Q1.A.1:- step wise backward.

1. 1st step:
    a. After 1st run with all columns I get the AIC of 513. Also when I look, the reading says if I remove "So" then the AIC will be 511. So the model is going drop "So" and run it.

```
> model = step(lm(Crime~.,data=data.sm,direction="backward"))
Start:  AIC=513
Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 +
    U2 + Wealth + Ineq + Prob

          Df Sum of Sq      RSS    AIC
- So       1        65 1365315 511.01
- LF       1      7553 1372802 511.26
- NW       1     12227 1377477 511.42
- Pop      1     21758 1387008 511.75
- Po2      1     28851 1394101 511.99
- Wealth   1     33688 1398938 512.15
- M.F      1     39812 1405062 512.36
<none>                 1365250 513.00
- U1       1     77378 1442628 513.60
- Po1      1    134191 1499441 515.41
- U2       1    174523 1539773 516.66
- M        1    184513 1549763 516.96
- Prob     1    222620 1587870 518.10
- Ed       1    407072 1772322 523.27
- Ineq     1    426138 1791388 523.77
```

2. See results below. Model is run without column "So". Now the AIC is 513. Also the below model says that If it drops Column "LF" the AIC will be 509.37. So the stepwise algo is going to drop column by column until There is no column that is potentially reduce the AIC of the Model.

```
Step:  AIC=511.01
Crime ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 +
    Wealth + Ineq + Prob

          Df Sum of Sq      RSS    AIC
- LF       1     10533 1375848 509.37
- NW       1     15482 1380797 509.54
- Pop      1     21846 1387161 509.75
- Po2      1     28932 1394247 509.99
- Wealth   1     36070 1401385 510.23
- M.F      1     41784 1407099 510.42
<none>                 1365315 511.01
- U1       1     91420 1456735 512.05
- Po1      1    134137 1499452 513.41
- U2       1    184143 1549458 514.95
- M        1    186110 1551425 515.01
- Prob     1    237493 1602808 516.54
- Ed       1    409448 1774763 521.33
- Ineq     1    502909 1868224 523.75
```

3. See below results of various iterations that the model has done. Dropping one by one

```
.Step:  AIC=509.37
Crime ~ M + Ed + Po1 + Po2 + M.F + Pop + NW + U1 + U2 + Wealth +
    Ineq + Prob

          Df Sum of Sq      RSS    AIC
- NW       1     11675 1387523 507.77
- Po2      1     21418 1397266 508.09
- Pop      1     27803 1403651 508.31
- M.F      1     31252 1407100 508.42
- Wealth   1     35035 1410883 508.55
<none>                 1375848 509.37
- U1       1     80954 1456802 510.06
- Po1      1    123896 1499744 511.42
- U2       1    190746 1566594 513.47
- M        1    217716 1593564 514.27
- Prob     1    226971 1602819 514.54
- Ed       1    413254 1789103 519.71
- Ineq     1    500944 1876792 521.96

Step:  AIC=507.77
Crime ~ M + Ed + Po1 + Po2 + M.F + Pop + U1 + U2 + Wealth + Ineq +
    Prob

          Df Sum of Sq      RSS    AIC
- Po2      1     16706 1404229 506.33
- Pop      1     25793 1413315 506.63
- M.F      1     26785 1414308 506.66
- Wealth   1     31551 1419073 506.82
<none>                 1387523 507.77
- U1       1     83881 1471404 508.52
- Po1      1    118348 1505871 509.61
```

```
- U2      1    201453 1588976 512.14
- Prob    1    216760 1604282 512.59
- M       1    309214 1696737 515.22
- Ed      1    402754 1790276 517.74
- Ineq    1    589736 1977259 522.41

Step:  AIC=506.33
Crime ~ M + Ed + Po1 + M.F + Pop + U1 + U2 + Wealth + Ineq +
    Prob

          Df Sum of Sq     RSS    AIC
- Pop     1     22345 1426575 505.07
- Wealth  1     32142 1436371 505.39
- M.F     1     36808 1441037 505.54
<none>                1404229 506.33
- U1      1     86373 1490602 507.13
- U2      1    205814 1610043 510.76
- Prob    1    218607 1622836 511.13
- M       1    307001 1711230 513.62
- Ed      1    389502 1793731 515.83
- Ineq    1    608627 2012856 521.25
- Po1     1   1050202 2454432 530.57

Step:  AIC=505.07
Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Wealth + Ineq + Prob

          Df Sum of Sq     RSS    AIC
- Wealth  1     26493 1453068 503.93
<none>                1426575 505.07
- M.F     1     84491 1511065 505.77
- U1      1     99463 1526037 506.24
- Prob    1    198571 1625145 509.20
- U2      1    208880 1635455 509.49
- M       1    320926 1747501 512.61
- Ed      1    386773 1813348 514.35
- Ineq    1    594779 2021354 519.45
- Po1     1   1127277 2553852 530.44

Step:  AIC=503.93
Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob

        Df Sum of Sq     RSS    AIC
<none>               1453068 503.93
- M.F   1    103159 1556227 505.16
- U1    1    127044 1580112 505.87
- Prob  1    247978 1701046 509.34
- U2    1    255443 1708511 509.55
- M     1    296790 1749858 510.67
- Ed    1    445788 1898855 514.51
- Ineq  1    738244 2191312 521.24
- Po1   1   1672038 3125105 537.93
```

