5030 Final Project

Analysis of Crime Rate

Instructor:Professor. Maria Caterina Bramati

Ziwei Yin(zy258@cornell.edu)

Hang Su(hs865@cornell.edu)

Gunjan Sood(gs652@cornell.edu)

Xihan Peng(xp49@cornell.edu)

❑ Abstract

This project is to analyze the crime rate present in different counties and variables that affect the crime rate and how it can be improved. After some exploratory analysis – the graphical and numerical summaries of the data and checking the relationship between the response and the predictors, it was seen that some of the variables and the errors were not associated linearly. Hence, some transformations were carried out in order to make the relationship linear. Log transformations of the variables provided the required results and hence the data modeling has been done using the log transformed variables.

Next, model diagnostics was done to find if there are any outliers in the model fitted after variable transformations and since one outlier was observed(with cook's distance>0.5), it was removed from the dataset (for simplicity) and the model was refitted to check if that had any high influence on the model or not. Also, the correlation plot among the variables showed that some of the variables were highly correlated. Hence, variable selection process was carried out in order to find the predictors which are significantly related to the response and removing the insignificant predictors from the model. Various methods like Best Subset, Forward Selection, Backward Selection, Shrinkage Methods were performed and different models were obtained from them based on AIC, BIC, adjusted R^2 and Mallow's Cp values.

The models obtained from above methods were then finally compared based on their MSE values obtained through Leave-One-Out Cross Validation and K-Fold Cross Validation procedures. And then the final model was chosen amongst them based on least MSE and significant p-values from the t-test, including log(pop), p18_25, ppoverty, punemployed, avg_income and log(tot_income) as the significant predictors.

❑ Data description

On June 13, 2016, a breaking news was reported that 49 people were killed and 53 injured by a 29-year-old security guard in an Orlando nightclub. This week, a professor in California was shot on campus. Crime rates are becoming increasingly frequent and therefore our team decided to analyze crime dataset and find factors affecting it in order to provide suggestions on crime rate reduction based on reliable statistical and analytical result.

Crime dataset was selected from the published book, Hands-On Programming with R, written by, Garrett Grolemund. URL of the data source can be found below:

https://github.com/rstudio/Intro/tree/master/data/counties.csv

The response variable in our analysis is crime, which represents "number of crimes per 1000 people". Dataset includes 15 variables including state, counties, area, pop, p18_25, p65, nphysician, phs, nhospbeds, pcollege, ppoverty, region, pumemployed, avg_income and total income.

The counties predictor in the dataset only contains few counties from each state, which is not comprehensive to make conclusions based on geographical influences. With limited geographical background information, this report focuses on the analysis of artificial reasons by searching factors that could be further improved. So, variable, county, state and region are dropped.

Table 1. Dataset Description

Variable	Measure
County	Name of County
state	Name of state
crime	**Number of Crime per 1000 people**
area	Land area of county
pop	Population of county
p18_25	% of people between 18 and 25years old
p65	% age greater than 65 years old
nphysician	Number of physicians
nhospbeds	Number of hospital beds
phs	% of people with high school degree
pcollege	% of people with college degree
ppoverty	% of people below poverty line
punemployed	% of people Unemployed
avg_income	Per capita income
tot_income	Total income
region	Region of County

❑ Preliminary transformations

Before the model selection was done, the dataset was verified and checked whether there is any missing observation in the dataset. No missing value were found in the dataset, hence all the observations in the original dataset were included to do further analysis.

