ShrinkageMethods:RidgeLASSOelnet

Marta Fajlhauer

3 June 2018

Comparision of Ridge, LASSO and Elastic net

We are ready now to compare OLS and Stepwise regression methods with more advanced regression techniques. So what is the difference between OLS and rest of the regression techniques presented in this project? We can decompose error into the error due to bias and error due to variance. The Gauss Markov theorem states that among all linear unbiased estimates OLS has the smallest variance. This implies that OLS estimates have the smallest mean squared error among all linear estimators with no bias. Can there be a biased estimator with a smaller MSE? In shrinkage method, we make up in variance. In the ridge regression, the assumptions are linearity, constant variance, independence. As we have already checked carefully constant variance assumption and linearity assumption is met and y values are independent. The big advantage is that normality not needs to be assumed.

Ridge

In regression modelling the values of coefficients are determined by fitting a polynomial to training data. This can be done by minimising the error function that measures the missfit between the function $y(x, \beta)$ for any value of β. It can be represented by

$$E(\beta) = \frac{1}{2}\sum_{n=1}^{N}[y(x_n, \beta) - \widehat{y_i}]$$

(x) where $\widehat{y_i}$ are predicted values. The next problem is selecting order of polynomial that we should take into the model. For each order of the polynomial we can evaluate the residual value of $E(\widehat{\beta})$. We need to be careful to not overfit the model and do not choose polynomial with too big order. One technique that is often use to control is to adding penalty element to the error function (x). in order to discourage the coefficients from reaching large values. The simplest penalty term takes the form of sum of squares of all the coefficients leading to a modified eror function.

$$E(\beta) = \frac{1}{2}\sum_{n=1}^{N}[y(x_n, \beta) - \widehat{y_i}] + \frac{\lambda}{2}\|\beta\|^2 = E_D(\beta) + \lambda E_w(\beta)$$

where

$$\|\beta\| = \beta^T\beta = \beta_0^2 + \beta_1^2 + \cdots + \beta_M^2$$

and the coefficient λ governs the relative improtance of the regularisation term compared with the sum-of-squares error. This is ridge regression equation.

Ridge regression proceeds by adding a small value λ to the diagonal elements of the correlation matrix. One of the main obstacles in ridge regression is choosing an appropriate value of λ.

Let $\mathbf{F}$ *represent the appropriate centered and scaled "X matrix" when the regression problem under study hasis in "correlation form". Thus, if the orginal model is*

$$Y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\boldsymbol{Z}_1 + \boldsymbol{\beta}_2\boldsymbol{Z}_2 + ... + \boldsymbol{\beta}_r\boldsymbol{Z}_r + \varepsilon$$

. The new centered and scaled predictor variables are

$$f_{ij} = \frac{\boldsymbol{Z}_{ju} - \overline{\boldsymbol{Z}}_j}{\sqrt{S_{jj}}}$$

where

$$\overline{\boldsymbol{Z}}_j$$

is the average of the $\boldsymbol{Z}_{ju}, u = 1, 2, ..., n$ *and*

$$\boldsymbol{S}_{jj} = \sum_u (\boldsymbol{Z}_{ju} - \overline{\boldsymbol{Z}_j})^2$$

. Thus

$$F = \begin{bmatrix} f_{11} & f_{21} & \dots & f_{r1} \\ \vdots & \vdots & \vdots & \vdots \\ f_{1u} & f_{2u} & \dots & f_{ru} \\ \vdots & \vdots & \vdots & \vdots \\ f_{1n} & f_{2n} & \dots & f_{rn} \end{bmatrix}$$

and $F^T F$ *is the correlation matrix of the Z.*[6]

The ridge regression estimates of the r elements of

$$\boldsymbol{\beta}_F = (\boldsymbol{\beta}_{1F}, \boldsymbol{\beta}_{2F}, \dots, \boldsymbol{\beta}_{rF})^T$$

are the elements of

$$\overrightarrow{b}_F(\lambda) = (b_{1F}(\lambda), b_{2F}(\lambda), \dots, b_{rF}(\lambda))^T$$

given by:

