STA 207 Project:Determinant of Crime rate in America

Heqiao Ruan
hruan@ucdavis.edu
Rafee Musabbir
rafeemusabbir@ucdavis.edu

March 21, 2018

1 Introduction

In this project we plan to do extensive data analysis by stepwise regression, ridge regression, lasso and glmnet and cross validation and PLS to perform the analysis on the crime dataset which depicts the crime-related statistics for 47 US states in 1960 and the goal is to relate crime rate to the other socio-economic variables. Here we use various approach to conduct the analysis.

2 Introduction to dataset

The dataset is crime-related statistics for 47 US states in 1960 are given and there are 47 observations and 13 features.
Column 1:Crime rate, number of offenses known to the police per 1000000 population.
Column 2:Age, number of males aged 24-24 per 100 of total population.
Column 3:Ed, mean number of years of schooling times 10 of the population, 25 years or older.
Column 4: PE, police expenditures- per capita expenditure on police protection by state and local governments in 1960.
Column 5: PE-1, police expenditures- per capita expenditure on police protection by state and local governments in 1959.
Column 6: LF, labor force participation rate per 1000 civilian urban males in the age group 14-24.
Column 7:M, number of males per 1000 females.
Column 8: Pop, state population size in 100000.
Column 9: NW, number of nonwhites per 1000.
Column 10: UE1, unemployment rate of urban males per 1000 in the age group 14-24.
Column 11: UE2, unemployment rate of urban males per 1000 in the age group 35-49.
Column 12: Wealth, median value of transferable goods and assets or family income(units 10 dollars).
Column 13: IncIneq, income inequality- number of families per 1000 earning below one-half of the median income.

3 Preliminary Analysis

First the response is crime rate and for detailed description of the variables in the dataset, see **Appendix A**. First we want to see the summary statistic of the dataset including and shown in **Table 1**. What's more, we plot the histogram of every variable and try various transformation techniques:**square, square root, log transformation** and choose the one that most like normal distributed, what's more log transformation is preferred in the response variable **Crime**(Graph 2.V)
Then from the plot we can see that log transformation is preferred for **UE.2,PE.1 and NW** while squareroot transformation is preferred for **POP** and square transformation is preferred

for **UE.1**. What's more we perform standardization to each variables so that everyone of them is in the same scale.
Then we explore the inter-correlation among the variables**(Graph 3)**. We can observe correlation present in this dataset, **wealth and PE,PE.1,ED** are highly correlated,**PE and PE.1** are highly correlated,**UE1 and UE2** are highly correlated. What's more we can see that the variance inflation factor of **(Table 4(I))** PE and PE.1 and Wealth are larger than 10 which means somewhat serious multicollinearity.
So as we can see that the correlation between PE and PE.1 are almost 1, we would like to choose only one of them in our downstream analysis and use PE.1 in our analysis then after dropping this we can see that no severe multicollinearity appear.**(Table 4(II))**.
Then we draw the pairwise scatterplot**(Graph 16)** and perform the lm model on the original data and the diagnostic plot is shown in **(Graph 17)**.

4 Model Building

4.1 Stepwise regression

Here we start from the full model and use **StepAIC** function to perform the model selection and see whether the model perform well.
Then from **Graph 5(I)** we can see that there's no obvious nonlinear pattern in the residual plot and the q-q plot shows no obvious dispersion from normal distribution and the summary of the model selected by stepwise regression is shown as **Graph 6**.
The model we select here is $log(crime)^* = -1.81 * 10e-15 + 0.367 * Age^* + 0.496 * Ed^* + 0.866 * log(PE.1)^* + 0.204 * log(UE2)^* + 0.403 * Wealth^* + 0.947 * IncInequ^*$, here $*$ means standardized variables.

4.2 Ridge regression

Here we use a shrinkage method using L2 penalty to alleviate the multicollinearity. After using the GCV criterion and cross validation, we get the optimal parameter k=1.2548**(Graph 7)**.
Then here from **(Graph 8)** we can see that there's no obvious nonlinear pattern in residual plot and only slight light tail appear in the qqplot which means the model assumption is decent.
Then the coefficient and estimation table is shown in **(Graph 9)** and the variance inflation factor is shown in **(Graph 10,Graph 5(II))** we can see that it is smaller than that in the stepwise regression and we can see that the ridge regression somewhat alleviate the multicollinearity.
Our final model get by ridge regression is $log(Crime)^* = 0.327 * age^* + 0.441 * Ed^* + 0.690 * log(PE.1)^* + 0.879 * LF^* - 0.0157 * M - 0.0569 * sqrt(Pop)^* + 0.187 * log(NW)^* - 0.131 * UE1^{2,*} + 0.324 * log(UE2)^* + 0.258 * Wealth^* + 0.589 * IncInequ^*$

4.3 Lasso

Here we use another regularization technique called lasso to do the variable selection and shrinkage. Then we use 10 fold cross validation and the MSE versus log-lambda curve is shown in **Graph 11**.Then our fitted model here become: $log(Crime)^* = 0.244 * age^* + 0.33 * Ed^* + 0.769 * log(PE.1)^* + 0.06 * LF^* + 0.195 * log(NW)^* - 0.013 * UE1^{2,*} + 0.166 * log(UE2)^* + 0.363 * IncInequ^*$ see also **Graph 12**.
Then we perform model diagnostic on lasso and we have seen that there are somewhat nonlinear pattern in the residual plot and the qq plot shows a little bit dispersion from the normal distribution. So here lasso may not be the optimal choice.

4.4 Partial Least Square

Here at first we apply PLS(partial least square) to build the model with 8 components and check the F statistic. Then we calculate the adjusted coefficient of determination and coefficient of determination and CV and the F statistic. The table is shown as below:

k	1	2	3	4	5	6	7	8	9	10
SSE	25.22	16.79	14.17	11.80	10.99	10.13	9.76	9.64	9.62	9.61
R-square	0.452	0.635	0.692	0.743	0.761	0.780	0.788	0.790	0.7908	0.7909
F(k)	37.08	22.07	7.97	8.42	3.02	3.40	1.46	0.490	0.073	0.0147
qF	4.057	4.062	4.067	4.073	4.079	4.085	4.091	4.098	4.105	4.113
CV	0.8062	0.7261	0.6958	0.6773	0.6758	0.6548	0.6457	0.6502	0.6512	0.6
Varpro	0.290	0.499	0.658	0.733	0.862	0.919	0.937	0.955	0.976	0.987

Then we can see that based on the CV criterion, we should choose 7 components which can explain 93.7% of the total variance and the loadings of the first 7 components is shown as **(Graph 14)**. Then our model here become

4.5 Model Selection

Here we use the 10-fold cross validation method(7 of the 10 samples are 5 while the other 3 samples are 4) and compare ridge, lasso, partial least square, stepwise regression. Here the CV error is given by $CV^{(-1)} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i\in I_j}(Y_i - x_i^T\hat{\beta^{(j)}})^2$ and we want to choose the model with the smallest one which means the best fitting and predictive ability for the downstream analysis.
Here the CV error are shown as below:

