Community and Crime

2019 FALL ESC Final Project

01

We are 1조 !

김윤환 엄상준
이솔희 전은지 최우현

Contents

00

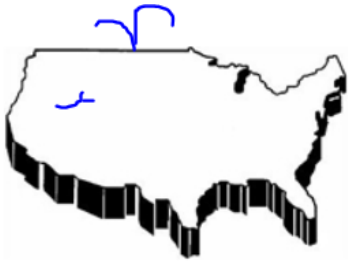
프로젝트 소개

00 프로젝트 소개

Attribute Information: (122 predictive, 5 non-predictive, 1 goal)
-- state: US state (by number) - not counted as predictive above, but if considered, should be consided nominal (nominal)
-- county: numeric code for county - not predictive, and many missing values (numeric)
-- community: numeric code for community - not predictive and many missing values (numeric)
-- communityname: community name - not predictive - for information only (string)
-- fold: fold number for non-random 10 fold cross validation, potentially useful for debugging, paired tests - not predictive (numeric)
-- population: population for community: (numeric - decimal)
-- householdsize: mean people per household (numeric - decimal)
-- racepctblack: percentage of population that is african american (numeric - decimal)
-- racePctWhite: percentage of population that is caucasian (numeric - decimal)
-- racePctAsian: percentage of population that is of asian heritage (numeric - decimal)
-- racePctHisp: percentage of population that is of hispanic heritage (numeric - decimal)
-- agePct12t21: percentage of population that is 12-21 in age (numeric - decimal)
-- agePct12t29: percentage of population that is 12-29 in age (numeric - decimal)
-- agePct16t24: percentage of population that is 16-24 in age (numeric - decimal)
-- agePct65up: percentage of population that is 65 and over in age (numeric - decimal)
-- numbUrban: number of people living in areas classified as urban (numeric - decimal)

⋮

-- LemasPctOfficDrugUn: percent of officers assigned to drug units (numeric - decimal)
-- PolicBudgPerPop: police operating budget per population (numeric - decimal)
-- ViolentCrimesPerPop: total number of violent crimes per 100K popuation (numeric - decimal) GOAL attribute (to be predicted)



Data Description

1989년 미국의 1994개 도시들에 대한 US Census

Goal

베이지안 method를 활용하여 범죄지도 완성하기

01

EDA 정리 및 요약

1. NA 제거

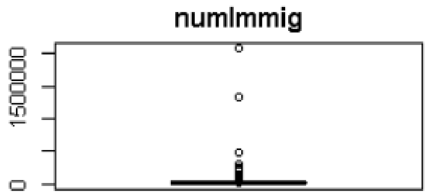
```
##                          n naratio nacatg
## 1              LemasSwornFT   0.845    Bad
## 2            LemasSwFTPerPop   0.845    Bad
## 3          LemasSwFTFieldOps   0.845    Bad
## 4       LemasSwFTFieldPerPop   0.845    Bad
## 5              LemasTotalReq   0.845    Bad
## 6          LemasTotReqPerPop   0.845    Bad
## 7           PolicReqPerOffic   0.845    Bad
## 8                PolicPerPop   0.845    Bad
## 9         RacialMatchCommPol   0.845    Bad
## 10            PctPolicWhite   0.845    Bad
## 11            PctPolicBlack   0.845    Bad
## 12             PctPolicHisp   0.845    Bad
## 13            PctPolicAsian   0.845    Bad
## 14            PctPolicMinor   0.845    Bad
## 15      OfficAssgnDrugUnits   0.845    Bad
## 16        NumKindsDrugsSeiz   0.845    Bad
## 17         PolicAveOTWorked   0.845    Bad
## 18                PolicCars   0.845    Bad
## 19            PolicOperBudg   0.845    Bad
## 20      LemasPctPolicOnPatr   0.845    Bad
## 21      LemasGangUnitDeploy   0.845    Bad
## 22          PolicBudgPerPop   0.845    Bad
```

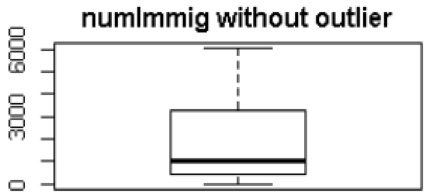
"변수 28개 제거"