$$\overrightarrow{b}_F(\lambda) = (F^T F + I_r \lambda)^{-1} F^T Y$$

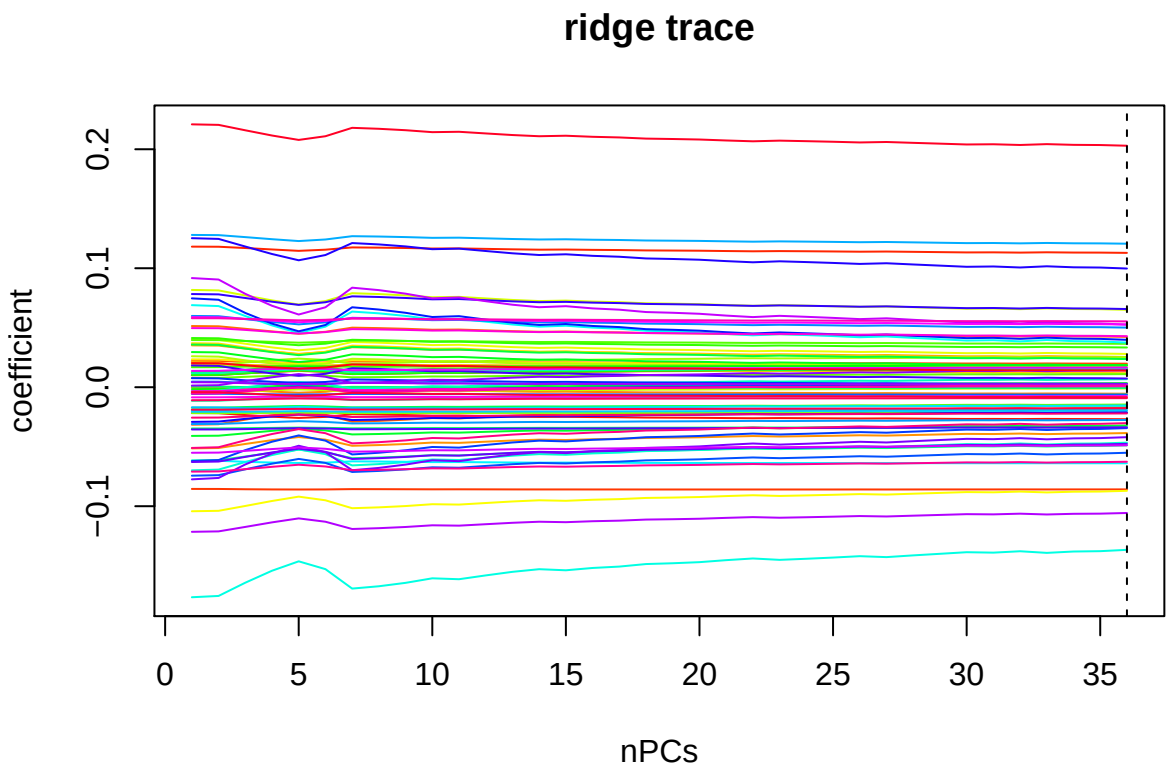
where λ is a positive number. Note that $\lambda = 0$ we obtain the least squares estimates. We can plot the coefficients $b_{jF}(\lambda)$ against λ and examine the resulting figure in ridge trace

A ridge trace is a plot that shows the ridge regression coefficients as the function of λ. In R there is a procedure that automatically selects the lowest value of λ that is calculated a general cross-validation procedure (GVC). The ridge regression doesn't set the coefficients exactly to zero unless theta = infinity. Therefore, the ridge regression cannot perform variable selection. Ridge regression performs well when there is a subset of true coefficients that are small or zero.

Once λ *has been selected (equal to* λ^**) the values* $b_j(\lambda^*)$ *are used in the prediction equation. The resulting equation is made up of estimates that are not least squares and are biased but that are more stable and provide a smaller overall means square error.* [6]

First, we will examine the correlation among the values in the data. From the correlation plots attached [Appendix 4], we can see that there are variables that are highly correlated with each other and variables that correlation between them is close to zero. We will compare ordinary least squares fit of the full model using multiple regression with other advanced regression techniques. The summary of the OLS is attached to the Appendix 8 and 10 but We know that for the best model, with the smallest AIC value, r squared is 0.57 with adjusted r squared 0.57 when ridge regression accuracy score is 0.63. There is an improvement in the model selection.

```
[1] 0.6298612
```



Can we achieve even better result using other advanced regression techniques?

