Homework 5

Chen Yi-Ju(Ernie)

2020/6/17

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

```
setwd("D:/ernie/self-study/GTxMicroMasters/Introduction to Analytics Modeling/week5/homework")
#loading library
library(tidyverse)
library(caret)
library(egg)
library(stargazer)
library(modelr)
library(glmnet)
library(foreach)
library(FrF2)
```

```
#loading data
crime <- read.table("uscrime.txt" , header = T)
```

1. Stepwise regression

```
#dividing into training and testing data sets
set.seed(101)
train.index <- createDataPartition(crime$Crime , p = 0.8 ,times = 1, list = F)
train <- crime[train.index,]
test <- crime[-train.index,]
```

First, I try using forward stepwise regression

```
#Forward stepwise
null = lm (Crime ~1, data = train) #setting upper bound
full = lm(Crime ~., data = train) #setting lower bound
```

Using the step function

```
#step function

forward.sel <- step(null,
                    scope = list(lower = null , upper = full),
```

```r
                    direction = "forward")
```

```
## Start:  AIC=468.47
## Crime ~ 1
##
##          Df Sum of Sq     RSS    AIC
## + Po1     1   2797294 3305694 446.56
## + Po2     1   2660798 3442189 448.13
## + Wealth  1   1343570 4759417 460.77
## + Prob    1   1342223 4760764 460.78
## + Pop     1    658975 5444012 466.01
## + Ed      1    641565 5461423 466.14
## + U2      1    472098 5630890 467.33
## + Time    1    359554 5743434 468.10
## <none>                6102988 468.47
## + Ineq    1    281710 5821277 468.63
## + M.F     1    275417 5827570 468.67
## + LF      1    105190 5997798 469.79
## + M       1     70170 6032817 470.02
## + So      1     23568 6079420 470.32
## + NW      1       815 6102172 470.46
## + U1      1       258 6102730 470.47
##
## Step:  AIC=446.56
## Crime ~ Po1
##
##          Df Sum of Sq     RSS    AIC
## + M       1    571704 2733990 441.15
## + Ineq    1    542382 2763312 441.57
## + M.F     1    261831 3043862 445.34
## + So      1    209343 3096350 446.00
## + NW      1    183579 3122115 446.33
## <none>                3305694 446.56
## + Prob    1    140296 3165398 446.86
## + Wealth  1     96884 3208810 447.40
## + Time    1     91357 3214337 447.46
## + Po2     1     87071 3218622 447.52
## + LF      1     55811 3249882 447.89
## + U2      1     17648 3288046 448.35
## + Pop     1     14427 3291267 448.39
## + Ed      1      4680 3301014 448.50
## + U1      1       673 3305021 448.55
##
## Step:  AIC=441.15
## Crime ~ Po1 + M
##
##          Df Sum of Sq     RSS    AIC
## + M.F     1    305040 2428949 438.54
## + Prob    1    180993 2552997 440.48
## + Ed      1    179759 2554231 440.50
## + Ineq    1    168644 2565346 440.67
## <none>                2733990 441.15
## + LF      1    120667 2613322 441.39
## + U2      1     80821 2653169 441.98
```

```
## + Po2     1     53415 2680575 442.38
## + U1      1     42891 2691099 442.53
## + Pop     1     11233 2722757 442.99
## + So      1     10914 2723076 442.99
## + Time    1      3044 2730945 443.11
## + Wealth  1      2248 2731742 443.12
## + NW      1         1 2733988 443.15
##
## Step:  AIC=438.54
## Crime ~ Po1 + M + M.F
##
##          Df Sum of Sq     RSS    AIC
## + Ineq    1    305223 2123726 435.30
## + Prob    1    143709 2285240 438.16
## + So      1    134036 2294913 438.32
## <none>                2428949 438.54
## + Time    1    121047 2307902 438.54
## + U2      1     94239 2334710 438.99
## + NW      1     70336 2358613 439.39
## + Pop     1     47568 2381381 439.77
## + Po2     1     32303 2396646 440.01
## + Ed      1     23032 2405917 440.17
## + Wealth  1     18526 2410423 440.24
## + LF      1      2694 2426255 440.49
## + U1      1        17 2428932 440.54
##
## Step:  AIC=435.3
## Crime ~ Po1 + M + M.F + Ineq
##
##          Df Sum of Sq     RSS    AIC
## + Ed      1    311535 1812191 431.11
## + Prob    1    245461 1878265 432.51
## + Wealth  1    193599 1930127 433.57
## <none>                2123726 435.30
## + Time    1     88270 2035456 435.64
## + U2      1     50524 2073202 436.36
## + NW      1     33305 2090421 436.68
## + LF      1     12170 2111556 437.08
## + Po2     1      8768 2114957 437.14
## + Pop     1      3135 2120591 437.24
## + So      1      2489 2121236 437.25
## + U1      1        42 2123684 437.30
##
## Step:  AIC=431.11
## Crime ~ Po1 + M + M.F + Ineq + Ed
##
##          Df Sum of Sq     RSS    AIC
## + U2      1    235507 1576684 427.68
## + Prob    1    196452 1615739 428.64
## + Time    1     94394 1717797 431.03
## + Wealth  1     92876 1719315 431.06
## <none>                1812191 431.11
## + U1      1     40346 1771845 432.23
## + Po2     1     33649 1778542 432.38
```

```
## + LF      1     21187 1791005 432.65
## + So      1     17499 1794692 432.73
## + Pop     1       735 1811456 433.10
## + NW      1       295 1811896 433.11
##
## Step:  AIC=427.68
## Crime ~ Po1 + M + M.F + Ineq + Ed + U2
##
##          Df Sum of Sq     RSS    AIC
## + Prob    1    170699 1405986 425.21
## + U1      1    169334 1407350 425.25
## + Wealth  1     95633 1481051 427.24
## <none>                1576684 427.68
## + Time    1     74739 1501946 427.79
## + Po2     1     35277 1541407 428.80
## + NW      1     25412 1551273 429.05
## + So      1     20860 1555824 429.16
## + LF      1      2016 1574668 429.63
## + Pop     1      1883 1574801 429.64
##
## Step:  AIC=425.21
## Crime ~ Po1 + M + M.F + Ineq + Ed + U2 + Prob
##
##          Df Sum of Sq     RSS    AIC
## + U1      1    160091 1245895 422.50
## + So      1    101888 1304098 424.28
## <none>                1405986 425.21
## + Wealth  1     62363 1343622 425.45
## + Po2     1     30649 1375337 426.36
## + Pop     1     28055 1377931 426.43
## + Time    1      1500 1404485 427.17
## + LF      1      1075 1404911 427.18
## + NW      1        71 1405915 427.21
##
## Step:  AIC=422.5
## Crime ~ Po1 + M + M.F + Ineq + Ed + U2 + Prob + U1
##
##          Df Sum of Sq     RSS    AIC
## <none>                1245895 422.50
## + So      1     47478 1198417 422.99
## + Wealth  1     31144 1214750 423.51
## + LF      1     25300 1220595 423.70
## + Po2     1     17337 1228558 423.95
## + Pop     1      4512 1241383 424.36
## + NW      1      4157 1241738 424.37
## + Time    1         0 1245895 424.50
```