- NA가 0으로 적혀 있는 LemasPctOfficDrugUn
- NA 비율이 80% 이상인 변수 22개
- 분석에 불 필요할 것으로 생각되는 변수들 (communityname, State, communityCode, countryCode, fold)
- Response Variable이 NA인 변수

2. Outlier 처리

```r
dtx_q1 <- c()
dtx_q3 <- c()
for(i in 1:ncol(dtx)) {
  dtx_q1[i] <-  quantile(dtx[,i])[1]
  dtx_q3[i] <-  quantile(dtx[,i])[3]
}

dtx_q <- as.data.frame((cbind(dtx_q1, dtx_q3)))

dtx_q <- dtx_q %>%
  mutate(dtx_out1 = dtx_q1 - 5*(dtx_q3-dtx_q1)) %>%
  mutate(dtx_out2 = dtx_q3 + 5*(dtx_q3-dtx_q1))
```





"사분위수 활용"

데이터의 분산을 고려하여

> Q3 + 5 * IQR

< Q1 - 5 * IQR

일 경우 Outlier 로 간주!

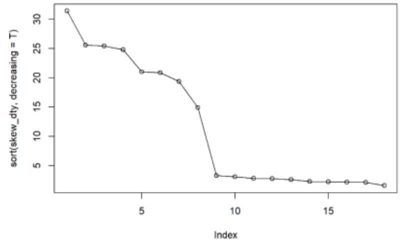
But

NumInShelters
NumStreet
는 삭제!

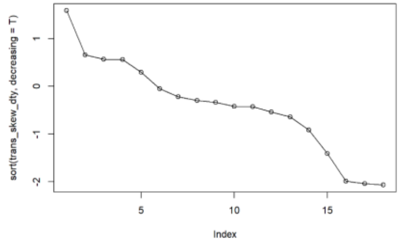
```
> stem(dtx[,92])

  The decimal point is 3 digit(s) to the right of the |

   0 | 0000000000000000000000000000000000000000000000000000000000000000000000000+1803
   1 | 001223367
   2 | 0248
   3 | 4
   4 | 067
   5 |
   6 |
   7 |
   8 |
   9 |
  10 |
  11 |
  12 |
  13 |
  14 |
  15 |
  16 |
  17 |
  18 |
  19 |
  20 |
  21 |
  22 |
  23 | 4
```

```
> stem(dtx[,93])

  The decimal point is 3 digit(s) to the right of the |

   0 | 000000000000000000000000000000000000000000000000000000000000000000000000000+1816
   1 | 16
   2 | 1
   3 | 1
   4 |
   5 |
   6 |
   7 |
   8 |
   9 |
  10 | 4
```

3. Variable Transformation

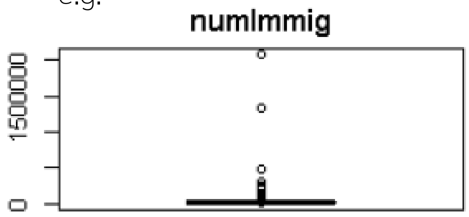


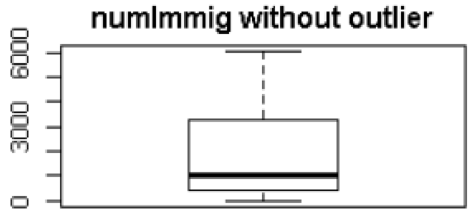
Y 변수




```r
trans_conti_dtx=conti_dtx
for(i in 1:ncol(conti_dtx)) {
  if(skew_dtx[i]>2){
    for(j in 1:dim(conti_dtx)[1]){
      if (conti_dtx[j,i]==0){
        conti_dtx[j,i]=0.03}
    }
    trans_conti_dtx[,i]=log(conti_dtx[,i])
    colnames(trans_conti_dtx)[i]=paste('log',colnames(trans_conti_dtx)[i])
  }
  if(skew_dtx[i]<(-2)){
    for(k in 1:dim(conti_dtx)[1]){
      if (conti_dtx[k,i]==0){
        conti_dtx[k,i]=0.03}
    }
    trans_conti_dtx[,i]=(conti_dtx[,i])^2
    colnames(trans_conti_dtx)[i]=paste('sq',colnames(trans_conti_dtx)[i])}
}
```

e.g.





