STAT 471/571/701 Modern Data Mining, HW 2

Group Member 1
Group Member 2
Group Member 3

Due: 9:00 AM, October 7, 2019

Contents

1 Overview

Multiple regression is one of the most popular methods used in statistics as well as in machine learning. We use linear models as a working model for its simplicity and interpretability. It is important that we use domain knowledge as much as we could to determine the form of the response as well as the function format for the factors. Then, when we have many possible features to be included in the working model it is inevitable that we need to choose a best possible model with a sensible criterion. `Cp`, `BIC` and regularizations are introduced. Be aware that if a model selection is done formally or informally, the inferences obtained with the final `lm()` fit may not be valid. Some asjustment will be needed. This last step is beyond the scope of this class. Check the current research line that Linda/Arun are working on.

1.1 Objectives

- Model building process
- Methods
  - Model selection
    - All subsets
    - Forward/Backward
  - Regularization
    - LASSO (L1 penalty)
    - Ridge (L2 penalty)
    - Elastic net
- Understand the criteia
  - `Cp`
  - Testing Errors
  - `BIC`
  - `K fold Cross Validation`
  - `LASSO`
- Packages
  - `lm()`, `Anova`
  - `regsubsets()`
  - `glmnet()` & `cv.glmnet()`

1.2 Instructions

- **Homework assignments can be done in a group consisting of up to three members**.
- **All work submitted should be completed in the R markdown format.** You can find a cheat sheet for R Markdown here. For those who have never used it before, we urge you to start this homework as soon as possible.
- **Submit the following files, one submission for each group:** (1) Rmd file, (2) a compiled PDF or HTML version, and (3) all necessary data files. You can directly edit this file to add your answers. If you intend to work on the problems separately within your group, compile your answers into one Rmd file before submitting. We encourage that you at least attempt each problem by yourself before working with your teammates. Additionally, ensure that you can 'knit' or compile your Rmd file. It is also likely that you need to configure Rstudio to properly convert files to PDF. **These instructions** should be helpful.
- In general, be as concise as possible while giving a fully complete answer. All necessary datasets are available in the "Data" folder or this homework folder on Canvas. Make sure to document your code with comments so the teaching fellows can follow along. R Markdown is particularly useful because it

follows a ‘stream of consciousness’ approach: as you write code in a code chunk, make sure to explain what you are doing outside of the chunk.

- A few good submissions will be used as sample solutions. When those are released, make sure to compare your answers and understand the solutions.

2 Review materials

- Study both R-tutorials
- Study lecture 3: Model selection
- Study lecture 4: Regularization
- Study lecture 2: Multiple regression

Review the code and concepts covered during lectures: multiple regression, model selection and penalized regression through elastic net.

3 Conceptual study

In this question, you will generate data (from a linear model) and perform variable selection using C_p, BIC, adjusted R^2 and lasso. You will also see that the summary from `lm()` can be misleading after model selection (as hinted in the overview). See ISLR, page 262, problem 8 for reference.

The following `r`-chunk generate data from a linear model with 10 features.

```r
n <- 100 ## sample size n
x <- rnorm(n)
eps <- rnorm(n)
xmat <- matrix(rep(x, 10), ncol = 10)
for(i in 1:10){
  xmat[,i] <- x^{i}
}
data <- data.frame(X = xmat, Y = eps)
```

`data` contains the response `Y` and feature matrix `X` (with 10 columns).

EDA: We start by exploring the structure of our data, confirming our intuitions from the code.

```
##      X.1   X.2     X.3     X.4      X.5      X.6       X.7      X.8
## 1 -0.881 0.776 -0.6830  0.6015 -0.52972  0.46651 -0.410834 3.62e-01
## 2 -0.344 0.118 -0.0406  0.0140 -0.00480  0.00165 -0.000567 1.95e-04
## 3  1.659 2.752  4.5647  7.5721 12.56089 20.83649 34.564379 5.73e+01
## 4 -0.388 0.150 -0.0583  0.0226 -0.00876  0.00340 -0.001317 5.10e-04
## 5  0.723 0.522  0.3774  0.2727  0.19706  0.14240  0.102904 7.44e-02
## 6  1.798 3.233  5.8119 10.4494 18.78721 33.77806 60.730533 1.09e+02
##         X.9     X.10       Y
## 1 -3.19e-01 2.81e-01 -0.8694
## 2 -6.71e-05 2.31e-05  0.4722
## 3  9.51e+01 1.58e+02 -0.0186
## 4 -1.98e-04 7.67e-05  0.0093
## 5  5.37e-02 3.88e-02 -2.2613
## 6  1.96e+02 3.53e+02  0.0176
```

(a) What is the TRUE model? (Write the true model relating Y to features in X).

