20181129
CSI: US CRIME DATA ANALYSIS

1조 과학수사대

김주은 김현지 이건
이상욱 정지혜

This is 발표자

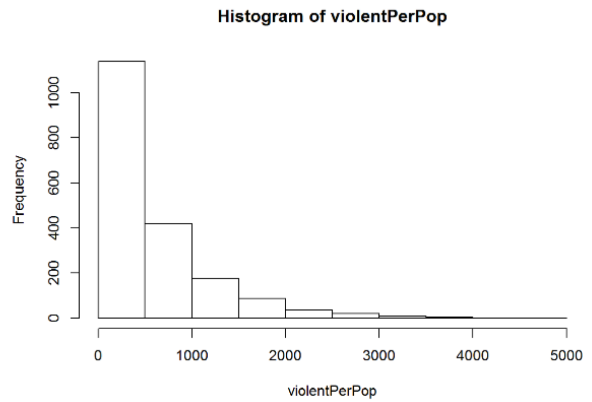
Contents

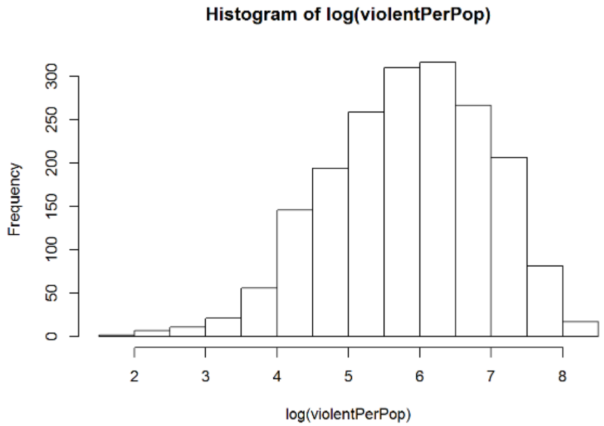
Preview: 지난주

- Categories : 19 Categories
- PCA로 변수 선택* With 40 Variables
- Stepwise Regression with adj.R-sq 0.68
- missing value imputation:
  arsons과 상관관계가 가장 높은 autoTheft 이용,
  4분위로 나눠 각 구간별 median 값으로 대체

끔-찍
차원축소가 불가피하
다..

1-(1) Variables Transformation





- 선형회귀모형을 사용할 것이기 때문에 target variable 의 분포를 대칭으로 만듦
- Train Set의 violentPerPop이 0인 Spencercity 삭제
- 나머지 설명변수들도 다음 기준을 우선으로 사용하여 변수 변환

To be symmetric, $H_U^p - M^p = M^p - H_L^p$

$$f(H_U) - M^p = M^p - f(H_L)$$

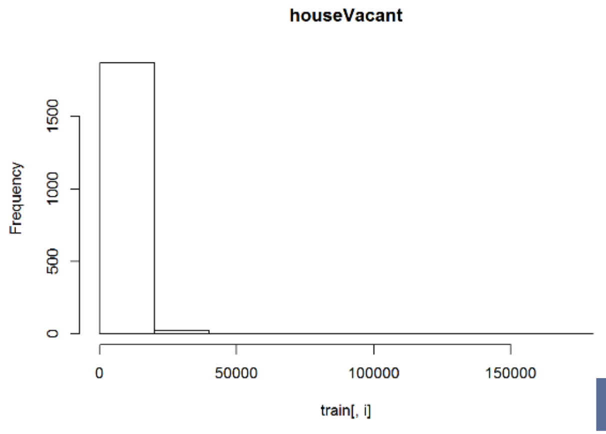
$$pM^{p-1}(H_U - M) + p(p-1)M^{p-2}(H_U - M)^2/2$$
$$= -pM^{p-1}(H_L - M) - p(p-1)M^{p-2}(H_L - M)^2/2$$

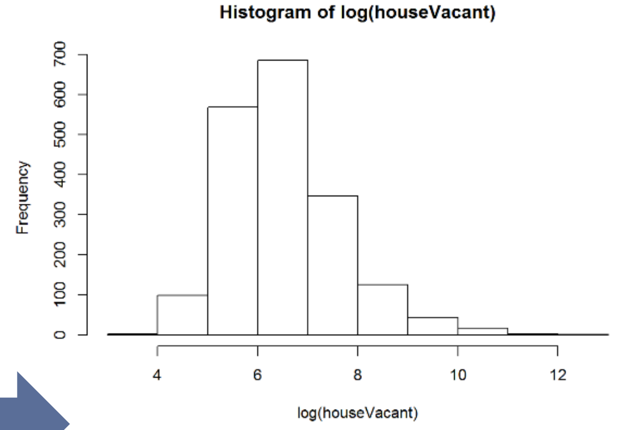
$$p \doteq 1 - \frac{2M\,[(H_U - M) + (H_L - M)]}{(H_U - M)^2 + (H_L - M)^2}$$

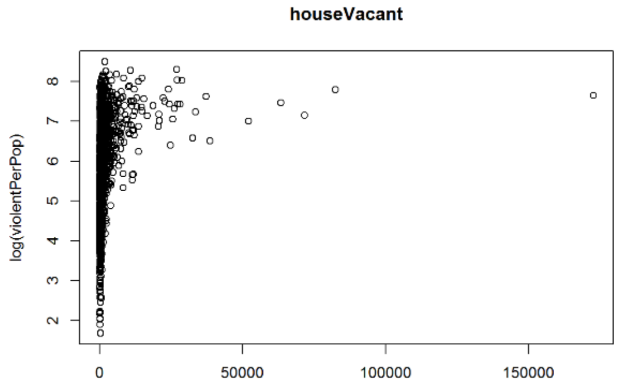
1-(1) Variables Transformation

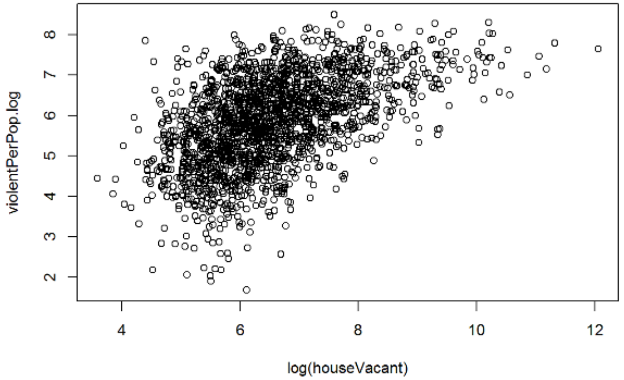
- Outlier를 줄이고, target과선형적 관계 가정을 만족하기 위해 X변수에 대해서도 변수 변환을 진행

