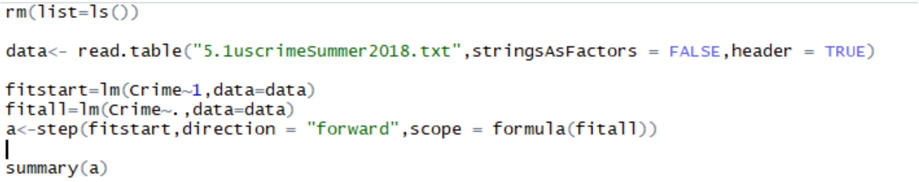
Homework 5

Question 11.1

Using the crime data set uscrime.txt from Questions 8.2, 9.1, and 10.1, build a regression model using: 1. Stepwise regression 2. Lasso 3. Elastic net For Parts 2 and 3, remember to scale the data first – otherwise, the regression coefficients will be on different scales and the constraint won't have the desired effect. For Parts 2 and 3, use the glmnet function in R

Answer:

For stepwise regression, this being a greedy algorithm, which means at each step it reevaluates itself and chooses the best path given what it's collected thus far. Stepwise regression has 3 main approaches for model selection , forward , backward or bidirectional. We'll look at forward selection, which starts out with no variable in the model. It will test an addition of each variable and chooses to keep any variable is the most statistically significant improvement to the fit. It will repeat this until no more improvement can be added.

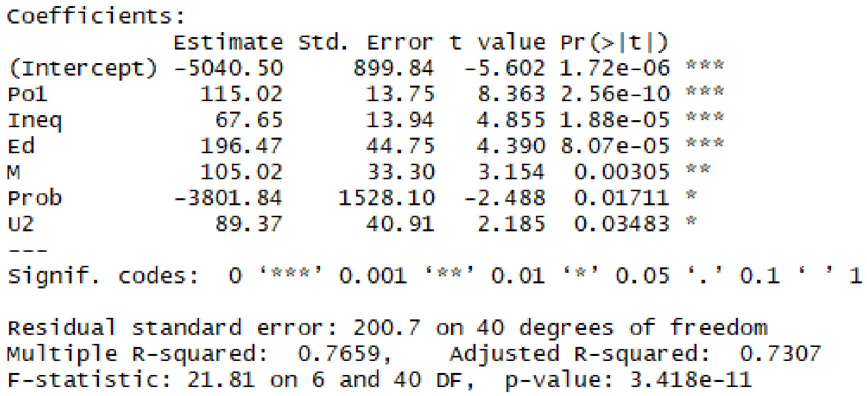


```
rm(list=ls())

data<- read.table("5.1uscrimeSummer2018.txt",stringsAsFactors = FALSE,header = TRUE)

fitstart=lm(Crime~1,data=data)
fitall=lm(Crime~.,data=data)
a<-step(fitstart,direction = "forward",scope = formula(fitall))

summary(a)
```

