ISYE 6501 HW8

Chengqi

2019/10/9

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.

Notes on R: • For the elastic net model, what we called λ in the videos, glmnet calls "alpha"; you can get a range of results by varying alpha from 1 (lasso) to 0 (ridge regression) [and, of course, other values of alpha in between]. • In a function call like glmnet(x,y,family="mgaussian",alpha=1) the predictors x need to be in R's matrix format, rather than data frame format. You can convert a data frame to a matrix using as.matrix – for example, x <- as.matrix(data[,1:n-1]) • Rather than specifying a value of T, glmnet returns models for a variety of values of T.

First of all, we load the data:

```
data <- read.table('C:\\Users\\huangchengqi\\Desktop\\MS SCE\\19Fall\\I
SYE6501\\hw8\\data 11.1\\uscrime.txt', header=TRUE)
```

We also have a data set to use the model to predict number of crimes:

```
test_data_set <-data.frame(M = 14,So = 0, Ed = 10.0, Po1 = 12.0, Po2 =
15.5,LF = 0.64, M.F = 94.0, Pop = 150, NW = 1.1, U1 = 0.12, U2 = 3.6, W
ealth = 3200, Ineq = 20.1, Prob = 0.040, Time = 39.0)
```

Stepwise regression

In order to do stepwise Regression, we have to fit an original model using all the predictors, and then we using AICc to choose the best combination of predictors:

```
model_1 <- lm(Crime~., data = data)

library(MASS)
stepAICc(model_1, direction='both', steps=1000)

## Start:  AIC=671.13
## Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 +
##     U2 + Wealth + Ineq + Prob + Time
```

```
##
##          Df Sum of Sq     RSS    AIC
## - So      1        29 1354974 512.65
## - LF      1      8917 1363862 512.96
## - Time    1     10304 1365250 513.00
## - Pop     1     14122 1369068 513.14
## - NW      1     18395 1373341 513.28
## - M.F     1     31967 1386913 513.74
## - Wealth  1     37613 1392558 513.94
## - Po2     1     37919 1392865 513.95
## <none>                1354946 514.65
## - U1      1     83722 1438668 515.47
## - Po1     1    144306 1499252 517.41
## - U2      1    181536 1536482 518.56
## - M       1    193770 1548716 518.93
## - Prob    1    199538 1554484 519.11
## - Ed      1    402117 1757063 524.86
## - Ineq    1    423031 1777977 525.42
##
## Step:  AIC=666.16
## Crime ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 +
##     Wealth + Ineq + Prob + Time
##
##          Df Sum of Sq     RSS    AIC
## - Time    1     10341 1365315 511.01
## - LF      1     10878 1365852 511.03
## - Pop     1     14127 1369101 511.14
## - NW      1     21626 1376600 511.39
## - M.F     1     32449 1387423 511.76
## - Po2     1     37954 1392929 511.95
## - Wealth  1     39223 1394197 511.99
## <none>                1354974 512.65
## - U1      1     96420 1451395 513.88
## + So      1        29 1354946 514.65
## - Po1     1    144302 1499277 515.41
## - U2      1    189859 1544834 516.81
## - M       1    195084 1550059 516.97
## - Prob    1    204463 1559437 517.26
## - Ed      1    403140 1758114 522.89
## - Ineq    1    488834 1843808 525.13
##
## Step:  AIC=661.87
## 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
## + Time    1     10341 1354974 512.65
## + So      1        65 1365250 513.00
## - 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
##
## Step:  AIC=657.87
## 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
## + LF      1     10533 1365315 511.01
## + Time    1      9996 1365852 511.03
## + So      1      3046 1372802 511.26
## - 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=654.18
## 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
## + NW      1     11675 1375848 509.37
## + So      1      7207 1380316 509.52
## + LF      1      6726 1380797 509.54
## + Time    1      4534 1382989 509.61
```

```
## - 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=650.88
## 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
## + Po2     1     16706 1387523 507.77
## + NW      1      6963 1397266 508.09
## + So      1      3807 1400422 508.20
## + LF      1      1986 1402243 508.26
## + Time    1       575 1403654 508.31
## - 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=647.99
## 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
## + Pop     1     22345 1404229 506.33
## + Po2     1     13259 1413315 506.63
## + NW      1      5927 1420648 506.87
## + So      1      5724 1420851 506.88
## + LF      1      5176 1421398 506.90
## + Time    1      3913 1422661 506.94
## - 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=645.43
## Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## <none>                1453068 503.93
## + Wealth  1     26493 1426575 505.07
## - M.F     1    103159 1556227 505.16
## + Pop     1     16697 1436371 505.39
## + Po2     1     14148 1438919 505.47
## + So      1      9329 1443739 505.63
## + LF      1      4374 1448694 505.79
## + NW      1      3799 1449269 505.81
## + Time    1      2293 1450775 505.86
## - 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

##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob,
##     data = data)
##
## Coefficients:
## (Intercept)            M           Ed          Po1          M.F
##    -6426.10        93.32       180.12       102.65        22.34
##          U1           U2         Ineq         Prob
##    -6086.63       187.35        61.33     -3796.03
```

