To: Professor Yuce
Applied Mathematics
New York City College of Technology

From: Brian Holliday
Applied Mathematics
New York City College of Technology

Subject: Project 7: Lasso Regression Crime Data

Figure:

Date: 5/22/20

Summary:

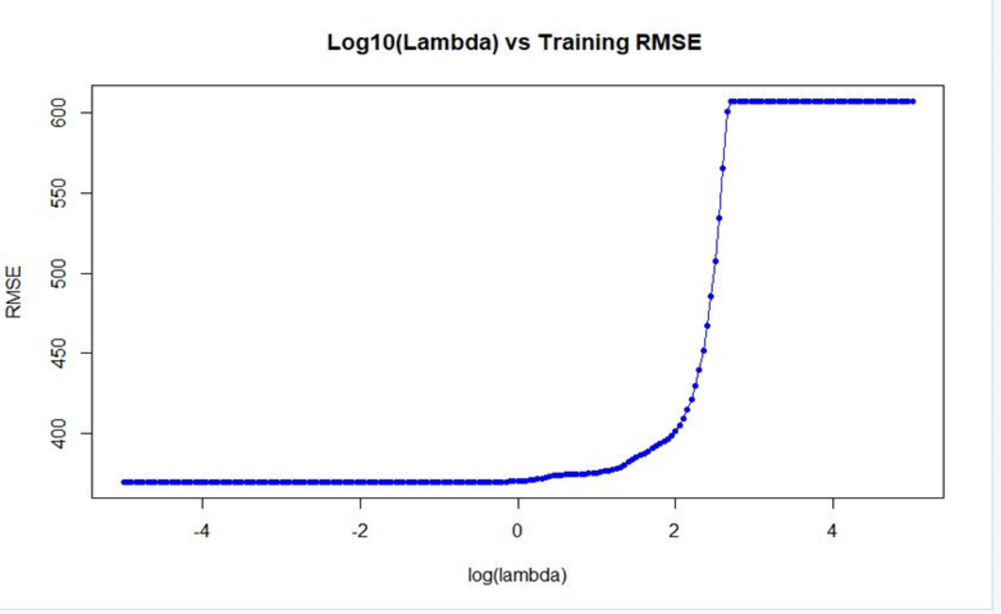
In this analysis we ran a lasso regression with data pertaining to crime rate. We used this regression to narrow down what factors are most associated with violet crime. Through our lasso regression we were able see that percentage of kids in family housing with two parents and percentage of kids born to parents never married are the two biggest factors according to this model. The intercept value is also important to this model. We went with a lambda value of 2.25 for our lasso regression.

Figure 1: Training Testing Split

```
> train_sample = sample(1:nrow(crime), floor(nrow(crime)*0.80))
> train_crime = crime[train_sample, ]
> test_crime = crime[-train_sample, ]
```

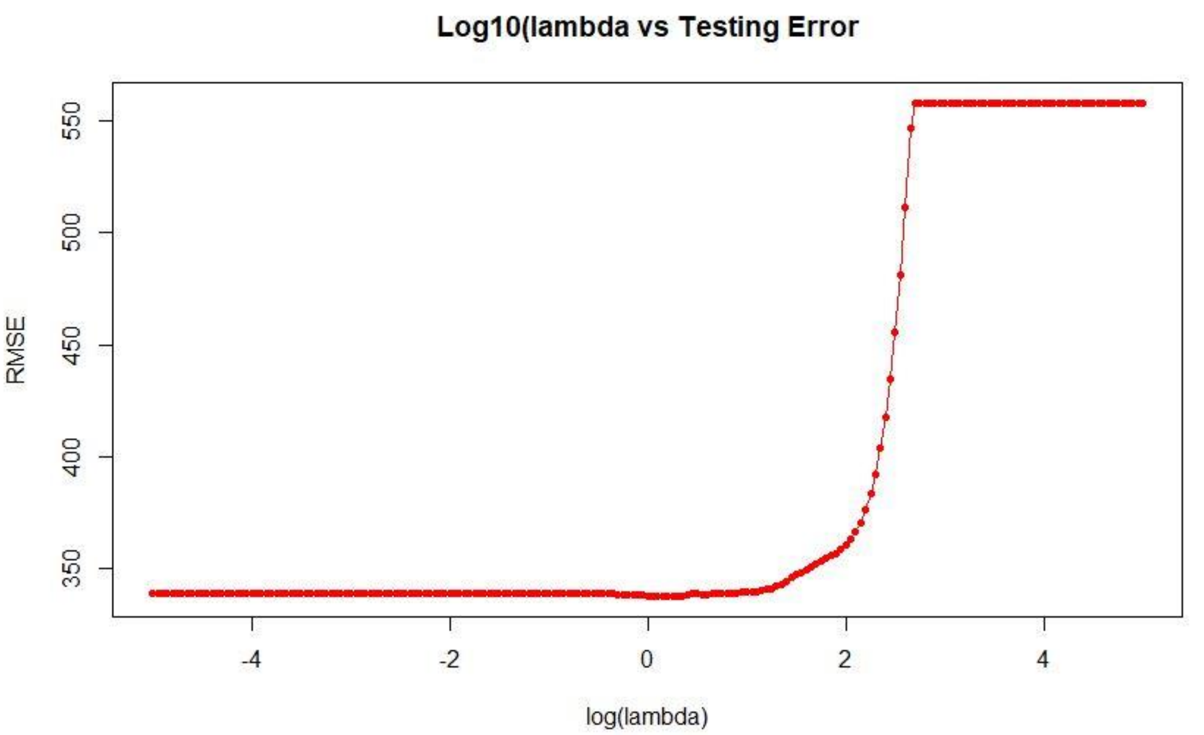
1.