Type	Stepwise Regression	Ridge	Lasso	PLS
MMSE	0.6289	0.3287	0.4210	0.6038

So here we find that the ridge regression is the best choice in terms of CV criterion.
Then our final model is the same as what we choose in ridge regression: $log(Crime)^* = 0.327 * age^* + 0.441 * Ed^* + 0.690 * log(PE.1)^* + 0.879 * LF^* - 0.0157 * M - 0.0569 * sqrt(Pop)^* + 0.187 * log(NW)^* - 0.131 * UE1^{2,*} + 0.324 * log(UE2)^* + 0.258 * Wealth^* + 0.589 * IncInequ^*$

5 Conclusion and Discussion

Here we have concluded that the ridge regression is the optimal one among the four types of fitted models with the best predictive ability. Here after comparing the sign of the coefficients in several models, we can see that they are consistent among these models. Then we can see that the crime rate is positively related to age, Ed, PE.1, LF, NW, UE2, Wealth, IncInequ and is negatively related to Pop, M and UE1.
Here we interpret this relationship. First the crime rate is positively related to the ratio of young men in the whole population and it matches our common sense. Then we can see that the crime rate is positively related to the police expenditure on police protection but it may not be a causal however, as we all know, higher crime rate will lead to higher police expenditure for police protection. Then we can see that the relation between crime rate and M,LF,Pop are not that significant(same scale, compare the coefficient), however we can see that the crime rate is negatively related to the population size and the number of men in the whole population and are positively related to the labor force participation rate per 1000 civilian urban males in the age group 14-24. Then we can see that crime rate is positively related to the number of nonwhites in the population which may somewhat match our common sense. What's more, the crime rate is positively related to the unemployment rate of urban males older than 35 which matches our common sense because men at that age are supposed to work. However, the crime rate is negatively related to the unemployment rate of urban males aged between 14 and 24 which is also explainable because young men at that age are supposed to be in the school and if they work when 15 or 16 years old, probably they have not received a decent education which potentially lead a higher crime rate.
What's more, the crime rate is positively related to the median value of transferable goods and assets or family income which is pretty interesting because we have found that even wealthy men tend to crime and we can see that the desire of some people can never be fulfilled.
Then we can see that the crime rate is positively related to the degree of income inequality which is pretty insightful because people are more worried about others being wealthier than themselves than worrying about their own poverty. So as government, we should provide help to those who fail to get a decent education.
So in conclusion, we can see that the crime rate are mainly related to the men's education level and the degree of inequality of economical status in this city as well as the unemployment rate of the men who are supposed to work to support their family in that age.
For the limitation we can see that even after transformation, there are still very slight dispersion from the normal assumption(either light or heavier tail). Another issue is that Lasso doesn't seem to be a particular optimal model here for the somewhat severe dispersion from the fundamental assumption. What's more, the CV criterion may be also limited in terms of comparing these kinds of models and other more delicate techniques are yet to be developed. The most important is that the number of observations in the dataset is not enough for us to conduct a very very compelling result and I think repeated measurements in the 47 states in US along the last a couple of years are required. For future work, maybe more delicate model selection techniques and longitudinal data analysis are required.