LASSO

LASSO stands for Least Absolute Selection and Shrinkage Operator. In this process, some of the coefficients are shrunk to zero exactly so it performs variable selection in the linear model. In Ridge regression, we introduced an idea of adding regularisation term to an error function to omit problem with over-fitting. One of the simplest forms of reguliser is

$$E_w(\beta) = \frac{1}{2}\beta^T\beta$$

. The LASSO regression has a form:

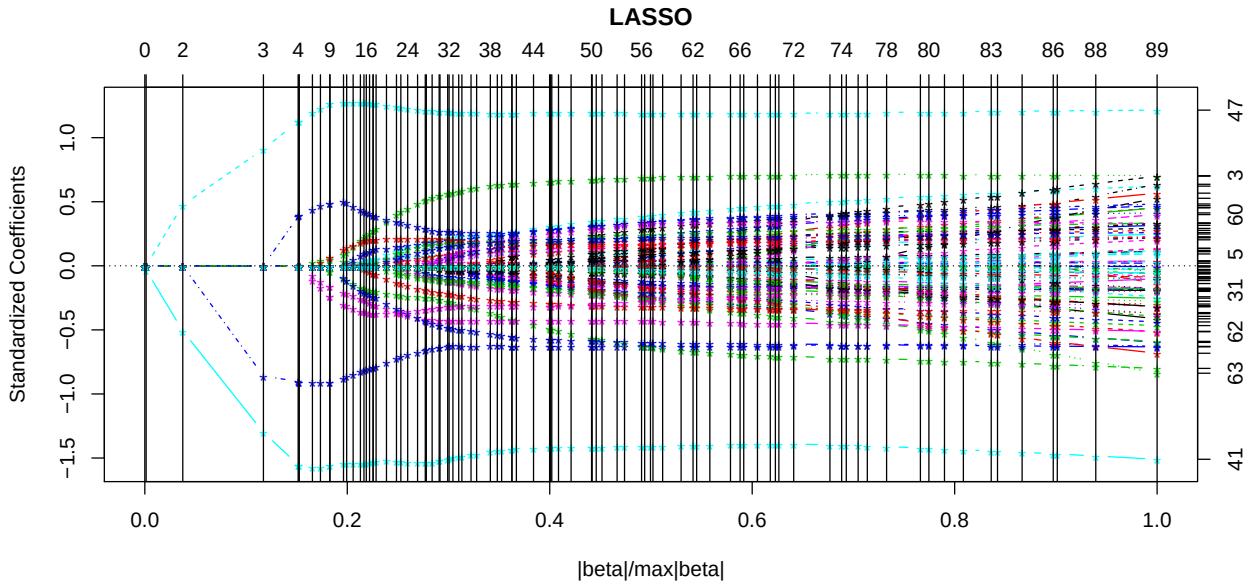
$$E_D(\beta) + \lambda E_w(\beta) = min(\beta)\frac{1}{2}\sum_{i=1}^{N} y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

As we increase lambda more coefficients shrink to zero. So, variables that are less likely to shrink to zero indicate variable more important in the model. The challenge of LASSO is to find a suitable value of the regularization coefficient λ. The disadvantage of LASSO is that if there is a group of explanatory variables that are correlated with each other LASSO takes only one variable out of the group into the model ignoring others. It takes only this variable that has the most significant impact on the model. However, if in the group will be more than 2 significant variables that are important the second one will be ignored by LASSO.

We can visualise LASSO path to check when new variables are introduced into the model [Figure 18]. In the plot, we can see: At the top are numbers of steps when different variables were introduced. 89 steps mean that there are 89 variables in the dataset as when reguliser is equal to one all variables are going into the model. On the left side of the box are variables numbers that were introduced to the model. $\frac{|\beta|}{max|\beta|}$ is the reguliser that has interval [0;1]. We can see how increasing value of β in comparison to the maximum value

allows more and more variables get into the model. On the right side of the box, we have the magnitude of the coefficient that goes into the model. In Appendix 11, we can see exactly which variable was introduced when. The new variables were coming in. PctKids2Par, PctIlleg, racePctWhite, PctPersDenseHous and so on