"Skewness 활용"

Skewness > 2

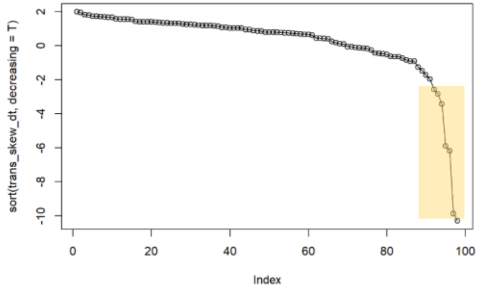
0을 0.03으로 대체 후 Log 변환

Skewness < -2

Square 변환

But

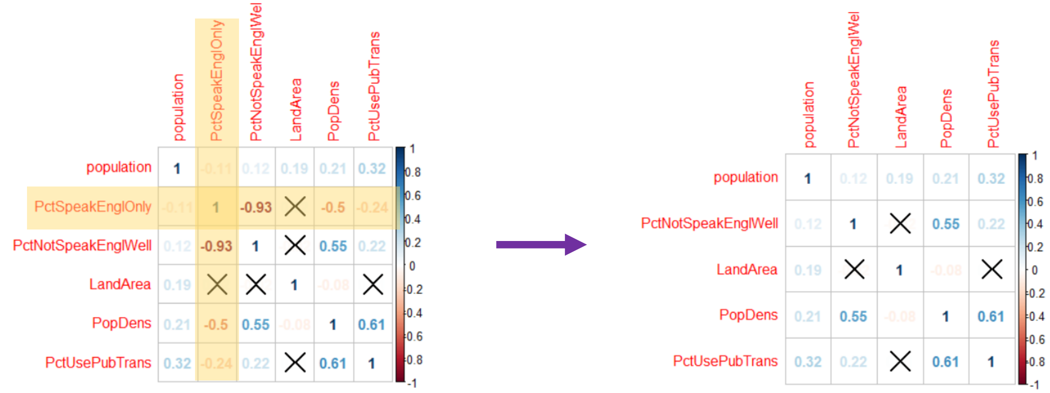
여전히 Skewness인
변수는 제거



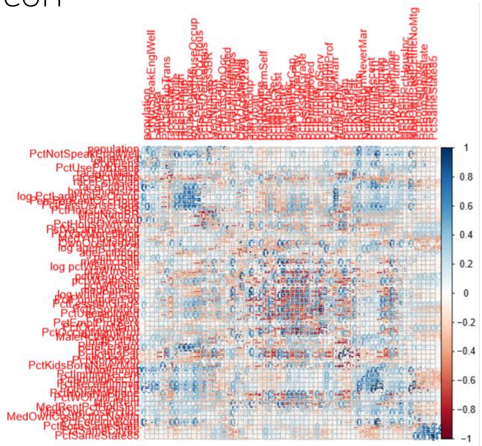

```
##        a
##  [1,] "log HispPerCap"        "27" "-10.2934214020652"
##  [2,] "log OwnOccQrange"      "82" "-9.87742635877414"
##  [3,] "log AsianPerCap"       "25" "-6.17693510840391"
##  [4,] "log blackPerCap"       "23" "-5.89841276251884"
##  [5,] "log OtherPerCap"       "26" "-3.41634952946269"
##  [6,] "log indianPerCap"      "24" "-2.83272299998629"
##  [7,] "sq PctHousOccup"       "72" "-2.57681111345612"
```

4. 변수 선택

1차) Group 내 corr



2-3차) Group 간 corr



최종) 49개 변수 선택

“Correlation 활용”

변수 Groping을 통해

높은 correlation을 갖는 변수를

제거하는 방식으로 차원 축소

others	race	house	age	urban	income
race income	economic	education	employment	marital state	family form
immigrant	ownership	rent	poverty	pop change	

[Description을 바탕으로 한 직관적 Grouping]