5. Finally, look above. The Final recommended model AIC is 503.94 and there is no column that can be removed to get a lower AIC. So this the best Model.

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

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)   905.09      28.52  31.731  < 2e-16 ***
M.F            65.83      40.08   1.642  0.10874
U1           -109.73      60.20  -1.823  0.07622 .
Prob          -86.31      33.89  -2.547  0.01505 *
U2            158.22      61.22   2.585  0.01371 *
M             117.28      42.10   2.786  0.00828 **
Ed            201.50      59.02   3.414  0.00153 **
Ineq          244.70      55.69   4.394 8.63e-05 ***
Po1           305.07      46.14   6.613 8.26e-08 ***
---
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
```

Conclusion:

6. Verify to ensure that this is the better model Rsquared .78. This better than the previous model that I ran by picking few factors which has Rsquared of .68. However my previous homework using PCA the Rsquared was .63. Though stepwise algorithm seems to get good R squared Iam concerned about over fitting since stepwise has included some factors that have higher P values.

Q1.A.1:- step wise Forward.

1. Slightly different model

```
> model.f =
step(lm(Crime~1,data=data.sm),direction="forward",scope=~M+So+Ed+Po1+Po2+LF+M.F+Pop+NW+U1+U2+Wealth+Ineq+Prob)
Start:  AIC=561.02
Crime ~ 1

          Df Sum of Sq     RSS    AIC
+ Po1      1   3253302 3627626 532.94
+ Po2      1   3058626 3822302 535.39
+ Wealth   1   1340152 5540775 552.84
+ Prob     1   1257075 5623853 553.54
+ Pop      1    783660 6097267 557.34
+ Ed       1    717146 6163781 557.85
+ M.F      1    314867 6566061 560.82
<none>                6880928 561.02
+ LF       1    245446 6635482 561.32
+ Ineq     1    220530 6660397 561.49
+ U2       1    216354 6664573 561.52
+ So       1     56527 6824400 562.64
+ M        1     55084 6825844 562.65
+ U1       1     17533 6863395 562.90
+ NW       1      7312 6873615 562.97

Step:  AIC=532.94
Crime ~ Po1

          Df Sum of Sq     RSS    AIC
+ Ineq     1    739819 2887807 524.22
+ M        1    616741 3010885 526.18
+ M.F      1    250522 3377104 531.57
+ NW       1    232434 3395192 531.82
+ So       1    219098 3408528 532.01
+ Wealth   1    180872 3446754 532.53
<none>                3627626 532.94
+ Po2      1    146167 3481459 533.00
+ Prob     1     92278 3535348 533.72
+ LF       1     77479 3550147 533.92
+ U2       1     17848 3609778 534.70
+ Pop      1      5666 3621959 534.86
+ U1       1      2878 3624748 534.90
+ Ed       1       767 3626859 534.93