- Log변환하려는 변수의 0값은 **0 → 0.01**로 대체하여 변환





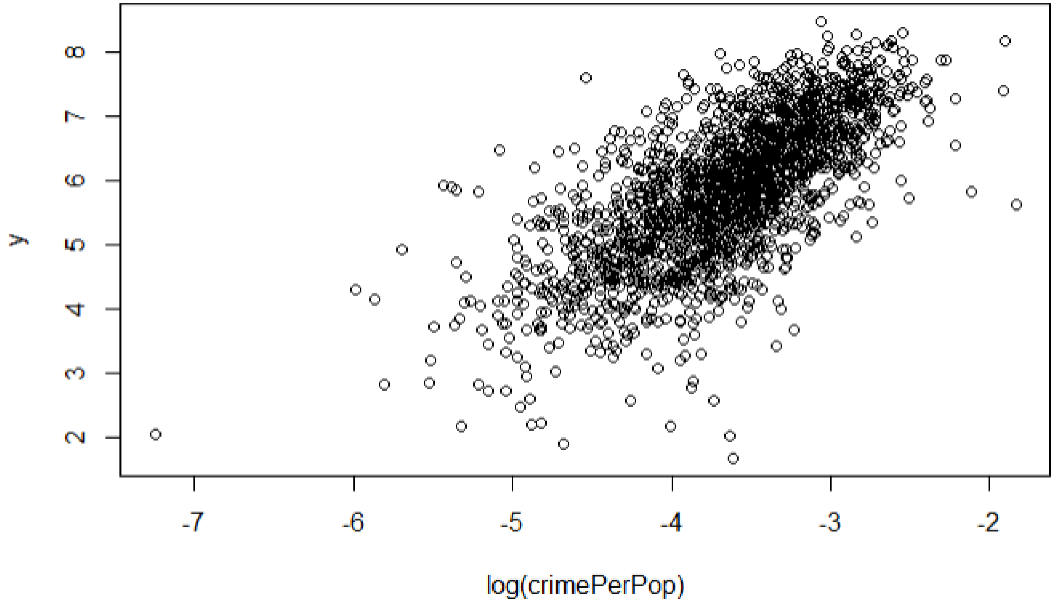




1-(2) New Variables

Data CLEANSING
PCA
Factor Analysis
Lasso
Further comments

- Burglaries, larcenies, autoTheft, arsons 등 범죄가 서로 상관성이 높음
- 단위 맞추기 위해 pop으로 나눠줌 => crimePerPop
- Plot(y, crimePerPop) 이 nonlinear해서 log를 취해 펴 줌 => crimePerPop.log



상당히 리-니어

1-(3) Dealing with OUTLIERS


```r
#### Outliers detection 함수 정의

# Z SCORE with threshold=2
isnt_out_z <- function(x, thres = 2, na.rm = TRUE) {
  abs(x - mean(x, na.rm = na.rm)) <= thres * sd(x, na.rm = na.rm)
}

#### Mahalanobis Distance
# Maha dist 정의
maha_dist <- . %>% select_if(is.numeric) %>%
  mahalanobis(center = colMeans(.), cov = cov(.))

isnt_out_maha <- function(tbl) {
  tbl %>% maha_dist() %>% isnt_out_z()
}
```

Recall:

Mahalanobis distance

- 같은 분포에서 생성된 두 점 $\boldsymbol{x}$, $\boldsymbol{y}$ 사이의 마할라노비스 거리(Mahalanobis distance)는 $\boxed{d_{mahala}(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{y})}}$ 이다. $\boldsymbol{\Sigma}$: covariance matrix of $\boldsymbol{x}$, $\boldsymbol{y}$
- 직관적인 의미를 살펴보자.

$$\{d_{mahala}(\boldsymbol{x},\boldsymbol{y})\}^2 = (\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{y})$$
$$= \left(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}-\boldsymbol{y})\right)^T \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}) \qquad Z = \frac{(X-Y)-0}{\sigma}$$

$\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}-\boldsymbol{y})$는 표준편차로 나누어 표준화(standardize)를 하는 것과 매우 비슷한 모양.

자료의 공분산 구조를 고려하여 거리를 잴 수 있게 거리함수를 잘 만든 것으로 이해할 수 있다.

우리는 '선형회귀모델' 을 사용할 것이므로,
모델에 큰 영향을 주는 아웃라이어 처리가 중요하다고 생각

일반적인 유클리드 거리를 이용하기보단 변수들끼리의 상관관계가 높기 때문에
마할라노비스 거리로 구한 z-score로 아웃라이어를 판단하기로 결정

➡ Mahalanobis 거리로 구한 z-score의 절댓값이 2 를 넘으면 아웃라이어로 판단

1-(4) Categorizing Variables

카테고리 없음을 포함, 총 12개의 category로 분류

1. 카테고리 없음	2. Household	3. Urban	4. Income	5. Education	6. Employment
• BlackPerCap • Pop • pctKidsBornNevrMarr • pctOfficDrugUnit • pctAllDivorc • persEmergShelt • pctBlack • Crime(new Variable)	• pctLargHousFam • pctLargHous • persPerFam • perHoush • pctAllDivorc	• popDensity • pctUrban	• medIncome • pctWwage • pctWfarm • pctWdiv • pctWsocsec • pctPubAsst • pctRetire • medFamIncome • perCapInc • pctPoverty • pctHousWOphone • pctHousWOplumb	• pctLowEdu • pctNotHSgrad • pctCollGrad	• pctUnemploy • pctEmploy • pctEmployProfServ • pctOccupManu • pctOccupMgmt
7. Immigrant	**8. Family**	**9. House Condition**	**10. Ownership**	**11. Vacancy**	**12. House Value**
• pctFgnImmig-3 • pctFgnImmig-5 • pctFgnImmig-8 • pctFgnImmig-10 • pctImmig-3 • pctImmig-5 • pctImmig-8 • pctImmig-10	• pct2Par • pctKids2Par • pctKids-4w2Par • Pct12-17w2Par	• pctPopDenseHous • pctSmallHousUnits • medNumBedrm	• pctPersOwnOccup • medGrossRent • medRentpctHousInc • medOwnCostpct • medOwnCostpctWO	• houseVacant • pctHousOccup • pctHousOwnerOccup • pctVacantBoarded • pctVacant6up	• ownHousMed • rentMed