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	others	race	House	Age	urban	Income	Race Incor	Economic	Education	Employme	Marital Sta	Family For	Immigrant	Ownership	Rent	Population Change	
2	populatior	racepctbla	household	log agePct	pctUrban	medIncom	OtherPerC	PctPopUn	PctLess9th	PctUnemp	MalePctNe	PctKids2Pa	NumImmi	PctWOFull	MedRentP	PctBornSameState	
3	LandArea	racePctWh	PersPerRe	agePct65up		log pctWFarmSelf			PctBSorM	PctEmploy	TotalPctDi	PctWorkM	PctImmigRecent		MedOwnC	PctSameCity85	
4	PopDens	racePctAsi	PctPersDenseHous			pctWInvInc				PctEmplProfServ		PctKidsBor	PctImmigRec10		MedOwnC	PctSameState85	
5	PctUsePub	racePctHis	PctHousLess3BR			pctWRetire				PctOccupManu			PctRecentImmig				
6			MedNumBR														
7			HousVacant														
8			PctHousOwnOcc														
9			PctVacantBoarded														
10			PctVacMore6Mos														
11			MedYrHousBuilt														

[최종 49개 변수]

02

모델링

02 Modeling

사용한 방법론들

Y 결측치에
Median 때림 (아야!)

Linear Regression
With Stepwise Selection

Bayesian
Regression

Median

PCA & LM

Lasso

1. Linear Regression with Stepwise Selection

Y 변수 별로 stepwise selection을 통해 변수 선택

e.g. Rages ~
population, racepctblck, racepctAsian, HousVacant, PctVacantBoarded, MedYrHousBuilt, medIncome, PctPopUnderPov, PctBSorMore, PctOccupManu, TotalPctDiv, PctImmigRec10, PctRecentImmig, PctWOFullplumb,
MedRentPctHousInc, MedOwnCostPctIncNoMtg

```
Call:
lm(formula = rapes ~ population + racepctblack + racePctWhite +
    PersPerRentOccHous + PctPersDenseHous + MedNumBR + PctVacantBoarded +
    MedYrHousBuilt + agePct65up + PctOccupManu + MalePctNevMarr +
    PctKids2Par + NumImmig + PctImmigRec10 + PctWOFullPlumb +
    MedRentPctHousInc + MedOwnCostPctIncNoMtg, data = new.dt.rapes)

Residuals:
    Min      1Q  Median      3Q     Max
-978.57 -373.32  -35.53  364.67 1200.05

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             644.49     198.77   3.242  0.00121 **
population               57.43      23.34   2.461  0.01398 *
racepctblack             65.42      24.20   2.703  0.00696 **
racePctWhite             69.94      29.86   2.342  0.01931 *
PersPerRentOccHous      -64.79      24.24  -2.673  0.00762 **
PctPersDenseHous         56.00      30.69   1.825  0.06827 .
MedNumBR2               159.05     197.41   0.806  0.42056
MedNumBR3                37.61     201.52   0.187  0.85199
MedNumBR4               189.72     231.65   0.819  0.41293
PctVacantBoarded         24.69      17.04   1.449  0.14744
MedYrHousBuilt           33.56      17.49   1.919  0.05523 .
agePct65up              -42.98      19.25  -2.233  0.02573 *
PctOccupManu             40.31      19.45   2.073  0.03837 *
MalePctNevMarr          -41.11      20.62  -1.994  0.04641 *
PctKids2Par             -85.21      31.24  -2.728  0.00647 **
NumImmig                -54.54      27.93  -1.952  0.05110 .
PctImmigRec10           -43.20      16.58  -2.606  0.00926 **
PctWOFullPlumb           30.75      16.62   1.850  0.06450 .
MedRentPctHousInc        43.24      16.70   2.590  0.00971 **
MedOwnCostPctIncNoMtg   -46.08      14.60  -3.155  0.00164 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 469.8 on 1311 degrees of freedom
Multiple R-squared:  0.1332,	Adjusted R-squared:  0.1206
F-statistic:  10.6 on 19 and 1311 DF,  p-value: < 2.2e-16
```

