STOR 767 HW 1

Brian N. White

10/6/2021

Problem 1: Regularization

```
crime.data_clean <- read.csv("CrimeData_clean.csv")
```

Our goal is to find the factors which relate to violent crime. This variable is included in crime as `crime.data$violentcrimes.perpop`.

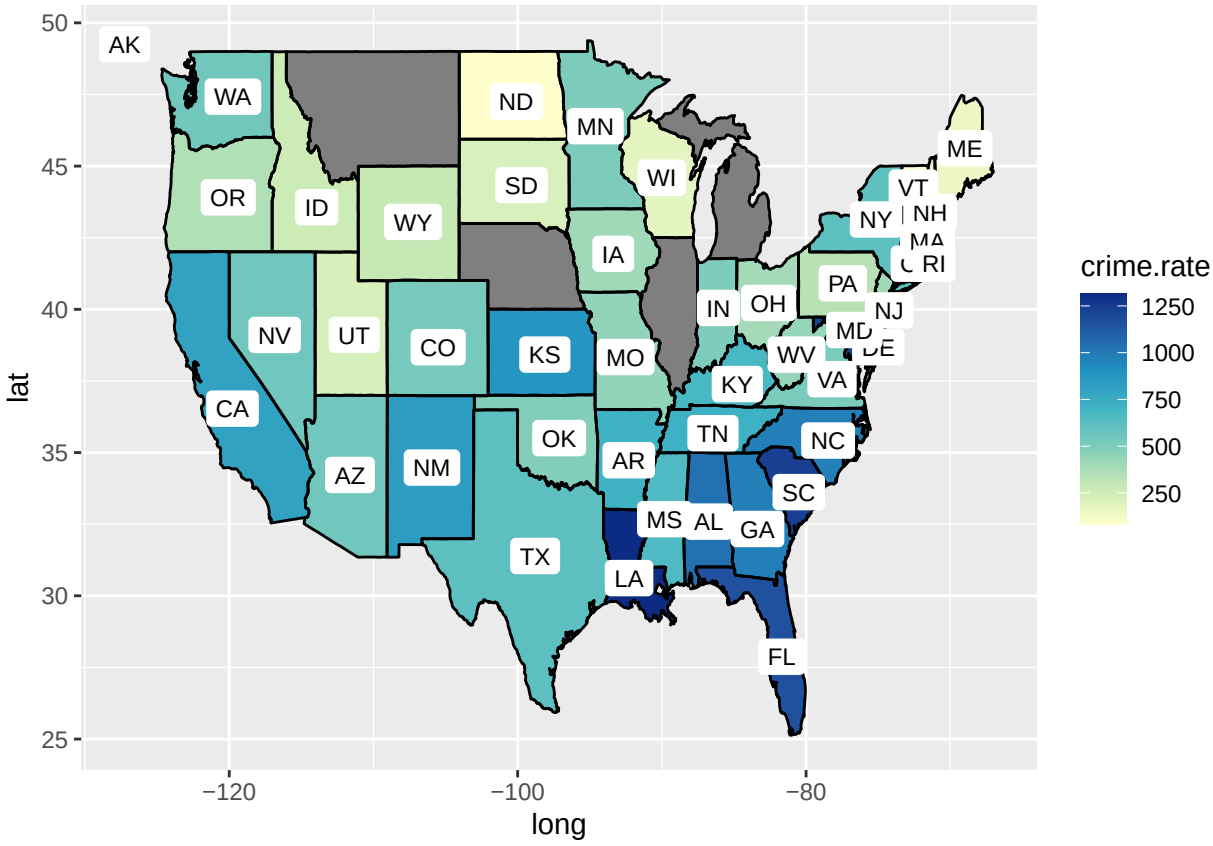
A) Exploratory data analysis (EDA)

- Show the heatmap with mean violent crime by state. You may also show a couple of your favorite summary statistics by state through the heatmaps.
- Write a brief summary based on your EDA.

