Elastic Net

Wesley_Tao

2018/3/30

Data preprocessing

```r
Census_data<-read.csv("../data/Census data with variables 2016.csv",header = T,sep=",",as.is = T)
rate_2014  <-read.table("../output/rate_2014.tsv",header=T)
# first row seems to be the description of the variable
var_des    <-Census_data[1,] # store description
Census_data<-Census_data[-1,-1] # rest of the data,drop the first column
head(Census_data[,c(1,2,3,4,5,6)])
```

```
##   GEO.id2       GEO.display.label HC01_VC03 HC02_VC03 HC03_VC03 HC04_VC03
## 2   01001 Autauga County, Alabama     42712       218     42712       (X)
## 3   01003 Baldwin County, Alabama    160301       451    160301       (X)
## 4   01005 Barbour County, Alabama     21476        68     21476       (X)
## 5   01007    Bibb County, Alabama     18496       145     18496       (X)
## 6   01009  Blount County, Alabama     46007       136     46007       (X)
## 7   01011 Bullock County, Alabama      8547        57      8547       (X)
```

```r
library(dplyr)
# combine the data of social capital index with cencus
# head(Census_data)
# head(rate_2014)
rate_2014$id          <- as.integer(rate_2014$id)
Census_data$GEO.id2   <- as.integer(Census_data$GEO.id2)
names(rate_2014)      <-c("GEO.id2","SK14")

newdata<-rate_2014 %>%
                 left_join(Census_data,by="GEO.id2") # we  are interested in sk2014

setdiff(rate_2014$GEO.id2,Census_data$GEO.id2) # code  2270 46113 are mismatched
```

```
## [1]  2270 46113
```

```r
# we dig into the cencus data we found that HC03_VC03 is the population
#                                    and HC03_VC(##) is the ## variable divided by population
# therefore we select those variables HC03_VC(##)
X_pattern       <-"HC03_VC[0-9]+"
newdata_colnames<-names(newdata)
X_index         <-grep(X_pattern,newdata_colnames)
head(newdata_colnames[X_index]) # but HC03_VC_03 is the population
```

```
## [1] "HC03_VC03" "HC03_VC04" "HC03_VC05" "HC03_VC06" "HC03_VC07" "HC03_VC08"
```

```r
X_index<-X_index[-1] # remove the first HC03_VC_03
y_index<-2
subset_index<-c(y_index,X_index) # combine with y
```

```r
# we also find some variables have (X) we have to get rid of them
head(newdata[,"HC03_VC118"])
```

```
## [1] "(X)" "(X)" "(X)" "(X)" "(X)" "(X)"
```

```r
newdata.subset<-apply(newdata[,subset_index],2,as.numeric)# these '(X)' will be automatically handled by
# head(newdata.subset)
# it seems that we also have some variables which are population labor force, we have to get rid of them
# rules are if value is larger than 200 then
log.index<-apply(newdata.subset,2,mean,na.rm=T)<200
new_index<-which(log.index)

cleaned_data<-newdata.subset[,new_index]
cleaned_data<-na.omit(cleaned_data) # omit na by row
```