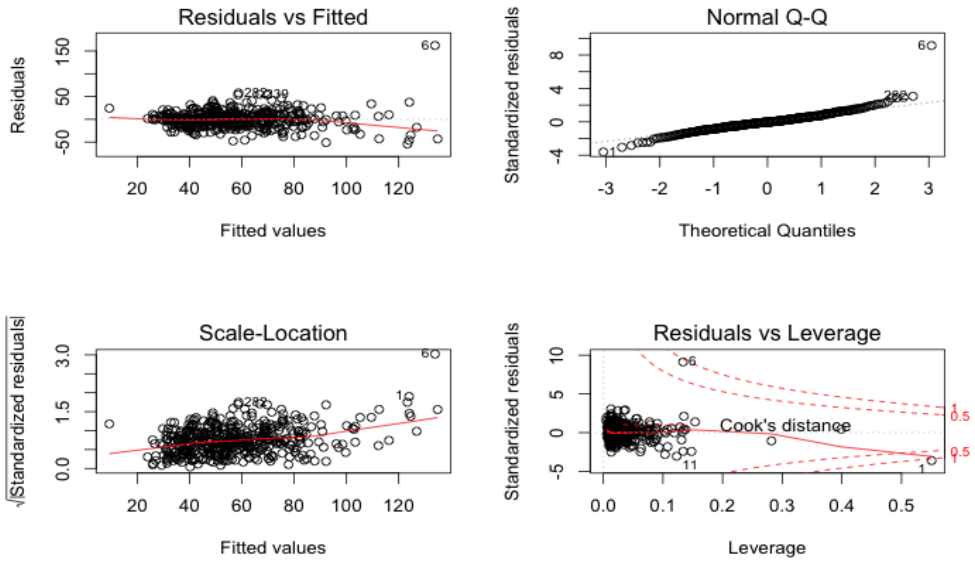
After that the correlation among the predictors were verified and it was observed that some of the predictors have very high correlation like 1. nhospbeds and pop 2. nphysician and pop 3. nphysician and nhospbeds 4. nphysician and nhospbeds. Therefore, nphysician was removed to avoid multicollinearity problem. And hence variable selection process was carried out later to identify significant and insignificant variables.

From the graphical CRplots and plots between residuals and predictors and QQplots it was verified if any transformation is needed or not. It was observed that some of the variables needed transformations and since log transformations looked appropriate they were transformed accordingly.

An initial linear regression model was fitted with crime as the response and area, pop, p18_25, p65, nhospbeds, phs, pcollege, ppoverty, punemployed, avg_income and tot_income as the predictors.

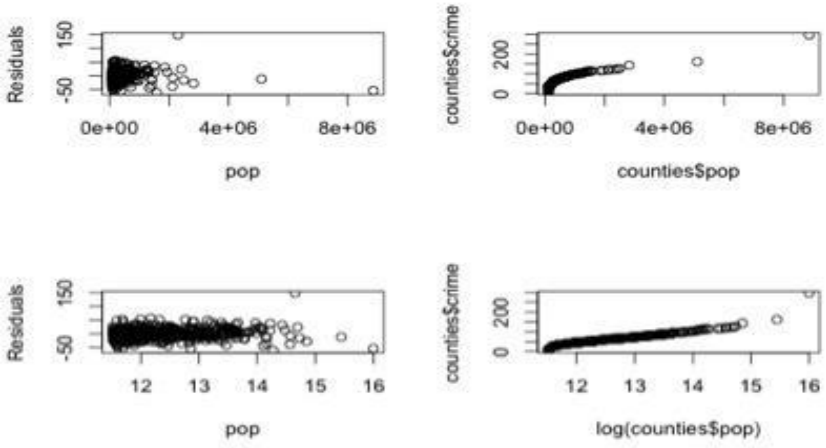
The diagnostics plots (Plot A) of the initial model. Both the residual plot and the QQplot look ok, but one should check the residual consistency and normality for each predictor individually and to see whether any transformation is needed.

Plot A: Diagnostics plots for the initial linear model



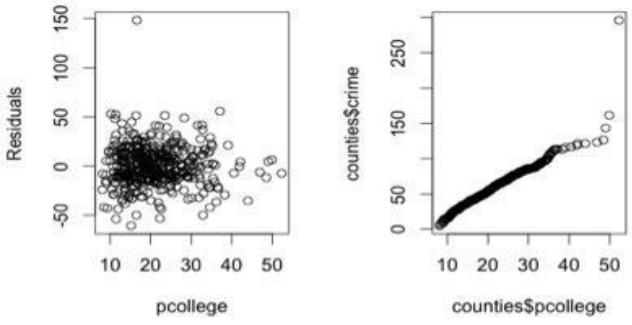
For pop, the residuals are more concentrative when pop is small than when pop is large. Also, the relationship between crime and pop is nonlinear. After trying several transformation methods, it was found that a log transformation for pop would improve this situation. For area, nhospbeds and tot_income, the situations are very similar to pop. Therefore, a log transformation was also applied to area, nhospbeds and tot_income. Plot B shows the residual plots and QQplots for pop before and after a log transformation.