Step:  AIC=524.22
Crime ~ Po1 + Ineq

          Df Sum of Sq     RSS    AIC
+ Ed       1    587050 2300757 515.53
+ M.F      1    454545 2433262 518.17
+ Prob     1    280690 2607117 521.41
+ LF       1    260571 2627236 521.77
+ Wealth   1    213937 2673871 522.60
+ M        1    181236 2706571 523.17
+ Pop      1    130377 2757430 524.04
<none>                2887807 524.22
+ NW       1     36439 2851369 525.62
+ So       1     33738 2854069 525.66
+ Po2      1     30673 2857134 525.71
+ U1       1      2309 2885498 526.18
+ U2       1       253 2887554 526.21

Step:  AIC=515.53
Crime ~ Po1 + Ineq + Ed

          Df Sum of Sq     RSS    AIC
+ M        1    239405 2061353 512.37
+ Prob     1    234981 2065776 512.47
+ M.F      1    117026 2183731 515.08
<none>                2300757 515.53
+ Wealth   1     79540 2221218 515.88
+ U2       1     62112 2238646 516.25
+ Po2      1     42584 2258174 516.66
+ Pop      1     39319 2261438 516.72
+ U1       1      7365 2293392 517.38
+ LF       1      7254 2293503 517.39
+ NW       1      4210 2296547 517.45
+ So       1      4135 2296622 517.45

Step:  AIC=512.37
Crime ~ Po1 + Ineq + Ed + M

          Df Sum of Sq     RSS    AIC
+ Prob     1    258063 1803290 508.08
+ U2       1    200988 1860365 509.55
+ Wealth   1    163378 1897975 510.49
<none>                2061353 512.37
+ M.F      1     74398 1986955 512.64
+ U1       1     50835 2010518 513.20
+ Po2      1     45392 2015961 513.32
+ NW       1     16488 2044865 513.99
+ Pop      1      8101 2053251 514.19
```

```
+ So      1      3189 2058164 514.30
+ LF      1      2988 2058365 514.30

Step:  AIC=508.08
Crime ~ Po1 + Ineq + Ed + M + Prob

          Df Sum of Sq     RSS    AIC
+ U2      1    192233 1611057 504.79
+ Wealth  1     86490 1716801 507.77
+ M.F     1     84509 1718781 507.83
<none>                1803290 508.08
+ U1      1     52313 1750977 508.70
+ Pop     1     47719 1755571 508.82
+ Po2     1     37967 1765323 509.08
+ So      1     21971 1781320 509.51
+ LF      1       990 1802301 510.06
+ NW      1       797 1802493 510.06

Step:  AIC=504.79
Crime ~ Po1 + Ineq + Ed + M + Prob + U2

          Df Sum of Sq     RSS    AIC
<none>                1611057 504.79
+ Wealth  1     59910 1551147 505.00
+ U1      1     54830 1556227 505.16
+ Pop     1     51320 1559737 505.26
+ M.F     1     30945 1580112 505.87
+ Po2     1     25017 1586040 506.05
+ So      1     17958 1593098 506.26
+ LF      1     13179 1597878 506.40
+ NW      1       359 1610698 506.78
```

Run The model: lower R2 than the rest

```
lm(formula = Crime ~ Po2 + LF + M.F + Pop + NW + U1 + Wealth,
    data = data.sm)

Residuals:
    Min      1Q  Median      3Q     Max
-544.94 -152.48   29.56  158.13  484.05

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   905.09      40.17  22.534  < 2e-16 ***
Po2           276.06      86.58   3.189  0.00282 **
LF             13.74      60.31   0.228  0.82097
M.F           137.02      67.14   2.041  0.04809 *
Pop            48.08      57.49   0.836  0.40800
NW             91.14      60.23   1.513  0.13827
U1            -32.57      54.29  -0.600  0.55207
Wealth        -36.81      96.31  -0.382  0.70439
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 275.4 on 39 degrees of freedom
Multiple R-squared:  0.5703,    Adjusted R-squared:  0.4931
F-statistic: 7.393 on 7 and 39 DF,  p-value: 1.186e-05
```

R code attached