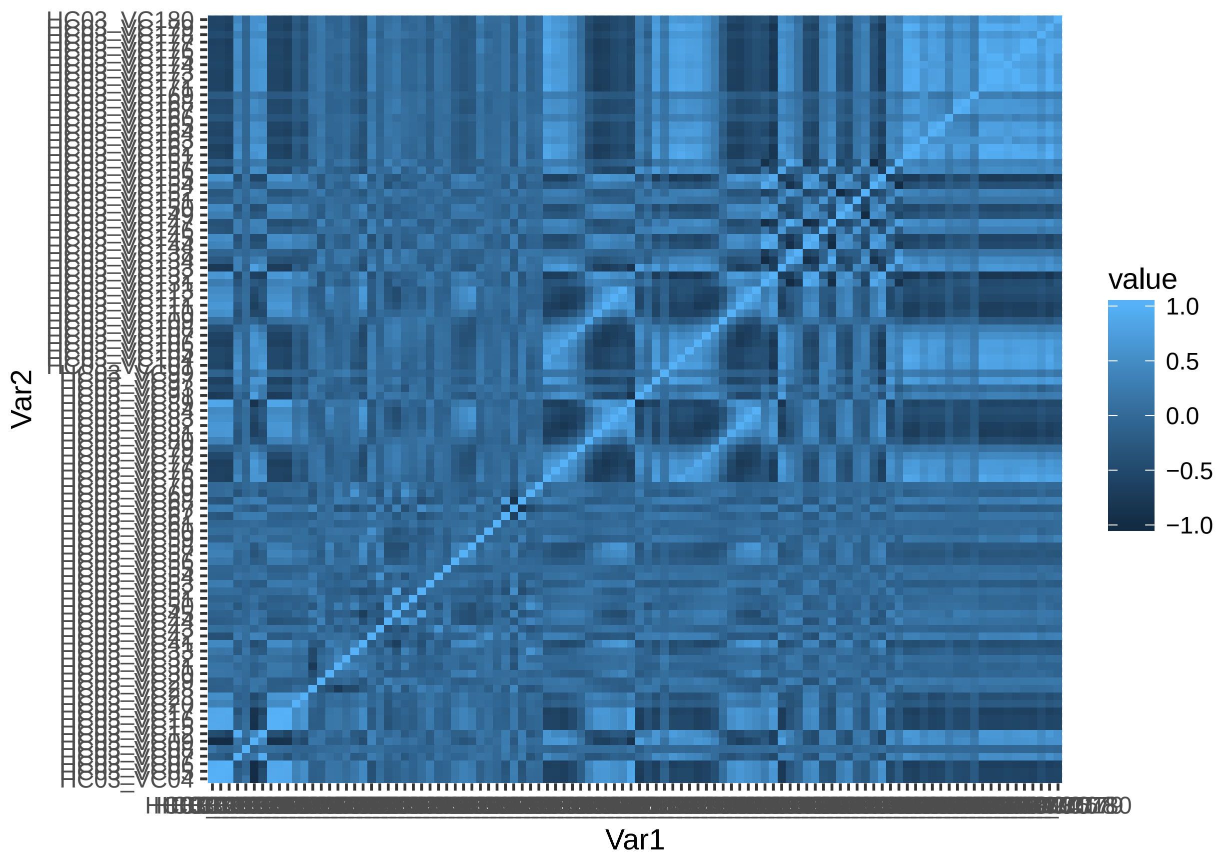
Train and test split

```r
# split the data for train and test
# 20% of the data as test data and 80 % as train data
set.seed(1)
total.num      <-nrow(cleaned_data)
test_index_row <-sample(seq(total.num),total.num*0.2)
train_index_row<-setdiff(seq(total.num),test_index_row)



X_test<-cleaned_data[test_index_row,-1]
y_test<-cleaned_data[test_index_row,1]

X_train<-cleaned_data[train_index_row,-1]
y_train<-cleaned_data[train_index_row,1]

length(test_index_row)+length(train_index_row)==total.num # make sure the number meet
```

```
## [1] TRUE
```

Heat map of variable correlation

```r
library(ggplot2)
library(reshape2)
cormat       <-round(cor(X_train),2)
melted_cormat<-melt(cormat) # transform to a narrow format

ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()
```




```
# it seems that this data set has a lot of collinearity, It might cause a lot of problem, we need regul
```