By construction, Y is normally distributed and independent from the distribution of all X_k's. The true linear model is given by

$$y_i = 0 + \sum_{k=1}^{10} 0 \cdot x_{ki} + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0,1)$$

In essence, Y has no linear relationship with respect to any X_k and has mean of 0.

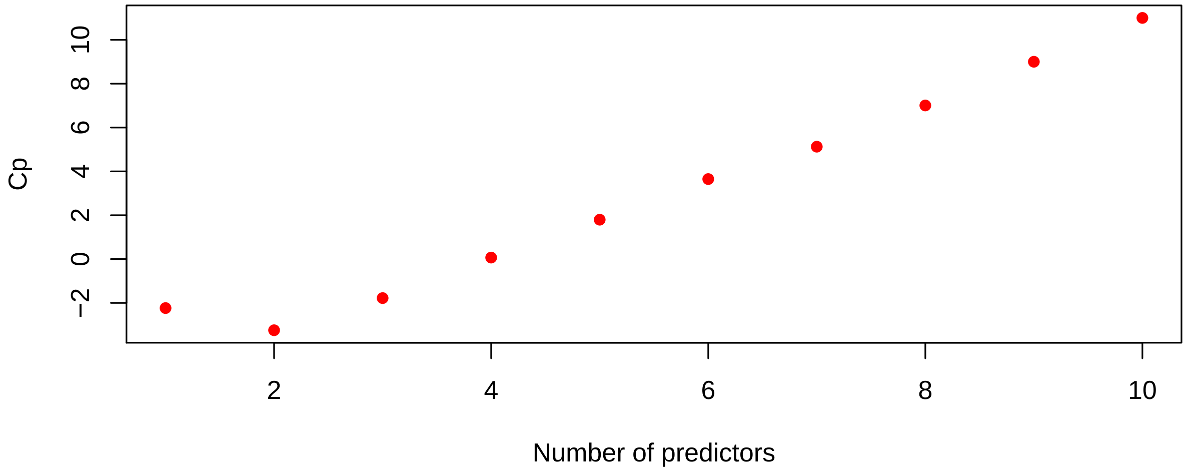
(b) Use the function `regsubsets()` to perform best subset selection in order to choose the best model. What is the best model obtained according to C_p, BIC and adjusted R^2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained.

We run `regsubsets` to obtain more information on how different models perform in terms of our criteria for model selection.

```
## Subset selection object
## Call: regsubsets.formula(Y ~ ., data, nvmax = 25, method = "exhaustive")
## 10 Variables  (and intercept)
##      Forced in Forced out
## X.1      FALSE      FALSE
## X.2      FALSE      FALSE
## X.3      FALSE      FALSE
## X.4      FALSE      FALSE
## X.5      FALSE      FALSE
## X.6      FALSE      FALSE
## X.7      FALSE      FALSE
## X.8      FALSE      FALSE
## X.9      FALSE      FALSE
## X.10     FALSE      FALSE
## 1 subsets of each size up to 10
## Selection Algorithm: exhaustive
##           X.1 X.2 X.3 X.4 X.5 X.6 X.7 X.8 X.9 X.10
## 1  ( 1 )  " " " " " " " " " " " " " " " " "*" " "
## 2  ( 1 )  "*" " " " " " " " " " " "*" " " " " " "
## 3  ( 1 )  "*" "*" " " " " " " " " " " " " "*" " "
## 4  ( 1 )  "*" "*" " " "*" " " " " "*" " " " " " "
## 5  ( 1 )  "*" "*" " " " " " " "*" " " "*" "*" " "
## 6  ( 1 )  "*" " " " " "*" " " "*" " " "*" "*" "*"
## 7  ( 1 )  " " " " "*" "*" "*" "*" "*" "*" "*" " "
## 8  ( 1 )  "*" " " "*" "*" "*" "*" "*" "*" "*" " "
## 9  ( 1 )  "*" "*" "*" "*" "*" "*" "*" "*" "*" " "
## 10  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
```