Plot B: Diagnostics plots for pop before and after log-transformation



For p18_25, p65, phs, pcollege, ppoverty, punemploye and avg_income, both the residual plot and the QQplot look ok, so no transformation was did. Plot C shows the residual plot and QQplot for pcollege.

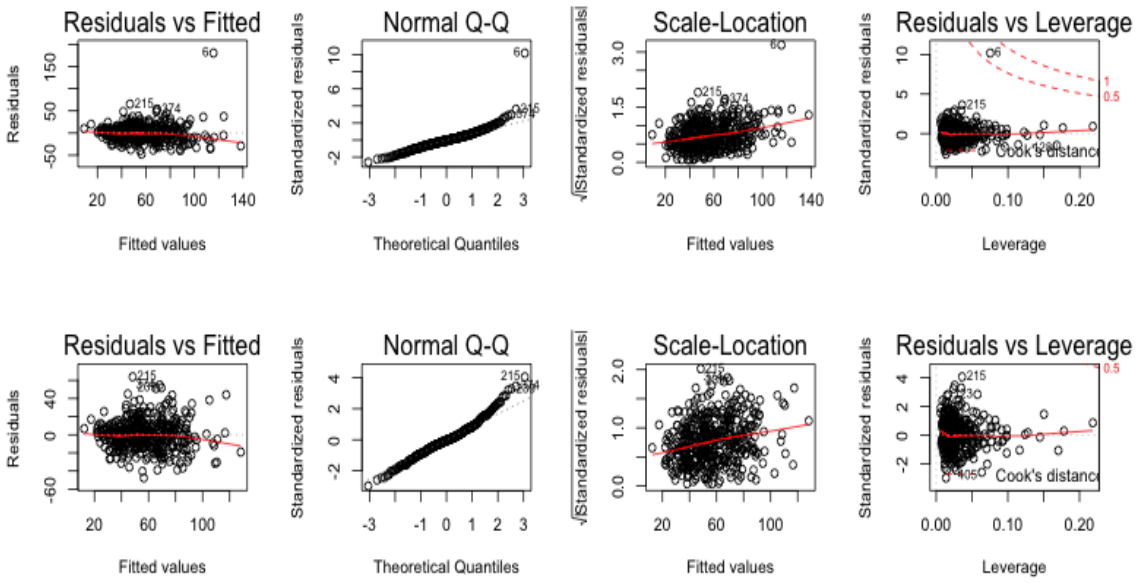
Plot C: Diagnostics plots for pcollege



After all the transformations were done linear regression model with crime as response and transformed variables: log(area), p18_25, p65, log(pop), log(nhospbeds), phs, pcollege, ppoverty, punemployed, avg_income, log(tot_income) as predictors was fitted. The diagnostics plots was again done for the new regression. It was observed that the sixth observation is an outlier, so for

simplicity removed the sixth observation from the dataset. Plot M shows the diagnostics plots for the transformed model before and after removing the outlier.

Plot D: Diagnostics plots for the transformed model before and after removing the outlier