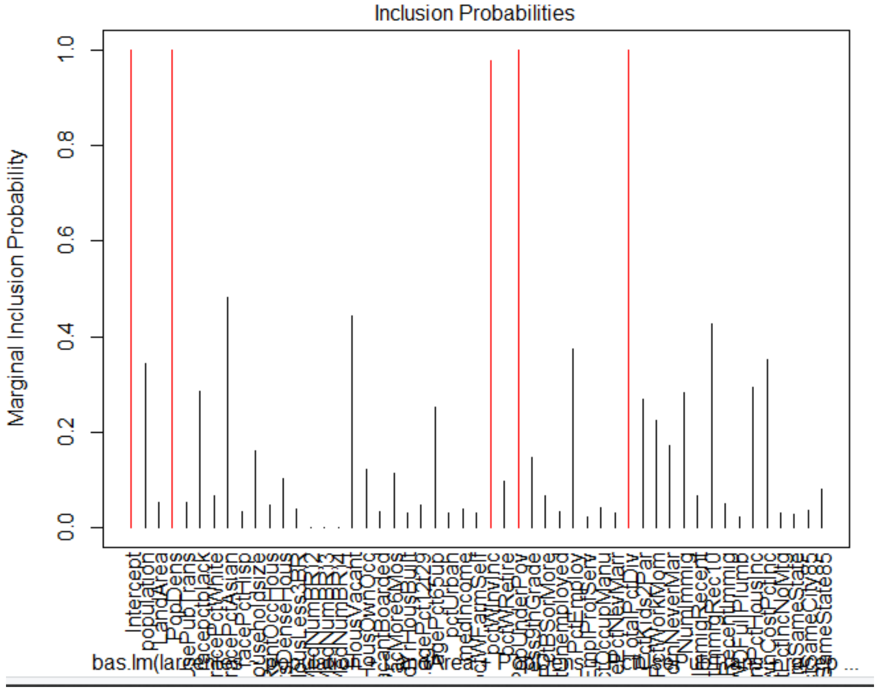
02 Modeling

2. Bayesian Regression

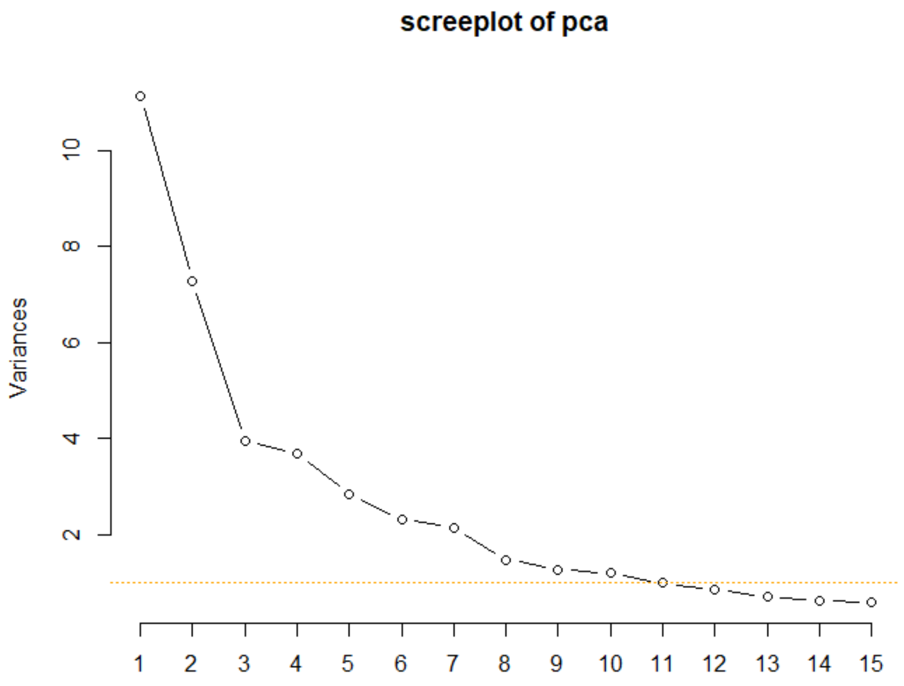
* Bas 라이브러리 사용

```
#larcenies
lm_larcenies <- bas.lm(larcenies ~population+LandArea+PopDens+PctUsePubTrans+racepctbl
                        prior='g-prior',
                        data=new.dt.larc,
                        method='MCMC',
                        MCMC.iterations=20000,
                        modelprior=uniform())

lm_larcenies
summary(lm_larcenies)

BPM_pred_larc =  predict(lm_larcenies, estimator="BPM", se.fit=TRUE)
bayes_var_larc<-lm_larcenies$namesx[BPM_pred_larc$bestmodel+1]
plot(lm_larcenies, which=4, ask=F)


#bayes test
BPM_pred_larc_test=predict(lm_larcenies, newdata = scaled.test.dtx,
                            estimator = 'BPM', se.fit=T)

mse_bay_larc<-mean((scaled.test.dty$larcPerPop-BPM_pred_larc_test$Ypred[1,])**2)

#linear model test
lm_larc_stand<-lm(larcenies~.,data = new.dt.larc)
lm_larc_stand<-lm(formula = larcenies ~ PopDens + racepctblack + racePctAsian +
                  householdsize + PctHousLess3BR + HousVacant + PctHousOwnOcc +
                  MedYrHousBuilt + agePct65up + pctWInvInc + pctWRetire + PctPopUnderPov +
                  PctBSorMore + PctUnemployed + PctEmploy + PctEmplProfServ +
                  TotalPctDiv + NumImmig + PctImmigRec10 + MedOwnCostPctInc,
                data = new.dt.larc)
summary(lm_larc_stand)


mse_lm_larc<-mean((scaled.test.dty$larcPerPop-predict(lm_larc_stand, newdata = scaled.test.dtx))**2)

#med imputation test
mse_med_larc<-mean((scaled.test.dty$larcPerPop-rep(median(as.numeric(scaled.train.dty$larcPerPop)),
                                                    dim(scaled.test.dty)[1]))**2)
```