We can see from the progress that AIC is decreasing.

```r
summary(forward.sel)
```

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

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -397.75 -106.78   14.48  141.65  507.68
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7306.48    1380.30  -5.293 1.02e-05 ***
## Po1            94.95      18.67   5.085 1.83e-05 ***
## M             100.62      37.41   2.690 0.011570 *
## M.F            29.67      15.61   1.901 0.066909 .
## Ineq           59.89      16.42   3.647 0.000998 ***
## Ed            185.06      57.91   3.195 0.003277 **
## U2            267.71      94.08   2.845 0.007919 **
## Prob        -3238.63    1642.54  -1.972 0.057924 .
## U1          -8256.55    4205.28  -1.963 0.058931 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 203.8 on 30 degrees of freedom
## Multiple R-squared:  0.7959, Adjusted R-squared:  0.7414
## F-statistic: 14.62 on 8 and 30 DF,  p-value: 1.924e-08
```

Testing model on test data

```
#Testing
res.forw <- test %>%
  add_predictions(.,forward.sel) %>%
  select('observations' = Crime, pred) %>%
  as.data.frame()
```

Summary of the results:

```
forward <- postResample(obs = res.forw$observations, pred = res.forw$pred)
forward
```

```
##        RMSE    Rsquared         MAE
## 192.5011883   0.6075089 164.3968602
```

We can see that the R squared value is 58% (approx.) Then we do the same thing with backwards stepwise

```
#Backwards stepwise
backward.sel<- step( full,
  scope = list(upper = full , lower = null),
  direction = "backward")
```

```
## Start:  AIC=431.79
## Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 +
##     U2 + Wealth + Ineq + Prob + Time
##
##          Df Sum of Sq     RSS    AIC
## - LF      1      2779 1106877 429.89
## - Time    1      2840 1106938 429.89
## - Pop     1     11369 1115467 430.19
## - NW      1     19010 1123108 430.45
## - Po2     1     23858 1127957 430.62
## - Wealth  1     31358 1135456 430.88
## - M.F     1     36810 1140908 431.07
```

```
## - So      1     40666 1144764 431.20
## - U1      1     49580 1153678 431.50
## <none>                1104098 431.79
## - Prob    1     86348 1190446 432.72
## - Po1     1    104236 1208334 433.31
## - U2      1    208412 1312510 436.53
## - M       1    221814 1325912 436.93
## - Ineq    1    276544 1380642 438.51
## - Ed      1    356788 1460886 440.71
##
## Step:  AIC=429.89
## Crime ~ M + So + Ed + Po1 + Po2 + M.F + Pop + NW + U1 + U2 +
##     Wealth + Ineq + Prob + Time
##
##          Df Sum of Sq     RSS    AIC
## - Time    1      3685 1110562 428.02
## - Pop     1     13888 1120765 428.37
## - Po2     1     21642 1128519 428.64
## - NW      1     26993 1133870 428.83
## - Wealth  1     30524 1137401 428.95
## - M.F     1     36283 1143160 429.14
## - U1      1     48877 1155754 429.57
## <none>                1106877 429.89
## - So      1     66054 1172932 430.15
## - Prob    1     83908 1190785 430.74
## - Po1     1    101689 1208566 431.31
## - U2      1    205953 1312830 434.54
## - M       1    234351 1341228 435.38
## - Ineq    1    274414 1381292 436.52
## - Ed      1    373656 1480533 439.23
##
## Step:  AIC=428.02
## Crime ~ M + So + Ed + Po1 + Po2 + M.F + Pop + NW + U1 + U2 +
##     Wealth + Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## - Pop     1     11190 1121753 426.41
## - NW      1     23364 1133926 426.83
## - Po2     1     30520 1141082 427.07
## - Wealth  1     30707 1141270 427.08
## - M.F     1     33147 1143710 427.16
## - U1      1     54601 1165164 427.89
## <none>                1110562 428.02
## - So      1     62395 1172957 428.15
## - Po1     1    122385 1232947 430.09
## - Prob    1    150014 1260576 430.96
## - U2      1    212176 1322739 432.83
## - M       1    259601 1370163 434.21
## - Ineq    1    271978 1382541 434.56
## - Ed      1    371935 1482497 437.28
##
## Step:  AIC=426.41
## Crime ~ M + So + Ed + Po1 + Po2 + M.F + NW + U1 + U2 + Wealth +
##     Ineq + Prob
```