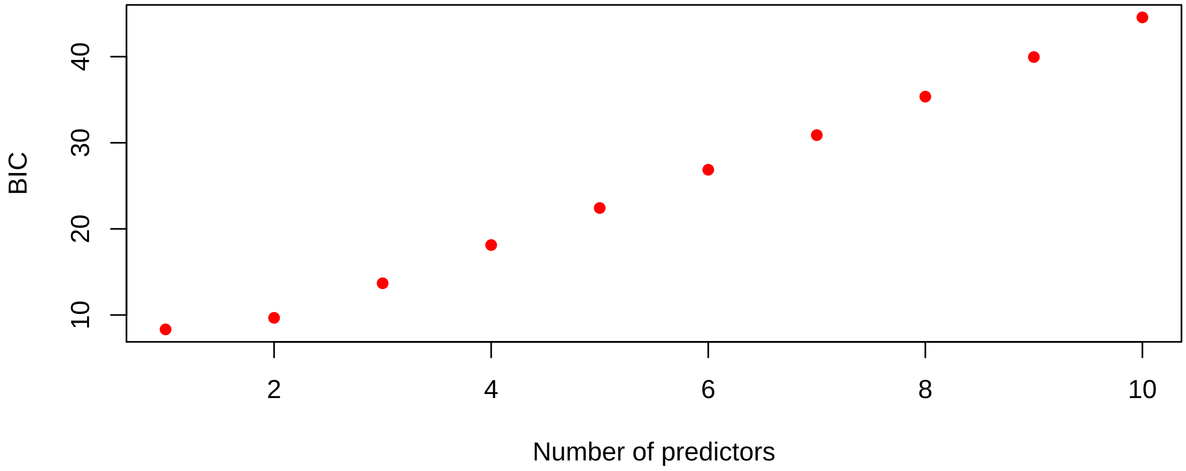
Now that we recorded `regsubsets`'s results, we proceed to plotting C_p, BIC, and adj. R^2 to select the best model. We are generally looking for C_p, BIC as small as possible, and although we will observe adj. R^2 and hope for a large value, we realize its flaws for model selection and will avoid biasing our judgement towards this metric.

C_p



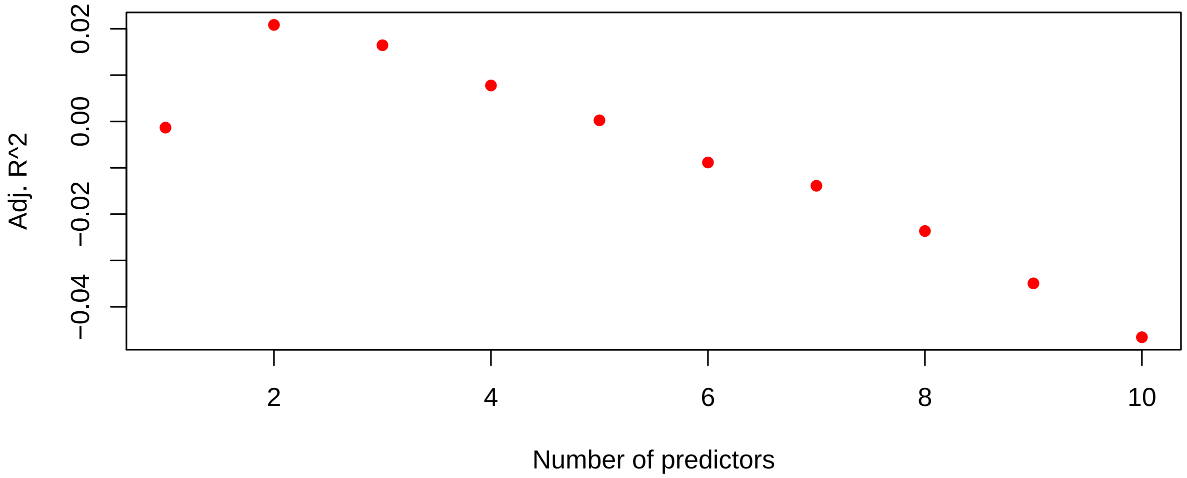
We note that C_p shows a moderate increasing pattern with respect to number of predictors. Although the minimum C_p usually corresponds to the smallest model, C_p shows some variability on the consistency of the upward trend. We now look for further confirmation from other Information Criteria.