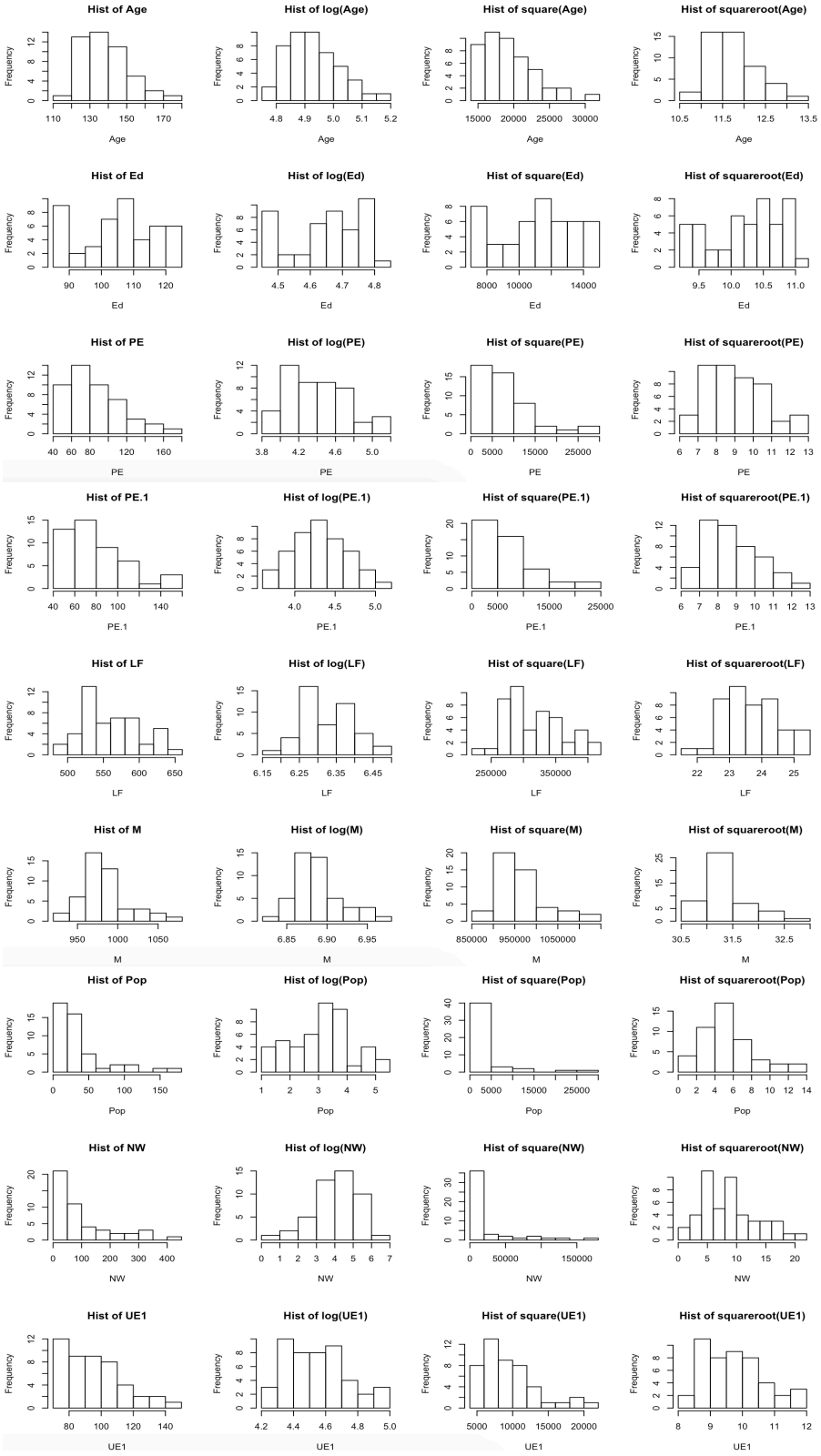
6 Appendix A:Tables and Graphs

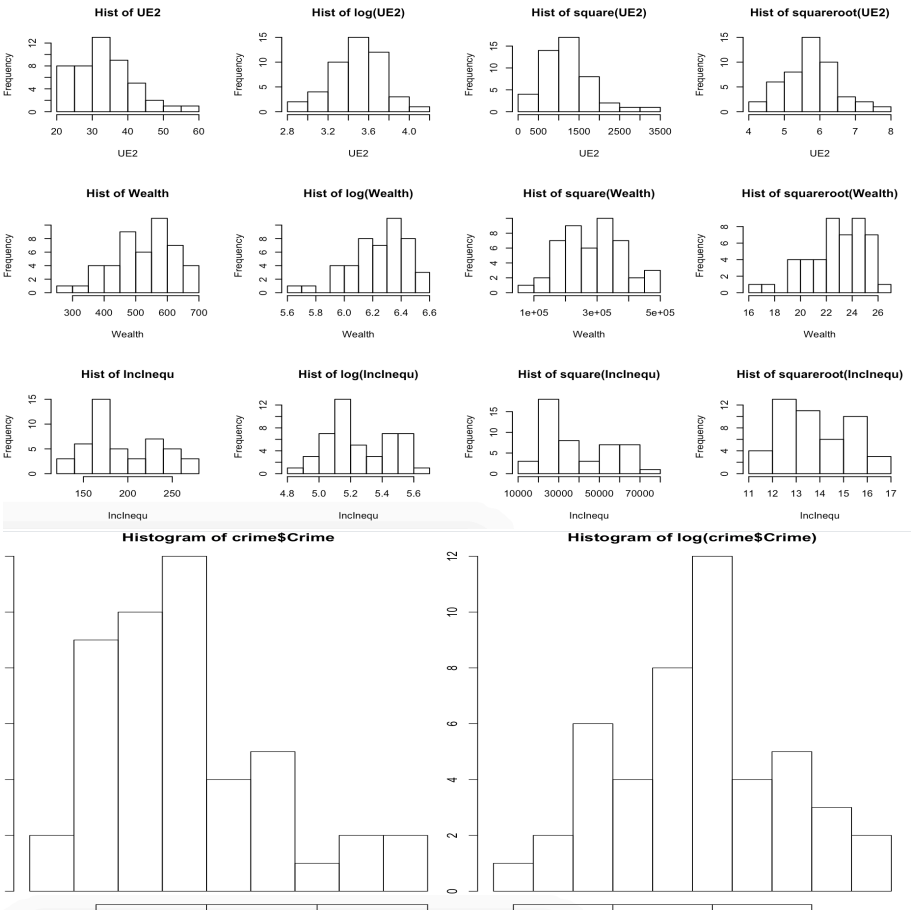
Table 1: Summary statistics:

	Crime	Age	Ed	PE	PE.1	LF	M	Pop
Min.	34.20000	119.0000	87.0000	45.0	41.00000	480.0000	934.0000	3.00000
1st Qu.	65.85000	130.0000	97.5000	62.5	58.50000	530.5000	964.5000	10.00000
Median	83.10000	136.0000	108.0000	78.0	73.00000	560.0000	977.0000	25.00000
Mean	90.50851	138.5745	105.6383	85.0	80.23404	561.1915	983.0213	36.61702
3rd Qu.	105.75000	146.0000	114.5000	104.5	97.00000	593.0000	992.0000	41.50000
Max.	199.30000	177.0000	122.0000	166.0	157.00000	641.0000	1071.0000	168.00000

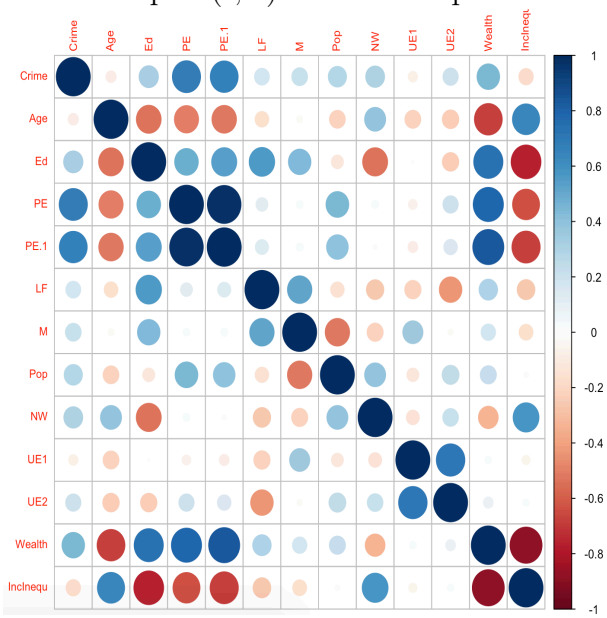
	NW	UE1	UE2	Wealth	IncInequ
Min.	2.0000	70.00000	20.00000	288.000	126.0
1st Qu.	24.0000	80.50000	27.50000	459.500	165.5
Median	76.0000	92.00000	34.00000	537.000	176.0
Mean	101.1277	95.46809	33.97872	525.383	194.0
3rd Qu.	132.5000	104.00000	38.50000	591.500	227.5
Max.	423.0000	142.00000	58.00000	689.000	276.0

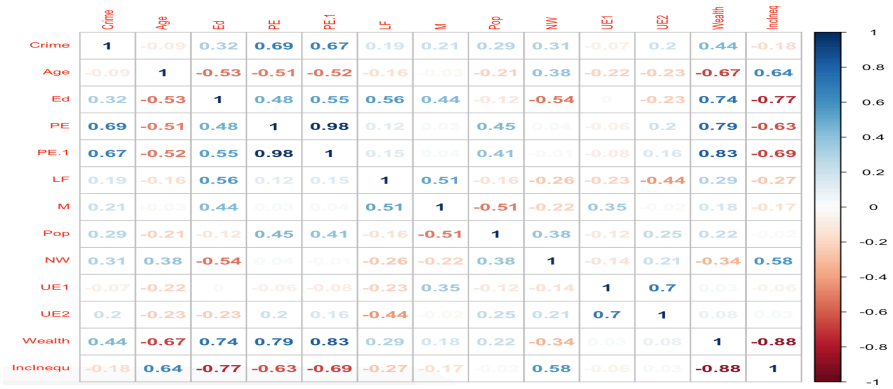
Graph 2(I,II,III,IV,V):Histograms of variables:





Graph 3(I,II):Correlation plot:





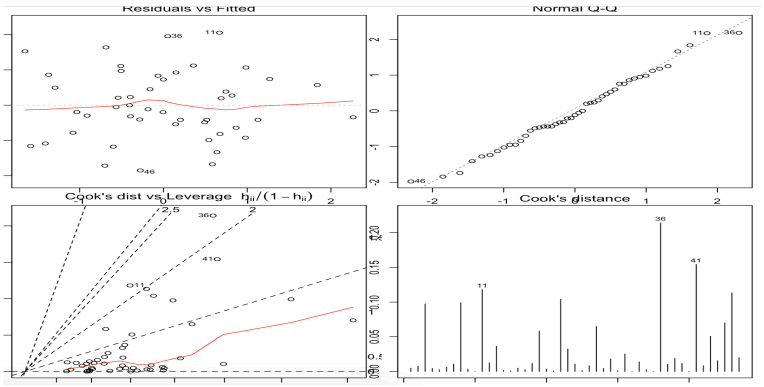
	Crime	Age	Ed	PE	PE.1	LF	M	Pop	NW	UE1	UE2	Wealth	IncIneq
Crime	1	-0.09	0.32	0.69	0.67	0.19	0.21	0.29	0.31	-0.07	0.2	0.44	-0.18
Age	-0.09	1	-0.53	-0.51	-0.52	-0.16	-0.05	-0.21	0.38	-0.22	-0.23	-0.67	0.64
Ed	0.32	-0.53	1	0.48	0.55	0.56	0.44	-0.12	-0.54	0	-0.23	0.74	-0.77
PE	0.69	-0.51	0.48	1	0.98	0.12	0.03	0.45	0.04	-0.06	0.2	0.79	-0.63
PE.1	0.67	-0.52	0.55	0.98	1	0.15	0.04	0.41	-0.01	-0.08	0.16	0.83	-0.69
LF	0.19	-0.16	0.56	0.12	0.15	1	0.51	-0.16	-0.26	-0.23	-0.44	0.29	-0.27
M	0.21	-0.05	0.44	0.03	0.04	0.51	1	-0.51	-0.22	0.35	-0.03	0.18	-0.17
Pop	0.29	-0.21	-0.12	0.45	0.41	-0.16	-0.51	1	0.38	-0.12	0.25	0.22	-0.05
NW	0.31	0.38	-0.54	0.04	-0.01	-0.26	-0.22	0.38	1	-0.14	0.21	-0.34	0.58
UE1	-0.07	-0.22	0	-0.06	-0.08	-0.23	0.35	-0.12	-0.14	1	0.7	0.03	-0.06
UE2	0.2	-0.23	-0.23	0.2	0.16	-0.44	-0.03	0.25	0.21	0.7	1	0.08	0.03
Wealth	0.44	-0.67	0.74	0.79	0.83	0.29	0.18	0.22	-0.34	0.03	0.08	1	-0.88
IncInequ	-0.18	0.64	-0.77	-0.63	-0.69	-0.27	-0.17	-0.05	0.58	-0.06	0.03	-0.88	1

Table 4(I,II): Eigen values of the inverse of the design matrix:

	Variables	Tolerance	VIF
	<chr>	<dbl>	<dbl>
1	Age	0.410	2.44
2	Ed	0.203	4.93
3	PE	0.0405	24.7
4	PE.1	0.0330	30.3
5	LF	0.384	2.60
6	M	0.259	3.87
7	Pop	0.359	2.79
8	NW	0.322	3.10
9	UE1	0.236	4.24
10	UE2	0.244	4.10
11	Wealth	0.0968	10.3
12	IncInequ	0.106	9.40

	Variables	Tolerance	VIF
	<chr>	<dbl>	<dbl>
1	Age	0.415	2.41
2	Ed	0.206	4.85
3	PE.1	0.173	5.77
4	LF	0.385	2.60
5	M	0.272	3.68
6	Pop	0.380	2.63
7	NW	0.323	3.10
8	UE1	0.238	4.20
9	UE2	0.246	4.06
10	Wealth	0.0977	10.2
11	IncInequ	0.106	9.40

Graph 5(I): Diagnostics for Stepwise Regression:



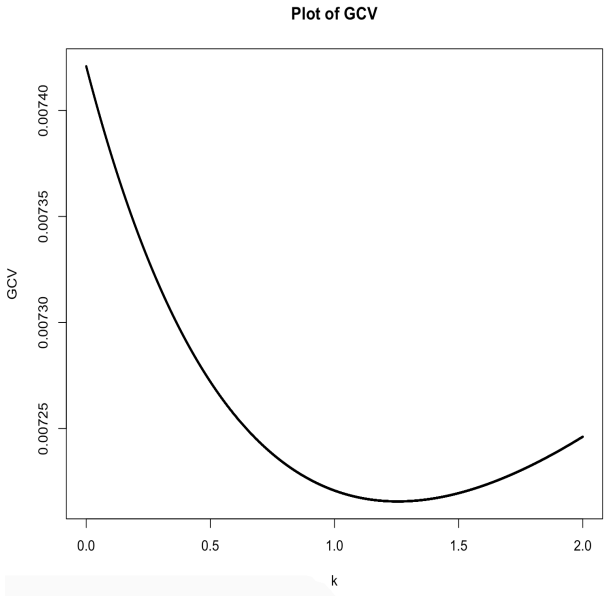
Graph 5(II): Variance Inflation Factor for stepwise regression.

	Variables	Tolerance	VIF
	<chr>	<dbl>	<dbl>
1	Age	0.481	2.08
2	Ed	0.326	3.07
3	PE.1	0.279	3.59
4	UE2	0.713	1.40
5	Wealth	0.107	9.32
6	IncInequ	0.182	5.51

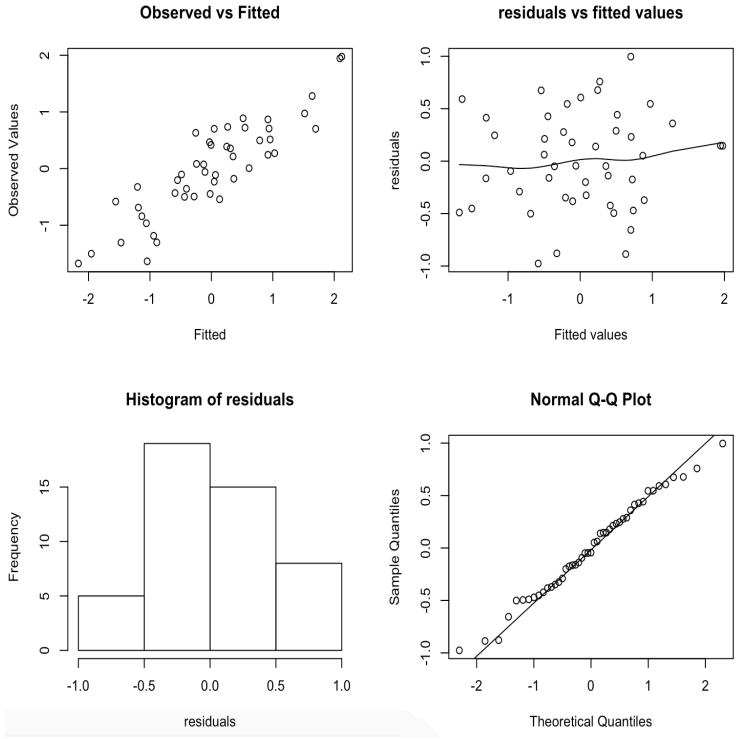
Graph 6: Summary of the selected model of stepwise regression:

```
Call:
lm(formula = Crime ~ Age + Ed + PE.1 + UE2 + Wealth + IncInequ,
    data = crime_std)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92650 -0.29577 -0.05818  0.36837  1.02899

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.810e-15  7.496e-02   0.000 1.000000
Age          3.671e-01  1.093e-01   3.359 0.001729 **
Ed           4.963e-01  1.327e-01   3.741 0.000574 ***
PE.1         8.663e-01  1.435e-01   6.035 4.21e-07 ***
UE2          2.043e-01  8.972e-02   2.277 0.028184 *
Wealth       4.030e-01  2.313e-01   1.742 0.089175 .
IncInequ     9.471e-01  1.778e-01   5.326 4.17e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5139 on 40 degrees of freedom
Multiple R-squared:  0.7704,    Adjusted R-squared:  0.7359
F-statistic: 22.37 on 6 and 40 DF,  p-value: 2.344e-11
```