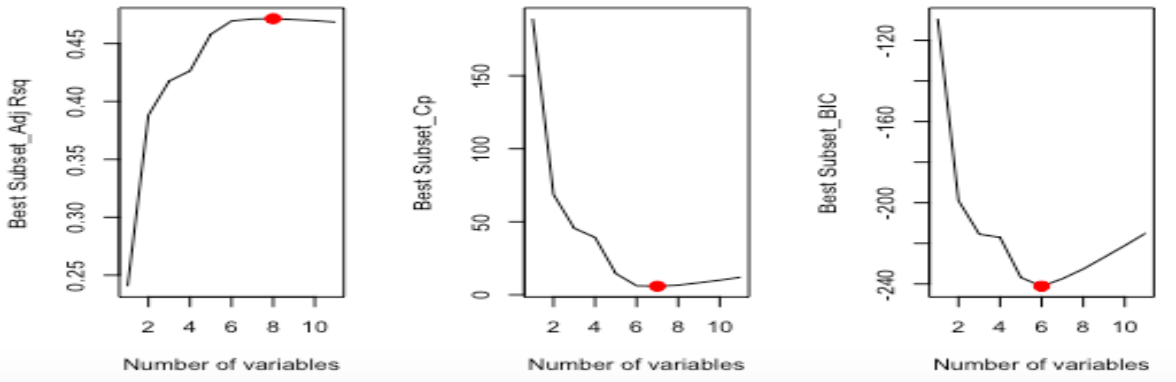
❑ Variable Selection

Besides Ridge Regression and Lasso Regression, Best subset, Forward and Backward selection were done to select desirable models based on the criteria of adjusted R^2, Mallow's Cp, AIC and BIC.

Firstly, best subset selection was applied using the regsubset function in R. For further observation and analysis,its adjusted $R^2$was checked. The model with highest R^2under Best Subset Selection was selected, which could explain most proportions of the response variable. It included 8 variables, which were log(area), log(pop), p18_25, log(nhospbeds), ppoverty, punemployment, avg_income, log(tot_income).Then checked the Mallow's Cp. A low Mallow's Cp indicates that the model is relatively precise or with smaller variance in estimating the true regression coefficients and predicting the response variable. So the one with lowest Mallow's Cp is selected as the best model. It is a 7-variable model with variables, log(area), log(pop), p18_25, ppoverty, punemployed, avg_income, log(tot_income). Besides of adjusted R^2 and Mallow's Cp, BIC was another important criterion, which was used to test the relative quality of different statistical models for a given data set. We selected the model with lowest BIC, which gave us the same 7-

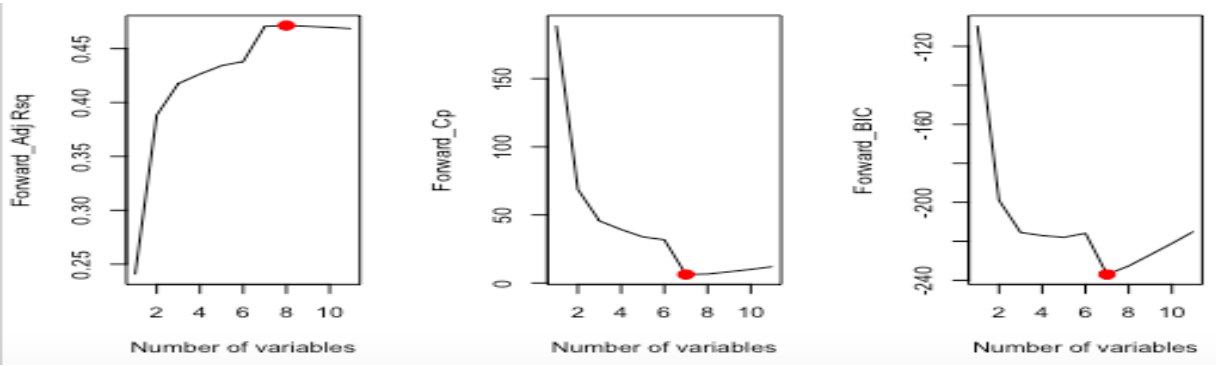
variable model including log(pop), p18_25, ppoverty, punemployed, avg_income and log(tot_income).

Plot E: Best Subset Variable Selection



Next forward stepwise selection approach was carried out. Again observed the model with largest R^2 and selected the 8-variable model including log(area), log(pop), p18_25, log(nhospbeds), ppoverty, pumemployment, avg_income and log(tot_income), which is the same as the model selected by criterion of adjusted R^2under best subset selction. The smallest Mallow's Cp and smallest BIC both selected the same 7-variable model including log(pop), p18_25, log(nhospbeds), ppoverty, pumemployment, avg_income and log(tot_income).

