# hw8

*Kunle Lawal, Anubhav Rana, Mihir Tulpule and Ali Lakdawala*

## Stepwise models

```
require(data.table)
```

```
## Loading required package: data.table
```

```
crimedata = read.table('uscrime.txt', header = TRUE)
library(MASS)

#Original Linear Model with all variables
model <- lm(Crime ~., data = crimedata)
summary(model)
```

```
## 
## Call:
## lm(formula = Crime ~ ., data = crimedata)
## 
## 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
```

```
#Model with only significant variables
model2 <- lm(Crime ~ Ed + Ineq + M +Po2 + Prob + U2, data = crimedata )
summary(model2)
```

```
##
## Call:
## lm(formula = Crime ~ Ed + Ineq + M + Po2 + Prob + U2, data = crimedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -550.12 -105.77    0.65  136.86  535.57
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5073.29     941.77  -5.387 3.43e-06 ***
## Ed            196.66      46.73   4.209 0.000141 ***
## Ineq           68.94      14.65   4.706 3.00e-05 ***
## M             105.04      34.77   3.021 0.004377 **
## Po2           120.66      15.47   7.801 1.47e-09 ***
## Prob        -3983.52    1592.33  -2.502 0.016552 *
## U2             96.35      42.58   2.263 0.029136 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209.5 on 40 degrees of freedom
## Multiple R-squared:  0.7448, Adjusted R-squared:  0.7065
## F-statistic: 19.46 on 6 and 40 DF,  p-value: 1.82e-10
```

```
#Stepwise with all
step_model <- step(model, direction = "both", trace = FALSE)
summary(step_model)
```

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

```r
#Stepwise with only significant
step_model2 <- step(model2, direction = "both", trace = FALSE)
summary(step_model2)
```

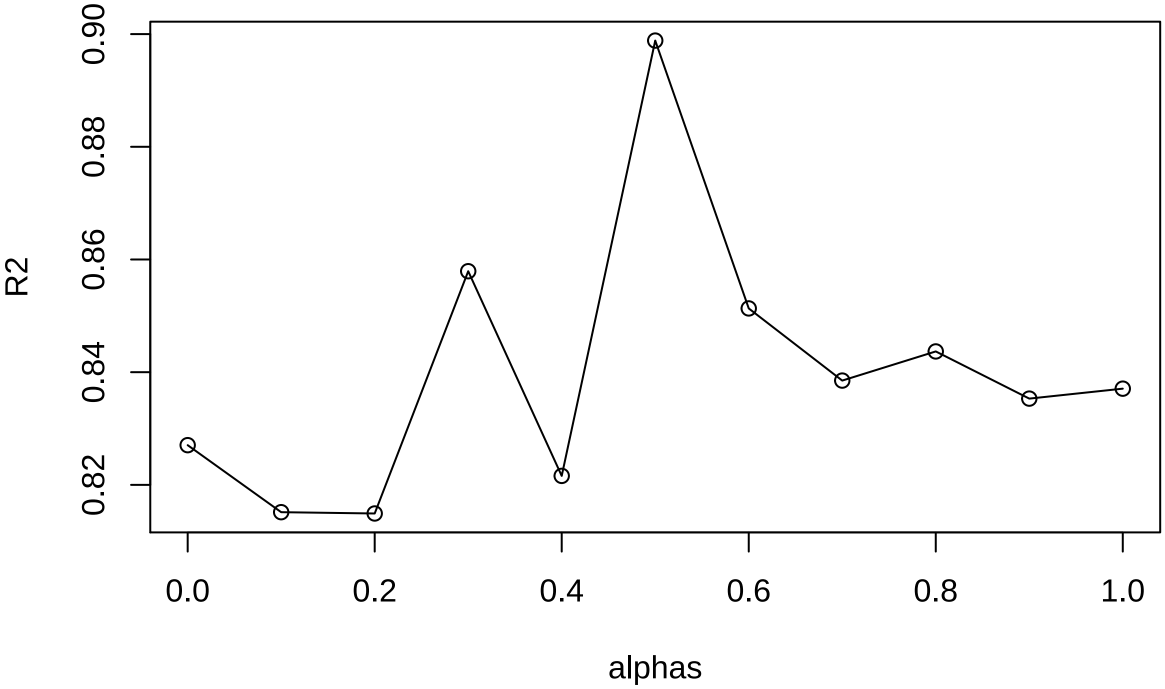
```
##
## Call:
## lm(formula = Crime ~ Ed + Ineq + M + Po2 + Prob + U2, data = crimedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -550.12 -105.77    0.65  136.86  535.57
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5073.29     941.77  -5.387 3.43e-06 ***
## Ed            196.66      46.73   4.209 0.000141 ***
## Ineq           68.94      14.65   4.706 3.00e-05 ***
## M             105.04      34.77   3.021 0.004377 **
## Po2           120.66      15.47   7.801 1.47e-09 ***
## Prob        -3983.52    1592.33  -2.502 0.016552 *
## U2             96.35      42.58   2.263 0.029136 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209.5 on 40 degrees of freedom
## Multiple R-squared:  0.7448, Adjusted R-squared:  0.7065
## F-statistic: 19.46 on 6 and 40 DF,  p-value: 1.82e-10
```

## Lasso Model & Elastic Net

```r
library(glmnet)
set.seed(1234)
scaleData <- scale(crimedata, center = TRUE, scale = TRUE)

x <- scaleData[,-16]
y <- scaleData[,16]

m <- nrow(x)
trn <- sample(1:m, size = round(m*0.7), replace = FALSE)
trainx <- x[trn,]
testx <- x[-trn,]


train_y <- y[trn]
test_y <- y[-trn]


fit.lasso <- glmnet(trainx, train_y, family = "gaussian", alpha = 1)
fit.elnet <- glmnet(trainx, train_y, family = "gaussian", alpha = 0.5)

R2 <- c()
for (i in 0:10) {
  elasticfit <- cv.glmnet(trainx, train_y, type.measure="mse",
                          alpha=i/10,family="gaussian")
  R2 = cbind(R2, elasticfit$glmnet.fit$dev.ratio[which(elasticfit$glmnet.fit$lambda == elasticfit$lambda
}
R2
```

```
##            [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
## [1,] 0.8270468 0.8151545 0.8149218 0.8579185 0.8215956 0.8988502 0.8513139
##            [,8]      [,9]     [,10]     [,11]
## [1,] 0.8385065 0.8436814 0.8353078 0.8370728
```

```
alphas <- (0:10)/10

plot(alphas, R2)
lines(alphas, R2)
```



In our original model with all the predictors adjusted the Multiple R-Squared is 0.8031 In original model with only significant predictors have the Multiple R-Squared with 0.744

For Stepwise model with all predictors we noticed R Squared Values of 0.789 For Stepwise model with just significant we notice the same R squared as the original model with significant predictors 0.744.

Then for Elastic net and Lasso models we plotted the r-squared values above for alphas 0 -1. The maximum R-Squared value for the Elastic models was higher than the maximum value for the Lasso model but only by a very small margin. The maximum value is around 0.899.

What we found is that Elastic and Lasso models were superior to the Stepwise model in terms of quality. We also notice that Stepwise regression and the orginal regression model with only significant predictors the R-squared value is the same. This makese sense as stepwise model chooses the top predictors and in this case they were all significant.

It is important to note that all three methods produced models that were superior to the original linear regression model in terms of r-squared values or complexity.