2. Principle Component Analysis

Data CLEANSING
PCA
Factor Analysis
Lasso
Further comments

12개 각 카테고리 내에서,
변수간 correlation을 고려하여 각 카테고리에서 변수들 선택

1. PC1을 사용한 카테고리:

- Employment, Family, Vacancy, Crime, House Condition, House Value, Income

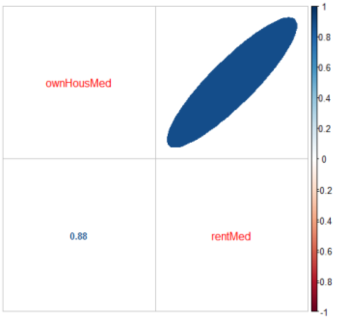
2. 원래 변수를 사용하기로 한 카테고리:

- Ownership : pctPersOwnOccup, medGrossRent, medRentpctHousInc 사용
- Household : pctLargHousFam.in 사용
- Urban : pctUrban(dummy variable), popDensity.log 사용
- Immigration : pctFgnImmig.10 사용
- Other : pctBlack.sqrt, pctAllDivorc, pctOfficDrugUnit(dummy variable) 사용

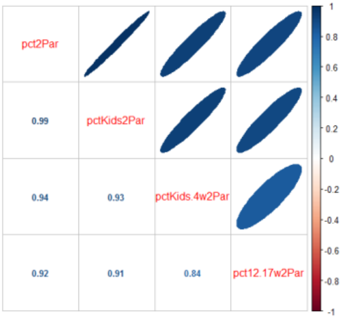
2. Principle Component Analysis

Data CLEANSING
PCA
Factor Analysis
Lasso
Further comments

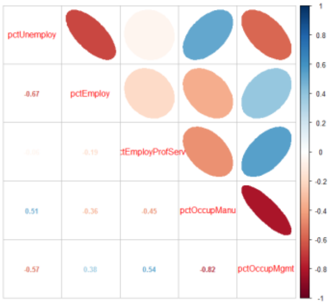
12개 각 카테고리 내에서,
변수간 **correlation**을 고려하여 각 카테고리에서 변수들 선택



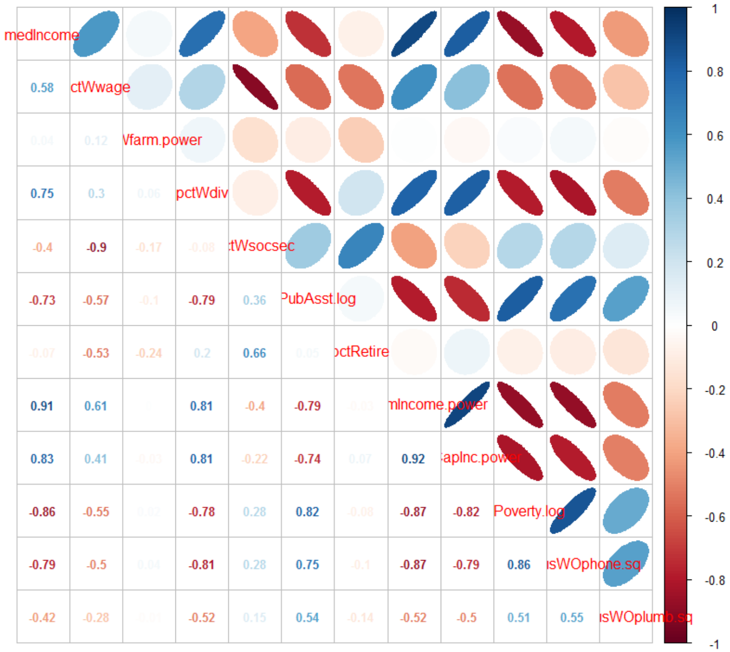
house value



family

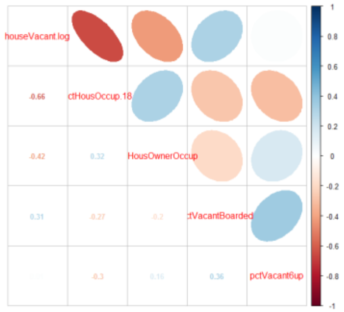


employment



income

house condition



vacancy

2. Principle Component Analysis

PCA를 이용한 결과, 총 1893 obs, of 20 variables

```
df <- data.frame(cbind(pctBlack.sqrt,pctAllDivorc,pctOfficDrugUnit,
                       pctPersOwnOccup,medGrossRent, medRentpctHousInc,
                       pctLargHousFam.in,
                       pctUrban,popDensity.log,
                       PC1, pctWwage, pctWfarm.power,
                       PC2, PC3, PC4, crimeperpop,
                       pctFgnImmig.10, PC6, PC7, violentPerPop.log
))
```

```
# MSE 추정하기  #134999.3
mean((exp(df2$violentPerPop.log) -exp(predict(crime.fit ,df2)))[-train]^2)

## [1] 134999.3
```