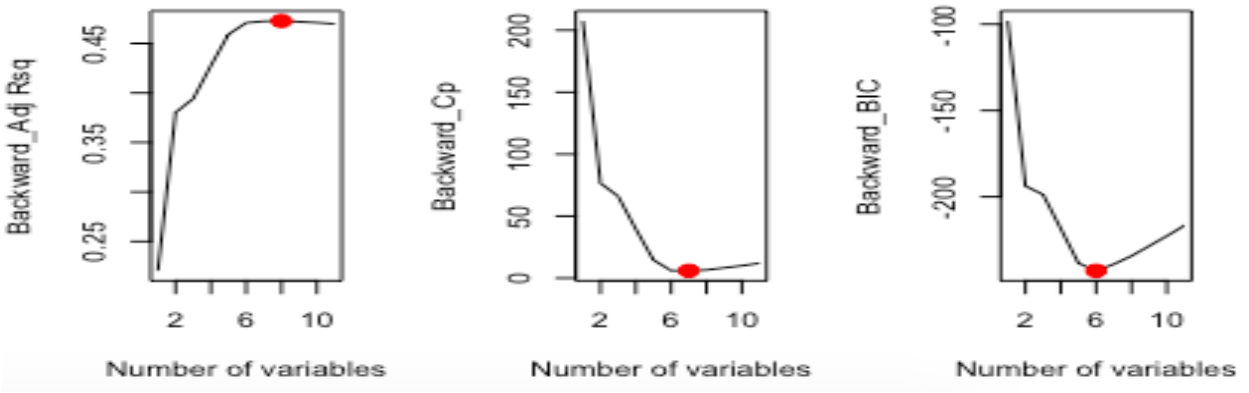
Plot F: Forward Stepwise Variable Selection



Backward Stepwise Selection was also took into consideration in order to select a desirable model. After similar, largest adjusted R^2gave us a model with 8 variables, log(area), log(pop), p18_25, log(nhospbeds), ppoverty, avg_income, log(tot_income) and punemployed. Similarly with best subset selection, the Mallow's Cp gave a 7-variable model including log(area), log(pop), p18_25, ppoverty, punemployed, avg_income, log(tot_income) and BIC analysis gave

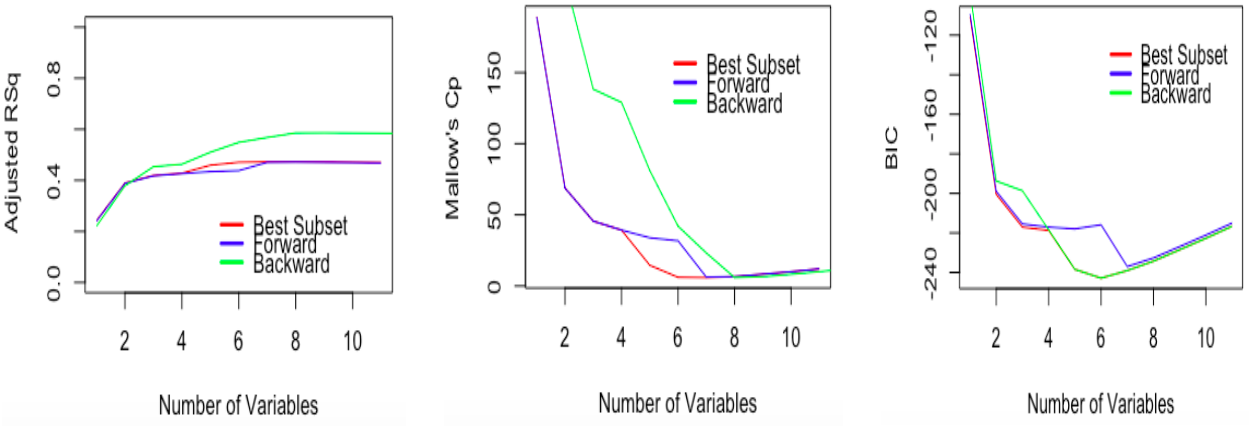
us a 6-variable model including log(pop), p18_25, ppoverty, pumemployed, avg_income and log(tot_income) with confidence.