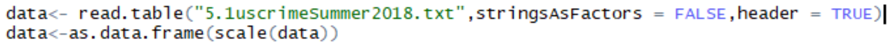


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5040.50     899.84  -5.602 1.72e-06 ***
Po1           115.02      13.75   8.363 2.56e-10 ***
Ineq           67.65      13.94   4.855 1.88e-05 ***
Ed            196.47      44.75   4.390 8.07e-05 ***
M             105.02      33.30   3.154  0.00305 ** 
Prob        -3801.84    1528.10  -2.488  0.01711 *  
U2             89.37      40.91   2.185  0.03483 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 200.7 on 40 degrees of freedom
Multiple R-squared:  0.7659,	Adjusted R-squared:  0.7307 
F-statistic: 21.81 on 6 and 40 DF,  p-value: 3.418e-11
```

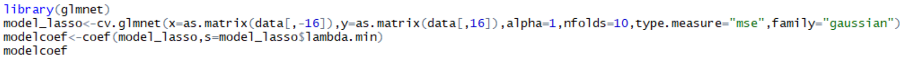
Observe that Po1,Ineq,Ed,M,Prob,U2, are all good variable to include in the model, and exclude all the other ones since the p value is quite low.

For Lasso, we'll load the data and scale it like so.

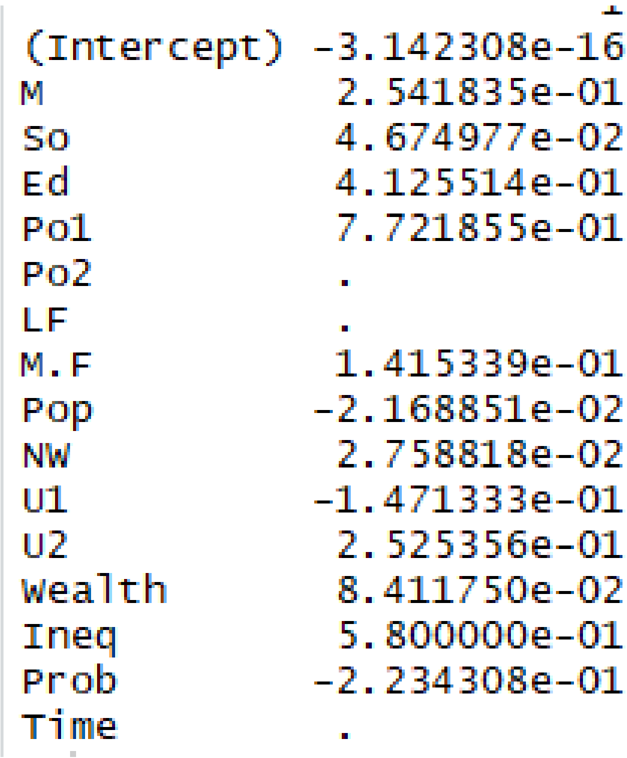


```
data<- read.table("5.1uscrimeSummer2018.txt",stringsAsFactors = FALSE,header = TRUE)
data<-as.data.frame(scale(data))
```

We'll do Lasso and have a k-fold cross validation with k=10. And we'll evaluate variable is the most important by their MSE, or Mean squared error value.

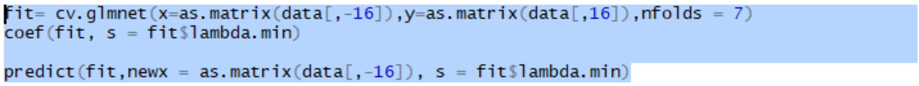


```
library(glmnet)
model_lasso<-cv.glmnet(x=as.matrix(data[,-16]),y=as.matrix(data[,16]),alpha=1,nfolds=10,type.measure="mse",family="gaussian")
modelcoef<-coef(model_lasso,s=model_lasso$lambda.min)
modelcoef
```



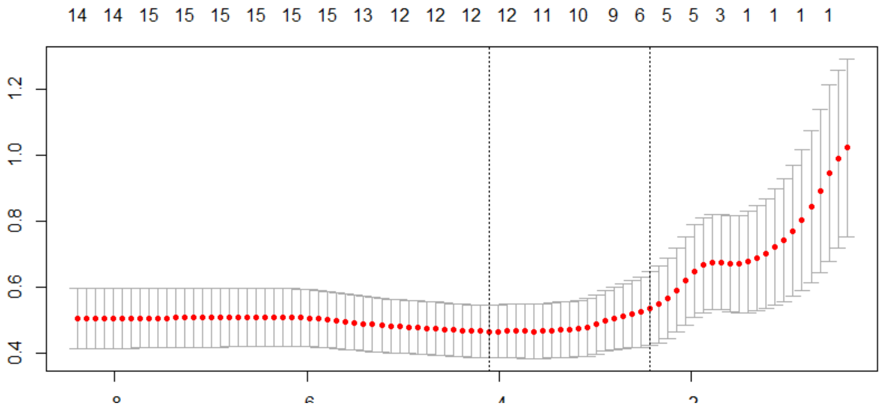
```
                        1
(Intercept) -3.142308e-16
M            2.541835e-01
So           4.674977e-02
Ed           4.125514e-01
Po1          7.721855e-01
Po2          .
LF           .
M.F          1.415339e-01
Pop         -2.168851e-02
NW           2.758818e-02
U1          -1.471333e-01
U2           2.525356e-01
Wealth       8.411750e-02
Ineq         5.800000e-01
Prob        -2.234308e-01
Time         .
```

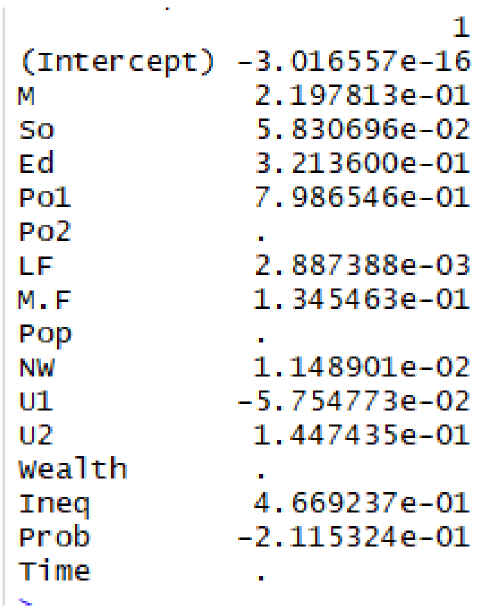
As we can see above, we got rid of P02, LF and also Time by using Lasso.