%>%

Mahalanobis distance 로 구한 Z스코어로
Outliers 제거 # 44 values

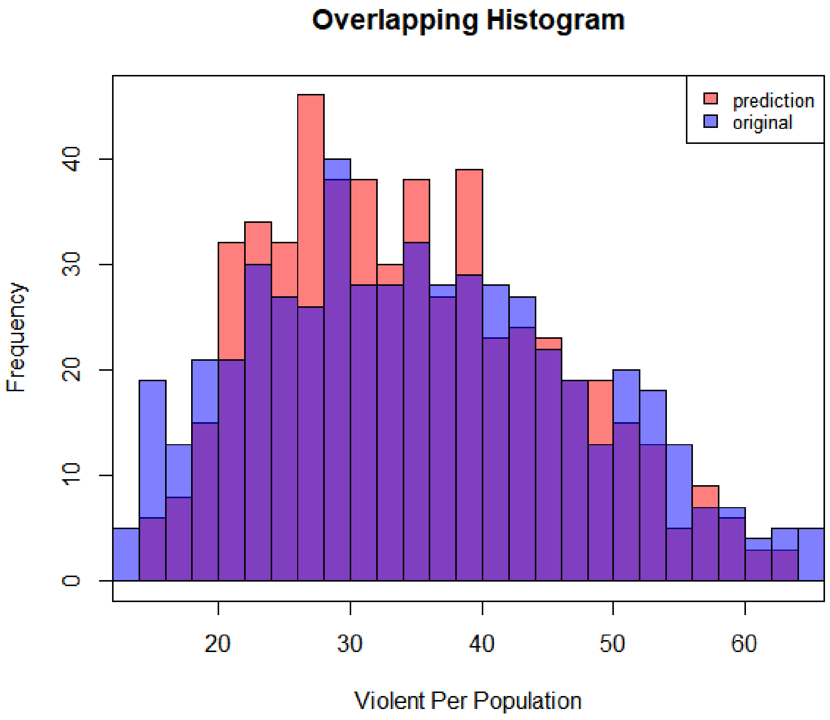
%>%

Training Set을 7:3으로 나누고
Linear Regression 적합

Test MSE: 134999.3

2. Principle Component Analysis

PCA를 이용한 결과, 총 1893 obs, of 20 variables



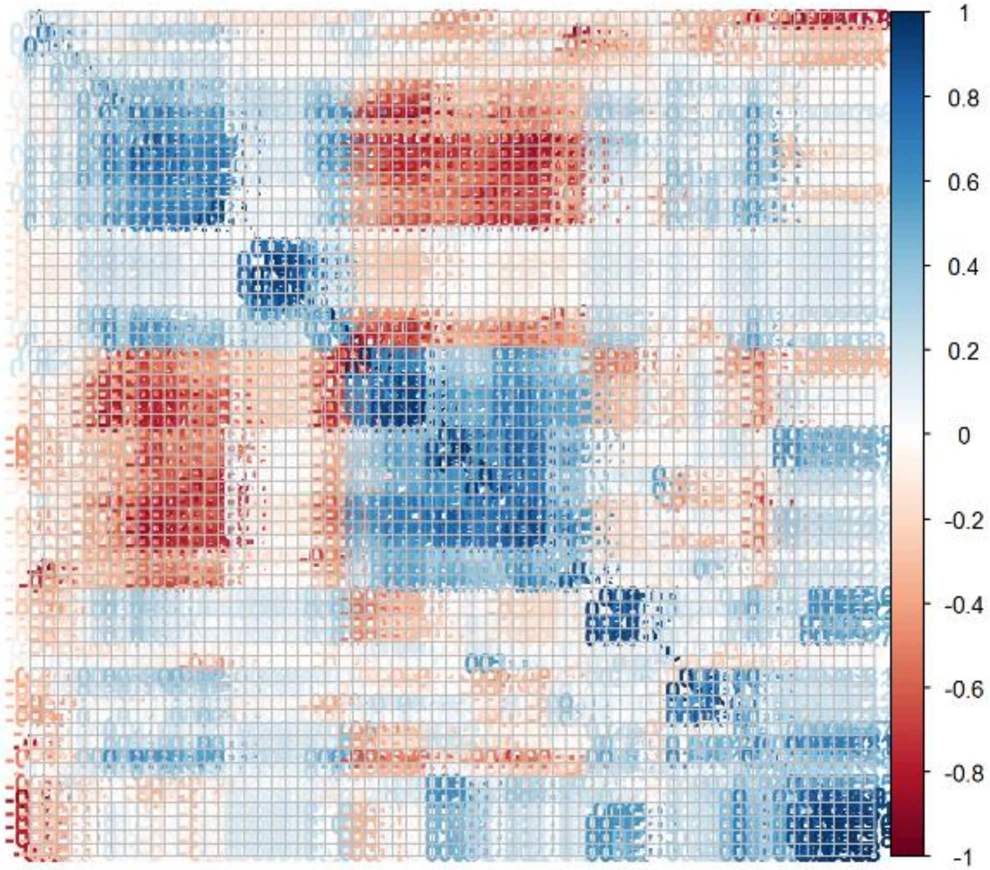
Multiple R-squared: 0.7077
Adjusted R-squared: 0.7033
Test MSE: 134999.3

```
## PC6                -0.04967    0.02394  -2.075  0.03823 *
## PC7                 0.16657    0.05190   3.209  0.00136 **
## pctOfficDrugUnit2   0.09830    0.05634   1.745  0.08131 .
## pctUrban2           0.18694    0.04786   3.906 9.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5971 on 1245 degrees of freedom
## Multiple R-squared:  0.7077,Adjusted R-squared:  0.7033
## F-statistic: 158.7 on 19 and 1245 DF,  p-value: < 2.2e-16
```

3. Factor Analysis

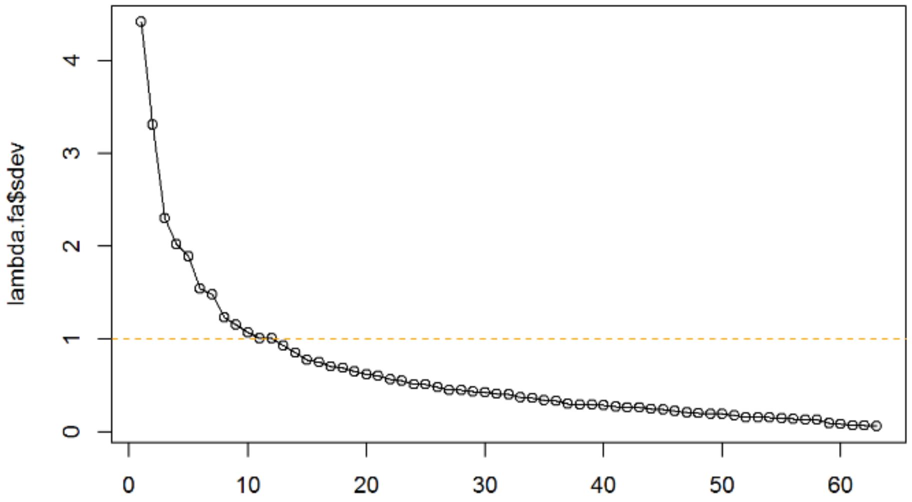
Data CLEANSING
PCA
Factor Analysis
Lasso
Further comments