```
##
##          Df Sum of Sq     RSS    AIC
## - NW      1     22200 1143953 425.17
## - Wealth  1     23535 1145288 425.22
## - Po2     1     27529 1149282 425.35
## <none>                1121753 426.41
## - So      1     69122 1190874 426.74
## - U1      1     74167 1195919 426.90
## - M.F     1     86539 1208291 427.31
## - Po1     1    112948 1234700 428.15
## - Prob    1    140019 1261771 428.99
## - U2      1    233510 1355263 431.78
## - M       1    267687 1389440 432.75
## - Ineq    1    275361 1397114 432.97
## - Ed      1    370216 1491969 435.53
##
## Step:  AIC=425.17
## Crime ~ M + So + Ed + Po1 + Po2 + M.F + U1 + U2 + Wealth + Ineq +
##     Prob
##
##          Df Sum of Sq     RSS    AIC
## - Wealth  1     19791 1163744 423.84
## - Po2     1     34993 1178946 424.35
## - So      1     51171 1195124 424.88
## <none>                1143953 425.17
## - U1      1     70867 1214820 425.52
## - M.F     1    103569 1247522 426.55
## - Po1     1    119245 1263198 427.04
## - Prob    1    182143 1326096 428.93
## - U2      1    213315 1357267 429.84
## - M       1    249110 1393063 430.85
## - Ineq    1    260618 1404571 431.18
## - Ed      1    408883 1552836 435.09
##
## Step:  AIC=423.84
## Crime ~ M + So + Ed + Po1 + Po2 + M.F + U1 + U2 + Ineq + Prob
##
##        Df Sum of Sq     RSS    AIC
## - Po2   1     34673 1198417 422.99
## <none>              1163744 423.84
## - So    1     64814 1228558 423.95
## - U1    1     83160 1246904 424.53
## - M.F   1    125510 1289254 425.83
## - Po1   1    129731 1293475 425.96
## - Prob  1    216973 1380717 428.51
## - U2    1    230709 1394453 428.89
## - M     1    234233 1397977 428.99
## - Ineq  1    310012 1473756 431.05
## - Ed    1    482074 1645818 435.36
##
## Step:  AIC=422.99
## Crime ~ M + So + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob
##
##        Df Sum of Sq     RSS    AIC
```

```
## - So    1     47478 1245895 422.50
## <none>              1198417 422.99
## - U1    1    105681 1304098 424.28
## - M.F   1    161272 1359689 425.91
## - Prob  1    206885 1405302 427.20
## - M     1    235510 1433926 427.98
## - U2    1    262291 1460708 428.70
## - Ineq  1    339458 1537875 430.71
## - Ed    1    451262 1649679 433.45
## - Po1   1    974280 2172697 444.19
##
## Step:  AIC=422.5
## Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob
##
##        Df Sum of Sq     RSS    AIC
## <none>              1245895 422.50
## - M.F   1    150124 1396019 424.94
## - U1    1    160091 1405986 425.21
## - Prob  1    161456 1407350 425.25
## - M     1    300427 1546322 428.93
## - U2    1    336243 1582138 429.82
## - Ed    1    424053 1669948 431.92
## - Ineq  1    552239 1798133 434.81
## - Po1   1   1073853 2319748 444.74
```

```r
summary(backward.sel)
```

```
##
## Call:
## lm.default(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq +
##     Prob, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -397.75 -106.78   14.48  141.65  507.68
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7306.48    1380.30  -5.293 1.02e-05 ***
## M             100.62      37.41   2.690 0.011570 *
## Ed            185.06      57.91   3.195 0.003277 **
## Po1            94.95      18.67   5.085 1.83e-05 ***
## M.F            29.67      15.61   1.901 0.066909 .
## U1          -8256.55    4205.28  -1.963 0.058931 .
## U2            267.71      94.08   2.845 0.007919 **
## Ineq           59.89      16.42   3.647 0.000998 ***
## Prob        -3238.63    1642.54  -1.972 0.057924 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 203.8 on 30 degrees of freedom
## Multiple R-squared:  0.7959, Adjusted R-squared:  0.7414
## F-statistic: 14.62 on 8 and 30 DF,  p-value: 1.924e-08
```

```r
#Testing
res.back <- test %>%
  add_predictions(.,backward.sel) %>%
  select('observations' = Crime, pred) %>%
  as.data.frame()
backward <- postResample(obs = res.back$observations, pred = res.back$pred)
backward
```

```
##        RMSE    Rsquared         MAE
## 192.5011883   0.6075089 164.3968602
```

In this case the backward stepwise regression did worse than the forward stepwise regression

Finally, we do stepwise regression from both sides, starting from no parameters (null) and all parameters(full)

```r
# Stepwise regression
step.both.1 <- step(null, scope = list(upper = full) , direction = "both")
```

```
## Start:  AIC=468.47
## Crime ~ 1
##
##          Df Sum of Sq     RSS    AIC
## + Po1     1   2797294 3305694 446.56
## + Po2     1   2660798 3442189 448.13
## + Wealth  1   1343570 4759417 460.77
## + Prob    1   1342223 4760764 460.78
## + Pop     1    658975 5444012 466.01
## + Ed      1    641565 5461423 466.14
## + U2      1    472098 5630890 467.33
## + Time    1    359554 5743434 468.10
## <none>                6102988 468.47
## + Ineq    1    281710 5821277 468.63
## + M.F     1    275417 5827570 468.67
## + LF      1    105190 5997798 469.79
## + M       1     70170 6032817 470.02
## + So      1     23568 6079420 470.32
## + NW      1       815 6102172 470.46
## + U1      1       258 6102730 470.47
##
## Step:  AIC=446.56
## Crime ~ Po1
##
##          Df Sum of Sq     RSS    AIC
## + M       1    571704 2733990 441.15
## + Ineq    1    542382 2763312 441.57
## + M.F     1    261831 3043862 445.34
## + So      1    209343 3096350 446.00
## + NW      1    183579 3122115 446.33
## <none>                3305694 446.56
## + Prob    1    140296 3165398 446.86
## + Wealth  1     96884 3208810 447.40
## + Time    1     91357 3214337 447.46
## + Po2     1     87071 3218622 447.52
## + LF      1     55811 3249882 447.89
## + U2      1     17648 3288046 448.35
## + Pop     1     14427 3291267 448.39
```