```
fit= cv.glmnet(x=as.matrix(data[,-16]),y=as.matrix(data[,16]),nfolds = 7)
coef(fit, s = fit$lambda.min)

predict(fit,newx = as.matrix(data[,-16]), s = fit$lambda.min)
```





```
                        1
(Intercept) -3.016557e-16
M            2.197813e-01
So           5.830696e-02
Ed           3.213600e-01
Po1          7.986546e-01
Po2          .
LF           2.887388e-03
M.F          1.345463e-01
Pop          .
NW           1.148901e-02
U1          -5.754773e-02
U2           1.447435e-01
Wealth       .
Ineq         4.669237e-01
Prob        -2.115324e-01
Time         .
```

Finally for Elastic net, We're able to see that Po2,Po1,and Time was excluded. And that seems to be the main for from using global regularize method vs the greedy method.

Question 12.1 Describe a situation or problem from your job, everyday life, current events, etc., for which a design of experiments approach would be appropriate.

Answer:

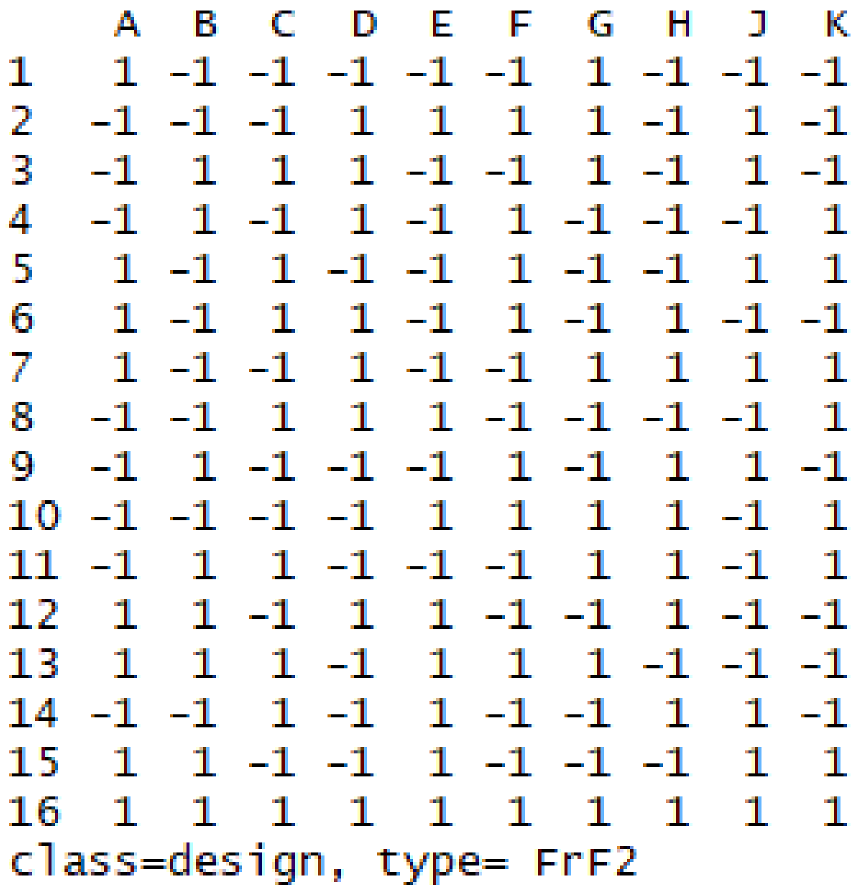
Say that I want to know the effectiveness of a certain pill that claims it will help people lose weight. So, I design an experiment where we gather a group of people, let's say 200 people, split them up evenly into two group, control group and treatment group. Where one group gets placebo and the other gets the treatment. Each person will not know if they are getting the real treatment or placebo. This design of experiment is aiming to describe the variation of the treatment, for example we can see if there exist any side-effect in the treatment, all while controlling for other variables or factors.

Question 12.2

To determine the value of 10 different yes/no features to the market value of a house (large yard, solar roof, etc.), a real estate agent plans to survey 50 potential buyers, showing a fictitious house with different combinations of features. To reduce the survey size, the agent wants to show just 16 fictitious houses. Use R's FrF2 function (in the FrF2 package) to find a fractional factorial design for this experiment: what set of features should each of the 16 fictitious houses have? Note: the output of FrF2 is "1" (include) or "-1" (don't include) for each feature.

Answer:

Since we are trying to reduce the number of rows in our fractional factorial design, we are going from 50 to 16 in order to save time. Each of the entry is binary, -1 or 1, although we can assign to mean different things, such as "off" or "on", "high" or "low" etc. For this problem, we'll use library(FrF2) and then type FrF2(16,10) and we'll have an output that will look like this. Since it's random, the output will look different each time you compile.