CORRELATION PLOT



추려낸 67개의 변수들중에서
범주형 변수를 제외한 63개의 변수들의 latent factor를 찾아 새로운
변수를 만들 목표로
factor analysis를 진행

SCREE PLOT



3. Factor Analysis

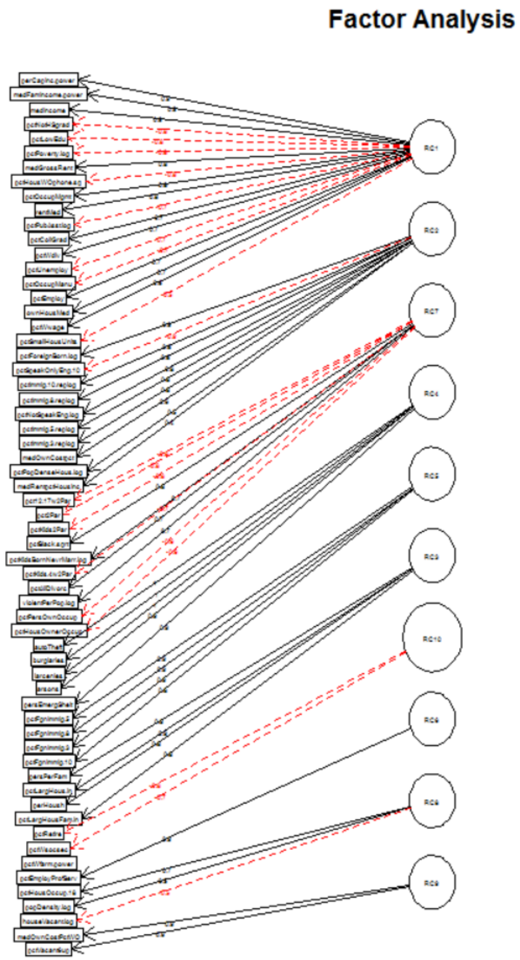
표준화한 변수들로 latent factor가 6~10개 인 경우를 살펴봄

새로운 변수는 어떻게 만드나?

Factor score(= estimates of unobserved variables)
사용할 수 있음

$$\hat{f} = \hat{Q}^T R^{-1} z$$

Factor score가 독립변수,
log(violentPerPop이 종속변수인 다중선형회귀적합



3. Factor Analysis

Variables with high uniqueness

```
Uniquenesses:
                 pop              perHoush         pctBlack.sqrt              pctUrban                medIncome
               0.030                 0.263                 0.355                 0.756                    0.074
            pctWwage       pctWfarm.power              pctWdiv             pctWsocsec           pctPubAsst.log
               0.361                 0.890                 0.156                 0.476                    0.208
           pctRetire  medFamIncome.power        perCapInc.power      blackPerCap.power          pctPoverty.log
               0.656                 0.062                 0.072                 0.946                    0.112
           pctLowEdu          pctNotHSgrad           pctCollGrad           pctUnemploy                pctEmploy
               0.334                 0.233                 0.325                 0.291                    0.462
 pctEmployProfServ          pctOccupManu          pctOccupMgmt           pctAllDivorc               persPerFam
               0.833                 0.439                 0.317                 0.407                    0.113
             pct2Par           pctKids2Par         pctKids.4w2Par         pct12.17w2Par pctKidsBornNevrMarr.log
               0.030                 0.030                 0.109                 0.148                    0.220
      pctFgnImmig.3         pctFgnImmig.5         pctFgnImmig.8        pctFgnImmig.10        pctImmig.3.replog
               0.249                 0.098                 0.053                 0.106                    0.110
  pctImmig.5.replog     pctImmig.8.replog    pctImmig.10.replog    pctNotSpeakEng.log      pctSpeakOnlyEng.10
               0.042                 0.030                 0.030                 0.188                    0.168
  pctLargHousFam.in       pctLargHous.in       pctPersOwnOccup   pctPopDenseHous.log       pctSmallHousUnits
               0.258                 0.294                 0.292                 0.215                    0.291
         medNumBedrm      houseVacant.log        pctHousOccup.18       pctHousOwnerOccup       pctVacantBoarded
               0.539                 0.555                 0.686                 0.360                    0.652
       pctVacant6up   pctHousWOphone.sq     pctHousWOplumb.sq            ownHousMed                  rentMed
               0.799                 0.188                 0.644                 0.280                    0.116
       medGrossRent    medRentpctHousInc         medOwnCostpct       medOwnCostPctWO           persEmergShelt
               0.127                 0.745                 0.521                 0.929                    0.154
 pctForeignBorn.log       popDensity.log      pctOfficDrugUnit            burglaries               larcenies
               0.032                 0.679                 0.759                 0.030                    0.040
           autoTheft                arsons    violentPerPop.log
               0.030                 0.192                 0.407
```