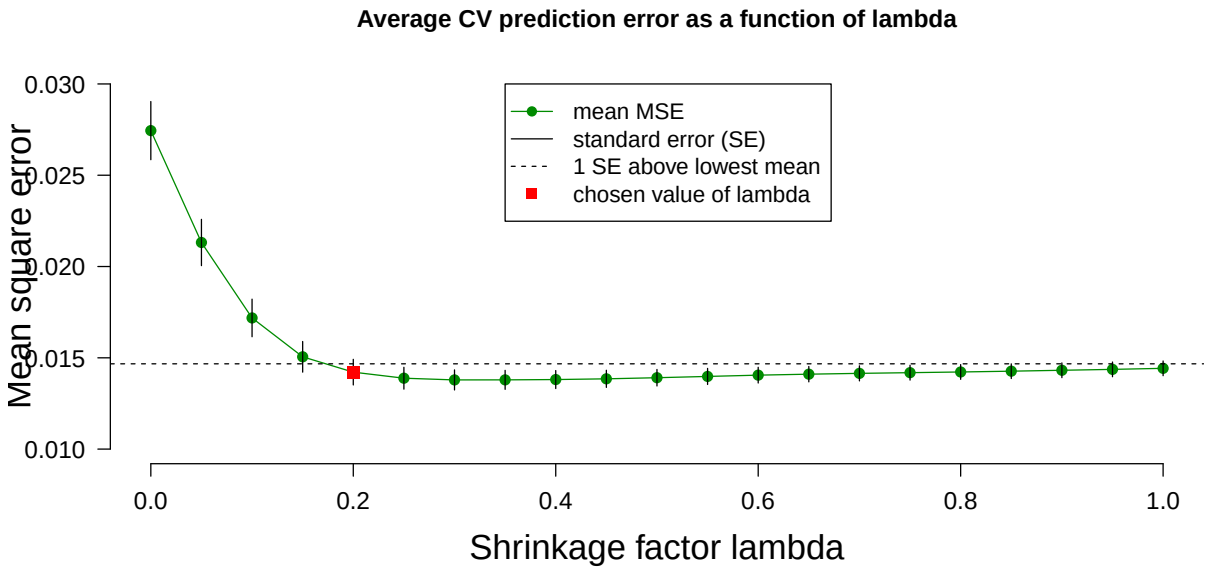
Figure 17 LASSO trace



The goal of modelling using LASSO is to decide on this value of λ that will give the best accuracy result in our model. First, we will use the randomly selected value of $\lambda = 0.375$. We can compare coefficients from LASSO with coefficients from OLS and see that they shrunk. That's one of the main goals of this approach. [Appendix 8, 10, 12]. We will now look for s value that will the best describe our model using 10 fold cross-validation. 10 cross-validation produces 10 different sizes for which the number of observation is displayed.

Figure 18 LASSO plot to choose the best λ

```
cvlab
  1   2   3   4   5   6   7   8   9  10
180 135 161 131 154 157 137 135 116 124
```



The graph shows mean MSE for each value of λ. The chosen value of λ is the first one that is below the distance of 1 SE above lowest mean. The first model corresponds to intercept. These are the coefficients for the chosen model with λ value of 0.2.

We are selecting variables: racepctblack, racePctWhite, pctWInvInc, FemalePctDiv, PctKids2Par, PctIlleg, PctNotSpeakEnglWell, PctPersDenseHous, MedRentPctHousInc, PctBornSameState, MedOwnCostPctInc-NoMtg into the model and the magnitude of the coefficients are saved in the Appendix 13.

To compare how well does the training model work on training and validation sets we compare MSE of the validation set with a mean of the training set. Since the mean MSE of validation set is in the interval (0.013, 0.027) that is constructed from values of the mean MSE for training sets. Very small difference between the estimator and what is estimated leads to the conclusion that this is a good model selection.

```
[1] 0.01550168
```

```
 [1] 0.01378886 0.01379238 0.01380939 0.01384950 0.01388288 0.01390947
 [7] 0.01398018 0.01405004 0.01410358 0.01414846 0.01418645 0.01421485
[13] 0.01422415 0.01426807 0.01431718 0.01436691 0.01442377 0.01505453
[19] 0.01718340 0.02131632 0.02744360
```

Elastic Net

When we are working with high dimensional data correlations between the variables can be high resulting in collinearity. The lasso penalty is not secure in the choice among a set of strong but correlated variables. As I already explained in the project LASSO fails to perform group selection. The ridge penalty tends to shrink the coefficients of correlated variables toward each other. We may want to include variables that are correlated with each other to the final modelling process. The compromise is the elastic net penalty.