Graph 7: Ridge coefficient versus GCV:



Graph 8: Ridge regression diagnostics:



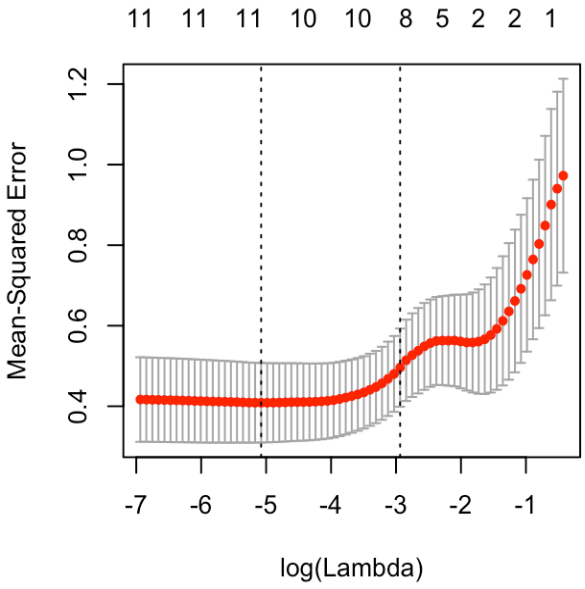
Graph 9: Ridge regression coefficient table and estimation:

	estimation	std_error
	-1.535339e-15	0.07548353
Age	3.270995e-01	0.11009174
Ed	4.412504e-01	0.14712647
PE.1	6.900085e-01	0.15057723
LF	8.788386e-02	0.11231313
M	-1.579924e-02	0.12989121
Pop	-5.696460e-02	0.11567178
NW	1.870250e-01	0.10813399
UE1	-1.311716e-01	0.11944912
UE2	3.236700e-01	0.12123126
Wealth	2.580289e-01	0.18064839
IncInequ	5.898477e-01	0.17399318

Graph 10: Variance inflation factor for ridge regression:

Age	1.9750571
Ed	3.5273766
PE.1	3.6947818
LF	2.0555652
M	2.7493482
Pop	2.1803440
NW	1.9054371
UE1	2.3250704
UE2	2.3949665
Wealth	5.3178780
IncInequ	4.9332675

Graph 11: LASSO MSE vs log(lambda):



Graph 12:Coefficient of Lasso:

```
12 x 1 sparse Matrix of class "dgCMatrix"
                    1
(Intercept)  .
Age          0.24445613
Ed           0.33056884
PE.1         0.76866809
LF           0.06352784
M            .
Pop          .
NW           0.19525084
UE1         -0.01349814
UE2          0.16612782
Wealth       .
IncInequ     0.36262584
```

Graph 13: Model diagnostic for Lasso:

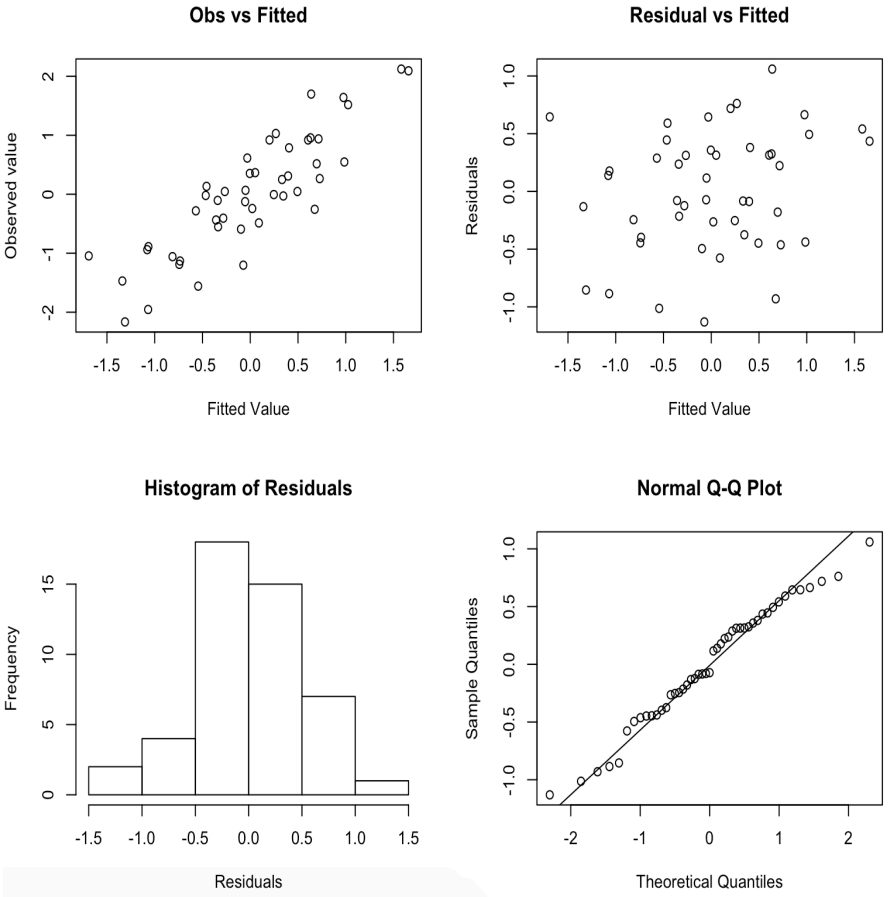
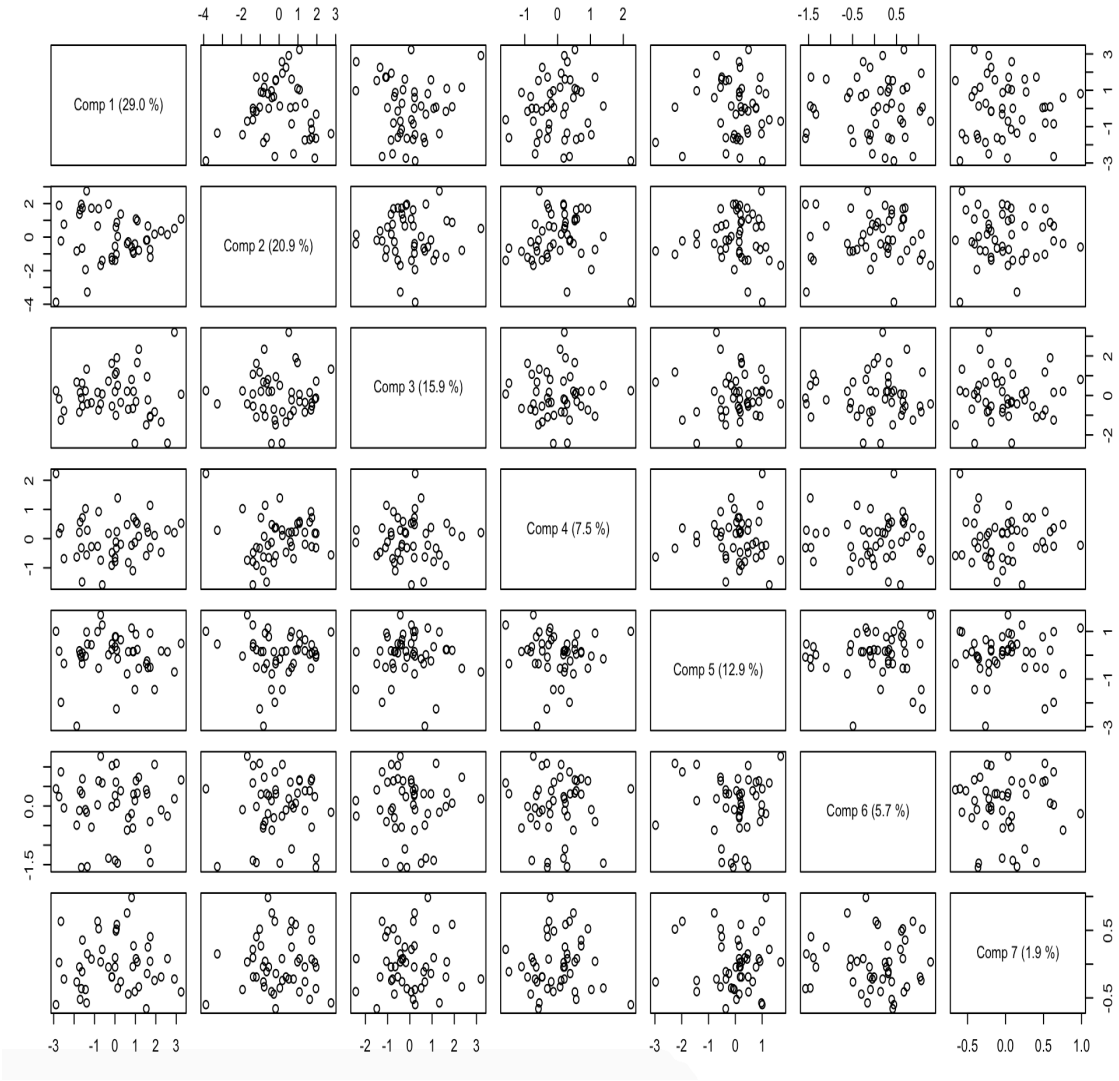