Figure 2: Lambda vs Training Error



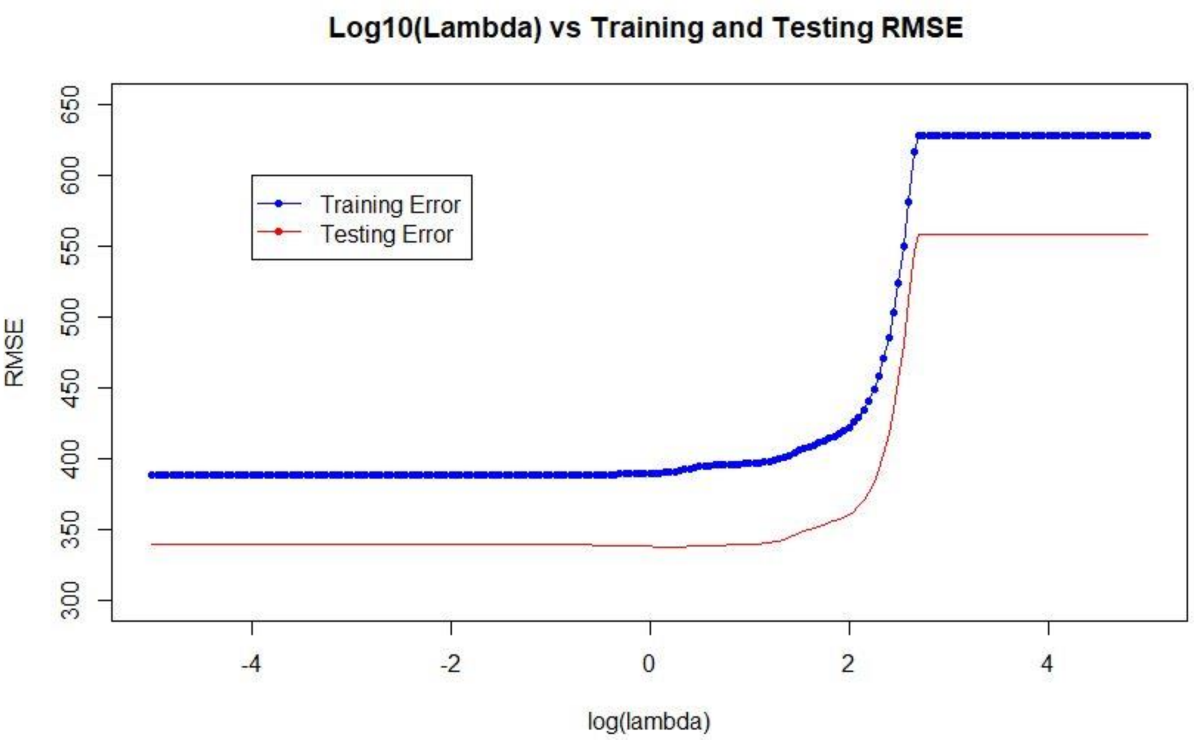
In this graph we have the lambda values vs the training error for our optimal model. We can see that from about -4 to 0 our model is at about 300 and then makes the transition to about 600 from 0 to about 2.5. Let's see if we see a similar pattern for the testing error.