K개 factor중 어느 것에도 크게 영향 받지 않음

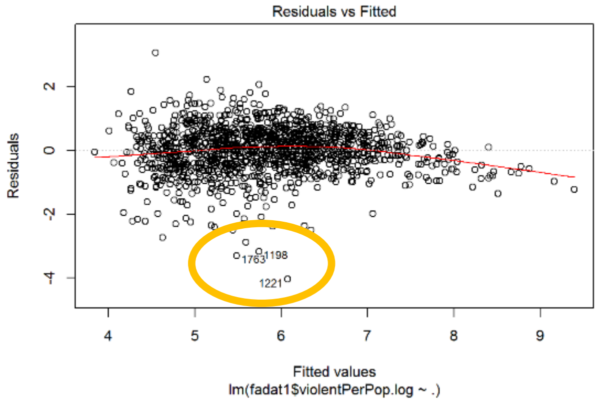
$$Y = B[X|F] + e$$

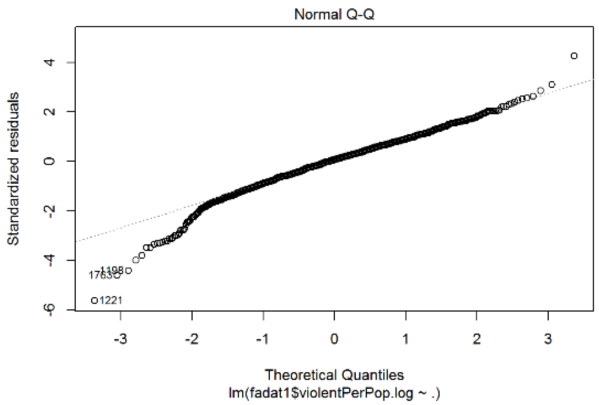
pctWfarm,blackPerCap,pctHousWoplumb + factor1~10

```
##
## Call:
## lm(formula = fadat1$violentPerPop.log ~ ., data = pcm11set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7152 -0.4064  0.0300  0.4241  2.8085
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.830757   0.018984 307.139  < 2e-16 ***
## pctWfarm.power     0.006362   0.021038   0.302  0.76240
## blackPerCap.power  0.017148   0.019648   0.873  0.38295
## pctHousWOplumb.sq -0.058403   0.024031  -2.430  0.01522 *
## RC1               -0.482884   0.022587 -21.379  < 2e-16 ***
## RC2                0.221688   0.019479  11.381  < 2e-16 ***
## RC7                0.688470   0.020332  33.861  < 2e-16 ***
## RC4                0.188589   0.019110   9.869  < 2e-16 ***
## RC9                0.106575   0.019236   5.540 3.65e-08 ***
## RC3                0.050381   0.019474   2.587  0.00979 **
## RC5               -0.049806   0.019329  -2.577  0.01008 *
## RC6               -0.103579   0.019444  -5.327 1.17e-07 ***
## RC10              -0.100655   0.019572  -5.143 3.12e-07 ***
## RC8               -0.134519   0.019461  -6.912 7.43e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.691 on 1311 degrees of freedom
## Multiple R-squared:  0.6293, Adjusted R-squared:  0.6256
## F-statistic: 171.2 on 13 and 1311 DF,  p-value: < 2.2e-16
```

3. Factor Analysis

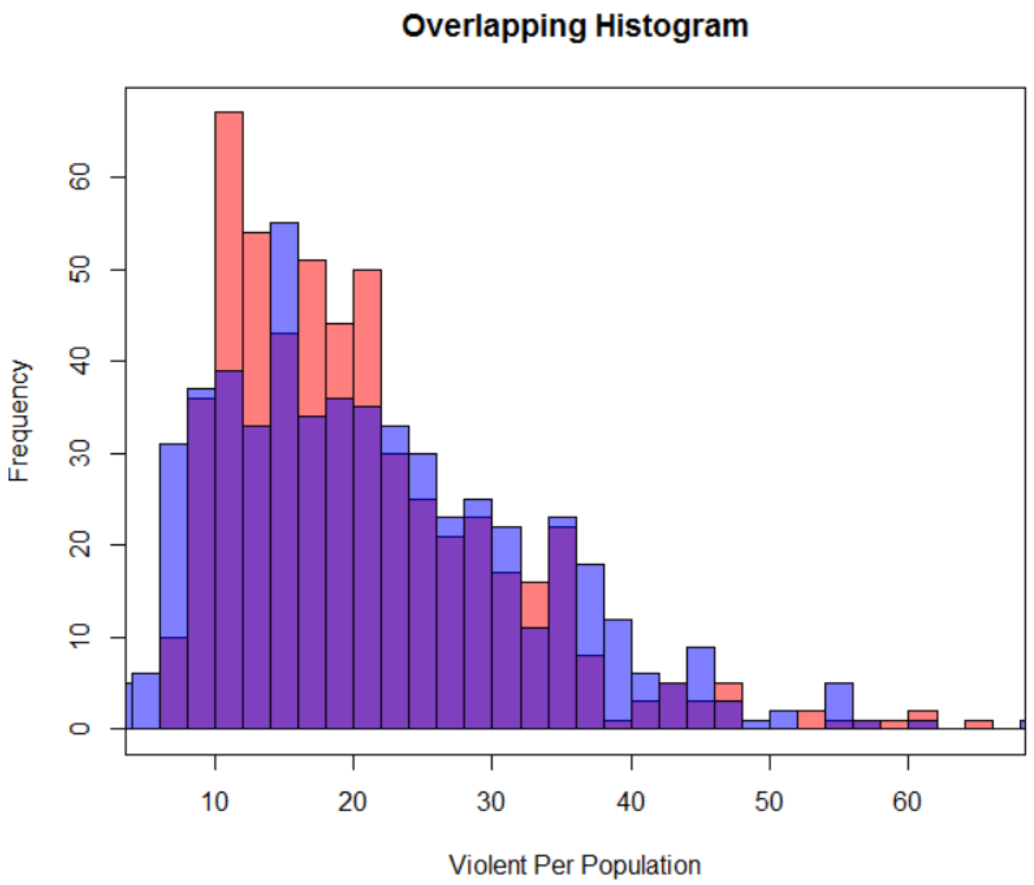
Outliers elimination






```
##
## Call:
## lm(formula = fadat1$violentPerPop.log ~ ., data = as.data.frame(pcm11set))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.91320 -0.35647  0.00018  0.38530  2.03171
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.83733    0.01710 341.367  < 2e-16 ***
## pctVacantBoarded     0.02599    0.02157   1.205    0.228
## pctHousWOplumb.sq   -0.03500    0.02146  -1.631    0.103
## RC1                 -0.39967    0.02111 -18.930  < 2e-16 ***
## RC2                  0.21692    0.01724  12.581  < 2e-16 ***
## RC7                  0.71587    0.01936  36.972  < 2e-16 ***
## RC4                  0.17688    0.01775   9.966  < 2e-16 ***
## RC5                  0.11751    0.01713   6.859 1.09e-11 ***
## RC3                  0.09745    0.01803   5.404 7.80e-08 ***
## RC10                -0.08123    0.01718  -4.729 2.51e-06 ***
## RC6                 -0.09553    0.01716  -5.567 3.17e-08 ***
## RC8                 -0.12435    0.01721  -7.224 8.75e-13 ***
## RC9                 -0.11570    0.01810  -6.391 2.32e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6084 on 1253 degrees of freedom
## Multiple R-squared:  0.6895, Adjusted R-squared:  0.6865
## F-statistic: 231.9 on 12 and 1253 DF,  p-value: < 2.2e-16
```