```
## + Ed      1      4680 3301014 448.50
## + U1      1       673 3305021 448.55
## - Po1     1   2797294 6102988 468.47
##
## Step:  AIC=441.15
## Crime ~ Po1 + M
##
##          Df Sum of Sq     RSS    AIC
## + M.F     1    305040 2428949 438.54
## + Prob    1    180993 2552997 440.48
## + Ed      1    179759 2554231 440.50
## + Ineq    1    168644 2565346 440.67
## <none>                2733990 441.15
## + LF      1    120667 2613322 441.39
## + U2      1     80821 2653169 441.98
## + Po2     1     53415 2680575 442.38
## + U1      1     42891 2691099 442.53
## + Pop     1     11233 2722757 442.99
## + So      1     10914 2723076 442.99
## + Time    1      3044 2730945 443.11
## + Wealth  1      2248 2731742 443.12
## + NW      1         1 2733988 443.15
## - M       1    571704 3305694 446.56
## - Po1     1   3298828 6032817 470.02
##
## Step:  AIC=438.54
## Crime ~ Po1 + M + M.F
##
##          Df Sum of Sq     RSS    AIC
## + Ineq    1    305223 2123726 435.30
## + Prob    1    143709 2285240 438.16
## + So      1    134036 2294913 438.32
## <none>                2428949 438.54
## + Time    1    121047 2307902 438.54
## + U2      1     94239 2334710 438.99
## + NW      1     70336 2358613 439.39
## + Pop     1     47568 2381381 439.77
## + Po2     1     32303 2396646 440.01
## + Ed      1     23032 2405917 440.17
## + Wealth  1     18526 2410423 440.24
## + LF      1      2694 2426255 440.49
## + U1      1        17 2428932 440.54
## - M.F     1    305040 2733990 441.15
## - M       1    614913 3043862 445.34
## - Po1     1   3341179 5770129 470.28
##
## Step:  AIC=435.3
## Crime ~ Po1 + M + M.F + Ineq
##
##          Df Sum of Sq     RSS    AIC
## + Ed      1    311535 1812191 431.11
## + Prob    1    245461 1878265 432.51
## + Wealth  1    193599 1930127 433.57
## <none>                2123726 435.30
```

```
## + Time    1     88270 2035456 435.64
## - M       1    156527 2280252 436.07
## + U2      1     50524 2073202 436.36
## + NW      1     33305 2090421 436.68
## + LF      1     12170 2111556 437.08
## + Po2     1      8768 2114957 437.14
## + Pop     1      3135 2120591 437.24
## + So      1      2489 2121236 437.25
## + U1      1        42 2123684 437.30
## - Ineq    1    305223 2428949 438.54
## - M.F     1    441620 2565346 440.67
## - Po1     1   3498295 5622021 471.27
##
## Step:  AIC=431.11
## Crime ~ Po1 + M + M.F + Ineq + Ed
##
##          Df Sum of Sq     RSS    AIC
## + U2      1    235507 1576684 427.68
## + Prob    1    196452 1615739 428.64
## - M.F     1     83400 1895591 430.87
## + Time    1     94394 1717797 431.03
## + Wealth  1     92876 1719315 431.06
## <none>                1812191 431.11
## + U1      1     40346 1771845 432.23
## + Po2     1     33649 1778542 432.38
## + LF      1     21187 1791005 432.65
## + So      1     17499 1794692 432.73
## + Pop     1       735 1811456 433.10
## + NW      1       295 1811896 433.11
## - M       1    242475 2054666 434.01
## - Ed      1    311535 2123726 435.30
## - Ineq    1    593726 2405917 440.17
## - Po1     1   3587215 5399406 471.69
##
## Step:  AIC=427.68
## Crime ~ Po1 + M + M.F + Ineq + Ed + U2
##
##          Df Sum of Sq     RSS    AIC
## + Prob    1    170699 1405986 425.21
## + U1      1    169334 1407350 425.25
## - M.F     1     32533 1609218 426.48
## + Wealth  1     95633 1481051 427.24
## <none>                1576684 427.68
## + Time    1     74739 1501946 427.79
## + Po2     1     35277 1541407 428.80
## + NW      1     25412 1551273 429.05
## + So      1     20860 1555824 429.16
## + LF      1      2016 1574668 429.63
## + Pop     1      1883 1574801 429.64
## - U2      1    235507 1812191 431.11
## - M       1    390760 1967444 434.32
## - Ed      1    496518 2073202 436.36
## - Ineq    1    665412 2242096 439.42
## - Po1     1   2946361 4523045 466.78
```

```
##
## Step:  AIC=425.21
## Crime ~ Po1 + M + M.F + Ineq + Ed + U2 + Prob
##
##          Df Sum of Sq     RSS    AIC
## + U1      1    160091 1245895 422.50
## - M.F     1     36742 1442727 424.22
## + So      1    101888 1304098 424.28
## <none>                1405986 425.21
## + Wealth  1     62363 1343622 425.45
## + Po2     1     30649 1375337 426.36
## + Pop     1     28055 1377931 426.43
## + Time    1      1500 1404485 427.17
## + LF      1      1075 1404911 427.18
## + NW      1        71 1405915 427.21
## - Prob    1    170699 1576684 427.68
## - U2      1    209754 1615739 428.64
## - M       1    347731 1753716 431.83
## - Ed      1    426031 1832017 433.54
## - Ineq    1    728593 2134579 439.50
## - Po1     1   2344120 3750105 461.48
##
## Step:  AIC=422.5
## Crime ~ Po1 + M + M.F + Ineq + Ed + U2 + Prob + U1
##
##          Df Sum of Sq     RSS    AIC
## <none>                1245895 422.50
## + So      1     47478 1198417 422.99
## + Wealth  1     31144 1214750 423.51
## + LF      1     25300 1220595 423.70
## + Po2     1     17337 1228558 423.95
## + Pop     1      4512 1241383 424.36
## + NW      1      4157 1241738 424.37
## + Time    1         0 1245895 424.50
## - M.F     1    150124 1396019 424.94
## - U1      1    160091 1405986 425.21
## - Prob    1    161456 1407350 425.25
## - M       1    300427 1546322 428.93
## - U2      1    336243 1582138 429.82
## - Ed      1    424053 1669948 431.92
## - Ineq    1    552239 1798133 434.81
## - Po1     1   1073853 2319748 444.74
```