Figure 3: Lambda vs Testing Error



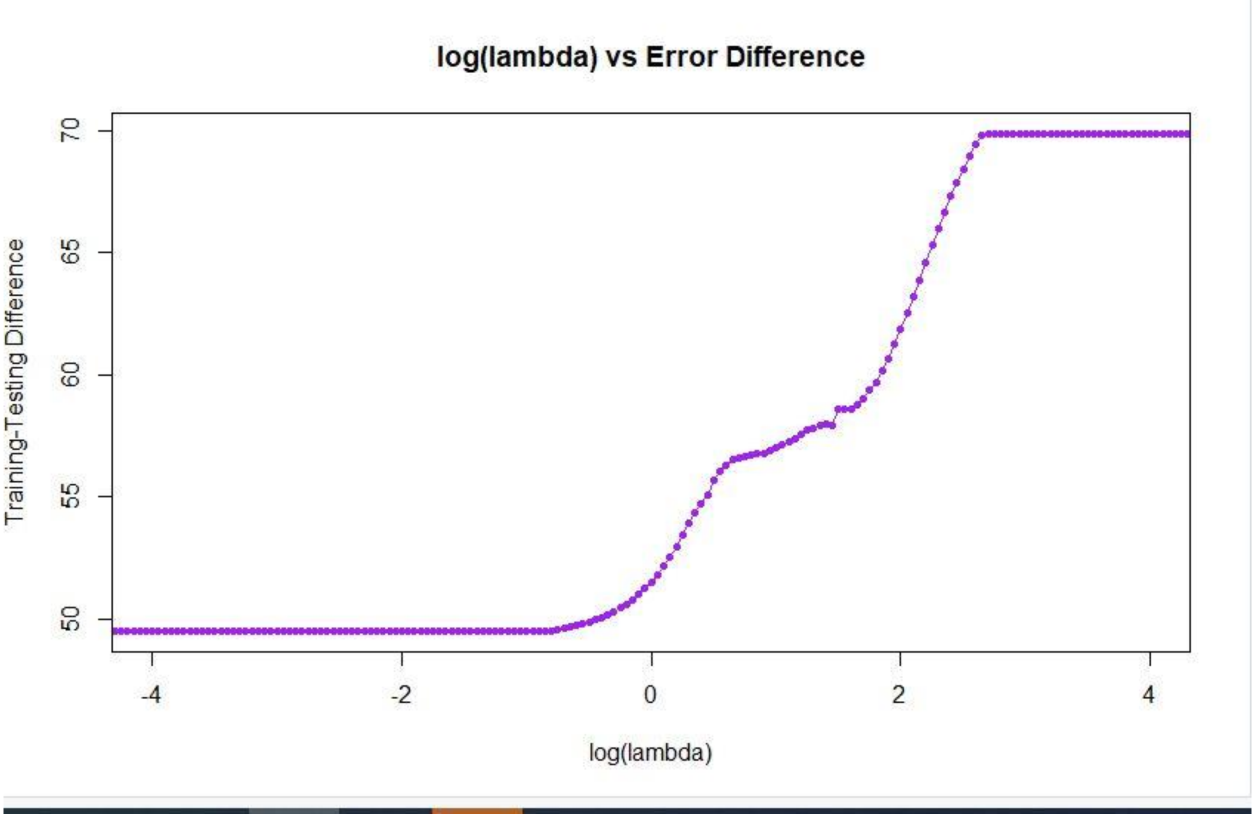
In this graph we have the lambda values vs the training error for our optimal model. We can see that from about -4 to 0 our model is at about 900 and then makes the transition to about 650 from 0 to about 2.5. We have a similar pattern in both models. The training error is less than the testing error, but the will both move closer to each other from 0 – 2.5 on the x-axis.

Figure 4: Lambda vs Training and Testing Error



We see that the lambda values are far apart and then coverage upon each from about 0 – 2.5. Let's check the difference graph of the RMSE of the training and testing before deciding on a interval for our lambda values.