Plot G: Backward Stepwise Variable Selection



To have a better understanding of the model selection result, we then did comparisons of model selections based on a certain criterion group by different model selection. We plotted graphs for model selection results for adjusted R^2, Mallow's Cp and BIC under best subset, forward and backward selection.

Plot H: Comparisons of Different Variable Selection Tools and Criterions



Graphs showed above gave us a better intuition of choosing which model under different criterions and model selections. But the result could be similar and the decision of selecting the best model became harder. So, next cross validation was done for more comprehensive analysis to obtain Cross validation errors as the CV errors give best idea about the test MSE.

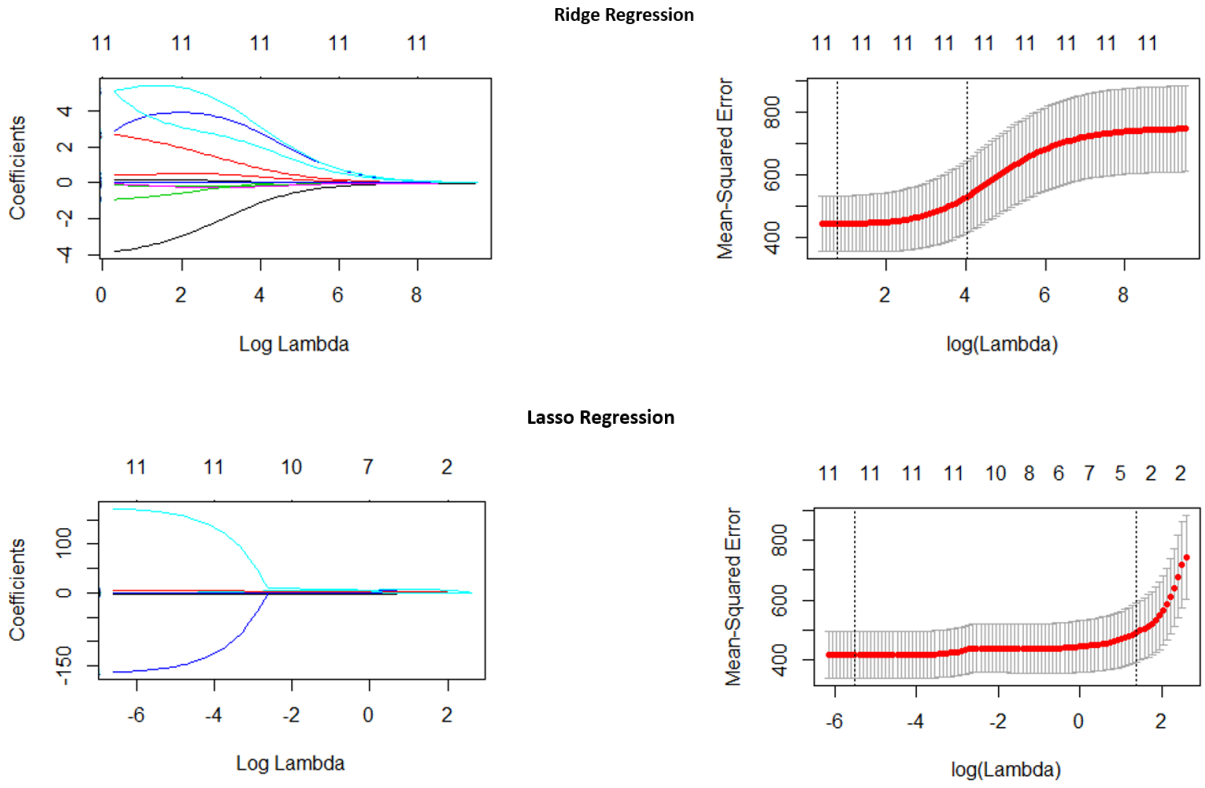
From analysis above, it was seen that the best subset selection and backward selection gave almost similar results. Forward selection and backward selection were easier for computation regarding

to large dataset. But the best subset selection is generally considered to be the finest as it looks into all the possible combination of the predictors and hence gives better results . However, forward and backward selections also have some drawbacks. One of the most remarkable one is the possibility of missing the "best" model, since we just add or drop one predictor at one time. So, best subset could be a better choice in general if number of predictors is not very large.