The heat map in question is output by the code-chunk below. A brief summary of my EDA based upon this map is as follows: The most obvious pattern in the heatmap is that crime rates are greatest in the South-Eastern part of the United States. The crime rate appears to ease up in the center of the country, increase somewhat on the West Coast and decrease substantially the further north one goes, particularly in the Mid-West.

```
data.s <- crime.data_clean %>%
  group_by(state) %>%
  summarise(
    mean.income=mean(med.income),
    income.min=min(med.income),
    income.max=max(med.income),
    crime.rate=mean(violentcrimes.perpop, na.rm=TRUE), #ignore the missing values
    n=n())

crime <- data.s[, c("state", "crime.rate")]

crime$region <- tolower(state.name[match(crime$state, state.abb)])
crime$center_lat  <- state.center$x[match(crime$state, state.abb)]
crime$center_long <- state.center$y[match(crime$state, state.abb)]

states <- map_data("state")
map <- merge(states, crime, sort=FALSE, by="region", all.x=TRUE)
map <- map[order(map$order),]

ggplot(map, aes(x=long, y=lat, group=group))+
```

```
  geom_polygon(aes(fill=crime.rate))+
  geom_path()+
  geom_label(data=crime,
             aes(x=center_lat, y=center_long, group=NA, label=state),
             size=3, label.size = 0) +
  scale_fill_distiller(palette = "YlGnBu", direction = 1)
```



B) We use a subset of the crime data discussed in class, but only look at Florida and California.

```
crime.fl.ca <- dplyr::filter(crime.data_clean, state %in% c("FL", "CA"))
```

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 1se or lambda min to answer the following questions where apply.

1. What is the model reported by LASSO?

The model reported by LASSO includes the features, with corresponding non-zero parameter estimates, output by the code-chunk below. Note that the value of lambda min is output as well.

```
#part 1
Y <- crime.fl.ca$violentcrimes.perpop
X.fl.ca <- model.matrix(violentcrimes.perpop~., data=crime.fl.ca)[, -1]
```

```
set.seed(10)
fit.fl.ca.cv <- cv.glmnet(X.fl.ca, Y, alpha=1, nfolds=10)
names(fit.fl.ca.cv); summary(fit.fl.ca.cv)
```

```
##  [1] "lambda"     "cvm"        "cvsd"       "cvup"       "cvlo"
##  [6] "nzero"      "call"       "name"       "glmnet.fit" "lambda.min"
## [11] "lambda.1se" "index"
```

```
##            Length Class  Mode
## lambda     100    -none- numeric
## cvm        100    -none- numeric
## cvsd       100    -none- numeric
## cvup       100    -none- numeric
## cvlo       100    -none- numeric
## nzero      100    -none- numeric
## call         5    -none- call
## name         1    -none- character
## glmnet.fit  12    elnet  list
## lambda.min   1    -none- numeric
## lambda.1se   1    -none- numeric
## index        2    -none- numeric
```

```
fit.fl.ca.cv$lambda.min
```

```
## [1] 25.97896
```

```
coef.min <- coef(fit.fl.ca.cv, s="lambda.min")
coef.min <- coef.min[which(coef.min !=0),]
coef.min
```

```
##         (Intercept)       race.pctblack    pct.farmself.inc         pct.inv.inc
##        4.951761e+03        1.693249e+01       -1.617929e+01       -8.892600e-01
##        asian.percap    male.pct.divorce    pct.kids2parents         pct.workmom
##        2.604897e-03        2.386895e+01       -1.681319e+01       -4.928049e+00
## num.kids.nvrmarried pct.kids.nvrmarried    pct.english.only     pct.house.occup
##        8.940535e-04        5.851934e+01       -3.127582e+00       -1.814786e+00
##    pct.house.vacant  med.yr.house.built   pct.house.nophone     num.in.shelters
##        1.149354e+01       -1.420177e+00        7.272686e+00        8.562481e-02
```