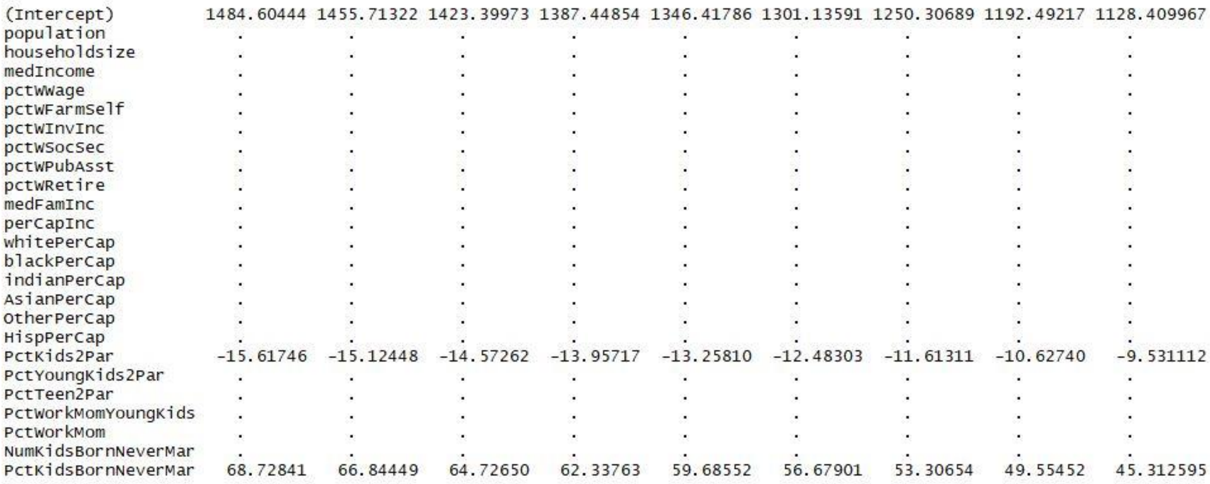
Figure 5: Lambda vs Training Testing Difference



If we follow our three-zone model, where in Zone 1, the models do not perform well because the testing error is much higher than the training error, this zone is overfit. In our graph this is from about [-5, 0] on the x-axis. There is an underfit in Zone 3, this is from about 2.5 to 5 on the x-axis. In the underfit the model is not complicated enough to get the correct set of variables. We will be looking at lambda values from about [1.95 – 2.5], because this is the zone in which the training and testing error are close to each other and they are not on the highest end of the error in this zone. This is when the RMSE has the greatest variance with respect to lambda.

3.

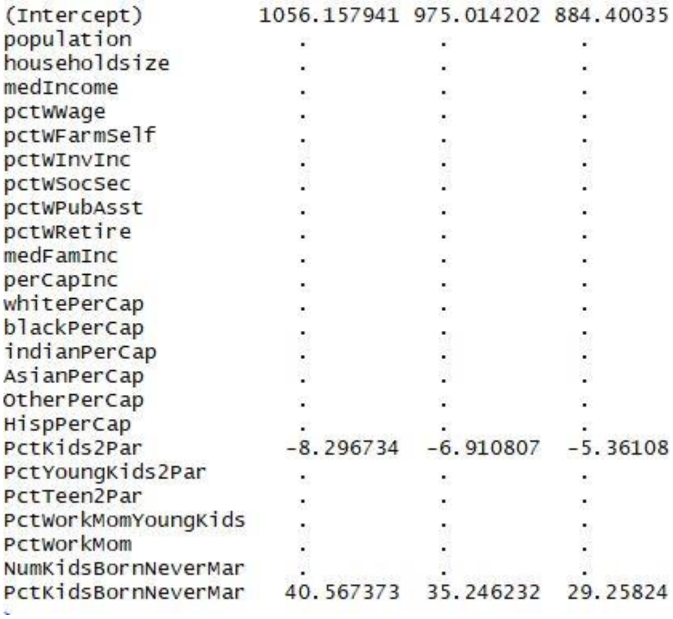
Figure 6: Lambda 1- 9



```
(Intercept)          1484.60444 1455.71322 1423.39973 1387.44854 1346.41786 1301.13591 1250.30689 1192.49217 1128.409967
population                .          .          .          .          .          .          .          .           .
householdsize             .          .          .          .          .          .          .          .           .
medIncome                 .          .          .          .          .          .          .          .           .
pctWWage                  .          .          .          .          .          .          .          .           .
pctWFarmSelf              .          .          .          .          .          .          .          .           .
pctWInvInc                .          .          .          .          .          .          .          .           .
pctWSocSec                .          .          .          .          .          .          .          .           .
pctWPubAsst               .          .          .          .          .          .          .          .           .
pctWRetire                .          .          .          .          .          .          .          .           .
medFamInc                 .          .          .          .          .          .          .          .           .
perCapInc                 .          .          .          .          .          .          .          .           .
whitePerCap               .          .          .          .          .          .          .          .           .
blackPerCap               .          .          .          .          .          .          .          .           .
indianPerCap              .          .          .          .          .          .          .          .           .
AsianPerCap               .          .          .          .          .          .          .          .           .
OtherPerCap               .          .          .          .          .          .          .          .           .
HispPerCap                .          .          .          .          .          .          .          .           .
PctKids2Par           -15.61746  -15.12448  -14.57262  -13.95717  -13.25810  -12.48303  -11.61311  -10.62740   -9.531112
PctYoungKids2Par          .          .          .          .          .          .          .          .           .
PctTeen2Par               .          .          .          .          .          .          .          .           .
PctWorkMomYoungKids       .          .          .          .          .          .          .          .           .
PctWorkMom                .          .          .          .          .          .          .          .           .
NumKidsBornNeverMar       .          .          .          .          .          .          .          .           .
PctKidsBornNeverMar    68.72841   66.84449   64.72650   62.33763   59.68552   56.67901   53.30654   49.55452   45.312595
```