0.5 이상의 변수가 생각보다 적어서
실제 사용시엔 0.5 미만의 변수도 선택함



3. PCA

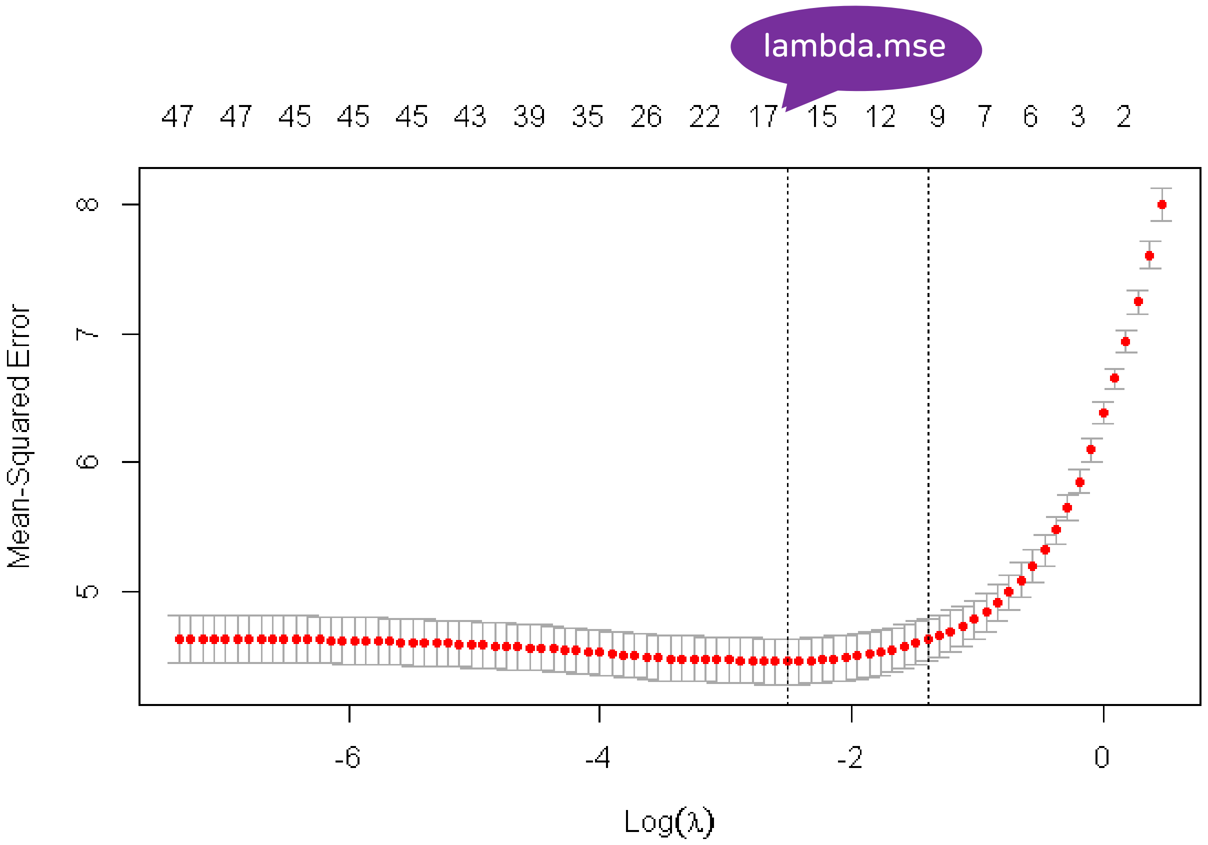


Component 10개 선택

```
     PC1      PC2      PC3      PC4      PC5      PC6      PC7      PC8 
1.003000 1.001178 1.002163 1.002354 1.003774 1.002005 1.001500 1.003328 
     PC9     PC10 
1.000948 1.002676 

Call:
lm(formula = murder ~ ., data = murder_pca_train)

Standardized Coefficients::
(Intercept)         PC1         PC2         PC3         PC4         PC5 
 0.00000000  0.59093655 -0.06528212  0.10348385 -0.01307535  0.22852739 
        PC6         PC7         PC8         PC9        PC10 
-0.13500549  0.05816879  0.03211252 -0.14949900 -0.02176842 
```

```
Call:
lm(formula = murder ~ ., data = murder_pca_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8676 -1.4366 -0.2704  1.5647  5.7369 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.42301    0.05997  -7.053 2.81e-12 ***
PC1          0.50257    0.01723  29.174  < 2e-16 ***
PC2         -0.06776    0.02100  -3.226  0.00129 ** 
PC3          0.14744    0.02885   5.111 3.67e-07 ***
PC4         -0.01901    0.02945  -0.646  0.51857    
PC5          0.38896    0.03449  11.278  < 2e-16 ***
PC6         -0.25144    0.03771  -6.668 3.79e-11 ***
PC7          0.11117    0.03868   2.874  0.00412 ** 
PC8          0.07426    0.04685   1.585  0.11318    
PC9         -0.37351    0.05055  -7.388 2.63e-13 ***
PC10        -0.05629    0.05237  -1.075  0.28263    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.093 on 1322 degrees of freedom
Multiple R-squared:  0.4592,	Adjusted R-squared:  0.4551 
F-statistic: 112.3 on 10 and 1322 DF,  p-value: < 2.2e-16

pca ~ y=murder mse : 4.504212