Elastic Net for variable selection

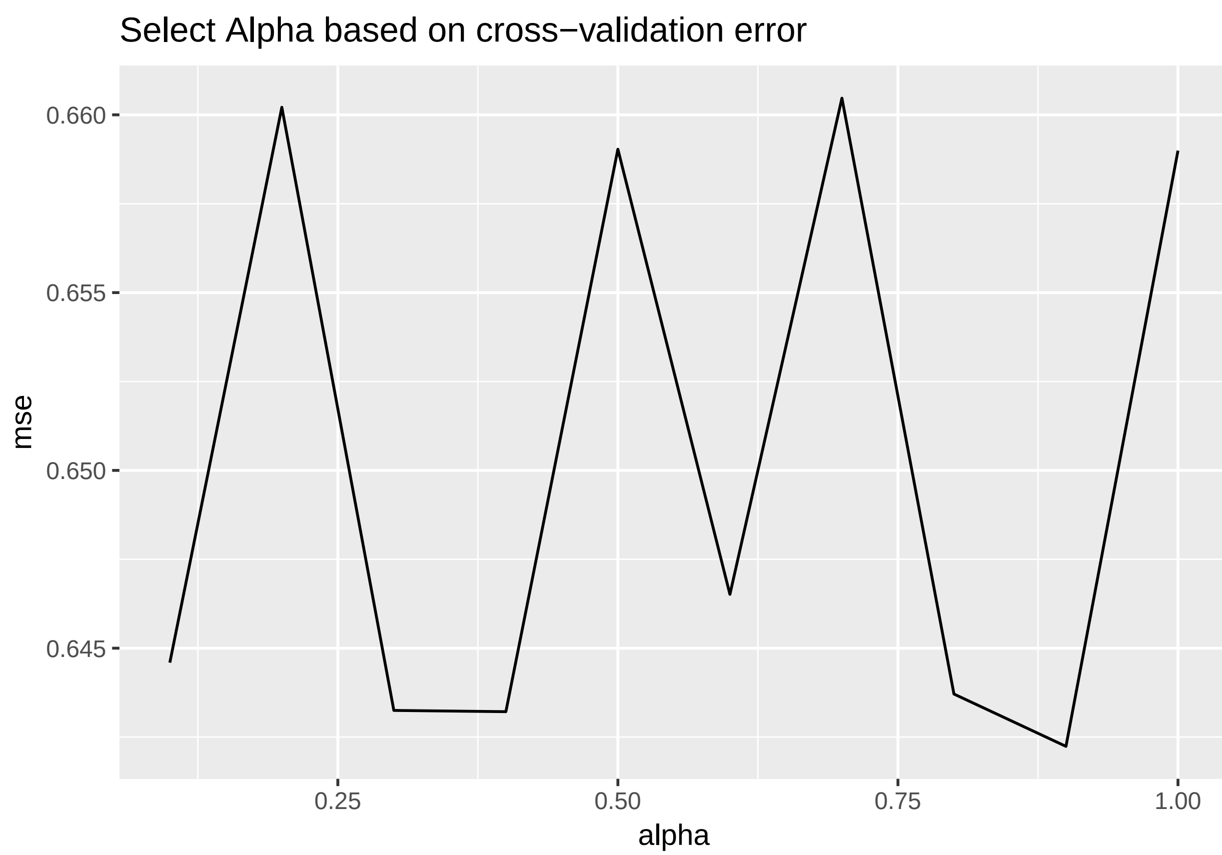
My original idea is to use lasso for variable selection, however , we can see clearly that some of the X variables are highly correlated thus regularizaiton of L2 penalty is needed. for more details about the loss function of the model:

$$\min_{\beta_0,\beta}\left(\frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2+\lambda[(1-\alpha)||\beta||_2^2/2+\alpha||\beta||_1)]\right)$$

There are two parameters, one is α another is λ . α is a compromise between Ridge($\alpha = 0$) and Lasso regression ($\alpha = 1$). We choose the ideal alpha based on 10-fold cross-validation. λ is the penalty we put on the parameter.