The stepwise regression indicates us to use the following predictors to fit the model:

```
model_stepswise <- lm(Crime~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob,
data = data)
summary(model_stepswise)

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

```
#See the prediction using the stepwise model
pred_1 <-predict(model_stepswise,test_data_set)
pred_1
```

```
##        1
## 1038.413
```

LASSO

Next we use LASSO to choose predictors

```
library('glmnet')
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

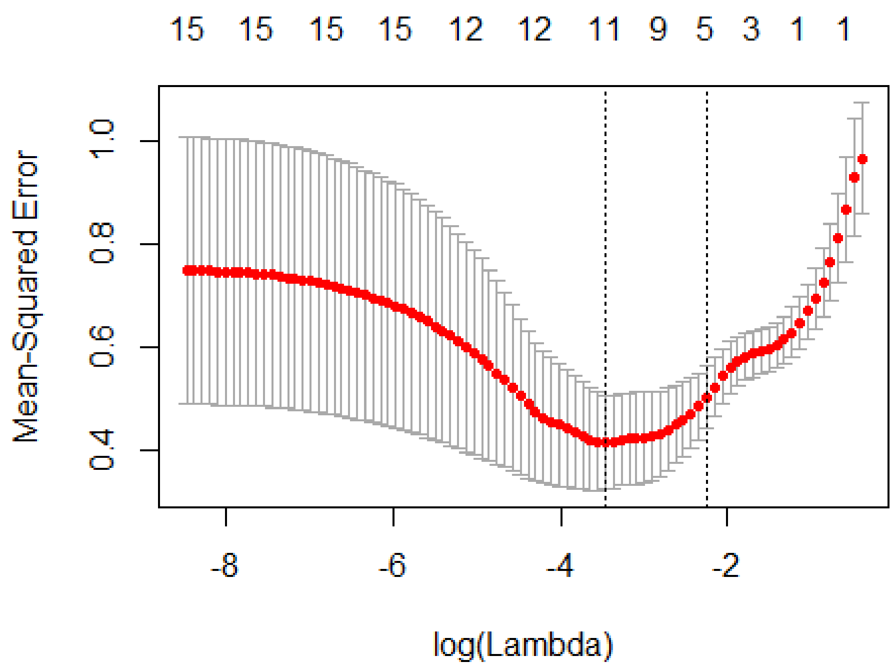
```
## Loaded glmnet 2.0-18
```

```
#We need to scale our data first before using LASSO
data2 <- scale(data)
X <- as.matrix(data2[,1:15])
Y <- as.matrix(data2[,16])
#At the same time we can do cross validation towards lasso models
Lasso <- cv.glmnet(X, Y, family = 'gaussian', alpha = 1, nfolds = 5, ty
pe.measure = "mse")
coef_Lasso <- coef(Lasso$glmnet.fit, s=Lasso$lambda.min)
coef_Lasso
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept) -2.983091e-16
## M            2.143137e-01
## So           5.938496e-02
## Ed           3.000769e-01
## Po1          8.020148e-01
## Po2          .
## LF           5.762247e-03
```

```
## M.F          1.288092e-01
## Pop          .
## NW           9.368004e-03
## U1          -3.577858e-02
## U2           1.207421e-01
## Wealth       .
## Ineq         4.538892e-01
## Prob        -2.079189e-01
## Time         .

plot(Lasso)
```