```

4. Lasso



최적 람다값으로 lambda.mse 선택

coef(cv_murder_lasso, s=cv_murder_lasso$lambda.min)

```
(Intercept)              -0.423381326
population                0.508425544
LandArea                  0.065466318
PopDens                   .
PctUsePubTrans            .
racepctblack              0.441841406
racePctWhite             -0.094762287
racePctAsian              .
racePctHisp               0.056529334
householdsize             .
PersPerRentOccHous        0.195251447
PctPersDenseHous          .
PctHousLess3BR            .
MedNumBR                 -0.020239569
HousVacant                0.022684460
PctHousOwnOcc             .
PctVacantBoarded          .
PctVacMore6Mos            .
MedYrHousBuilt            .
log.agePct12t29           .
agePct65up                .
pctUrban                  .
medIncome                 .
log.pctWFarmSelf          .
pctWInvInc               -0.178136508
pctWRetire               -0.015239269
PctPopUnderPov            0.007787827
PctLess9thGrade           .
PctBSorMore               .
PctUnemployed             .
PctEmploy                 .
PctEmplProfServ           .
PctOccupManu              .
MalePctNevMarr            .
TotalPctDiv               0.354592689
PctKids2Par              -0.572010437
PctWorkMom               -0.166451362
PctKidsBornNeverMar       .
NumImmig                  0.163679955
PctImmigRecent            .
PctImmigRec10             0.035248734
PctRecentImmig            .
PctWOFullPlumb            .
MedRentPctHousInc         .
MedOwnCostPctInc          .
MedOwnCostPctIncNoMtg     .
PctBornSameState          .
PctSameCity85             .
PctSameState85            .
```