```r
step.both.2 <- step(full, scope = list(upper = full), direction = "both")
```

```
## Start:  AIC=431.79
## Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 +
##     U2 + Wealth + Ineq + Prob + Time
##
##          Df Sum of Sq     RSS    AIC
## - LF      1      2779 1106877 429.89
## - Time    1      2840 1106938 429.89
## - Pop     1     11369 1115467 430.19
## - NW      1     19010 1123108 430.45
## - Po2     1     23858 1127957 430.62
```

```
## - Wealth  1     31358 1135456 430.88
## - M.F     1     36810 1140908 431.07
## - So      1     40666 1144764 431.20
## - U1      1     49580 1153678 431.50
## <none>                1104098 431.79
## - Prob    1     86348 1190446 432.72
## - Po1     1    104236 1208334 433.31
## - U2      1    208412 1312510 436.53
## - M       1    221814 1325912 436.93
## - Ineq    1    276544 1380642 438.51
## - Ed      1    356788 1460886 440.71
##
## Step:  AIC=429.89
## Crime ~ M + So + Ed + Po1 + Po2 + M.F + Pop + NW + U1 + U2 +
##     Wealth + Ineq + Prob + Time
##
##          Df Sum of Sq     RSS    AIC
## - Time    1      3685 1110562 428.02
## - Pop     1     13888 1120765 428.37
## - Po2     1     21642 1128519 428.64
## - NW      1     26993 1133870 428.83
## - Wealth  1     30524 1137401 428.95
## - M.F     1     36283 1143160 429.14
## - U1      1     48877 1155754 429.57
## <none>                1106877 429.89
## - So      1     66054 1172932 430.15
## - Prob    1     83908 1190785 430.74
## - Po1     1    101689 1208566 431.31
## + LF      1      2779 1104098 431.79
## - U2      1    205953 1312830 434.54
## - M       1    234351 1341228 435.38
## - Ineq    1    274414 1381292 436.52
## - Ed      1    373656 1480533 439.23
##
## Step:  AIC=428.02
## Crime ~ M + So + Ed + Po1 + Po2 + M.F + Pop + NW + U1 + U2 +
##     Wealth + Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## - Pop     1     11190 1121753 426.41
## - NW      1     23364 1133926 426.83
## - Po2     1     30520 1141082 427.07
## - Wealth  1     30707 1141270 427.08
## - M.F     1     33147 1143710 427.16
## - U1      1     54601 1165164 427.89
## <none>                1110562 428.02
## - So      1     62395 1172957 428.15
## + Time    1      3685 1106877 429.89
## + LF      1      3625 1106938 429.89
## - Po1     1    122385 1232947 430.09
## - Prob    1    150014 1260576 430.96
## - U2      1    212176 1322739 432.83
## - M       1    259601 1370163 434.21
## - Ineq    1    271978 1382541 434.56
```

```
## - Ed      1    371935 1482497 437.28
##
## Step:  AIC=426.41
## Crime ~ M + So + Ed + Po1 + Po2 + M.F + NW + U1 + U2 + Wealth +
##     Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## - NW      1     22200 1143953 425.17
## - Wealth  1     23535 1145288 425.22
## - Po2     1     27529 1149282 425.35
## <none>                1121753 426.41
## - So      1     69122 1190874 426.74
## - U1      1     74167 1195919 426.90
## - M.F     1     86539 1208291 427.31
## + Pop     1     11190 1110562 428.02
## - Po1     1    112948 1234700 428.15
## + LF      1      5661 1116091 428.21
## + Time    1       987 1120765 428.37
## - Prob    1    140019 1261771 428.99
## - U2      1    233510 1355263 431.78
## - M       1    267687 1389440 432.75
## - Ineq    1    275361 1397114 432.97
## - Ed      1    370216 1491969 435.53
##
## Step:  AIC=425.17
## Crime ~ M + So + Ed + Po1 + Po2 + M.F + U1 + U2 + Wealth + Ineq +
##     Prob
##
##          Df Sum of Sq     RSS    AIC
## - Wealth  1     19791 1163744 423.84
## - Po2     1     34993 1178946 424.35
## - So      1     51171 1195124 424.88
## <none>                1143953 425.17
## - U1      1     70867 1214820 425.52
## + NW      1     22200 1121753 426.41
## - M.F     1    103569 1247522 426.55
## + LF      1     13467 1130486 426.71
## + Pop     1     10027 1133926 426.83
## - Po1     1    119245 1263198 427.04
## + Time    1       258 1143695 427.16
## - Prob    1    182143 1326096 428.93
## - U2      1    213315 1357267 429.84
## - M       1    249110 1393063 430.85
## - Ineq    1    260618 1404571 431.18
## - Ed      1    408883 1552836 435.09
##
## Step:  AIC=423.84
## Crime ~ M + So + Ed + Po1 + Po2 + M.F + U1 + U2 + Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## - Po2     1     34673 1198417 422.99
## <none>                1163744 423.84
## - So      1     64814 1228558 423.95
## - U1      1     83160 1246904 424.53
```