```
############### Assignment 1 #################

###############################################
############ Read and Scale data #############
###############################################

data <- read.table("http://www.statsci.org/data/general/uscrime.txt",sep="\t", header=TRUE)
head(data)
data.s = as.data.frame(scale(data[,c(1,3:15)]))
data.sm = cbind(s[,1],data[,2],s[,3:14],data[,16])
colnames(data.sm)[1] = "M"
colnames(data.sm)[2] = "So"
colnames(data.sm)[15] = "Crime"

###############################################
############ Run the Model ###################
###############################################

model = step(lm(Crime~.,data=data.sm,direction="backward"))

#### using the factors run the LM model ####

model.b = lm(Crime~M.F+U1+Prob+U2+M+Ed+Ineq+Po1,data=data.sm)

#data.f = subset(data.sm, select=c("M","Ed","Po1","M.F", "U1","U2","Ineq","Prob"))
#data.pca = subset(data.sm, select=c(-Crime))
#
#model.pca = prcomp(data.pca,center=TRUE,scale=TRUE)
```

```
#
#plot(model.pca,type="l")
#summary(model.pca)
##############################################
########### calculate R2  ###################
##############################################

sstot = sum(data$Crime - mean(data$Crime))^2
totsse <- 0
for (i in nrow(data.sm)) {
mod_step_i = lm(Crime~M.F+U1+Prob+U2+M+Ed+Ineq+Po1,data=data.sm[-1,])
pred_i = predict(mod_step_i,newdata = data.sm[i,])
totsse = totsse + ((pred_i - data[i,16])^2)
}

r2mod = 1 - totsse / sstot

r2mod

##############################################
########### Run the Model forward  #########
##############################################

    ########### Run the Model forward  #########
    ##############################################

    model.f =
    step(lm(Crime~1,data=data.sm),direction="forward",scope=~M+So+Ed+Po1+Po2+LF+M.F+Pop+NW+U1+U2+Wealth+Ineq+Prob)


    model.bnew = lm(Crime~Po2+LF+M.F+ Pop+NW+U1+Wealth,data=data.sm)