In this figure we have the coefficient values for lambda values 1 – 9.

Figure 7: Lambda 10 -12



```
(Intercept)          1056.157941 975.014202 884.40035
population                .          .          .
householdsize             .          .          .
medIncome                 .          .          .
pctWWage                  .          .          .
pctWFarmSelf              .          .          .
pctWInvInc                .          .          .
pctWSocSec                .          .          .
pctWPubAsst               .          .          .
pctWRetire                .          .          .
medFamInc                 .          .          .
perCapInc                 .          .          .
whitePerCap               .          .          .
blackPerCap               .          .          .
indianPerCap              .          .          .
AsianPerCap               .          .          .
OtherPerCap               .          .          .
HispPerCap                .          .          .
PctKids2Par            -8.296734  -6.910807  -5.36108
PctYoungKids2Par          .          .          .
PctTeen2Par               .          .          .
PctWorkMomYoungKids       .          .          .
PctWorkMom                .          .          .
NumKidsBornNeverMar       .          .          .
PctKidsBornNeverMar    40.567373  35.246232  29.25824
```

In figures 6 and 7, we have all the coefficients for our twelve lambda values.

Figure 8: Lambda Values

```
      [,1]
 [1,] 1.95
 [2,] 2.00
 [3,] 2.05
 [4,] 2.10
 [5,] 2.15
 [6,] 2.20
 [7,] 2.25
 [8,] 2.30
 [9,] 2.35
[10,] 2.40
[11,] 2.45
[12,] 2.50
```

2

Given the evidence from figure 6 and 7, we could see as we move from lambda 1 – 12 the coefficient values become smaller as we move through each lambda value. We see a clear pattern in our coefficients as well. This is that we only have three significant factors. For the optimal model, I would choose Model 7, with lambda value 2.25. This is because if we look at the criteria, we have for a good model, we want a model in which the training and testing error are close to each other and they are not on the highest end of the error in this zone. This would most likely be happening in middle of our lambda values because it wouldn't be on either end of an under or overfit and RMSE would be in the middle this high- and low-end values.

4.

The top three factors for our model are the intercept, Percentage of kids in family housing with two parents and percentage of kids born to parents never married. Two parents' houses have a negative impact on crime rate, while the kids born to parents are were never married will have a positive impact on crime rate. In other words, according to this model places with kids born into situations with unmarried parents will have more crime and places with kids born into two parent households will have less crime. The intercept probably has the most influence on our model with a value of about 1025. This may tell use that this maybe the minimum threshold for crime per 100,000 people for our model. Next, is the percentage of kids born to parents never married. This has a coefficient of about 53 and it increases our crime rate. Last is the percentage of kids in family housing with two parents. This has a negative impact on our crime rate with a coefficient of about -11.