snewt
ISYE6501x - HW#5
Due June 21, 2018

Question 11.1

Stepwise regression

```
> library(tidyverse)
> library(caret)
> library(leaps)
> library(MASS)
> full.model<-lm(crime$Crime~.,data=crime)
> summary(full.model)

Call:
lm(formula = crime$Crime ~ ., data = crime)

Residuals:
    Min      1Q  Median      3Q     Max
-395.74  -98.09   -6.69  112.99  512.67

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.984e+03  1.628e+03  -3.675 0.000893 ***
M            8.783e+01  4.171e+01   2.106 0.043443 *
So          -3.803e+00  1.488e+02  -0.026 0.979765
Ed           1.883e+02  6.209e+01   3.033 0.004861 **
Po1          1.928e+02  1.061e+02   1.817 0.078892 .
Po2         -1.094e+02  1.175e+02  -0.931 0.358830
LF          -6.638e+02  1.470e+03  -0.452 0.654654
M.F          1.741e+01  2.035e+01   0.855 0.398995
Pop         -7.330e-01  1.290e+00  -0.568 0.573845
NW           4.204e+00  6.481e+00   0.649 0.521279
U1          -5.827e+03  4.210e+03  -1.384 0.176238
U2           1.678e+02  8.234e+01   2.038 0.050161 .
Wealth       9.617e-02  1.037e-01   0.928 0.360754
Ineq         7.067e+01  2.272e+01   3.111 0.003983 **
Prob        -4.855e+03  2.272e+03  -2.137 0.040627 *
Time        -3.479e+00  7.165e+00  -0.486 0.630708
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 209.1 on 31 degrees of freedom
Multiple R-squared:  0.8031,  Adjusted R-squared:  0.7078
F-statistic: 8.429 on 15 and 31 DF,  p-value: 3.539e-07
```

So we've got a model where the significant predictors are M, Ed, Ineq, and Prob. It's a pretty decent model to start with, ***R^2 = .8031***. *F(15, 31)* = 8.429, *p* = 3.539e-07 (*p* << .001).

Now we're going to do the actual stepwise regression.

```
> step.model<-stepAIC(full.model, direction = "both",trace = FALSE)
> summary(step.model)

Call:
lm(formula = crime$Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq +
    Prob, data = crime)

Residuals:
    Min      1Q  Median      3Q     Max
-444.70 -111.07    3.03  122.15  483.30

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6426.10    1194.61  -5.379 4.04e-06 ***
M              93.32      33.50   2.786  0.00828 **
Ed            180.12      52.75   3.414  0.00153 **
Po1           102.65      15.52   6.613 8.26e-08 ***
M.F            22.34      13.60   1.642  0.10874
U1          -6086.63    3339.27  -1.823  0.07622 .
U2            187.35      72.48   2.585  0.01371 *
Ineq           61.33      13.96   4.394 8.63e-05 ***
Prob        -3796.03    1490.65  -2.547  0.01505 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 195.5 on 38 degrees of freedom
Multiple R-squared:  0.7888,  Adjusted R-squared:  0.7444
F-statistic: 17.74 on 8 and 38 DF,  p-value: 1.159e-10
```

So our stepwise model gets actually more significant predictors: M, Ed, Po1, U2, Ineq, and Prob. We've got less of a good model though, **R^2 = .7888,** and *F*(8,38) = 17.74, *p* = 1.159e-10 (*p* << .001), but it's still significant.

To get here, we followed the steps on http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/

snewt
ISYE6501x - HW#5
Due June 21, 2018

Lasso

We need package glmnet to get started.
> mcrime <- as.matrix(crime[,-1])

So we made it a matrix (and got rid of the M variable?)

```
> lassocrime <- glmnet(mcrime,crime$Crime,family="mgaussian",alpha=1)
> summary(lassocrime)
          Length Class     Mode
a0         39    -none-    numeric
beta      585    dgCMatrix S4
df         39    -none-    numeric
dim         2    -none-    numeric
lambda     39    -none-    numeric
dev.ratio  39    -none-    numeric
nulldev     1    -none-    numeric
npasses     1    -none-    numeric
jerr        1    -none-    numeric
offset      1    -none-    logical
call        5    -none-    call
nobs        1    -none-    numeric
```

Our computer restarted, and we had to redo R again ---

> mcrimedata<-as.matrix(crimedata[,1:15])

snewt
ISYE6501x - HW#5
Due June 21, 2018

First, I know we needed crime data as a matrix.