    summary(model.bnew)
```

Q1.B:- Lasso:-

Step 1:- Run with Lasso model with Alpha as 1. Suggestion is to drop Po2 and LF. See below the Coefficients.

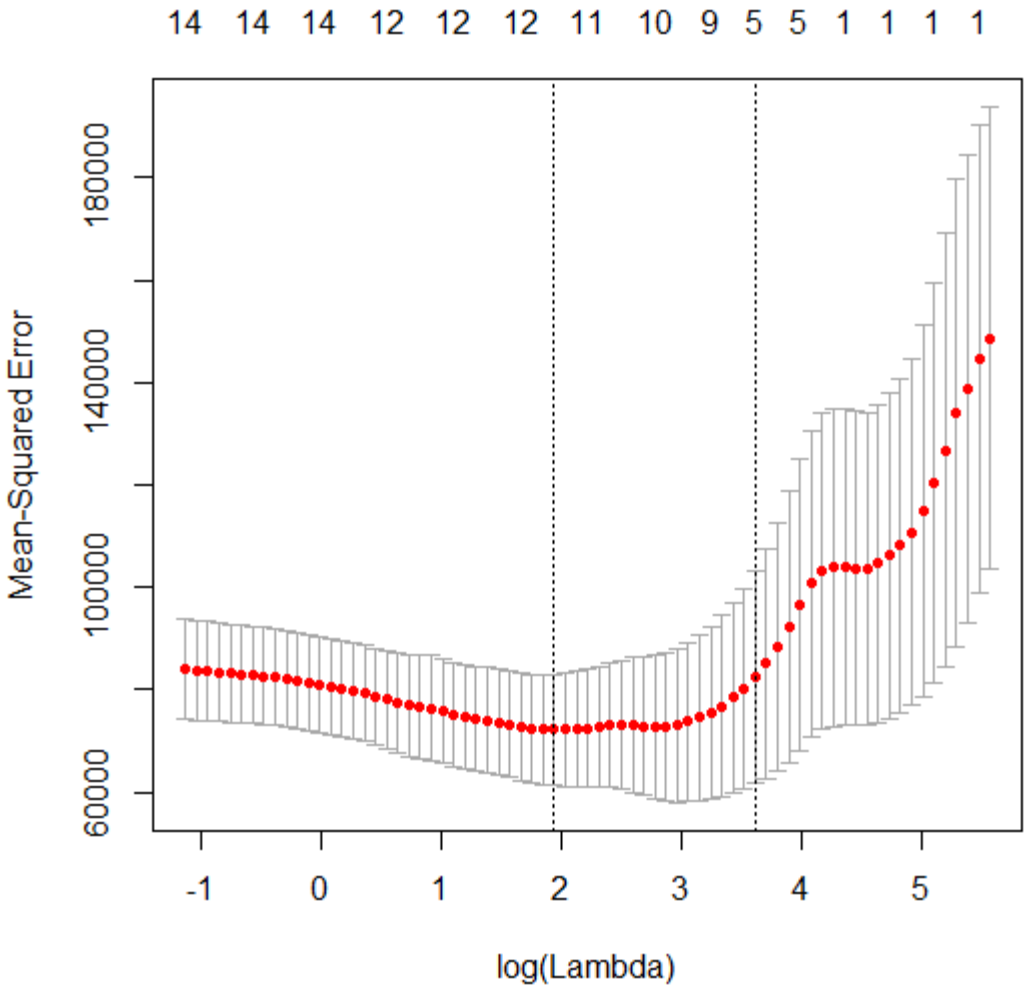
```
303. coef(model.lasso,s=model.lasso$lambda.min)
304. 15 x 1 sparse Matrix of class "dgCMatrix"
305.                       1
306. (Intercept) 891.225281
307. M            95.361253
308. So           40.713238
309. Ed          152.117061
310. Po1         299.772697
311. Po2           .
312. LF            .
313. M.F          55.894107
314. Pop          -3.334867
315. NW            8.922502
316. U1          -50.127785
317. U2           88.788601
318. Wealth       22.358885
319. Ineq        211.923204
     Prob        -85.091950
```

Step 2:- Run LM with those factors

summary(model.l)