The Ridge and Lasso models were fit in attempt to develop a model that makes the RSS as small as possible and removing the insignificant variables from the model. They both penalize the number of predictors and lasso model also performs the variable selection.

Figures below shows plot log(lambda) against coefficients showing how they are being shrunk towards 0 for Ridge and lasso regression respectively. And cross validation mse's of Ridge and Lasso.

Plot I: Ridge and Lasso Regression



From the Lasso regression, it can be seen that most of the coefficients were shrunk to 0 and hence the remaining predictors obtained after shrinkage were: p18_25, log(pop), log(nhospbeds), pppoverty.

❑ Cross Validation

Based on variable selection analysis above, following 7 models were selected for cross validation analysis to compare their test MSE in order to find out the best final model.

Model 1. Lasso Model

Crime ~ log(area) + p18_25 + log(nhospbeds) + pcollege + ppoverty + punemployed + log(tot_income)

Model 2. Ridge Regression

Crime ~ log(area) + p18_25 + p65 + log(pop) + log(nhospbeds) + phs + pcollege + ppoverty + punemployed + avg_income + log(tot_income)

Model 3. Best Subset/Backward Selection - Adjusted R^2

Crime ~ log(area) + log(pop) + log(nhospbeds) + p18_25 + ppoverty + punemployed + avg_income + log(tot_income)

Model 4. Best Subset/Backward Selection - Mallow's Cp

Crime ~ log(area) + log(pop) + p18_25 + ppoverty + punemployed + avg_income + log(tot_income)

Model 5. Best Subset/Backward Selection - BIC

Crime ~ log(pop) + p18_25 + ppoverty + punemployed + avg_income + log(tot_income)

Model 6. Forward Selection - Adjusted R^2

Crime ~ log(area) + log(pop) + log(nhospbeds) + p18_25 + ppoverty + punemployed + avg_income + log(tot_income)

Model 7. Forward Selection - Mallow's Cp / BIC

Crime ~ log(pop) + log(nhospbeds) + p18_25 + ppoverty + punemployed + avg_income + log(tot_income)

Leave-One-Out Cross Validation (LOOCV)

Because the dataset is not very huge with a large number of observations and variables, it was practical to do the LOOCV test.

LOOCV is a method used to estimate the test error by splitting the data into two parts multiple times. One is the validation set with a single observation and another one is the training set with remaining observations. The final test MSE could be reached by taking the average of MSEs got from repeating above process.

For LOOCV, nearly the entire data set was taken into consideration with only separating one observation out as the test set. So, there is nearly no randomness in splitting. In order to ensure the integrity of cross validation of models, LOOCV was used to check all models obtained above.

Table2: MSE of different models by using LOOCV

LOOCV	
Lasso Model	355.9835
Ridge Regression	338.4752
Best Subset Model/ Backward - Adjusted R^2	333.843
Best Subset Model/Backward - Cp	333.9345
Best Subset Model/Backward - BIC	334.105
Forward Model - Adjusted R^2	333.843
Forward Model - Cp/BIC	333.5206

Model selected by Mallow's Cp and BIC criterions under forward selection gave the smallest MSE around 335.5.

Model Obtained:

Crime ~ log(pop) + log(nhospbeds) + p18_25 + ppoverty + punemployed + avg_income + log(tot_income)

K-Fold Cross Validation

However, LOOCV has issues of low bias and high variance. So, outputs using it could have high correlation since training set selected every time is almost the same as the full model.

To get rid off the high variance given by LOOCV test, K-Fold Cross Validation was selected to conduct further analysis on the model selection, which could give good bias-variance trade-off and eliminate correlation of outputs in processes.

Table 3: MSE of different models by using K-Fold Cross Validation

K- Fold Cross Validation	
Lasso Model	356.3485
Ridge Regression	334.2299
Best Subset Model/ Backward - Adjusted R^2	331.9782
Best Subset Model/Backward - Cp	332.5523
Best Subset Model/Backward - BIC	333.6209
Forward Model - Adjusted R^2	331.9782
Forward Model - Cp/BIC	332.2918

Based on K-Fold Cross Validation, the model selected based on adjusted R^2 under best subset selection has smallest MSE compared with others.Surprisingly, the model given by adjusted R^2under forward selection had the same test MSE since those two criterions gave exactly same models.