2. What is the model after running OLS?

The model determined after running OLS on the features selected by LASSO in part 1, along with their corresponding parameter estimates, are output in a summary by the code-chunk below.

```
#part 2
var.min <- rownames(as.matrix(coef.min))
lm.input <- as.formula(paste("violentcrimes.perpop", "~", paste(var.min[-1], collapse = "+")))

fit.min.lm <- lm(lm.input, data=crime.fl.ca)
lm.output <- coef(fit.min.lm)
summary(fit.min.lm)
```

```
##
## Call:
## lm(formula = lm.input, data = crime.fl.ca)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1066.86  -177.95   -22.71   145.24  1949.88
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.181e+04  6.352e+03   1.860  0.06376 .
## race.pctblack         2.229e+01  3.428e+00   6.501 2.73e-10 ***
## pct.farmself.inc     -6.013e+01  4.385e+01  -1.371  0.17120
## pct.inv.inc          -4.781e+00  3.160e+00  -1.513  0.13119
## asian.percap          7.932e-03  2.530e-03   3.136  0.00186 **
## male.pct.divorce      3.896e+01  1.193e+01   3.264  0.00120 **
## pct.kids2parents     -1.303e+01  4.769e+00  -2.733  0.00660 **
## pct.workmom          -8.398e+00  3.696e+00  -2.272  0.02367 *
## num.kids.nvrmarried   1.592e-03  2.584e-03   0.616  0.53826
## pct.kids.nvrmarried   3.374e+01  1.798e+01   1.876  0.06146 .
## pct.english.only     -5.699e+00  2.158e+00  -2.641  0.00865 **
## pct.house.occup      -7.865e+00  3.979e+00  -1.977  0.04886 *
## pct.house.vacant      1.238e+01  1.094e+01   1.131  0.25876
## med.yr.house.built   -4.546e+00  3.187e+00  -1.427  0.15460
## pct.house.nophone     1.141e+01  1.088e+01   1.049  0.29489
## num.in.shelters       1.339e-01  8.088e-02   1.655  0.09878 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 350.7 on 352 degrees of freedom
## Multiple R-squared:  0.7333, Adjusted R-squared:  0.722
## F-statistic: 64.54 on 15 and 352 DF,  p-value: < 2.2e-16
```

3. What is your final model, after excluding high p-value variables? You will need to use model selection method to obtain this final model. Make it clear what criterion/criteria you have used and justify why they are appropriate.

From the summary table given in part 2 there are 7 features with significant t-tests. The features include 'race.pct.black', 'asian.percap', the divorce percentage among adult males, the percentage of kids living with two parents, the percentage of working moms, the percentage of people that only speak English and the percentage of houses occupied in the area. Thus, the final model is a linear model that includes these features.

In sum, this model was arrived at via the following two-step procedure. First, the full model with all possible features was reduced to a far smaller model via the feature selection performed by LASSO. Then, a linear model was fit from this reduced set of features and those features with insignificant t-tests were discarded (i.e. relaxed LASSO).

C) Now, instead of Lasso, we want to consider how changing the value of alpha (i.e. mixing between Lasso and Ridge) will affect the model. Cross-validate between alpha and lambda, instead of just lambda. 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 parsimoniousness.

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

I performed 10-fold CV over α and λ in the code-chunk below. The range of α considered was 0 to 1 in increments of 0.1. The range of λ is determined by the cv.glmnet command. My final elastic net model selected the 17 features output last by the code-chunk below. For this model, $\alpha = 0.5$ and $\lambda = 75.38206$. The prediction error is 122595.2

```
#part 1
set.seed(10)

#each entry of this vector stores the min mean CV error over lambda for a fixed alpha
model_results <- vector()

for(i in seq(0, 1, by=0.1)){
a <- cv.glmnet(X.fl.ca, Y, alpha=i, nfolds=10)
model_results <- c(model_results, min(a$cvm))}

model_results
```

```
##  [1] 148003.6 155930.2 154515.2 159298.8 156030.5 144743.4 168450.8 153424.5
##  [9] 150033.4 159023.9 151510.4
```

```
#estimated optimal prediction error
min(model_results)
```

```
## [1] 144743.4
```

```
#alpha=0.5 should be used
which.min(model_results)
```

```
## [1] 6
```

```
#the corresponding lambda min for this model is lambda=75.38206
cv.glmnet(X.fl.ca, Y, alpha=0.5, nfolds=10)$lambda.min
```

```
## [1] 75.38206
```

```
#elastic-net model alpha=0.5, lambda=75.38206
coef.min_elastic <- coef(cv.glmnet(X.fl.ca, Y, alpha=0.5, nfolds=10) , s="lambda.min")
coef.min_elastic <- coef.min_elastic[which(coef.min_elastic !=0),]
coef.min_elastic
```

```
##           (Intercept)         race.pctblack      pct.farmself.inc
##          6.310918e+03          1.567674e+01         -1.467001e+01
##           pct.inv.inc          asian.percap      male.pct.divorce
##         -1.441554e+00          1.879695e-03          2.095154e+01
##      pct.kids2parents pct.youngkids2parents           pct.workmom
##         -1.379802e+01         -2.849145e+00         -3.562882e+00
##   num.kids.nvrmarried   pct.kids.nvrmarried      pct.english.only
##          1.112345e-03          5.803571e+01         -2.298921e+00
##       pct.house.occup      pct.house.vacant    med.yr.house.built
##         -1.949967e+00          1.311227e+01         -2.138837e+00
##     pct.house.nophone       num.in.shelters       pct.foreignborn
##          7.911264e+00          7.566635e-02          5.059440e-01
```

```
#produce the prediction error for the chosen elastic net model
d <- cv.glmnet(X.fl.ca, Y, alpha=0.5, nfolds=10)
mean((Y-predict(d, s=d$lambda.min, X.fl.ca))^2)
```

```
## [1] 113364.8
```