$$E_w(\beta) = \sum_{j=1}^{p} (\alpha|\beta_j| + (1-\alpha)\beta_j^2)$$

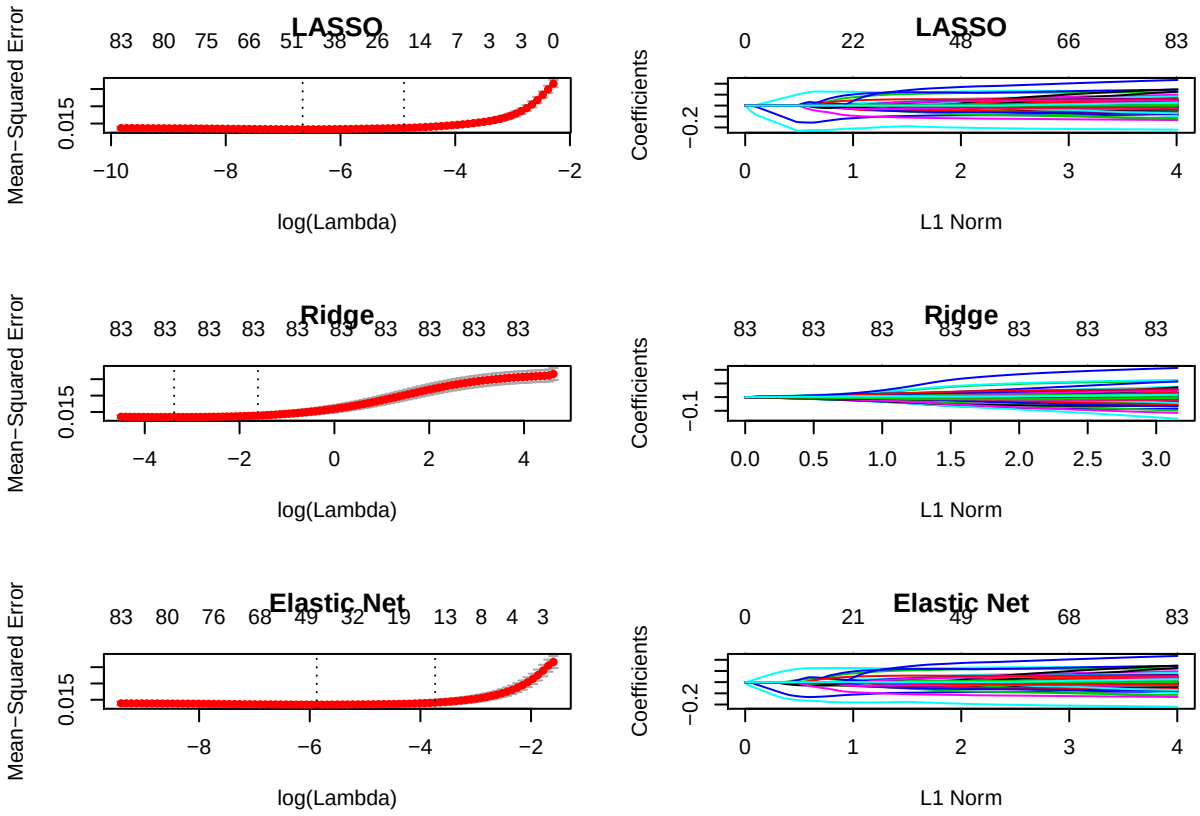
The equation assures that highly correlated features are averaged and there is a sparse solution in coefficients of these averaged features.

Comparision of Advanced regression techniques

After learning about different advanced regression techniques it is good to see the application on our dataset. First, we can visualise trace plot for different models and cross-validation curve (red dotted line) and upper