Model Obtained:

Crime ~ log(area) + log(pop) + log(nhospbeds) + p18_25 + ppoverty + punemployed + avg_income + log(tot_income)

❑ Final Model

Model selection ended up with two models (Model 3 and Model 7) with relatively small mean squared errors. A t-test was applied to each model, and the results indicated that log(pop), p18_25, ppoverty, punemployed, avg_income and log(tot_income) are statistically significant predictors for crime. Therefore, the final model was fitted with crime as response and log(pop), p18_25, ppoverty, punemployed, avg_income and log(tot_income) as predictors:

crime = 804.3 – 179.1log(pop) + 0.8392p18_25 + 4.146ppoverty + 1.421punemployed – 0.007869avg_income + 186.9log(tot_income)

The coefficients of log(pop) and avg_income are negative, which means that a percentage increase of population or an increase in per capita income will decrease the number of crime per 1000 people. If the other predictors remain constant, 1 percentage increase of population will decrease the number of crime per 1000 people by 1.791; if the other predictors remain constant, 1000 unit increase in per capita income will decrease the number of crime per 1000 people by 7.869.
The coefficients of p18_25, punemployment, ppoverty and log(tot_income) are positive, which means that an increase in the percentage between 18 and 25 years old people, an increase in the percentage of unemployed people, an increase in the percentage below poverty line or a percentage increase of total income will increase the number of crime per 1000 people. If the other predictors remain constant, 1 unit increase in the percentage between 18 and 25 years old people will increase the number of crime per 1000 people by 0.8392; if the other predictors remain constant, 1 unit increase in the percentage of unemployed people will increase the number of crime per 1000 people by 1.421; if the other predictors remain constant, 1 unit increase in the percentage below poverty line will increase the number of crime per 1000 people by 4.146; if the other predictors remain constant, 1 percentage increase of total income will increase the number of crime per 1000 people by 1.869.
Forecasting: The final model was used to predict the number of crime per 1000 people for another county out-of-sample. The predicted number of crime per 1000 people is 47.61243 and the prediction interval for the number of crime per 1000 people is (6.821819, 88.40304). The confidence interval for the number of crime per 1000 people is (45.02297, 50.20188).

❑ Conclusion

To fit a model that explains and predicts Crime in a county, several linear models have been assessed. These models include but not limited to: best subset selection, Lasso, ridge regression, Forward, Backward. It was found that the Best Subset Model and Forward stepwise model outperformed all other models and based on the t-test result of different predictors , the most significant predictors were obtained and a linear model was fitted with them :

crime~log(pop)+p18_25+ppoverty+punemployed+avg_income+log(tot_income)

Hence, based on the analysis above it was found that the crime is mainly affected by the factors like population, poverty unemployment, average income, total income of the people. Our suggestion is to decrease the crime rate as below.

Suggestion:

1. Improve the economy, create job opportunity and reduce the unemployment rate.
2. Provide better education to the people between the age group of 18 to 25 years.
3. Increase the minimum wage of the people in order to reduce the people below the poverty line.

References

Wikipedia, The Free Encyclopedia, s.v. "*Cross-Validation (statistics)*", accessed December 4, 2016, https://en.wikipedia.org/wiki/Cross-validation_(statistics)

Ralph Ellis, Ashley Frantz, Faith Karimi and Eliott C. McLaughlin, "*Orlando shooting: 49 killed, shooter pledged ISIS allegiance*," CNN (2016), accessed December 4, 2016, http://www.cnn.com/2016/06/12/us/orlando-nightclub-shooting/

Garrett Grolemund, "*Introduction to Data Science with R*", https://github.com/rstudio/Intro/blob/master/data/counties.csv, 2014

Juleyka Lantigua-Williams, "*Raise the Minimum Wage, Reduce Crime?"*, The Atlantic (2015), http://www.theatlantic.com/politics/archive/2016/05/raise-the-minimum-wage-reduce-crime/480912/