```
320. Call:
321. lm(formula = Crime ~ M + So + Ed + Po1 + M.F + Pop + NW + U1 +
322.     U2 + Wealth + Ineq + Prob, data = data.sm)
323.
324. Residuals:
325.     Min      1Q  Median      3Q     Max
326. -434.18 -107.01   18.55  115.88  470.32
327.
328. Coefficients:
329.             Estimate Std. Error t value Pr(>|t|)
330. (Intercept)   897.29      51.91  17.286  < 2e-16 ***
331. M             112.71      49.35   2.284  0.02876 *
332. So             22.89     125.35   0.183  0.85621
333. Ed            195.70      62.94   3.109  0.00378 **
334. Po1           293.18      64.99   4.511 7.32e-05 ***
335. M.F            48.92      48.12   1.017  0.31656
336. Pop           -33.25      45.63  -0.729  0.47113
337. NW             19.16      57.71   0.332  0.74195
338. U1            -89.76      65.68  -1.367  0.18069
339. U2            140.78      66.77   2.108  0.04245 *
340. Wealth         83.30      95.53   0.872  0.38932
341. Ineq          285.77      85.19   3.355  0.00196 **
342. Prob          -92.75      41.12  -2.255  0.03065 *
343. ---
344. Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
345.
346. Residual standard error: 202.6 on 34 degrees of freedom
347. Multiple R-squared:  0.7971,    Adjusted R-squared:  0.7255
348. F-statistic: 11.13 on 12 and 34 DF,  p-value: 1.52e-08
```

Step 3:- non zero coefficients stays at 14 when lambda reached 1. R2 is .7971.
It is better than Stepwise.

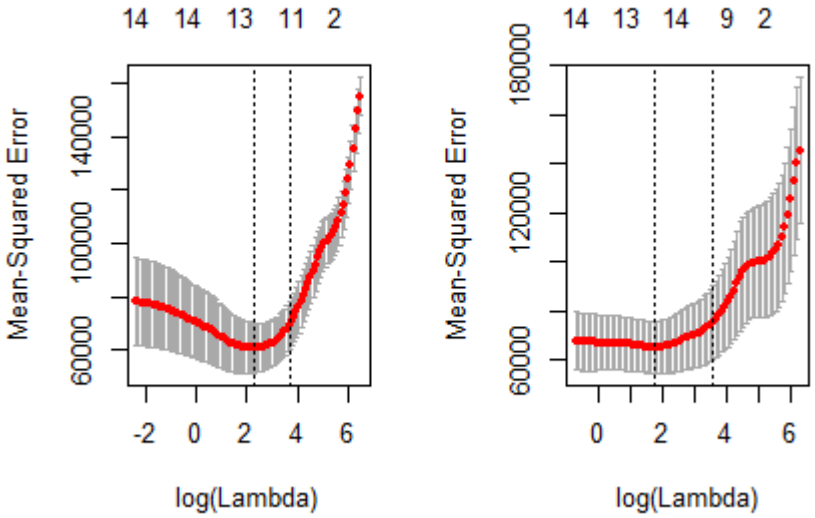


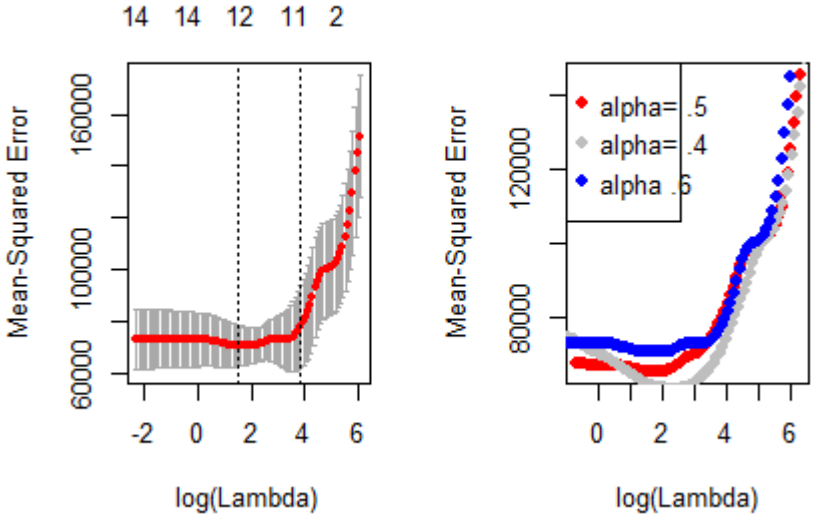
Q1.C:- Run Elastic to balance: with alpha .5

```
model.elas = cv.glmnet(x=as.matrix(data.sm[,-15]),y=as.matrix(data.sm$Crime),alpha=.5,nfolds=5,
+ type.measure="mse",family="gaussian")
> coef(model.elas,s=model.elas$lambda.min)
15 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) 894.23767
M           106.28676
So           31.86433
Ed          181.09367
Po1         290.12273
Po2           .
LF            .
M.F          53.66392
Pop         -23.39678
NW           19.11828
U1          -80.09361
U2          126.53988
Wealth       65.75459
Ineq        257.79262
Prob        -91.73531
> model.e = lm(Crime~M+So+Ed+Po1+M.F+Pop+NW+U1+U2+Wealth+Ineq+Prob,data=data.sm)
> summary(model.e)

Call:
lm(formula = Crime ~ M + So + Ed + Po1 + M.F + Pop + NW + U1 +
    U2 + Wealth + Ineq + Prob, data = data.sm)