> glmnet(mcrimedata,crimedata$Crime,family = "mgaussian",standardize = TRUE)

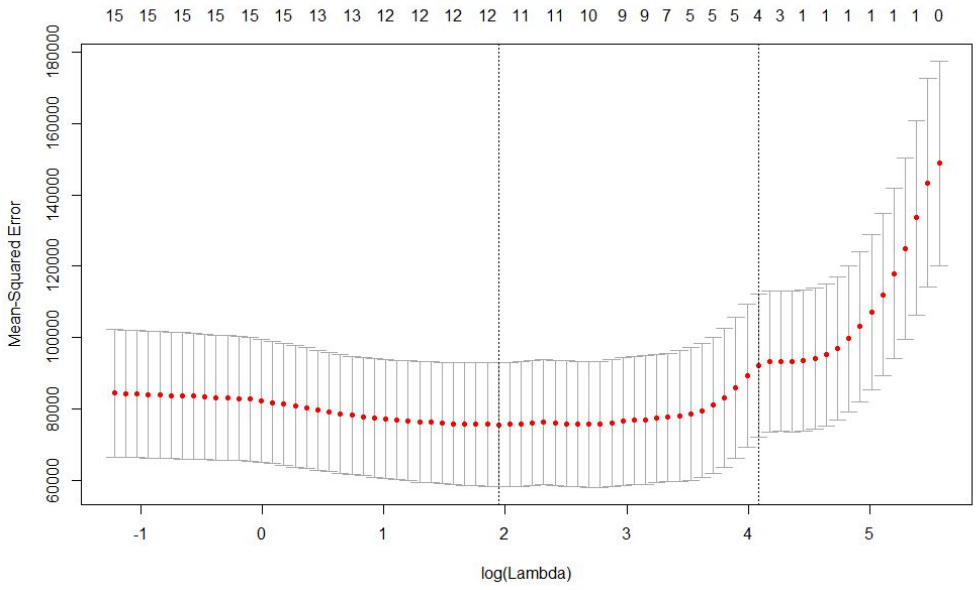
Call: glmnet(x = mcrimedata, y = crimedata$Crime, family = "mgaussian", standardize = TRUE)

	Df	%Dev	Lambda
[1,]	0	0.00000	263.10000
[2,]	1	0.08027	239.70000
[3,]	1	0.14690	218.40000
[4,]	1	0.20220	199.00000
[5,]	1	0.24820	181.30000
[6,]	1	0.28630	165.20000
[7,]	1	0.31800	150.60000
[8,]	1	0.34430	137.20000
[9,]	1	0.36610	125.00000
[10,]	1	0.38420	113.90000

It keeps going, for 99 items.

How about a cross validated one:

> cv.out <- cv.glmnet(mcrimedata,crimedata$Crime,alpha=1)
> View(cv.out)
> plot(cv.out)



```
> best_lam
[1] 6.988045
> cv.out$lambda.1se
[1] 59.38115
```