```
## + Wealth  1     19791 1143953 425.17
## + NW      1     18456 1145288 425.22
## + LF      1      9981 1153763 425.50
## + Pop     1      3786 1159958 425.71
## - M.F     1    125510 1289254 425.83
## + Time    1         4 1163740 425.84
## - Po1     1    129731 1293475 425.96
## - Prob    1    216973 1380717 428.51
## - U2      1    230709 1394453 428.89
## - M       1    234233 1397977 428.99
## - Ineq    1    310012 1473756 431.05
## - Ed      1    482074 1645818 435.36
##
## Step:  AIC=422.99
## Crime ~ M + So + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## - So      1     47478 1245895 422.50
## <none>                1198417 422.99
## + Po2     1     34673 1163744 423.84
## + NW      1     25310 1173107 424.15
## - U1      1    105681 1304098 424.28
## + Wealth  1     19471 1178946 424.35
## + LF      1      5786 1192631 424.80
## + Time    1      2408 1196009 424.91
## + Pop     1      1977 1196440 424.92
## - M.F     1    161272 1359689 425.91
## - Prob    1    206885 1405302 427.20
## - M       1    235510 1433926 427.98
## - U2      1    262291 1460708 428.70
## - Ineq    1    339458 1537875 430.71
## - Ed      1    451262 1649679 433.45
## - Po1     1    974280 2172697 444.19
##
## Step:  AIC=422.5
## Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## <none>                1245895 422.50
## + So      1     47478 1198417 422.99
## + Wealth  1     31144 1214750 423.51
## + LF      1     25300 1220595 423.70
## + Po2     1     17337 1228558 423.95
## + Pop     1      4512 1241383 424.36
## + NW      1      4157 1241738 424.37
## + Time    1         0 1245895 424.50
## - M.F     1    150124 1396019 424.94
## - U1      1    160091 1405986 425.21
## - Prob    1    161456 1407350 425.25
## - M       1    300427 1546322 428.93
## - U2      1    336243 1582138 429.82
## - Ed      1    424053 1669948 431.92
## - Ineq    1    552239 1798133 434.81
## - Po1     1   1073853 2319748 444.74
```

```r
summary(step.both.1)
```

```
##
## Call:
## lm.default(formula = Crime ~ Po1 + M + M.F + Ineq + Ed + U2 +
##     Prob + U1, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -397.75 -106.78   14.48  141.65  507.68
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7306.48    1380.30  -5.293 1.02e-05 ***
## Po1            94.95      18.67   5.085 1.83e-05 ***
## M             100.62      37.41   2.690 0.011570 *
## M.F            29.67      15.61   1.901 0.066909 .
## Ineq           59.89      16.42   3.647 0.000998 ***
## Ed            185.06      57.91   3.195 0.003277 **
## U2            267.71      94.08   2.845 0.007919 **
## Prob        -3238.63    1642.54  -1.972 0.057924 .
## U1          -8256.55    4205.28  -1.963 0.058931 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 203.8 on 30 degrees of freedom
## Multiple R-squared:  0.7959, Adjusted R-squared:  0.7414
## F-statistic: 14.62 on 8 and 30 DF,  p-value: 1.924e-08
```

```r
summary(step.both.2)
```

```
##
## Call:
## lm.default(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq +
##     Prob, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -397.75 -106.78   14.48  141.65  507.68
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7306.48    1380.30  -5.293 1.02e-05 ***
## M             100.62      37.41   2.690 0.011570 *
## Ed            185.06      57.91   3.195 0.003277 **
## Po1            94.95      18.67   5.085 1.83e-05 ***
## M.F            29.67      15.61   1.901 0.066909 .
## U1          -8256.55    4205.28  -1.963 0.058931 .
## U2            267.71      94.08   2.845 0.007919 **
## Ineq           59.89      16.42   3.647 0.000998 ***
## Prob        -3238.63    1642.54  -1.972 0.057924 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 203.8 on 30 degrees of freedom
## Multiple R-squared:  0.7959, Adjusted R-squared:  0.7414
## F-statistic: 14.62 on 8 and 30 DF,  p-value: 1.924e-08
```

```
#Testing
res.both.1 <- test %>%
  add_predictions(.,step.both.1) %>%
  select('observations' = Crime, pred) %>%
  as.data.frame()
res.both.2 <- test %>%
  add_predictions(.,step.both.2) %>%
  select('observations' = Crime, pred) %>%
  as.data.frame()
```

```
#Prediction
stepwise.fromNull <- postResample(obs = res.both.1$observations, pred = res.both.1$pred)
stepwise.fromFull <- postResample(obs = res.both.2$observations, pred = res.both.2$pred)
```

```
stepwise.fromNull
```

```
##        RMSE    Rsquared         MAE
## 192.5011883   0.6075089 164.3968602
```

```
stepwise.fromFull
```

```
##        RMSE    Rsquared         MAE
## 192.5011883   0.6075089 164.3968602
```

Below is the results on test data of all 4 methods:

```
data.frame(forward,backward,stepwise.fromFull,stepwise.fromNull)
```

```
##              forward    backward stepwise.fromFull stepwise.fromNull
## RMSE     192.5011883 192.5011883       192.5011883       192.5011883
## Rsquared   0.6075089   0.6075089         0.6075089         0.6075089
## MAE      164.3968602 164.3968602       164.3968602       164.3968602
```

2. Lasso

Then we try the Lasso method: Slitting data

```
set.seed(101)
train.index <- createDataPartition(crime$Crime , p = 0.8 ,times = 1, list = F)
train <- crime[train.index,]
test <- crime[-train.index,]
```

Building model