BIC



Our plot of BIC exhibits a consistent increasing trend with respect to number of predictors. Again, the univariate model minimizes BIC, which reinforces our observations from before.

Adjusted R^2



We find that adj. R^2 tends to attain its maximum at a mid-range number, but we notice that the improvement in absolute terms is not very significant. We favor the two previous criteria to inform our decision.

We have that both C_p and BIC strongly point towards the smallest model. Thus, we conclude that a univariate model is our best choice. From the summary of `regsubsets`, we know the best univariate model constists of only X_1. We proceed to fitting such model and reporting the summary.

```
##
## Call:
## lm(formula = Y ~ X.1, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5527 -0.4985  0.0528  0.6026  2.0420
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.2161     0.0985    2.19    0.031 *
## X.1           0.1044     0.1128    0.92    0.357
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.985 on 98 degrees of freedom
## Multiple R-squared:  0.00865,    Adjusted R-squared:  -0.00146
## F-statistic: 0.856 on 1 and 98 DF,  p-value: 0.357
```

(c) Describe as accurate as possible what C_p and BIC are estimating?

Both C_p and BIC are estimators for the so-called Prediction of Testing Error (TE). That is, the expected MSE to be obtained from fitting the same model to out-of-sample observations. As this is a theoretical quantity, we rely on information criteria or a "testing dataset" (set of observations from same population but not included in model fitting) to estimate TE.

(d) Fit a lasso model to the simulated data. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates, and discuss the results obtained.

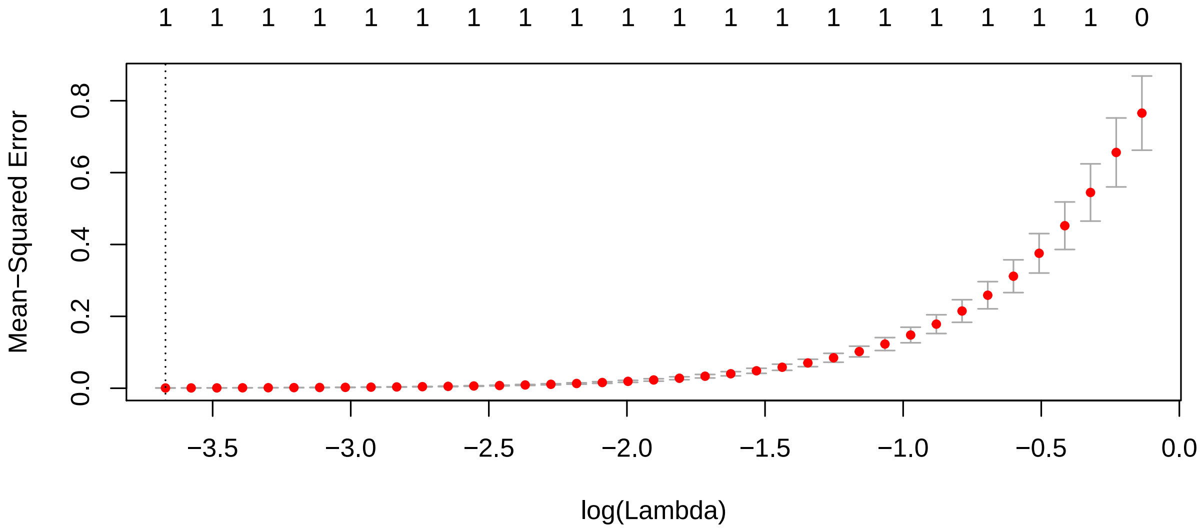
We start by preparing our data for LASSO. We then fit an initial LASSO regression with $\lambda = 100$ and report the coefficients:

```
Y <- data[,1]
X <- model.matrix(Y~., data)[,-1]

#LASSO
fit.lambda <- glmnet(X, Y, alpha=1, lambda = 100)
coef(fit.lambda)
```