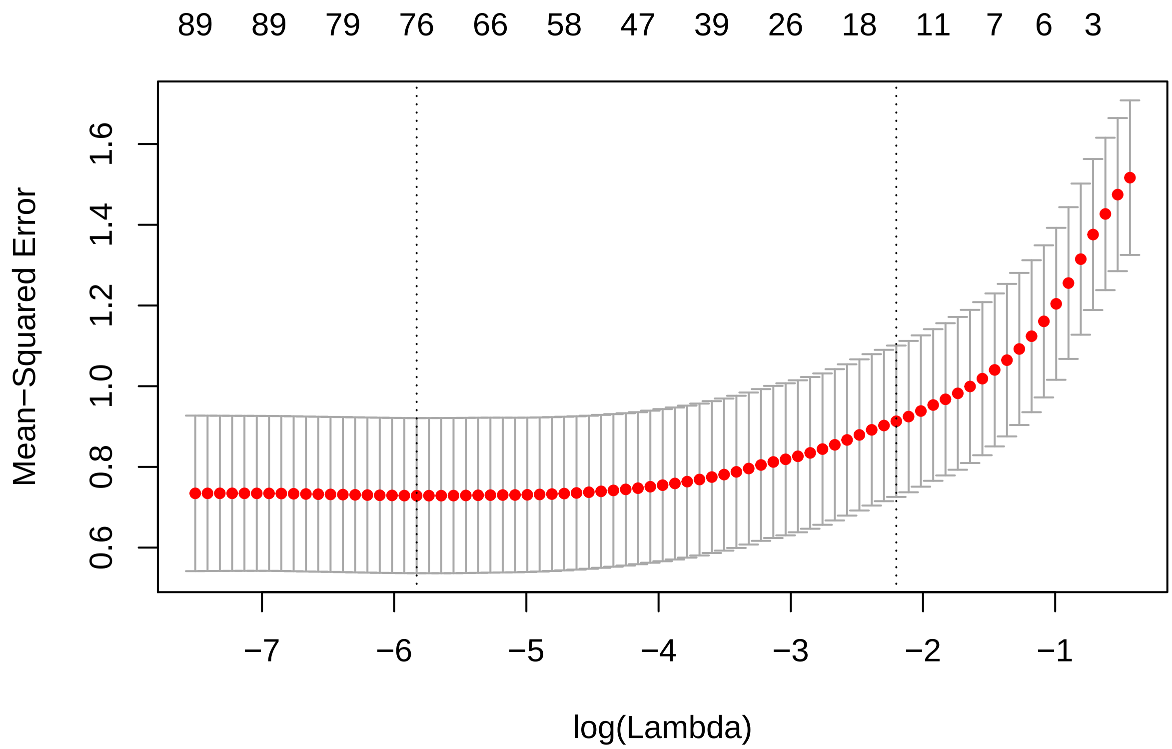
```
library(glmnet)
# 5-fold Cross validation for each alpha = 0, 0.1, ... , 0.9, 1.0
set.seed(4)
alpha_seq<-seq(10)/10

result<-data.frame(alpha=alpha_seq,mse=rep(NA,length(alpha_seq)))# generate a data frame to store the r

for(i in 1:length(alpha_seq)){
  this.alpha <-alpha_seq[i]
  fit.reg.cv <- cv.glmnet(X_train, y_train,nfold=10, type.measure="mse", alpha=this.alpha,
```

```r
                                family="gaussian")
  this.lambda<-fit.reg.cv$lambda.min # we choose lambda with the minimun mse

  yhat        <-predict(fit.reg.cv, s=this.lambda, newx=X_train) # get pred over the train set
  mse         <- mean((y_train - yhat)^2) #compute mse
  result[i,2]<-mse #

}

# plot mse over alpha
ggplot(data = result,aes(x=alpha,y=mse))+
  geom_line()+
  labs(title="Select Alpha based on cross-validation error")
```



it seems that lasso perform better than ridge regression. based on the pics above We are going to select alpha=0.9

```r
set.seed(1)
fit.ela.cv <- cv.glmnet(X_train, y_train, type.measure="mse", alpha=0.9,
                                family="gaussian",nfolds = 5)
plot(fit.ela.cv)
```




```r
yhat <- predict(fit.ela.cv, s=fit.ela.cv$lambda.1se, newx=X_test)
test_mse <- mean((y_test - yhat)^2)
```

```r
test_mse
```

```
## [1] 0.613089
```