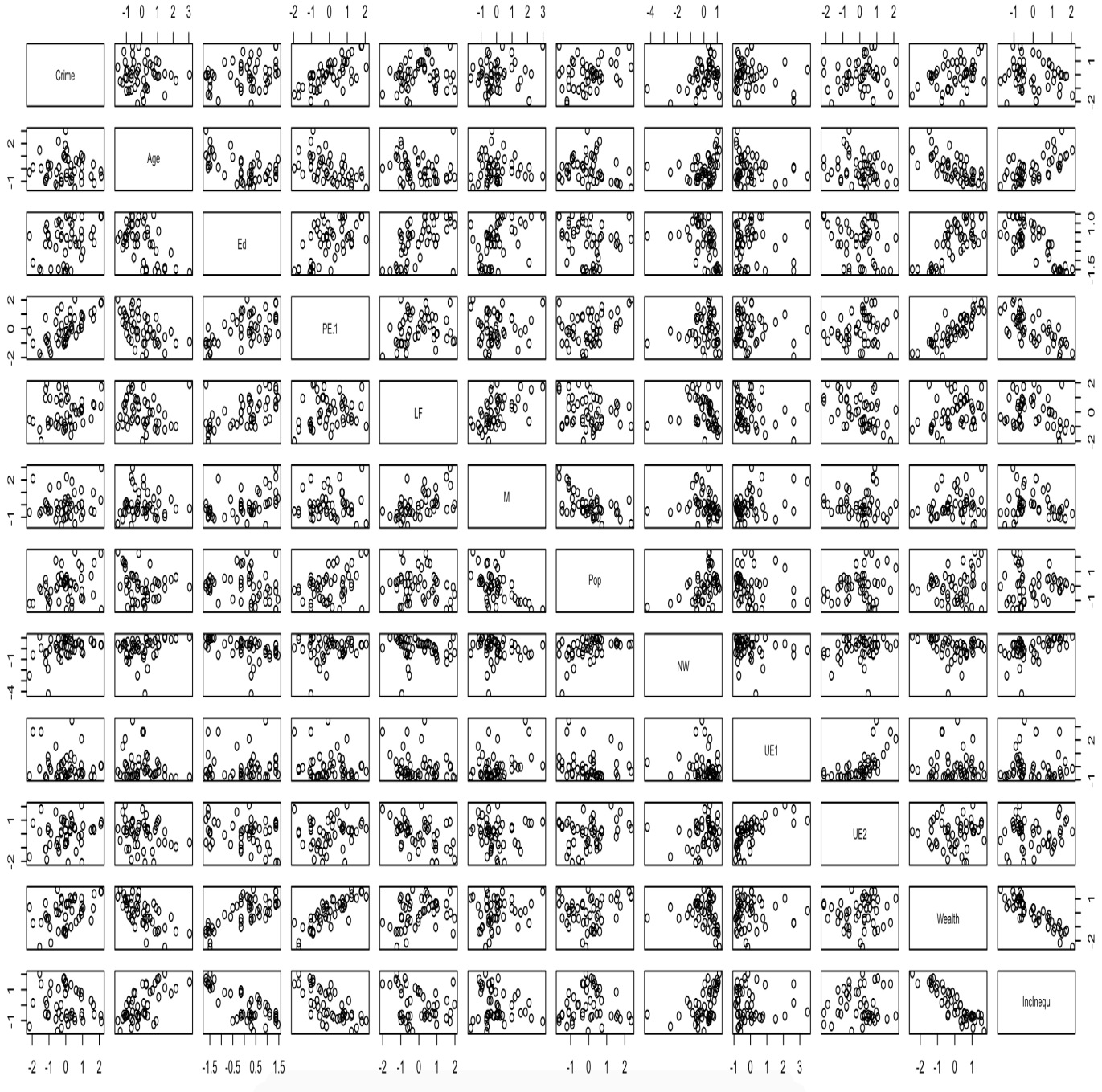
Table 14: Loadings of PLS under chosen number of components:

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7
Age	-0.36879709	0.40520860	0.305120816	0.15524608	0.19317997	-0.381217943	-0.2936111
Ed	0.40614824	-0.40555865	0.337740944	-0.06521700	0.20545524	0.139018412	0.2166889
PE.1	0.60293707	-0.02370157	-0.019382995	0.31487510	0.04621064	-0.141454678	0.1632475
LF	0.21163552	-0.15077678	0.430395779	-0.82163779	0.35585916	0.284600344	-0.5211128
M	0.09595830	-0.09887740	0.738143952	-0.33608706	-0.39566775	0.189155020	-0.1981444
Pop	0.25729486	0.29269976	-0.630283648	-0.06977315	0.15984282	0.148903228	-0.5236562
NW	0.11657059	0.64891826	-0.106815597	-0.31951963	-0.16719050	-0.001293357	0.4647805
UE1	-0.05583545	-0.12504848	0.036509759	0.17978720	-0.90466834	0.558780534	0.1432412
UE2	0.09297922	0.14162046	-0.191728532	0.66820480	-0.68402654	0.567010317	-0.5754360
Wealth	0.56362584	-0.28385225	0.008879955	0.08577938	-0.08192454	-0.164686517	0.1188680
IncInequ	-0.43704932	0.47097883	-0.033415129	-0.12536577	0.09102583	0.359246702	0.2067510

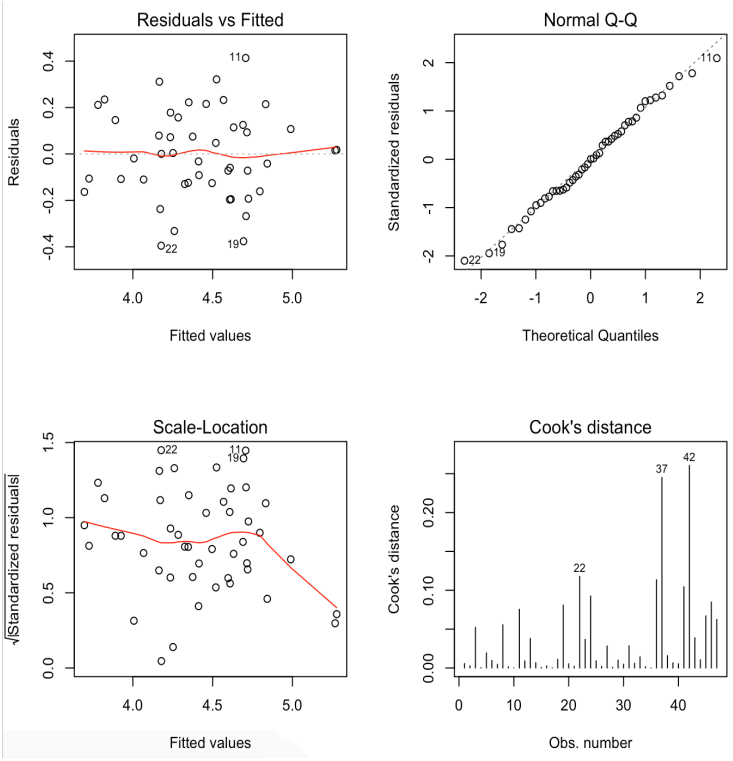
Graph 15: Scores plot for PLS:



Graph 16: Pairwise scatterplot:



Graph 17:Initial lm model diagnostics:



6.1 Appendix B: R code

```
library(MASS)
library(glmnet)
library(pls)
library(DAAG)
library(e1071)
crime<-read.csv("crime.csv",header=TRUE)
n<-nrow(crime)
#summary statistics:
apply(crime,2,summary)
#Preprocess:
fit0<-lm(Crime~.,data=crime)
par(mfrow=c(2,2))
plot(fit0,which=1)
plot(fit0,which=2)
plot(fit0,which=3)
plot(fit0,which=4)
#Pairwise scatterplot:
plot(crime)
#Histogram and possible transformation:
par(mfrow=c(3,4))
```

```
crime<-data.frame(crime)
sapply(colnames(crime[,2:4]),function(x){
  hist(crime[[x]],main=sprintf("Hist of %s",x),xlab=x)
  hist(log(crime[[x]]),main=sprintf("Hist of log(%s)",x),xlab=sprintf("%s",x))
  hist(crime[[x]]^2,main=sprintf("Hist of square(%s)",x),xlab=sprintf("%s",x))
  hist(sqrt(crime[[x]]),main=sprintf("Hist of squareroot(%s)",x),xlab=sprintf("%s",x))
})
par(mfrow=c(3,4))
sapply(colnames(crime[,5:7]),function(x){
  hist(crime[[x]],main=sprintf("Hist of %s",x),xlab=x)
  hist(log(crime[[x]]),main=sprintf("Hist of log(%s)",x),xlab=sprintf("%s",x))
  hist(crime[[x]]^2,main=sprintf("Hist of square(%s)",x),xlab=sprintf("%s",x))
  hist(sqrt(crime[[x]]),main=sprintf("Hist of squareroot(%s)",x),xlab=sprintf("%s",x))
})
par(mfrow=c(3,4))
sapply(colnames(crime[,8:10]),function(x){
  hist(crime[[x]],main=sprintf("Hist of %s",x),xlab=x)
  hist(log(crime[[x]]),main=sprintf("Hist of log(%s)",x),xlab=sprintf("%s",x))
  hist(crime[[x]]^2,main=sprintf("Hist of square(%s)",x),xlab=sprintf("%s",x))
  hist(sqrt(crime[[x]]),main=sprintf("Hist of squareroot(%s)",x),xlab=sprintf("%s",x))
})
par(mfrow=c(3,4))
sapply(colnames(crime[,11:13]),function(x){
  hist(crime[[x]],main=sprintf("Hist of %s",x),xlab=x)
  hist(log(crime[[x]]),main=sprintf("Hist of log(%s)",x),xlab=sprintf("%s",x))
  hist(crime[[x]]^2,main=sprintf("Hist of square(%s)",x),xlab=sprintf("%s",x))
  hist(sqrt(crime[[x]]),main=sprintf("Hist of squareroot(%s)",x),xlab=sprintf("%s",x))
})
par(mfrow=c(1,2))
hist(crime$Crime)
hist(log(crime$Crime))
#Then we do transformations: log(UE2),log(PE.1),squareroot(Pop),log(NW),square(UE1)
crime$UE2<-log(crime$UE2)
crime$PE.1<-log(crime$PE.1)
crime$Pop<-sqrt(crime$Pop)
crime$NW<-log(crime$NW)
crime$UE1<-(crime$UE1)^2
crime$Crime<-log(crime$Crime)
crime_std<-as.data.frame(scale(crime))
#pairwise scatterplot:
pairs(crime_std)
#Draw the correlation plot after transformation:
```

```
library(corrplot)
par(mfrow=c(1,1),mar=c(1,1,1,1))
corrplot(cor(crime_std),method="circle",pch=1,tl.cex=0.75)
corrplot(cor(crime_std),method="number",pch=1,tl.cex=0.75)
#Variance Inflation factor:
fit0<-lm(Crime~.,data=crime_std)
ols_vif_tol(fit0)
crime_std<-as.data.frame(as.matrix(crime_std[,-4]))
fit0<-lm(Crime~.,data=crime_std)
ols_vif_tol(fit0)
#Stepwise regression:
fit1<-lm(Crime~.,data=crime_std)
fit1_AIC<-stepAIC(fit1,trace=FALSE)
par(mfrow=c(2,2))
plot(fit1_AIC,which=1)
plot(fit1_AIC,which=2)
plot(fit1_AIC,which=6)
plot(fit1_AIC,which=4)
ols_vif_tol(fit1_AIC)
#Ridge regression:
X<-cbind(rep(1,n),as.matrix(crime_std[,-1]))
ridge1<-lm.ridge(Crime~.,data=crime_std,lambda=seq(0,2,0.0001))
plot(seq(0,2,0.0001),ridge1$GCV,main="Plot of GCV",xlab="k",ylab="GCV",cex=0.2)
k_select<-1.2548
fit_ridge<-lm.ridge(Crime~.,data=crime_std,lambda=k_select)
beta<-matrix(coef(fit_ridge))
fv<-X%*%beta
res<-crime_std$Crime-fv
par(mfrow=c(2,2))
plot(crime_std$Crime,fv,main="Observed vs Fitted",xlab="Fitted",ylab="Observed Values"
plot(fv,res,main="residuals vs fitted values",xlab="Fitted values",ylab="residuals")
lines(smooth.spline(fv,res,spar=1.06))
hist(res,main="Histogram of residuals",xlab="residuals")
qqnorm(res)
qqline(res)
#Coefficients estimation for ridge regression:
Y<-as.matrix(crime_std$Crime)
D<-t(X)%*%X
H<-X%*%solve(D+k_select*diag(ncol(X)))%*%t(X)
beta_e<-solve(D+k_select*diag(ncol(X)))%*%t(X)%*%Y
residual<-Y-X%*%beta_e
dematrix<-diag(ncol(H))-H
```

```
sigma2<-sum(residual^2)/sum(diag(dematrix%*%dematrix))
cov_beta<-sigma2*solve(D+k_select*diag(ncol(X)))%*%D%*%solve(D+k_select*diag(ncol(X)))
std_error<-sqrt(diag(cov_beta))
estimation<-beta_e
data.frame(estimation,std_error)
#VIF values:
vif_cov<-solve(D+k_select*diag(ncol(X)))%*%D%*%solve(D+k_select*diag(ncol(X)))
vif_ridge<-diag(D)*diag(vif_cov)
as.matrix(vif_ridge)
coef_ridge<-as.matrix(coef(fit_ridge))
sort(coef_ridge,decreasing=TRUE)
#Lasso:
fit_lasso<-cv.glmnet(X,Y,intercept=FALSE)
plot(fit_lasso)
fit_lasso$lambda.min
coef(fit_lasso)
#lasso model diagnostic:
Ylasso<-predict(fit_lasso,newx=X)
par(mfrow=c(2,2))
plot(Ylasso,Y,xlab="Fitted Value",ylab="Observed value",main="Obs vs Fitted")
plot(Ylasso,Y-Ylasso,xlab="Fitted Value",ylab="Residuals",main="Residual vs Fitted")
lines(smooth.spline(Ylasso,Y-Ylasso,spar=1.3))
hist(Y-Ylasso,main="Histogram of Residuals",xlab="Residuals")
qqnorm(Y-Ylasso)
qqline(Y-Ylasso)
#Partial Least Square:
set.seeds(1999)
fit_pls<-plsr(Crime~.,10,data=crime_std,validation="CV")
summary(fit_pls)
sse<-rep(0,10)
for(i in 1:10){
  sse[i]<-sum((residuals(fit_pls)[(47*i-46):(47*i)])^2)
}
ssto<-sum((crime_std$Crime)^2)
R_square<-rep(0,10)
for(i in 1:10){
  R_square[i]<-1-sse[i]/ssto
}
k<-c(1:10)
data.frame(k,sse,R_square)
F_k<-rep(0,10)
F_k[1]<-(nrow(crime_std)-1-1)*R_square[1]/(1-R_square[1])
```

```
for(i in 2:10){
  F_k[i]<-(nrow(crime_std)-i-1)*(R_square[i]-R_square[i-1])/(1-R_square[i])
}
qF<-rep(0,10)
for(i in 1:10){
  qF[i]<-qf(0.95,1,nrow(crime_std)-i-1)
}
data.frame(k,sse,R_square,F_k,qF)
plot(fit_pls,plottype="scores",comp=1:7)
loadings(fit_pls)[,1:7]
fit_pls_final<-plsr(Crime~.,7,data=crime_std,validation="CV")

#Model selection by leave one cross validation error:
cv<-function(traindata,testdata,method){
  if(method=="stepregr"){
    fit<-lm(Crime~.,data=traindata)
    stepfit<-stepAIC(fit,trace=FALSE)
    res<-testdata$Crime-predict(stepfit,testdata)
    return(sum(res^2)/nrow(testdata))
  }
  else if(method=="ridge"){
    rid<-lm.ridge(Crime~.,data=traindata,lambda=seq(0,30,0.001))
    k_select<-0.001*which.min(rid$GCV)
    fit<-lm.ridge(Crime~.,data=traindata,lambda=k_select)
    beta<-matrix(coef(fit))
    X<-cbind(rep(1,nrow(testdata)),testdata[,-1])
    fv<-X%*%beta
    res<-testdata$Crime-fv
    return(sum(res^2)/nrow(testdata))
  }
  else if(method=="lasso"){
    X<-cbind(rep(1,nrow(traindata),traindata[,-1]))
    Y<-as.matrix(traindata[,1])
    fit<-cv.glmnet(X,Y,intercept=FALSE)
    X_pred<-cbind(rep(1,nrow(testdata)),testdata[,-1])
    Y_pred<-as.matrix(testdata[,1])
    yfit<-predict(fit,X_pred)
    res<-Y_pred-yfit
    return(sum(res^2)/nrow(testdata))
  }
  else if(method=="pls"){
    fit_pls=plsr(Crime~.,10,data=traindata,validation="CV")
```

```
    k_select=which.min(fit_plr$validation$PRESS)
    fit=plsr(Crime~.,k_select,data=traindata,validation="CV")
    n<-nrow(testdata)
    fv<-predict(fit,newdata=testdata)[((k_select-1)*n+1):(k_select*n)]
    res<-testdata$Crime-fv
    return(sum(res^2)/nrow(testdata))
  }
}
getCV<-function(method,datanew,nfolder){
  n<-nrow(crime)
  cv_value<-rep(0,nfolder)
  n=floor(nrow(datanew)/nfolder)
  for(i in 1:nfolder){
    testdata=datanew[((i-1)*n+1):(i*n),]
    traindata=datanew[((i-1)*n+1):(i*n),]
    cv_value[i]<-CV(traindata,testdata,method)
  }
  print(method)
  return(sum(cv_value)/nfolder)
}
sampl<-sample(1:n,n,relplace=FALSE)
datanew<-crime_std[sampl,]
method=c("stepreg","ridge","lasso","pls")
datanew<-sample
CV_result<-sapply(method,function(x) getCV(x,datanew,10))
```