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                 s0
## (Intercept) 0.0165
## X.1         0.0000
## X.2         .
## X.3         .
## X.4         .
## X.5         .
## X.6         .
## X.7         .
## X.8         .
## X.9         .
## X.10        .
```

As can be observed, a LASSO fit with penalty $\lambda = 100$ essentially returns all variable coefficients as 0. This is indeed consistent with the true model. Now, we will use K-fold cross-validation with $K = 10$ to obtain an estimate for the Testing Error-minimizing λ.



```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                    1
## (Intercept) 0.000481
## X.1         0.970849
## X.2         .
## X.3         .
## X.4         .
## X.5          .
## X.6          .
## X.7          .
## X.8          .
## X.9          .
## X.10         .
```

We note that teh CV algorithm picks an extremely small lambda if we's like to minimize Testing MSE, such that the penalty in the LASSO fit is almost null. Thus, the resulting model is similar to those obtained by selecting the best model of one variable in our previous model selections. Note though that the LASSO coefficients are biased, so we move on to fit an OLS regression with the selected LASSO variable and report the results.

```
## 
## Call:
## lm(formula = lm.input, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.5527 -0.4985  0.0528  0.6026  2.0420 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept)   0.2161     0.0985    2.19    0.031 *
## X.1           0.1044     0.1128    0.92    0.357  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.985 on 98 degrees of freedom
## Multiple R-squared:  0.00865,    Adjusted R-squared:  -0.00146 
## F-statistic: 0.856 on 1 and 98 DF,  p-value: 0.357
```

Note that although the coefficients are different in absolute terms, their sign is consistent. Thus, our results agree with the fact that LASSO tends to preserve the direction of the relation betwene variables.

(e) **Summary after selection:** Recall that if the null hypothesis is true, then the p-value is supposed to be less than 0.05 about 5% of the time. This means that if we repeat the experiment 100 times then in about 5 of these experiments we see a p-value less than 0.05. In this question, you will explore the validity of summary table after selection. For simplicity, we restrict to choosing a single feature model.

- Generate 100 datasets using the `r`-chunk above. You can do this by wrapping the above code into

```r
## Remove eval = F when working on homework.
nexperiment <- 100 ## number of experiments
pmat <- matrix(0, nrow = 100, ncol = 3)
## `pmat` is for saving output from questions below.
colnames(pmat) <- c("Cp", "BIC", "Adj_R2")
for(idx in 1:nexperiment){
  set.seed(630 + idx)

  #generating data
  n <- 100 ## sample size n
  x <- rnorm(n)
  eps <- rnorm(n)
  xmat <- matrix(rep(x, 10), ncol = 10)
  for(i in 1:10){
   xmat[,i] <- x^{i}
  }
  data <- data.frame(X = xmat, Y = eps)

  #selecting best univariate model
  fit.exh <- regsubsets(Y ~., data, nvmax = 1, method = "exhaustive")
  f.e <- summary(fit.exh)
  univar <- colnames(f.e$which)[f.e$which][2]

  #fitting model
  lm.input <- as.formula(paste("Y", "~", univar))
  fit.loop <- lm(lm.input, data)

  #extracting p-value
  p <- anova(fit.loop)$'Pr(>F)'[-2]
  pmat[idx,] <- rep(p, 3)
}
```

- use the function `regsubsets()` to find the best one variable model (using `nvmax = 1` argument) in each of the 100 datasets according to C_p, BIC and adjusted R^2.
- find and save the p-value for each of the three selected models in the matrix `pmat` defined as
- You must now have 100 different p-values for each selection procedure. Find the proportion of times the p-value turned out to be less than 0.05 for C_p, BIC and adjusted R^2 separately.

We have obtained our matrix of p-values. Now, we will compute the proportion of times such p-values where less than 0.05.