3. Factor Analysis



Outlier 제거한 data에서
Training Set을 7:3으로 나눈 후
Uniqueness 큰 변수 3개와
나머지 변수들의 factor scores로

linear Regression 적합

Multiple R-squared: 0.6895
Adjusted R-squared: 0.6865
Test MSE: 134850

```
#test MSE 추정
  p1 <-predict(pcmlm11,newdata=as.data.frame(pcm11set.val))
  testMSE <- mean((exp(fa.val$violentPerPop.log) - exp(p1)) ^2)

print(testMSE)

```
```

```
[1] 134850
```
```

3. Factor Analysis

Data CLEANSING

PCA

Factor Analysis

Lasso

Further comments

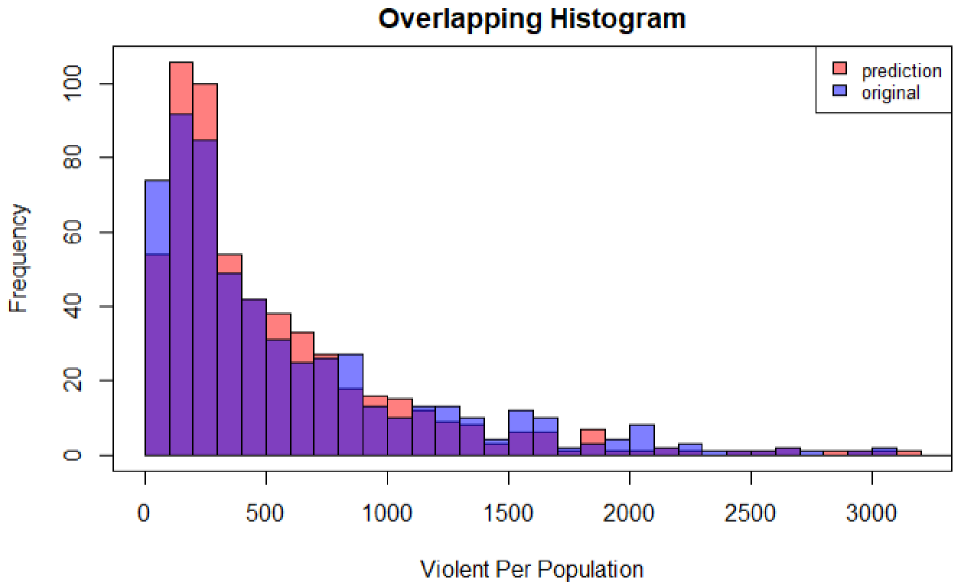
FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6	FACTOR7	FACTOR8	FACTOR9	FACTOR10
소득	문화적다양성	가구크기	범죄	이민자	직업전문성	가정환경	거주밀도	주거환경	고정수입
perCapInc.power	pctForeignBorn.log	pctSmallHousUnits	autoTheft	pctFgnImmig.5	pctOccupMgmt	pct12.17w2Par	pctHousOccup.18	pctVacant6up	pctRetire
medFamIncome.power	pctSpeakOnlyEng.10	persPerFam	burglaries	pctFgnImmig.8	pctCollGrad	pct2Par	popDensity.log	medOwnCostPctWO	pctWsocsec
medIncome	pctNotSpeakEng.log	pctLargHous.in	larcenies	pctFgnImmig.3	pctOccupManu	pctKids2Par	houseVacant.log		pctWwage
pctNotHSgrad	pctImmig.10.replog	perHoush	arsons	pctFgnImmig.10	pctEmployProfServ	pctBlack.sqrt			pctEmploy
pctPoverty.log	pctImmig.8.replog	pctLargHousFam.in	persEmergShelt	pctImmig.8.replog		pctKidsBornNevrMarr.log			
pctLowEdu	pctImmig.5.replog			pctImmig.5.replog		pctKids.4w2Par			
medGrossRent	pctImmig.3.replog			pctImmig.3.replog		pctAllDivorc			
pctHousWOphone.sq	medOwnCostpct					houseVacant.log			
pctOccupMgmt	pctPopDenseHous.log					pctPersOwnOccup			
rentMed						pctHousOwnerOccup			
pctPubAsst.log									

	coeficiensts
(Intercept)	5.83732681
pctVacantBoarded	0.02390518
pctHousWOplumb.sq	-0.03352498
RC1	-0.39742378
RC2	0.21563016
RC7	0.71648283
RC4	0.17726665
RC5	0.11919533
RC3	0.10012109
RC10	-0.08096968
RC6	-0.09245622
RC8	-0.12832502
RC9	-0.11410398

