HW5

Question 1

Using the crime data set from Homework 3, 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.*

Answer:

1. Stepwise regression

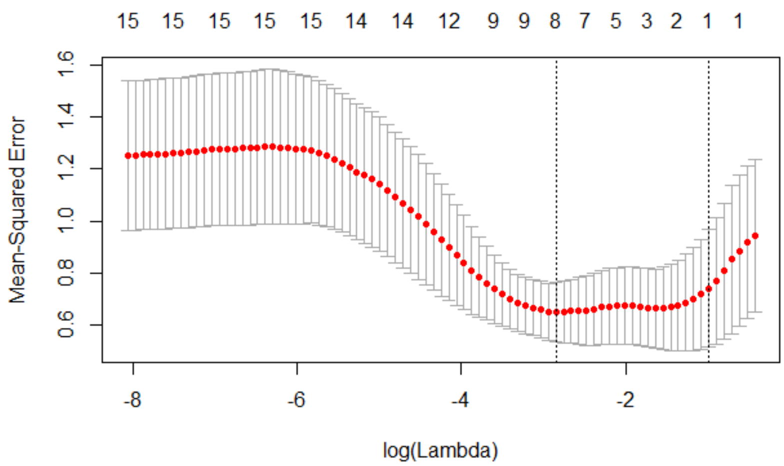
```
# Clear environment
rm(list = ls())

# Read data
data <- read.table("uscrime.txt", stringsAsFactors = FALSE, header = TRUE)

# Install packages
install.packages("MASS") # for setpwise regression
library(MASS)

# Setting the random number generator seed so that our results are reproducible
set.seed(1)

# Run stepwise regression model
datalm <- lm(Crime~M+So+Ed+Po1+Po2+LF+M.F+Pop+NW+U1+U2+Wealth+Ineq+Prob+Time, data)
datalmstep <- stepAIC(datalm, direction = "both")
datalmstep$anova
```

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 +
    U2 + Wealth + Ineq + Prob + Time

Final Model:
Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob


      Step Df    Deviance Resid. Df Resid. Dev      AIC
1                                31    1354946 514.6488
2     - So  1    28.57405        32    1354974 512.6498
3   - Time  1 10340.66984        33    1365315 511.0072
4     - LF  1 10533.15902        34    1375848 509.3684
5     - NW  1 11674.63991        35    1387523 507.7655
6    - Po2  1 16706.34095        36    1404229 506.3280
7    - Pop  1 22345.36638        37    1426575 505.0700
8 - Wealth  1 26493.24677        38    1453068 503.9349
```

⇨ Results as above Final Model.

2. Lasso

```
# Install packages
install.packages("glmnet") # for LASSO and Elastic net
library(glmnet)

# Scale the data
datascale<-scale(data)

# set x & y for Lasso
x <- as.matrix(datascale[,1:15])
y <- datascale[,16]

# Split data into train (2/3) and test (1/3) sets
train_rows <- sample(1:47, .66*47)
x.train <- x[train_rows, ]
x.test <- x[-train_rows, ]

y.train <- y[train_rows]
y.test <- y[-train_rows]

# Run Lasso
Model_Lasso <- glmnet(x.train, y.train, alpha = 1, family = "mgaussian")

# Use cross validation to select lambda
cv.Model_Lasso <- cv.glmnet(x.train, y.train, alpha=1)
plot(cv.Model_Lasso)
```



(best.lambda <- cv.Model_Lasso$lambda.min)

#[1] 0.05777771

coef of the best model

coef(cv.Model_Lasso, s = "lambda.min")