```
## [1] 0.15
```

- (**Bonus Question**) Comment on why the proportion is above or below the expected 5%.

Notice that when the number of variables is fixed, C_p, BIC, and adj. R^2 are all optimized when SSR is minimized, and so they will always agree on the best model to select. Differences among the three will only occur when comparing models of different sizes.

4 Case study 1: `Auto` data

This will be the last part of the Auto data from ISLR. The original data contains 408 observations about cars. It has some similarity as the Cars data that we use in our lectures. To get the data, first install the package `ISLR`. The data set `Auto` should be loaded automatically. We use this case to go through methods learned so far.

You can access the necessary data with the following code:

```
# check if you have ISLR package, if not, install it
if(!requireNamespace('ISLR')) install.packages('ISLR')
auto_data <- ISLR::Auto
```

Final modelling question: We want to explore the effects of each feature as best as possible.

You may explore the possibility of variable transformations. We normally do not suggest to transform x for the purpose of interpretation. You may consider to transform y to either correct the violation of the linear model assumptions or if you feel a transformation of y makes more sense from some theory. In this case we suggest you to look into `GPM=1/MPG`. Can you provide some background knowledge to support the notion: it makes more sense to mode `GPM`? You may also explore by adding interactions and higher order terms. The model(s) should be as *parsimonious* (simple) as possible, unless the gain in accuracy is significant from your point of view. Use Mallow's C_p or BIC to select the model.

- Describe the final model and its accuracy. Include diagnostic plots with particular focus on the model residuals.
- Summarize the effects found.
- Predict the mpg of a car that is: built in 1983, in the US, red, 180 inches long, 8 cylinders, 350 displacement, 260 as horsepower, and weighs 4,000 pounds. Give a 95% CI.
- Any suggestions as to how to improve the quality of the study?

5 Case Study 2: What can be done to reduce the crime rates?

5.1 Part I: EDA

Crime data continuation: We continue to use the crime data analyzed in the lectures. We first would like to visualize how crime rate (violentcrimes.perpop) distributes by states. The following r-chunk will read in the entire crime data into the r-path and it also creates a subset.

```r
crime.all <- read.csv("CrimeData.csv", stringsAsFactors = F, na.strings = c("?"))
crime <- dplyr::filter(crime.all, state %in% c("FL", "CA"))
```

Show a heat map displaying the mean violent crime by state. You may also show a couple of your favorite summary statistics by state through the heat maps. Write a brief summary based on your findings.

5.2 Part II: LASSO selection

Our goal for the rest of the study is to find the factors that are related to violent crime. We will only use communities from two states FL and CA to assure the maximum possible number of variables.

1. Prepare a set of sensible factors/variables that you may use to build a model. You may show the R-chunk to show this step. Explain what variables you may have excluded in the study and why? Or what other variables you have created to be included in the study.

Then use LASSO to choose a reasonable, small model. Fit an OLS model with the variables obtained. The final model should only include variables with p-values < 0.05. Note: you may choose to use "lambda 1st" or "lambda min" to answer the following questions where applicable.

2. What is the model reported by LASSO?

3. What is the model after running OLS? Comment on the difference between the equation from questions (2) and that from the OLS here.

4. What is your final model, after excluding high p-value variables?

a) What is your process of getting this final model?
b) Write a brief report based on your final model.

5.3 Part III: Elastic Net

Now, instead of LASSO, we want to consider how changing the value of α (i.e. mixing between LASSO and Ridge) will affect the model. Cross-validate between α and λ, instead of just λ. Note that the final model may have variables with p-values higher than 0.05; this is because we are optimizing for accuracy rather than parsimony.

1. What is your final elastic net model? What were the α and λ values? What is the prediction error?

2. Use the elastic net variables in an OLS model. What is the equation, and what is the prediction error?

3. Summarize your findings, with particular focus on the difference between the two equations.

5.4 Summary

Write a brief summary: 1) Summarize the crime situation in general in United States. 2) Based on the analyse done, can you make some suggestions to local officials/policy holders how to reduce the crime rates. 3) How to improve the study.