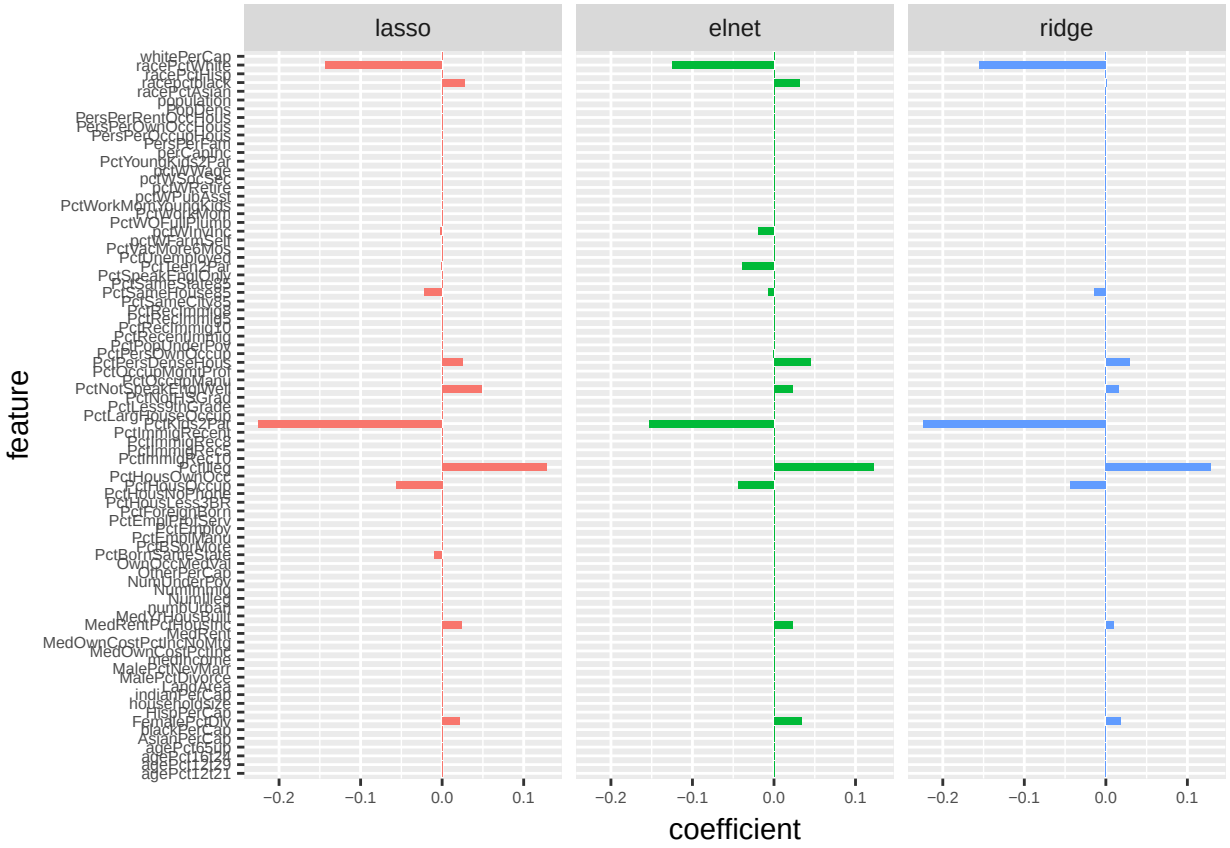
and lower standard deviation curves along the lambda sequence (error bars). Two selected lambdas are indicated by the vertical dotted lines.

Figure 19: trace plots and cross-validation curves



We can also visualise what variables are selected during feature selection model for LASSO and Elastic Net. There are 3 variables more selected to LASSO regression which are the percentage of public assistance, PctVacMore6Mos and NumbUrban. There are six variables that are selected as very important variables by both models which are: RacePctWhite, RacePctBlack, PctNotSpeakEnglishWell, PctKids2Par, PctIlleg and Pct House Occup and rest of the variables selected by both of the models are less significant namely: PctWInvInc, PctTeen2Par, PctSameHouse85, PctPersDenseHouse, PctBornSameState, MedRentPctHouseInc, FemalePctDiv.

Figure 20: Importance of variables in Ridge, LASSO and Elastic Net models



We can compare models to select the best one. With the highest r-squared of close to 0.68 and the lowest MAE - Mean Absolute Error (0.72), the elastic net seems to be the best model for this problem. This is a big improvement from OLS approach where was the r-squared value of 0.57.

In the table below we have:

- **Mean Absolute Error (MAE)**: It is a mean value of the sum of absolute residual values. It measures the average size of the error between the actual and predicted value where all individuals have equal weight.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_i - \hat{y}_i|$$

  . So smaller the residuals, smaller the MAE and better prediction outcome. elnet has the smallest Min(MAE) and the model with the smallest MAE is taken into account.

- **Root mean squared error (RMSE)**: RMSE the square root of the average of squared differences between prediction and actual observation.

$$RMSE = sqrt\frac{1}{n}\sum_{j=1}^{n}(y_i - \hat{y}_i)^2$$

  the RMSE gives a relatively high weight to large errors. MAE is constant and RMSE increases as the variance associated with the frequency distribution of error magnitudes also increase. Similarly to MAE, the best model is the one with the smallest RMSE.