```
16 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) -0.03481449
M            0.19085890
So           .
Ed           .
Po1          0.56457180
Po2          .
LF           0.05981152
M.F          0.05555324
Pop          0.07305772
NW           .
U1           .
U2           .
Wealth       0.18538482
Ineq         0.23343055
Prob        -0.10017579
Time         .
```

⇨ Lambda & coefs (total 8 valuables are chosen) of the best Lasso model as above

Check MSE based on test set

yhat <- predict(cv.Model_Lasso, s=cv.Model_Lasso$lambda.min, newx=x.test)

mse <- mean((y.test - yhat)^2)

⇨ MSE is 0.5174414 for this Lasso model

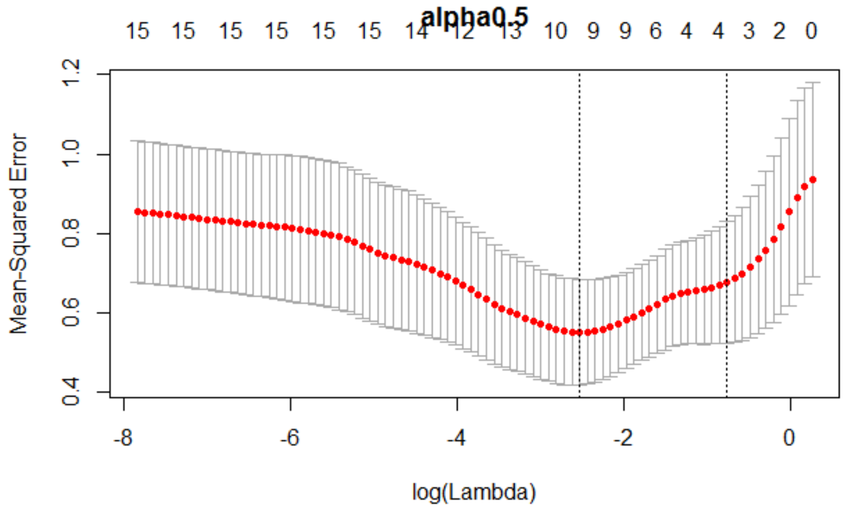
3. Elastic net

Run Elastic net, alpha=0.5

Model_EN0.5 <- glmnet(x.train, y.train, alpha = 0.5, family = "mgaussian")

Use cross validation to select lambda

cv.Model_EN0.5 <- cv.glmnet(x.train, y.train, alpha=0.5) plot(cv.Model_EN0.5, main="alpha0.5")



(best.lambda <- cv.Model_EN0.5$lambda.min)

#[1] 0.07964786

coef(cv.Model_Lasso, s = "lambda.min")

```
16 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) -0.03481449
M            0.19085890
So           .
Ed           .
Po1          0.56457180
Po2          .
LF           0.05981152
M.F          0.05555324
Pop          0.07305772
NW           .
U1           .
U2           .
Wealth       0.18538482
Ineq         0.23343055
Prob        -0.10017579
Time         .
```

Check MSE based on test set

yhatEN <- predict(cv.Model_EN0.5, s=cv.Model_EN0.5$lambda.min, newx=x.test)

mseEN <- mean((y.test - yhatEN)^2)

mseEN

⇨ MSE is 0.5319272 for this Elastic Net model. This MSE is larger than the MSE of previous Las so model. Suggest that Lasso model works better.

Question 2

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

Answer:

When design a new tablet, what would be a proper color for the tablet? DOE could be conducted find a better answer whether the color should be black or dark gray. We can make samples and for some representative potential customers to choose from.

Question 3

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? Note: the output of FrF2 is "1" (include) or "-1" (don't include) for each feature.

Answer:

```
# Clear environment
rm(list = ls())

# install package
install.packages("FrF2")
library(FrF2)

# fractional factorial design, 10 factors (features), 16 runs(fictitious houses)
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
```

⇨ Result of fractional factorial design for this experiment as above

Question 4

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

b. Geometric

c. Poisson

d. Exponential

e. Weibull

Answer:

a: rolling dice

b: How many times to roll a dice until it shows 1

c: How many people arrival at a bus stop during certain time period

d: Time interval between people arrival at bus stop

e: Turn on a tablet and plug with ac adapter, how long until the tablet be defective?