098           2.518501           3.791354
```

*VIF after removing selected variables*

```
##           pctUrban          medIncome          pctLowEdu        pctUnemploy
##           1.504212           6.363429           2.815199           3.316024
##          pctEmploy pctKidsBornNevrMarr       pctHousOccup   pctHousOwnerOccup
##           2.837485           2.920242           1.365158           3.249822
##   pctVacantBoarded        pctVacant6up         popDensity      pctForeignBorn
##           1.889971           1.722032           2.323287           2.582596
```

## Modeling

With the problem of collinearity solved, we go on to building our linear regression models. We have chose to use **LASSO regression** for model building for several reasons:

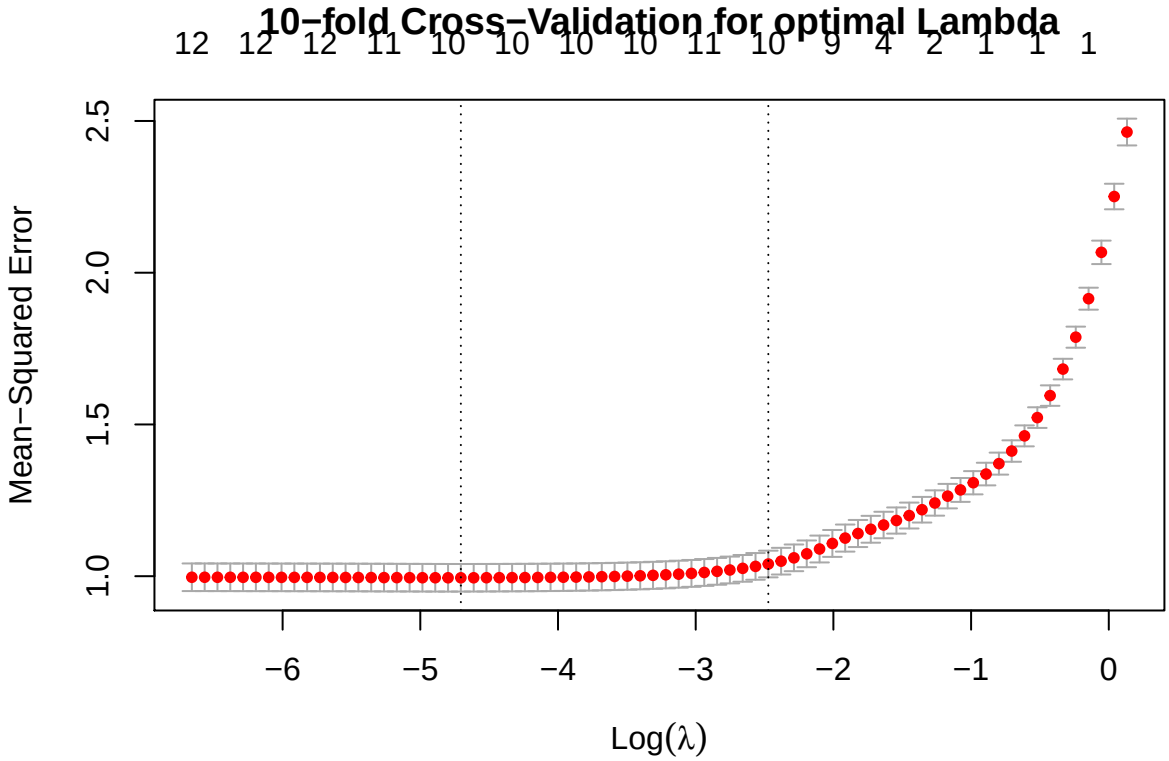
1) LASSO regression is able to perform **variable selection**, this means that only variables that helps predict our outcome variables will be used in our model and the rest will be removed from the model (coefficient equals zero). This helps the model to have less predictor variables, be less complex and have stronger explanation power.

2) LASSO regression have a **penalized likelihood strategy**. This means that our model is not overfitted to our trained data, allowing it to work as well with future data. However, this will also result in our model having higher MSE with our trained data compared to a normal linear model, but we believe that it is a worthy trade-off between bias and variance.

Compared to the other models, Stepwise regression with AIC and/or BIC or Ridge regression, both lacks an aspect that LASSO regression is able to provide. Hence, we decide on LASSO regression as our choice of model.

### LASSO regression model for violent Crimes

While building LASSO regression model, we used 10-fold Cross Validation to find the optimal $\lambda$. The graph below shows $log(\lambda)$ versus Mean-Squared Error of the model with the top row showing the number of variables in the model. To make sure our model's explanation power, we will keep the number of variables of our model equal to or below 10.




As we can see from the graph, the optimal value of $\lambda$ is `0.0090462`.

With this value of $\lambda$, we have our LASSO regression model. The coefficients for each variables are:

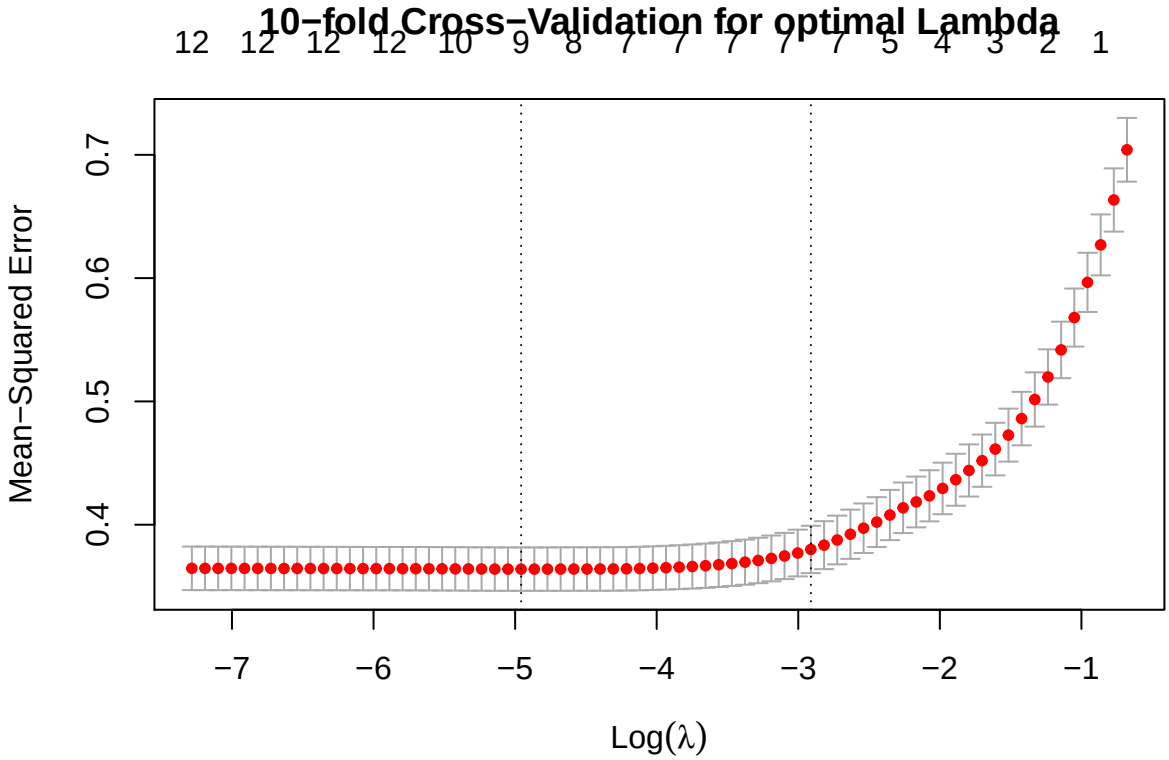
```
##                              1
## (Intercept)      1.062977e+01
```