- **R-squared** measures how much variability in the data is explained by the model. So we know that for these selected variables explained 68% number of violent crimes per population.

$$R^2 = \frac{SumSquaredRegressionerror}{SumSquaredTotalError}$$

```
Call:
summary.resamples(object = results)

Models: elnet, ridge, lasso
Number of resamples: 30

MAE
            Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
elnet 0.07206638 0.08154293 0.08692103 0.08651187 0.08994061 0.10131315
ridge 0.07941956 0.08406998 0.08628255 0.08656081 0.08894831 0.09500617
lasso 0.07437205 0.08180110 0.08485058 0.08517579 0.08905376 0.09799299
      NA's
elnet    0
ridge    0
lasso    0

RMSE
            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
elnet 0.09696319 0.1126408 0.1186252 0.1188824 0.1238172 0.1392964    0
ridge 0.10817436 0.1151878 0.1187702 0.1192015 0.1225558 0.1330894    0
lasso 0.09628154 0.1135001 0.1191357 0.1187937 0.1254055 0.1356064    0

Rsquared
           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
elnet 0.4087235 0.4804743 0.5270551 0.5220207 0.5557413 0.6779807    0
ridge 0.4100411 0.4848545 0.5178770 0.5188117 0.5451701 0.6342425    0
lasso 0.4344633 0.4874135 0.5244186 0.5211939 0.5474685 0.6052880    0
```

So we have 68 percent variability explained in the data which suggest that there is indeed positive correlation between number of violent crimes in a given region and variables that LASSO and ElasticNet feature selection methods highlighted as important namely: racepctblack, racePctWhite, pctWInvInc, FemalePctDiv, PctTeen2Par, PctKids2Par, PctIlleg, PctNotSpeakEnglWell, PctPersDenseHous, PctHousOccup, MedRentPctHousInc, PctBornSameState, PctSameHouse85.

I have included a table that represents which variables were used to Backwards Stepwise Regression before and after logarithmic transformation of the response variable as well as LASSO and Elastic Net. I didn't include Ridge regression as Ridge regression fails to perform selection process. There are 8 variables selected by different feature selection methods and 6 variables selected by 3 out of 4 models. Some of the variables are measures of family life such as PctKids2Par, PctTeen2Par and FamlePctDiv. Other variables related to hard conditions of immigrants in the USA such as PctIlleg, PctNotSpeakEnglwell, PctBornSameState. On the other hand, there are also variables measuring poverty level of the society such as pctWPubAsst, PctPersDenseHous, PctHouseOccup, PctWInvInc. There are also: RacePctWhite, RacePctBlack and PctVacMore6Months

Figure 21: Variables going into the model after running different feature selection methods

Variable	Back	Back Log	LASSO	Elastic Net
racepctblack	✓	✓	✓	✓
racePctWhite	✓	✓	✓	✓
pctWInvInc	✓	✓	✓	✓
PctKids2Par	✓	✓	✓	✓
PctIlleg	✓	✓	✓	✓
PctNotSpeakEnglWell	✓	✓	✓	✓
PctPersDenseHous	✓	✓	✓	✓
PctHousOccup	✓	✓	✓	✓
PctBornSameState	✓	✓	✓	✓
PctVacMore6Mos	✓	✓	✓	
pctWPubAsst	✓	✓	✓	
FemalePctDiv		✓	✓	✓
PctTeen2Par	✓		✓	✓
OtherPerCap	✓	✓		
PctWSocSec	✓	✓		
NumUnderPov	✓	✓		
PctEmplManu	✓	✓		
PersPerFam	✓	✓		
PctImmigRecent	✓	✓		
PctImmigRec10	✓	✓		
MedOwnCostPctIncNoMtg	✓	✓		
PctSameHouse85			✓	✓
MedRentPctHousInc			✓	✓
agePct12t21	✓			
pctWWage	✓			
PctEmploy	✓			
PctWorkMom	✓			
NumImmig	✓			
PctSpeakEnglOnly	✓			
PersPerRentOccHous	✓			
PctHousOwnOcc	✓			
PctHousNoPhone	✓			
OwnOccMedVal	✓			
MedOwnCostPctInc	✓			
PctForeignBorn	✓			
householdsize		✓		
agePct12t29		✓		
medIncome		✓		
PerCapInc		✓		
whitePerCap		✓		
HispPerCap		✓		
PctEmplProfServ		✓		
NumIlleg		✓		
PctWOFullPlumb		✓		
LandArea		✓		
PopDens		✓		
NumUrban			✓	