Now we have the chosen predictors. Using these predictors to fit the regression model:

```
model_Lasso <- lm(Crime ~ M + So + Ed + Po1 + M.F + NW + U2 + Ineq + Pr
ob, data = data)
summary(model_Lasso)

##
## Call:
## lm(formula = Crime ~ M + So + Ed + Po1 + M.F + NW + U2 + Ineq +
##     Prob, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -415.05 -122.24    0.05  114.69  557.46
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5709.8007  1172.1390  -4.871 2.10e-05 ***
## M              87.4640    38.9342   2.246 0.030730 *
## So             98.9665   119.8910   0.825 0.414395
## Ed            176.1445    55.4389   3.177 0.002998 **
## Po1           112.1986    16.4389   6.825 4.85e-08 ***
## M.F            13.7469    13.2263   1.039 0.305384
## NW              0.3986     5.5821   0.071 0.943465
## U2             74.9516    43.7048   1.715 0.094719 .
## Ineq           59.5760    16.5608   3.597 0.000935 ***
## Prob        -4482.9501  1735.7313  -2.583 0.013892 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 204.5 on 37 degrees of freedom
## Multiple R-squared:  0.7752, Adjusted R-squared:  0.7205
## F-statistic: 14.17 on 9 and 37 DF,  p-value: 1.541e-09
```

```
pred_2 <- predict(model_Lasso, test_data_set)
pred_2
```

```
##        1
## 1203.153
```

Elastic Net

since alpha can be any value between 0 and 1, we try different alpha

```
set.seed(5000)
Dev <- vector()
for (i in 0:10) {
  Elastic = cv.glmnet(X,Y,alpha=i/10,family="gaussian",nfolds=5,type.me
asure = "mse")
  Dev[i+1] = Elastic$glmnet.fit$dev.ratio[which(Elastic$glmnet.fit$lamb
da == Elastic$lambda.min)]
  }
View(Dev)
#The largest Dev appears when i=8, which means alpha = 0.7
Elastic_model = cv.glmnet(X,Y,alpha=0.7,family="gaussian",nfolds=5,type.
measure = "mse")
coef_Elastic <- coef(Elastic_model$glmnet.fit, s=Elastic_model$lambda.m
in)
coef_Elastic
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept) -3.109819e-16
## M            2.074104e-01
## So           6.496341e-02
## Ed           2.954245e-01
## Po1          7.178419e-01
```

```
## Po2          5.435303e-02
## LF           1.131839e-02
## M.F          1.407239e-01
## Pop          .
## NW           2.597821e-02
## U1          -5.832268e-02
## U2           1.431788e-01
## Wealth       .
## Ineq         4.238232e-01
## Prob        -2.143996e-01
## Time         .
```

```
model_Elastic <- lm(Crime ~ M + So+ Ed + Po1 + LF + M.F + NW + U1 + U2
+ Ineq + Prob, data = data)
summary(model_Elastic)
```

```
##
## Call:
## lm(formula = Crime ~ M + So + Ed + Po1 + LF + M.F + NW + U1 +
##     U2 + Ineq + Prob, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -443.2 -101.4    4.1  120.5  486.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6434.101   1253.263  -5.134 1.07e-05 ***
## M              84.825     39.221   2.163  0.03747 *
## So             36.573    139.615   0.262  0.79489
## Ed            186.954     58.101   3.218  0.00278 **
## Po1            99.463     18.338   5.424 4.44e-06 ***
## LF           -264.646   1339.041  -0.198  0.84447
## M.F            25.438     17.353   1.466  0.15159
## NW              1.265      5.783   0.219  0.82814
## U1          -6050.130   3977.786  -1.521  0.13725
## U2            179.349     78.140   2.295  0.02783 *
## Ineq           58.402     16.962   3.443  0.00151 **
## Prob        -4222.327   1740.886  -2.425  0.02059 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 202.9 on 35 degrees of freedom
## Multiple R-squared:  0.7906, Adjusted R-squared:  0.7248
## F-statistic: 12.01 on 11 and 35 DF,  p-value: 6.965e-09
```

```
pred_3 <- predict(model_Elastic, test_data_set)
pred_3
```

```
##        1
## 964.3991
```