```
#modeling
lasso <- glmnet(x = scale(as.matrix(train[,-16])),
                y =scale(as.matrix(train[,16]))   ,
                family = "gaussian" ,
                alpha = 1)
```

Finding the best lambda value through cv.glmnet

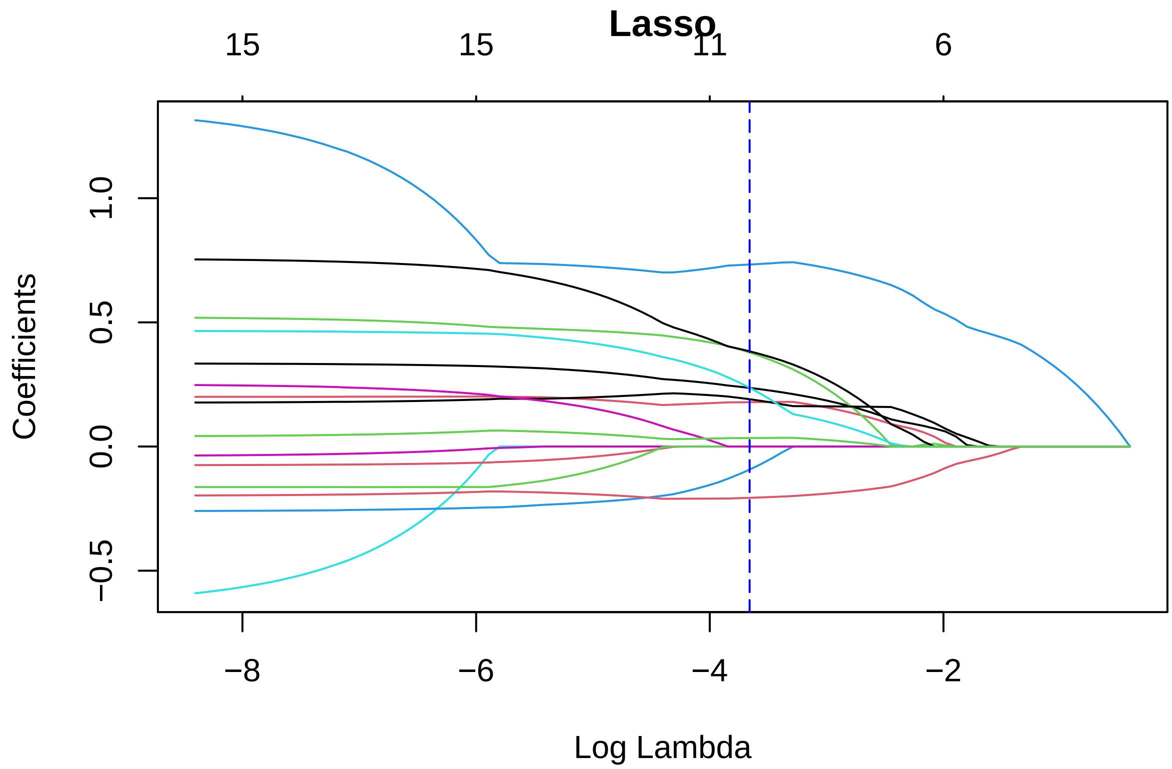
```
cv.lasso <- cv.glmnet(x = scale(as.matrix(train[,-16])),
                      y =scale(as.matrix(train[,16]))   ,
                      family = "gaussian" ,
                      alpha = 1)
```

```r
best.lambda = cv.lasso$lambda.min
best.lambda
```

```
## [1] 0.02575229
```

Plotting

```r
plot(lasso, xvar='lambda', main="Lasso")
abline(v=log(best.lambda), col="blue", lty=5.5 )
```



Choosing the coefficients based on the lambda chosen

```r
#choosing coefficients
coef(cv.lasso, s = "lambda.min")
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept) -3.139861e-16
## M            2.361438e-01
## So           1.785041e-01
## Ed           3.787988e-01
## Po1          7.328666e-01
## Po2          .
## LF           .
## M.F          1.904635e-01
## Pop          .
## NW           .
## U1          -9.300471e-02
```

```
## U2           2.369465e-01
## Wealth       .
## Ineq         3.833617e-01
## Prob        -2.065999e-01
## Time         3.417036e-02
```

```
select.ind = which(coef(cv.lasso, s = "lambda.min") != 0)
select.ind = select.ind[-1]-1 # remove Intercept
important <- colnames(train[select.ind])
important# which one is important
```

```
##  [1] "M"    "So"   "Ed"   "Po1"  "M.F"  "U1"   "U2"   "Ineq" "Prob" "Time"
```

```
#Regression model
lasso.reg <- lm(Crime~., data = train[,c(important,"Crime")])
summary(lasso.reg)
```

```
##
## Call:
## lm.default(formula = Crime ~ ., data = train[, c(important, "Crime")])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -376.11  -95.74   15.06  122.03  526.81
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7382.332   1534.545  -4.811 4.65e-05 ***
## M              88.500     40.827   2.168  0.03884 *
## So            143.343    132.641   1.081  0.28906
## Ed            193.504     59.391   3.258  0.00294 **
## Po1            90.907     19.450   4.674 6.76e-05 ***
## M.F            32.222     16.919   1.904  0.06717 .
## U1          -6809.787   4489.095  -1.517  0.14049
## U2            239.632     99.391   2.411  0.02272 *
## Ineq           50.769     18.820   2.698  0.01170 *
## Prob        -3720.703   2103.563  -1.769  0.08783 .
## Time            1.696      7.142   0.237  0.81406
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 206.7 on 28 degrees of freedom
## Multiple R-squared:  0.804,  Adjusted R-squared:  0.734
## F-statistic: 11.49 on 10 and 28 DF,  p-value: 1.683e-07
```

Using test data

```
res.lasso <- test %>%
  add_predictions(.,lasso.reg) %>%
  select('observations' = Crime, pred) %>%
  as.data.frame()
```

Regression results on test data

```
Lasso.regression <- postResample(obs = res.lasso$observations, pred = res.lasso$pred)
Lasso.regression
```

```
##        RMSE    Rsquared         MAE
## 215.4929042   0.5093659 174.1936776
```

3. Elastic net

```r
#divide data
set.seed(101)
train.index <- createDataPartition(crime$Crime , p = 0.8 ,times = 1, list = F)
train <- crime[train.index,]
test <- crime[-train.index,]
```

```r
#Finding suitable aplpha
a <- seq(0.05, 0.95, 0.05)
search <- foreach(i = a, .combine = rbind) %dopar% {
cv.elastic <- cv.glmnet(x = scale(as.matrix(train[,-16])),
                  y =scale(as.matrix(train[,16]))  ,
                  family = "gaussian" ,
                  nfold = 10,
                  type.measure = "deviance",
                  paralle = TRUE,
                  alpha = i)
  data.frame(cvm = cv.elastic$cvm[cv.elastic$lambda == cv.elastic$lambda.1se],
             lambda.1se = cv.elastic$lambda.1se,
             alpha = i)
}
```

```
## Warning: executing %dopar% sequentially: no parallel backend registered
```