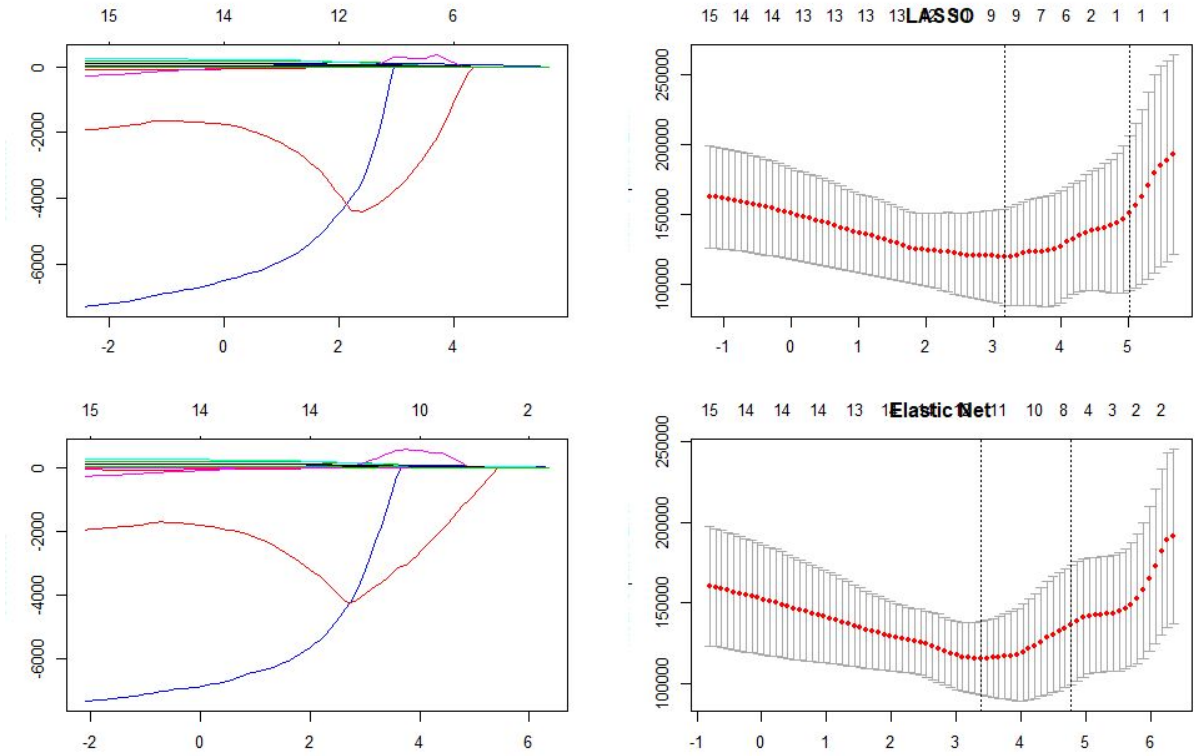
Okay let's try another approach from
https://www4.stat.ncsu.edu/~post/josh/LASSO_Ridge_Elastic_Net_-_Examples.html

```
> n<-nrow(mcrimedata)
> trainrows<-sample(1:n, .66*n)
> x<-mcrimedata
> y<-crimedata$Crime
> x.train<-x[trainrows,]
> x.test<-x[-trainrows,]
> y.train<-y[trainrows,]
Error in y[trainrows, ] : incorrect number of dimensions
> length(y)
[1] 47
> y.train<-y[trainrows]
> y.test<-y[-trainrows]
```

So now we have our test, train sets

```
> fit.lasso <- glmnet(x.train, y.train, family="gaussian", alpha=1)
> fit.elnet <- glmnet(x.train, y.train, family="gaussian", alpha=.5)
> for (i in 0:10) {
+     assign(paste("fit", i, sep=""), cv.glmnet(x.train, y.train, type.measure="mse",
+                                               alpha=i/10,family="gaussian"))
> par(mfrow=c(2,2))
> plot(fit.lasso, xvar="lambda")
> plot(fit10, main="LASSO")
>
> plot(fit.elnet, xvar="lambda")
> plot(fit5, main="Elastic Net")
```

snewt
ISYE6501x - HW#5
Due June 21, 2018



Elastic Net

Plots above. Script also above. Anything with label elnet was for elastic net, lasso was lasso.

Question 12.1

A situation for which a design of experiments would be appropriate: for serious? Ah, sure. Okay. My partner and I are looking to try to either build or buy a house within the next five years. There are arguments to be made for modular homes, for stick building, for staying in rural Virginia, for moving to my grew-up home of Atlanta, for moving to a beach somewhere. Some variables could include combined income at location, expected cost of living adjustment, price of home, number of bedrooms, process of acquiring land. We could eventually ANOVA those items to try and find the best situation in terms of price alone.

Question 12.2

First I'd like to point out that this experiment is a mess already, with a sample size of only 50 potential buyers you can have 10 fake houses, hard max.

Okay so we want to use the FrF2 package, function to find a fractional factorial design, with what set of features?

So we have nruns = 16 (a power of 2, as required) for 16 houses, we have 10 features so nfactors = 10.

```
>install.packages("FrF2")
>library(FrF2)
> FrF2(16,10)
   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
```

That's our 16 houses to show.

Question 13.1

Examples of distributions of data:

A. Binomial
    a. I record the likelihood that the light in my office is on or off as I walk in every hour on the hour from 6 am to 3 pm on weekdays, repeated for a year
B. Geometric
    a. 80% of cabs are already occupied when it is raining. It's raining. You still try to hail a cab as each one passes you. X is the number of cabs you try before successfully hailing one.
C. Poisson
    a. Exam grades or lab grades at the university of georgia in a chemistry lab for one student's masters' thesis
D. Exponential
    a. Affordability of solar panel technologies & their implementations - it's getting more and more affordable as technology methods become more

and more widely disseminated throughout the bright minds who can put it to use.

E. Weibull
    a. Weird but true I used to work for a medical device manufacturer, and we made suture there. We measured the tensile strength of suture, and we expected that to be a Weibull distribution - how much force was required before the suture broke. It's a time to failure measure.