Residuals:
     Min       1Q   Median       3Q      Max
-434.18 -107.01   18.55  115.88  470.32
```

```
379.
380. Coefficients:
381.             Estimate Std. Error t value Pr(>|t|)
382. (Intercept)   897.29      51.91  17.286  < 2e-16 ***
383. M             112.71      49.35   2.284  0.02876 *
384. So             22.89     125.35   0.183  0.85621
385. Ed            195.70      62.94   3.109  0.00378 **
386. Po1           293.18      64.99   4.511 7.32e-05 ***
387. M.F            48.92      48.12   1.017  0.31656
388. Pop           -33.25      45.63  -0.729  0.47113
389. NW             19.16      57.71   0.332  0.74195
390. U1            -89.76      65.68  -1.367  0.18069
391. U2            140.78      66.77   2.108  0.04245 *
392. Wealth         83.30      95.53   0.872  0.38932
393. Ineq          285.77      85.19   3.355  0.00196 **
394. Prob          -92.75      41.12  -2.255  0.03065 *
395. ---
396. Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
397.
398. Residual standard error: 202.6 on 34 degrees of freedom
399. Multiple R-squared:  0.7971,    Adjusted R-squared:  0.7255
400. F-statistic: 11.13 on 12 and 34 DF,  p-value: 1.52e-08
```

Try alpha = .4
Try alpha = .6
The R2 from the models were .780 and .78077. with alpha .6 is .795.
Alpha=.6 of elastic model seems more appropriate.





Q2:- We sell insurance policies.(Auto Home and commercial). We have strong underwriting rules.

a. We want to provide recommendation for lower coverage for our agents when the underwriting rule is Fired and if we get a good customer we should show options to up sell a product or cross sell a product. We can also see if the customer is waiting on a screen to provide pops or displays to give him various Options.
b. Also when we measure the success of an a cross sell or upsell we have to make sure to understand the Underwriters bias. If they are promoting restaurant business up selling could be more so it is wiser to see How upselling and cross selling are happening for other business like Institutional policy and classify big Agents vs career and small agents
c. Similarly the blocking factor is promotional sale. If I take away the restaurant business or elite agents then it is easy to analyze other factors that influence upselling and cross selling. And tehn compare restaurant business by it self and top agents activity by itself.

Q3:-

```
> home_design = FrF2(nruns=16,nfactors=10)
> home_design
```

```
403.     A  B  C  D  E  F  G  H  J  K
404. 1   1  1  1 -1  1  1  1 -1 -1 -1
405. 2  -1  1 -1  1 -1  1 -1 -1 -1  1
406. 3   1 -1 -1  1 -1 -1  1  1  1  1
407. 4  -1  1  1 -1 -1 -1  1  1 -1  1
408. 5  -1  1  1  1 -1 -1  1 -1  1 -1
409. 6  -1 -1  1 -1  1 -1 -1  1  1 -1
410. 7  -1  1 -1 -1 -1  1 -1  1  1 -1
411. 8  -1 -1 -1 -1  1  1  1  1 -1  1
412. 9  -1 -1 -1  1  1  1  1 -1  1 -1
413. 10  1 -1  1  1 -1  1 -1  1 -1 -1
414. 11  1  1 -1 -1  1 -1 -1 -1  1  1
415. 12  1 -1  1 -1 -1  1 -1 -1  1  1
416. 13  1 -1 -1 -1 -1 -1  1 -1 -1 -1
417. 14 -1 -1  1  1  1 -1 -1 -1 -1  1
418. 15  1  1  1  1  1  1  1  1  1  1
419. 16  1  1 -1  1  1 -1 -1  1 -1 -1
420. class=design, type= FrF2
```

Question 4:-
Binomial :- Number of homeruns across the baseball Season
Geometric:- How many times I have to roll the dice before I get 6.
Poisson:- Number warranty claim that we get over software deployment.
Exponential distribution:- Time interval between the txt message I receive.
Weibull:- The time between the getting 6 when I roll the Dice