5. Model MSE 비교

Imputation

	Bayes	lm	median	pca	Lasso
Rapes	1 229585.1	233093.7	245029.5	227433.9	226243.5
Robb	1 373348.2	378676.1	339494.7	367232.2	368062.8
Assault	1 385216.7	387188.8	381820.3	382104.1	380456.7
Burg	1 367396.6	372042.3	405636.3	362480.3	359970
Larg	1 313305	316558	425239.2	327773.4	309277.7
auto	1 377780.9	384298.1	374954.4	373611.8	374410.5
arsons	1 279218.1	276164	275986.5	274905	273179.2

* Murder은 데이터 결측치로 예상되는 0이 1000개 이상이라 제외함

02 Modeling

6. Y 변수 Imputation

murdPerPc	rapes	rapesPerPc	robberies	robbbPerP	assaults	assaultPerl	burglaries	burglPerPc	larcenies	larcPerPop	autoTheft	autoTheftP	arsons	arsonsPerP	ViolentCrir
0.03	2	2	3	1870	339	1048	127	146	218	44	138	263	50	281	1170
0.03	3	844	300	666	203	20	668	855	846	247	275	92	3	896	195
8.3	133	274	317	403	83	1216	374	1863	389	1783	89	1298	56	1258	681
4.63	153	1079	59	2002	371	977	281	377	1278	1805	442	1071	24	1509	1231
0.03	99	150	403	1000	421	570	41	1155	683	2200	107	1533	41	1187	704
13.13	86	1144	416	438	64	685	694	2195	531	1328	66	690	173	207	1217
0.03	87	944	90	629	345	1492	543	1468	737	1439	208	914	164	10	109
26.88	30	164	285	1295	14	2120	327	780	1133	1911	22	2171	45	294	393
3.11	74	1579	357	721	70	1274	586	427	397	1908	121	1508	140	359	1662
44.42	54	1427	211	123	309	292	597	624	261	1865	426	460	51	1394	818
6.54	91	1435	323	208	232	1634	675	274	383	1417	114	1150	173	386	1750
27.26	132	26	178	1362	418	2060	213	993	467	1199	479	2072	115	1420	352
2.19	2	2	154	1458	123	1189	383	1634	1392	437	244	1562	154	246	1873
5.02	79	278	238	614	239	414	177	2054	951	724	108	1979	157	940	1097
0.03	49	140	4	1702	403	1072	904	1695	909	852	168	167	101	597	86
2.39	31	1037	358	802	82	1430	511	354	97	1432	262	2009	179	773	1721
26.59	73	812	396	148	458	2001	907	349	562	1099	36	425	112	1256	642
12.89	167	1169	271	903	384	999	308	428	1068	1245	471	1148	122	753	1561
0.03	3	1445	2	2	123	302	722	1430	1105	1700	287	756	101	754	338
0.03	4	1140	122	51	491	1259	308	259	1233	1460	525	1220	50	65	1439
0.03	167	620	4	1059	363	341	321	1859	139	1433	456	432	140	380	611

03

지도 시각화

04

한계점

한계점

1) 모델들의 R-square 값들이 굉장히 안 좋음.

e.g.) rapes의 multivariate r-square는 0.1

-> 회귀 모델의 가정이 틀렸을 것.
-> y 변수들의 Skewness 처리를 해줬지만 정규분포화 되지 않는다.
-> Box-cox 등을 이용해서 Y 변수 처리를 새로 해줘야 할 듯

2) murder도 0인 애들이 결측치일 것.

-> 다른 변수들과 같이 없는 애들은 결측치일 확률이 높으므로 그런 애들은 결측치로 처리해서 Imputation

Q&A