But based on the graph above, we can select top 10 variables which might have statistical significant influence over the social capital index

```r
# we are interested in those variabels
head(data.frame(n=fit.ela.cv$nzero,lambda=fit.ela.cv$lambda),14)
```

```
##      n    lambda
## s0   0 0.6478059
## s1   1 0.5902566
## s2   1 0.5378199
## s3   3 0.4900414
## s4   5 0.4465075
## s5   5 0.4068410
## s6   5 0.3706984
## s7   6 0.3377666
## s8   6 0.3077603
## s9   6 0.2804197
## s10  6 0.2555080
## s11  7 0.2328094
## s12  7 0.2121272
## s13  9 0.1932824
```

Based on the table above, I am going to select top 8 variables lambda=0.1932824

```r
# use full data
final_model<-glmnet(cleaned_data[,-1], cleaned_data[,1], alpha=0.9, lambda=0.1932824,
                    family="gaussian")
head(final_model$beta,10) # as we can see many variables are shrink to seros
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                     s0
## HC03_VC04  .
## HC03_VC05  .
## HC03_VC06  .
## HC03_VC07 -0.05212367
## HC03_VC08  .
## HC03_VC09  .
## HC03_VC12  .
## HC03_VC15  .
## HC03_VC16  .
## HC03_VC17  .
```

we are interested in those non-zero varariables

```r
myvar<-which(final_model$beta!=0)
# final_model$beta
x_names<-colnames(cleaned_data)[-1]
data.frame(code=x_names[myvar],coe=final_model$beta[myvar])
```

```
##        code           coe
## 1  HC03_VC07 -5.212367e-02
## 2  HC03_VC23  1.934705e-02
## 3  HC03_VC50  1.126535e-02
## 4  HC03_VC69  9.183745e-02
## 5 HC03_VC101 -4.350882e-05
## 6 HC03_VC131  2.422074e-03
## 7 HC03_VC132  1.600232e-03
## 8 HC03_VC134 -3.506262e-03
## 9 HC03_VC164 -3.522046e-02
```

we are also interested what are these

```r
dex_index<-which(colnames(var_des) %in% x_names[myvar])
t(var_des[,dex_index])
```

```
##            1
## HC03_VC07  "Percent; EMPLOYMENT STATUS - Population 16 years and over - In labor force - Civilian la
## HC03_VC23  "Percent; EMPLOYMENT STATUS - Own children of the householder 6 to 17 years - All parents
## HC03_VC50  "Percent; INDUSTRY - Civilian employed population 16 years and over - Agriculture, foresti
## HC03_VC69  "Percent; CLASS OF WORKER - Civilian employed population 16 years and over - Self-employe
## HC03_VC101 "Percent; INCOME AND BENEFITS (IN 2016 INFLATION-ADJUSTED DOLLARS) - With Food Stamp/SNAP
## HC03_VC131 "Percent; HEALTH INSURANCE COVERAGE - Civilian noninstitutionalized population - With heal
## HC03_VC132 "Percent; HEALTH INSURANCE COVERAGE - Civilian noninstitutionalized population - With heal
## HC03_VC134 "Percent; HEALTH INSURANCE COVERAGE - Civilian noninstitutionalized population - No health
## HC03_VC164 "Percent; PERCENTAGE OF FAMILIES AND PEOPLE WHOSE INCOME IN THE PAST 12 MONTHS IS BELOW T
```

Conclusion

We found several varibles which are highly correlated to the social capital index, they are listed below.

1. Employment status (+)

2. Unemployed rate (-)
3. Health Insurance Coverage (+)
4. No Health cover (-)
5. Below Poverty (-)