```
    A  B  C  D  E  F  G  H  J  K
1   1 -1 -1 -1 -1 -1  1 -1 -1 -1
2  -1 -1 -1  1  1  1  1 -1  1 -1
3  -1  1  1  1 -1 -1  1 -1  1 -1
4  -1  1 -1  1 -1  1 -1 -1 -1  1
5   1 -1  1 -1 -1  1 -1 -1  1  1
6   1 -1  1  1 -1  1 -1  1 -1 -1
7   1 -1 -1  1 -1 -1  1  1  1  1
8  -1 -1  1  1  1 -1 -1 -1 -1  1
9  -1  1 -1 -1 -1  1 -1  1  1 -1
10 -1 -1 -1 -1  1  1  1  1 -1  1
11 -1  1  1 -1 -1 -1  1  1 -1  1
12  1  1 -1  1  1 -1 -1  1 -1 -1
13  1  1  1 -1  1  1  1 -1 -1 -1
14 -1 -1  1 -1  1 -1 -1  1  1 -1
15  1  1 -1 -1  1 -1 -1 -1  1  1
16  1  1  1  1  1  1  1  1  1  1
class=design, type= FrF2
```

Question 13.1 For each of the following distributions, give an example of data that you would expect to follow this distribution (besides the examples already discussed in class).

Answer:

a. Binomial The probability of a player winning X many matches given he wins with p in a poker game will follow a Binomial Distribution.

b. Geometric

In the game of Magic the Gathering card game, a player draws 7 cards in the beginning of the game and draws 1 cards per turn, and an essential part of the game is to have land cards, in order to cast spells. P=.3, out of the 7 cards, how many of them are lands cards?

c. Poisson

A typical fast food worker makes a burger every 3 mins on average. Finding the probability that the worker finishes it in 4 mins would make /lambda equal to 3, and the distribution would follow Poisson.

d. Exponential

The time it takes before a person's next phone call.

e. Weibull

Weibull distribution is analytical tool for modeling the strength of material. So, we could be interested when a certain compound could be hazardous. Or the decay of a material before it becomes hazardous, for example the steel pipes that are holding up roller coaster rides.