4. LASSO

LASSO by R

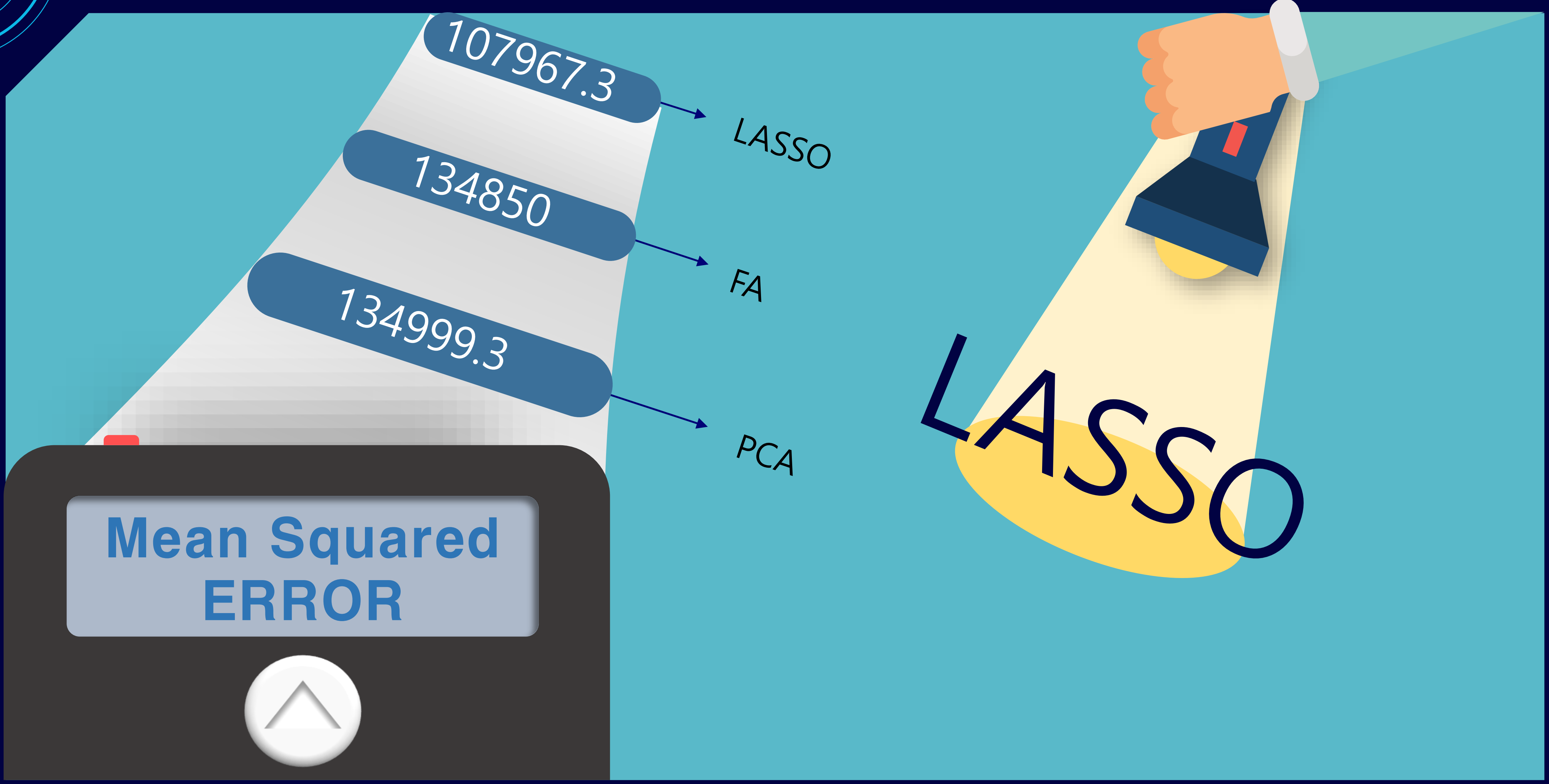
```
grid = 10^seq(10, -2, length = 100)
lasso3.mod = glmnet(H3[train,], y3[train], alpha = 1,
                    lambda = grid, thresh = 1e-12)
lasso3.pred = predict(lasso3.mod, s=lambda3, newx=H3[test,])
lasso3.mse = mean((exp(lasso3.pred)-exp(y3[test]))^2)
lasso3.mse #107967.3


## [1] 107967.3
```



Original VS Deleting outliers

	Original	STD * 2	STD * 3
Variable Selection	30/63	32/63	30/63
MSE	**107738.5**	**121522.7**	**107967.3**

5.What's our final model?



5. Conclusion

Data CLEANSING

PCA

Factor Analysis

Lasso

Further comments

최종 LASSO model에서 살아남은 계수들 : 34개

(Intercept)	9.734476e+00
perHoush	.
pctBlack.sqrt	1.121895e-01
pctUrban	1.797068e-01
medIncome	.
pctWwage	-1.486886e-02
pctWfarm.power	.
pctWdiv	-1.766099e-02
pctWsocsec	.
pctPubAsst.log	8.497500e-02
pctRetire	.
medFamIncome.power	1.386489e+02
perCapInc.power	2.368337e+03
blackPerCap.power	2.712239e-02
pctPoverty.log	.
pctLowEdu	.
pctNotHSgrad	1.255177e-03
pctCollGrad	.
pctUnemploy	.
pctEmploy	.
pctEmployProfServ	-6.793263e-04
pctOccupManu	.
medGrossRent	1.037022e-04
medRentpctHousInc	7.773475e-03
medOwnCostpct	.
medOwnCostPctWO	-2.746061e-02
pctForeignBorn.log	3.408196e-02
popDensity.log	7.027587e-03
pctOfficDrugUnit	9.121336e-02
crimePerPop.log	5.725530e-01
pctEmergShelt	1.579419e+01

pctAllDivorc	1.396606e-02
persPerFam	.
pct2Par	.
pctKids2Par	-3.772591e-03
pctKids.4w2Par	.
pct12.17w2Par	.
pctKidsBornNevrMarr.log	7.735463e-02
pctFgnImmig.3	.
pctFgnImmig.5	-1.830349e-04
pctFgnImmig.8	-2.500270e-05
pctFgnImmig.10	-1.865596e-03
pctImmig.3.replog	2.983148e-03
pctImmig.5.replog	.
pctImmig.8.replog	.
pctImmig.10.replog	.
pctNotSpeakEng.log	.
pctSpeakOnlyEng.10	-2.020939e-22
pctLargHousFam.in	.
pctLargHous.in	.
pctPersOwnOccup	.
pctPopDenseHous.log	1.669934e-01
pctSmallHousUnits	4.556269e-04
medNumBedrm	-1.593891e-02
pctHousOccup.18	-5.857344e-38
pctHousOwnerOccup	.
pctVacantBoarded	5.397142e-03
pctVacant6up	.
pctHousWOphone.sq	.
pctHousWOplumb.sq	-4.724866e-02
ownHousMed	4.838086e-07
rentMed	.

Data CLEANSING

PCA

Factor Analysis

Lasso

Further comments

Thank you