```r
search
```

```
##          cvm lambda.1se alpha
## 1  0.5587357 0.68086028  0.05
## 2  0.5808445 0.37362166  0.10
## 3  0.5676798 0.27336621  0.15
## 4  0.5441202 0.20502465  0.20
## 5  0.6236860 0.19756234  0.25
## 6  0.5512257 0.23885760  0.30
## 7  0.5344820 0.11715695  0.35
## 8  0.6009797 0.12347646  0.40
## 9  0.4988893 0.10975685  0.45
## 10 0.5523747 0.09000572  0.50
## 11 0.5616747 0.09855656  0.55
## 12 0.5203781 0.09034351  0.60
## 13 0.5306673 0.09152482  0.65
## 14 0.5865062 0.09327349  0.70
## 15 0.5881139 0.08705526  0.75
## 16 0.7048490 0.08957160  0.80
## 17 0.6121278 0.10154288  0.85
## 18 0.5221102 0.07254605  0.90
## 19 0.4977953 0.06872784  0.95
```

```r
cv <- search[search$cvm == min(search$cvm), ]
cv
```

```
##          cvm lambda.1se alpha
## 19 0.4977953 0.06872784  0.95
```

```r
#modeling
elastic <- glmnet(x = scale(as.matrix(train[,-16])),
                 y =scale(as.matrix(train[,16]))  ,
                 family = "gaussian" ,
                 alpha = cv$alpha,
                 lambda = cv$lambda.1se)
```

```r
#choosing coefficients
coef(elastic, s = "lambda.min")
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept) -1.342940e-16
## M            1.515020e-01
## So           1.299846e-01
## Ed           1.351864e-01
## Po1          6.842464e-01
## Po2          .
## LF           .
## M.F          1.610136e-01
## Pop          .
## NW           .
## U1           .
## U2           6.369060e-02
## Wealth       .
## Ineq         1.878501e-01
## Prob        -1.768478e-01
## Time         1.715607e-02
```

```r
select.ind2 = which(coef(elastic, s = "lambda.min") != 0)
select.ind2 = select.ind[-1]-1 # remove Intercept
important2 <- colnames(train[select.ind2])
important2
```

```
## [1] "M"      "So"     "Ed"     "LF"     "NW"     "U1"     "Wealth" "Ineq"
## [9] "Prob"
```

```r
#Regression model
elastic.reg <- lm(Crime~., data = train[,c(important,"Crime")])
```

```r
summary(elastic.reg)
```

```
##
## Call:
## lm.default(formula = Crime ~ ., data = train[, c(important, "Crime")])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -376.11  -95.74   15.06  122.03  526.81
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7382.332   1534.545  -4.811 4.65e-05 ***
## M              88.500     40.827   2.168  0.03884 *
## So            143.343    132.641   1.081  0.28906
## Ed            193.504     59.391   3.258  0.00294 **
```

```
## Po1            90.907     19.450   4.674 6.76e-05 ***
## M.F            32.222     16.919   1.904  0.06717 .
## U1          -6809.787   4489.095  -1.517  0.14049
## U2            239.632     99.391   2.411  0.02272 *
## Ineq           50.769     18.820   2.698  0.01170 *
## Prob        -3720.703   2103.563  -1.769  0.08783 .
## Time            1.696      7.142   0.237  0.81406
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 206.7 on 28 degrees of freedom
## Multiple R-squared:  0.804,  Adjusted R-squared:  0.734
## F-statistic: 11.49 on 10 and 28 DF,  p-value: 1.683e-07
```

```
res.elastic <- test %>%
  add_predictions(.,elastic.reg) %>%
  select('observations' = Crime, pred) %>%
  as.data.frame()
```

```
Elastic.regression <- postResample(obs = res.lasso$observations, pred = res.lasso$pred)
Elastic.regression
```

```
##        RMSE    Rsquared         MAE
## 215.4929042   0.5093659 174.1936776
```

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:

One design of experiments I would implement is that of target marketing for followers of a facebook Fanpage. By using factoral design, we could target potential followers more effectively.

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.

```
FrF2(16,nfactors = 10,
     factor.names =  c("Large yard" , "solar roof" , "double restrooms" , "Garage" , "pool" ,"lawn",
                       "security system" , "Smart house system" ,"Elevator","Walk-in closet") )
```

```
##    Large.yard solar.roof double.restrooms Garage pool lawn security.system
## 1           1         -1               -1     -1   -1   -1               1
## 2          -1          1                1     -1   -1   -1               1
## 3          -1         -1                1      1    1   -1              -1
## 4           1         -1                1     -1   -1    1              -1
## 5          -1          1               -1      1   -1    1              -1
## 6          -1         -1                1     -1    1   -1              -1
## 7          -1         -1               -1     -1    1    1               1
## 8          -1         -1               -1      1    1    1               1
## 9           1         -1                1      1   -1    1              -1
## 10          1          1                1      1    1    1               1
```

```
## 11          1          1                1     -1    1    1              1
## 12          1          1               -1     -1    1   -1             -1
## 13          1         -1               -1      1   -1   -1              1
## 14         -1          1               -1     -1   -1    1             -1
## 15          1          1               -1      1    1   -1             -1
## 16         -1          1                1      1   -1   -1              1
##    Smart.house.system Elevator Walk.in.closet
## 1                  -1       -1             -1
## 2                   1       -1              1
## 3                  -1       -1              1
## 4                  -1        1              1
## 5                  -1       -1              1
## 6                   1        1             -1
## 7                   1       -1              1
## 8                  -1        1             -1
## 9                   1       -1             -1
## 10                  1        1              1
## 11                 -1       -1             -1
## 12                 -1        1              1
## 13                  1        1              1
## 14                  1        1             -1
## 15                  1       -1             -1
## 16                 -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).

a. Binomial

The chances of strucking the grand prize of a lottery.

b. Geometric

How many times does an average fielder successfully make a catch befor making an error.

c. Poisson

How many people are just late for a class, just late defined as within 5 minutes after class starts.

d. Exponential

How long does it take between students who are late for class.

e. Weibull

How long does it take for a computer's battery to wear out if continuously charged.