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

The features selected by elastic net in part 1 are used in an OLS model as seen in the following code-chunk. None of these features are removed, even if their t-tests are insignificant, as I am optimizing for accuracy. These 7 features are race.pctblack, asian.percap, male.pct.divorce, pct.kids2parents, pct.workmom, pct.english.only, pct.house.occup. The prediction error for the final OLS model is 117001.9

```
#part 2
var.min_elastic <- rownames(as.matrix(coef.min_elastic))
lm.input2 <- as.formula(paste("violentcrimes.perpop", "~", paste(var.min_elastic[-1], collapse = "+")))

fit.min.lm2 <- lm(lm.input2, data=crime.fl.ca)
lm.output2 <- coef(fit.min.lm2)
summary(fit.min.lm)
```

```
##
## Call:
## lm(formula = lm.input, data = crime.fl.ca)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1066.86  -177.95   -22.71   145.24  1949.88
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.181e+04  6.352e+03   1.860  0.06376 .
## race.pctblack        2.229e+01  3.428e+00   6.501 2.73e-10 ***
## pct.farmself.inc    -6.013e+01  4.385e+01  -1.371  0.17120
## pct.inv.inc         -4.781e+00  3.160e+00  -1.513  0.13119
## asian.percap         7.932e-03  2.530e-03   3.136  0.00186 **
## male.pct.divorce     3.896e+01  1.193e+01   3.264  0.00120 **
## pct.kids2parents    -1.303e+01  4.769e+00  -2.733  0.00660 **
## pct.workmom         -8.398e+00  3.696e+00  -2.272  0.02367 *
## num.kids.nvrmarried  1.592e-03  2.584e-03   0.616  0.53826
## pct.kids.nvrmarried  3.374e+01  1.798e+01   1.876  0.06146 .
## pct.english.only    -5.699e+00  2.158e+00  -2.641  0.00865 **
## pct.house.occup     -7.865e+00  3.979e+00  -1.977  0.04886 *
## pct.house.vacant     1.238e+01  1.094e+01   1.131  0.25876
## med.yr.house.built  -4.546e+00  3.187e+00  -1.427  0.15460
## pct.house.nophone    1.141e+01  1.088e+01   1.049  0.29489
## num.in.shelters      1.339e-01  8.088e-02   1.655  0.09878 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 350.7 on 352 degrees of freedom
## Multiple R-squared:  0.7333, Adjusted R-squared:  0.722
## F-statistic: 64.54 on 15 and 352 DF,  p-value: < 2.2e-16
```

```r
lm_final_elastic <- lm(violentcrimes.perpop~ race.pctblack + asian.percap + male.pct.divorce + pct.kids

#the prediction error
mean((Y-predict(fit.min.lm2))^2)
```

```
## [1] 117001.9
```

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

In sum, I tuned an elastic net model over a grid of α and λ. The range of the λ values was pre-determined by the cv.glmnet command whereas the range of the α values was 0 to 1 by 0.1. The optimal choice of α and λ was determined to be $\alpha = 0.5$ and $\lambda = 75.38206$. The predictions errors for the elastic net and OLS models were, respectively, 122595.2 and 117001.9. Thus, the OLS model outperformed the elastic-net model w.r.t to this measure. The OLS and elastic-net model possess identical features; however, these feature are obtained in different manners. Elastic net minimizes RSS with a convex combination of L_1 and L_2 penalties. OLS is a special case of this regime with $\lambda = 0$.