```
## pctUrban              1.619179e-01
## medIncome            -1.696868e-01
## pctLowEdu             1.272936e-02
## pctUnemploy           9.584025e-02
## pctEmploy             .
## pctKidsBornNevrMarr   9.567759e-01
## pctHousOccup         -1.677777e-06
## pctHousOwnerOccup    -1.203414e-03
## pctVacantBoarded      8.318167e-02
## pctVacant6up         -8.306831e-02
## popDensity            .
## pctForeignBorn        9.601780e-03
```

The Mean-Squared Error for our model is `0.9809507`.

**LASSO regression model for non-violent crimes**

Similar to the LASSO model predicting violent crimes, we used 10-fold Cross Validation to find the optimal $\lambda$ for our non-violent crime model as well. The graph below shows $log(\lambda)$ versus Mean-Squared Error of the non-violent crime model with the top row showing the number of variables in the model.




As we can see from the graph, the optimal value of $\lambda$ is `0.0070303`.

With this value of $\lambda$, we have our LASSO regression model. The coefficients for each variables are:

```
##                               1
## (Intercept)          1.814836e+01
## pctUrban             6.115513e-02
## medIncome           -3.618093e-01
## pctLowEdu            .
## pctUnemploy          .
```

```
## pctEmploy            .
## pctKidsBornNevrMarr  3.057318e-01
## pctHousOccup        -8.189186e-07
## pctHousOwnerOccup   -4.321287e-03
## pctVacantBoarded     2.910231e-02
## pctVacant6up        -6.716375e-02
## popDensity           .
## pctForeignBorn       .
```

The Mean-Squared Error for our model is `0.3591024`.